A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among Australians

Author

Ng, SK

Published

2015

Journal Title

Statistics in Medicine

Version

Submitted Manuscript (SM)

DOI

10.1002/sim.6542

Statistics in Medicine

Research Article

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among Australians

S. K. Ng[a]*

Multimorbidity is present in more than one quarter of the population in Australia and its prevalence increases with age. Greater multimorbidity burden among individuals is always associated with poor health-related outcomes, including quality of life, health service utilization, and mortality, among others. It is thus significant to identify the heterogeneity in multimorbidity patterns in the community and determine the impact of multimorbidity on individual health outcomes. In this paper, I propose a two-way clustering framework to identify clusters of most significant non-random comorbid health conditions and disparities in multimorbidity patterns among individuals. This framework can establish a clustering-based approach to determine the association between multimorbidity patterns and health-related outcomes, and to calculate a multimorbidity score for each individual. The proposed method is illustrated using simulated data and a national survey data set of mental health and wellbeing in Australia.

Keywords: multimorbidity; mixture models; multivariate generalized Bernoulli distribution; national survey data; cluster analysis; EM algorithm

1. Introduction

In Australia, "multimorbidity" (the co-occurrence of two or more medical conditions within one person [1]) is present in more than one quarter of the population and its prevalence increases with age [2]. Based on the 2004-2005 Australian National Health Survey (NHS), 80% of Australians aged 65 years or older have at least three or more chronic conditions. This population is projected to increase from 2.6 million in 2004 to over 6.5 million by 2051 [3]. High prevalence rates of multimorbidity are also observed in other countries, such as Canada and the USA [4]. One of the major goals in multimorbidity or comorbidity (defined as the occurrence of medical conditions in addition to an index condition of interest [5]) research is to reveal groups of medical conditions where the co-occurrence of conditions within a group is not random [6, 7]. The aim is to identify clinically relevant medical conditions that exhibit a potential sharing of risk

[a] *School of Medicine and Menzies Health Institute Queensland, Griffith University, Meadowbrook, QLD 4131, Australia*
[*] *Correspondence to: Shu Kay Angus Ng, School of Medicine and Menzies Health Institute Queensland, Griffith University, Meadowbrook, QLD 4131, Australia.*

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

Statistics
in Medicine

S. K. Ng

factors or common pathophysiological causes [8]. This information could improve the evidence-base by helping physicians and patients to prioritize therapies and goals [9]. Another major goal is to investigate the impact of multimorbidity on individual health-related outcomes. In this context, it has been shown that coexisting morbidities are associated with many adverse health outcomes, including quality of life, health service utilization, and mortality, among others [3, 10, 11]. Multimorbidity thus constitutes a serious burden and challenge on the health care system, including the treatment and care of diseases, the utilization of resources and the associated costs; see, for example, [12]. Although many long-term national surveys have been conducted worldwide in order to determine the impact and magnitude of health problems in terms of multimorbidity, as well as the role of health programs and health care providers [13], there is a fundamental lack of knowledge about how to appropriately measure multimorbidity within one person and quantify the heterogeneity in multimorbidity patterns among individuals.

According to a recent review [14], most existing measures of multimorbidity for individuals use the sum of the number of diagnosed diseases without any weighting [10, 15] or numerical indices, such as the Charlson Comorbidity Index [16]. These numerical indices do not account for multimorbidity by chance [17] and often require clinical judgment for gathering information on each medical condition item, thus imposing concerns about interrater reliability and the need to detail the multiple-item questionnaire before the study starts. Moreover, these indices were originally developed and validated for specific diseases, such as the Kaplan Index for diabetes [18], or for specific outcomes such as the Charlson index for prediction of mortality (although the Charlson index has been adapted for use with other outcomes including disability and length of stay [19]). As large-scale population health surveys are now commonly used worldwide for studying multimorbidity of medical conditions of a general population, the capability of these numerical indices to be adapted for use with national survey data is highly questionable.

In this paper, I seek to address the aforementioned challenges by developing a new theoretical framework, using a two-way clustering approach to achieve a dual purpose of identifying groups of comorbid conditions after adjusting for multimorbidity by chance and clustering individuals on the basis of their multimorbidity patterns. The impact of multimorbidity on health-related outcomes is then investigated by studying the disparities in the identified multimorbidity clusters and their association with the health-related outcomes. The major advantages of the proposed method are: (a) it uses only the basic binary (present or not present) indicator variable for each condition; (b) it does not rely on clinical judgments on conditions, and thus it is particularly suitable for national surveys where it is implausible to assess each participant; and (c) it provides additional information on the heterogeneity of multimorbidity among individuals.

The proposed method is based on a two-way clustering approach, where firstly groups of comorbid health conditions are formed using a “clumping” clustering algorithm [7], and then a mixture model-based approach is adopted to cluster individuals according to the groups of comorbid health conditions formed (see Figure 1). For the first clustering task, the intent is to identify groups of comorbid conditions that exhibit significant pairwise multimorbidity in addition to random “coincidental multimorbidity” [6], where the latter is the expected proportion of co-occurrence of conditions when the conditions are completely independent of one another [7]. As the multimorbidity measure used in the first clustering task accounts for coincidental multimorbidity, the groups of comorbid conditions identified will not depend on prevalence rates in the population under study. The second clustering task adopts a mixture model of multivariate generalized Bernoulli distributions to cluster the respondents into different groups that correspond to different patterns of comorbid health conditions. Here a generalized Bernoulli distribution corresponds to a multinomial distribution consisting of one draw on two or more categories. Mixture models with different types of Bernoulli components have been considered in other fields. For example, univariate multinomial mixture models have been widely used in the context of text clustering [20, 21]. A multivariate multinomial mixture has been adopted to obtain a consensus clustering of ensembles [22], and mixtures of three-variate Bernoulli distributions have been used to analyze random-graph networks [23]. However, this paper is the first to use mixture models of multivariate generalized Bernoulli distributions in the development of a new two-way clustering framework for studying the heterogeneity in multimorbidity patterns among individuals.

The primary product of the proposed two-way clustering method is the membership of individuals in clusters of

different multimorbidity patterns. This clustering result can be used directly to study the impact of morbidity on health-related outcomes, or to improve prediction of outcomes by adjusting for confounding of multimorbidity patterns. In some applications, it is also valuable to have an index value of multimorbidity assigned for each individual [14]. I therefore present how the proposed clustering method can be used to calculate a multimorbidity score for each individual.

The rest of the paper is organized as follows. In Section 2, I provide the theoretical concepts of the two-way clustering framework and show how the two clustering tasks are achieved using two different clustering methods. A clustering-based approach is also established to formulate a multimorbidity score for each individual. In Section 3, a simulation study is presented to assess the performance of the mixture model-based method for clustering multimorbidity groups. Section 4 presents the application of the proposed two-way clustering method to a national survey data set. Section 5 provides concluding remarks and discusses the use of the proposed method to perform a discriminant analysis.

2. Two-way Clustering Framework

Let n and p denote the numbers of respondents (subjects) and health conditions, respectively. The data can thus be represented by an $n \times p$ matrix, where the observed value for the ith column and the jth row of the data matrix is one or zero, indicating the presence or absence of the ith health condition for the jth respondent $(i = 1, \ldots, p;\ j = 1, \ldots, n)$. With the proposed two-way clustering approach, the first stage involves clustering the p conditions into non-overlapping groups of comorbid conditions (Figure 1(a)). On the basis of individual patterns in these groups of comorbid conditions, the second stage is to cluster the n respondents into clusters that correspond to different patterns of comorbid health conditions (Figure 1(b)). The source code in R for performing the proposed two-way clustering framework is available on request from the corresponding author.

2.1. *Clustering of health conditions*

I now extend the three-step clustering method of *Ng et al.* [7] to cluster the p health conditions into non-overlapping groups of health conditions, where all health conditions belonging to a group significantly coexist with one another. The first step adopts the asymmetric Somers' D statistic to quantify the degree of non-random multimorbidity between any two health conditions,

$$\text{Somers' } D = \frac{A - B}{\min(W_r, W_c)}, \tag{1}$$

where A and B are, respectively, the numbers of concordant and discordant pairs among all possible pairs of respondents $(n(n-1)/2)$. In (1), $W_r = A + B + T_r$ and $W_c = A + B + T_c$, where T_r and T_c are the numbers of tied pairs on each of the two health conditions alone, respectively [7]. With p health conditions, there are a total of $n_p = p(p-1)/2$ pairwise comorbidity measures in terms of the asymmetric Somers' D statistic (1) that can be represented by an $p \times p$ symmetric matrix M. The second step assesses the significance of these Somers' D statistics using the Benjamini-Hochberg procedure [24] to control the false discovery rate (FDR) at level α, which means that the expected proportion of false positives among the identified significant Somers' D statistics is at most α. In the third step, a clumping clustering technique is used to search in the matrix M for identifying groups of comorbid health conditions, in which all members of a group significantly coexist with one another [7]. This three-step method obtains overlapping clusters of health conditions, where the strength of multimorbidity among health conditions in a cluster is given by the averaged pairwise Somers' D statistics, as

$$\text{strength (cluster)} = \frac{\sum_{k=1}^{n_p} I(k \in \text{ cluster})\, SD_k}{\sum_{k=1}^{n_p} I(k \in \text{ cluster})}, \tag{2}$$

where $I(k \in \text{ cluster})$ is an indicator function that equals one if the kth pair of health conditions belong to the cluster $(k = 1, \ldots, n_p)$ and SD_k is the asymmetric Somers' D statistic (1) for the kth pair of health conditions. Overlapping

Statistics
in Medicine

S. K. Ng

clusters of health conditions identified above are easy to interpret. However, for the subsequent clustering of respondents, I propose the following procedure to obtain unique non-overlapping groups of comorbid conditions:

(a) 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;
(b) Repeat (a) for the next cluster and name it as a group if it is not "singular" (singular cluster is defined as a cluster consists of a single health condition);
(c) If a group is formed in (b), remove its member health conditions in all subsequent clusters with smaller strength or singular clusters;
(d) Repeat (b) and (c) until all clusters are visited;
(e) Put the condition in a singular cluster into a pre-defined group where more than half of the member conditions are significantly comorbid with the condition;
(f) Name those remaining singular clusters as a singular group.

To illustrate its applicability, the above procedure was applied to the Australian National Survey of Mental Health and Wellbeing (SMHWB) 2007 data [25] considered in *Ng et al.* [7], where twenty-two overlapping clusters and a singular cluster of mental and physical health conditions were identified. The conversion of these twenty-three clusters into nine non-overlapping groups of health conditions is presented in Table 1. Some of these groups show relevant similarities to those detected in the recently published systematic review on multimorbidity patterns [26]; for example, Group 1 presents as the "cardiovascular and metabolic diseases" and Group 3 as the "mental health disorders".

2.2. *Clustering of respondents*

Let q denote the number of non-overlapping groups of comorbid health conditions obtained in the first clustering task. To cluster the n respondents, I now convert the original $n \times p$ data matrix into an $n \times q$ matrix that presents individual patterns of multimorbidity given by the q groups of comorbid conditions. Precisely, let $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{qj})^T$ contain the q categorical indicator variables for the jth respondent $(j = 1, \ldots, n)$, where y_{ij} is the indicator variable for the ith group, taking on d_i distinct labels $(i = 1, \ldots, q)$, as

$$y_{ij} = \begin{cases} 1 & \text{if absence of conditions in the } i\text{th group} \\ 2 & \text{if presence of one condition in the } i\text{th group} \\ 3 & \text{if presence of more than one condition in the } i\text{th group} \end{cases} \tag{3}$$

and the superscript T denotes vector transpose. Depending on individual patterns of multimorbidity in the ith group of conditions, the number of distinct labels d_i equals to either two or three, as defined in (3). The clustering of $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$ is implemented using a finite mixture model, which is a highly popular approach for a wide range of applications in the clustering of multivariate data [27, 28]. With this approach, the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are assumed to have come from a mixture of a finite number, say g, of components in some unknown proportions $\pi_1, \ldots, \pi_g$ that sum to one. The mixture density of $\boldsymbol{y}_j$ can then be written as

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h f_h(\boldsymbol{y}_j; \boldsymbol{\theta}_h) \qquad (j = 1, \ldots, n), \tag{4}$$

where $\boldsymbol{\Psi}$ is the vector of unknown parameters containing the mixing proportions $\pi_1, \ldots, \pi_{g-1}$, and the vectors of component parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$. Each component-density $f_h(\boldsymbol{y}_j; \boldsymbol{\theta}_h)$ is given by a multivariate generalized Bernoulli distribution consisting of one draw on d_i labels with probabilities $\theta_{i1h}, \ldots, \theta_{id_ih}$ for each group $i = 1, \ldots, q$, and where $\theta_{id_ih} = 1 - \sum_{l=1}^{d_i-1} \theta_{ilh}$. Assuming that the categorical variables $y_{1j}, \ldots, y_{qj}$ are independent of each other, the mixture

model (4) is given by

$$f(\boldsymbol{y}_j;\boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h f_h(\boldsymbol{y}_j;\boldsymbol{\theta}_h) = \sum_{h=1}^{g} \pi_h \prod_{i=1}^{q} \prod_{l=1}^{d_i} \theta_{ilh}^{I(y_{ij},l)} \qquad (j = 1,\ldots,n), \tag{5}$$

where $I(y_{ij}, l)$ is an indicator function which is equal to one if $y_{ij} = l$ and is zero otherwise $(l = 1,\ldots,d_i)$. Although the independence assumption of categorical variables in (5) may not be true, it often performs well in practice for handling multivariate categorical variables [22, 29]. This is because it usually requires fewer parameters to be estimated than more complicated alternative methods that try to model interactions between the categorical variables.

The identifiability of mixture models (that is, the parametric model (5) is unique up to a permutation of the component labels) with generalized Bernoulli distributions has been discussed [30]. While it is well known that mixtures of univariate Bernoulli distributions are not identifiable [23], mixtures of multivariate generalized Bernoulli distributions are generically identifiable under the condition

$$q \geq 2\log_d g + 1; \tag{6}$$

see [30]. That is, the identifiability of mixture models (5) is ensured when the number of variates q is sufficiently large. For example, with the number of distinct labels $d = 3$ (see (3)) and the number of components $g = 5$, mixture models (5) are identifiable when the number of variates q is greater than 3. A greater number of components g or less distinct labels d will need a larger value of q for identifiability. For model estimation, it is also necessary for the dimension of the parameter space for $\boldsymbol{\Psi}$ being not larger than that of the observed sample space. That is,

$$g\sum_{i=1}^{q}(d_i - 1) + g - 1 \leq S - 1, \tag{7}$$

where S is the number of distinct observed patterns of $\boldsymbol{y}_j$, which has an upper bound of $\prod_{i=1}^{q} d_i$ corresponding to the dimension of the distribution space. In practice, the category $y_{ij} = 3$ corresponding to the presence of more than one condition may not have sufficient observed counts for some health condition groups $i \in \{1,\ldots,q\}$ to achieve consistent estimation of θ_{i3h}. To this end, if the observed counts of $y_{ij} = 3$ are fewer than 0.5% of n respondents, then this category will be relabeled as $y_{ij} = 2$, forming a combined category with $y_{ij} = 2$.

The unknown parameter vector $\boldsymbol{\Psi}$ can be estimated by the maximum likelihood (ML) method via the expectation-maximization (EM) algorithm [31, 32]. The E- and M-steps of the EM algorithm for the mixture model (5) are in closed form and are provided in Web Appendix A. The EM algorithm is implemented from a variety of initial values for the parameter vector $\boldsymbol{\Psi}$ in an attempt to locate all local maxima of the likelihood function [32]. With model (5), initial values of $\pi_1,\ldots,\pi_{g-1}$ and $\theta_{i1h},\ldots,\theta_{i(d_i-1)h}$ for $h = 1,\ldots,g$ and $i = 1,\ldots,q$ are generated randomly and independently from a uniform distribution $U(0,1)$, while keeping the constraint that $\sum_{h=1}^{g}\pi_h = 1$ and $\sum_{l=1}^{d_i}\theta_{ilh} = 1$ for $i = 1,\ldots,q$ with all π_h and θ_{ilh} being positive. In this study, the number of components in the mixture model is assessed using the Bayesian information criterion (BIC); see, for example, [27, 28]. The standard errors of the estimates of $\boldsymbol{\Psi}$ are obtained by the bootstrap resampling method with replacement, where the number of bootstrap replications is taken to be 100 [27].

Let $\hat{\boldsymbol{\Psi}}$ denote the ML estimate of $\boldsymbol{\Psi}$ so obtained. The mixture approach provides a probabilistic clustering of n respondents in terms of the estimated posterior probabilities of component membership:

$$\tau_h(\boldsymbol{y}_j;\hat{\boldsymbol{\Psi}}) = \frac{\hat{\pi}_h f_h(\boldsymbol{y}_j;\hat{\boldsymbol{\theta}}_h)}{\sum_{l=1}^{g}\hat{\pi}_l f_l(\boldsymbol{y}_j;\hat{\boldsymbol{\theta}}_l)} \qquad (h = 1,\ldots,g;\ j = 1,\ldots,n). \tag{8}$$

An outright clustering of n respondents into g nonoverlapping clusters is obtained by assigning each respondent to the component to which it has the highest estimated posterior probability of belonging [27].

Statistics
in Medicine

S. K. Ng

2.3. Formulation of individual multimorbidity scores

Clustering of respondents based on (8) can be used directly to study the impact of multimorbidity on health-related outcomes of individuals. Alternatively, a multimorbidity score can be formulated based on the clustering result. One such scoring method is to add the cluster-specific products of the posterior probability of membership (8) and a measure of multimorbidity accounting for the groups of comorbid health conditions. That is,

$$I_j = \sum_{h=1}^{g} \tau_h(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) m_h \qquad (j = 1, \ldots, n), \tag{9}$$

where the cluster-specific multimorbidity measure for the hth cluster m_h is given by

$$m_h = \sum_{i=1}^{q} s_i \sum_{l=1}^{d_i} \hat{\theta}_{ilh} w_l \qquad (h = 1, \ldots, g), \tag{10}$$

where s_i is a predefined severity index for the ith comorbid condition group and w_l is a weighting coefficient for the lth estimated probability $\hat{\theta}_{ilh}$ corresponding to the level of multimorbidity represented by the d_i labels $(i = 1, \ldots, q;\ l = 1, \ldots, d_i)$. In this paper, I assume all comorbid condition groups are of equal severity ($s_i = 1\ \forall i$) and $w_l = l - 1$ such that a larger weighting coefficient is adopted to reflect a higher level of multimorbidity corresponding to the presence of more than one condition in a comorbid condition group; see (3). That is, on the basis of (9) and (10), I propose a clustering-based multimorbidity scoring procedure as

$$I_j = \sum_{h=1}^{g} \tau_h(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}) \sum_{i=1}^{q} \sum_{l=1}^{d_i} \hat{\theta}_{ilh}(l-1) \qquad (j = 1, \ldots, n). \tag{11}$$

In some applications (such as the study of comorbidity corresponding to an index condition of interest), it could be appropriate to adopt a larger value of s_i for the ith comorbid condition group consisting of the index condition.

3. Simulation Experiments

In this section, I investigate the performance of mixture models of multivariate generalized Bernoulli distributions (5) for clustering multivariate categorical data within the setting of a multimorbidity study. Let us consider simulated data involving $q = 6$ groups of comorbid health conditions from n respondents, with $g = 3$ components corresponding to distinct patterns of multimorbidity among the respondents. The numbers of distinct labels are assumed to be $(3, 3, 2, 2, 2, 2)$ for the $q = 6$ groups, respectively, and the values of θ_{ilh} $(i = 1, \ldots, q;\ l = 1, \ldots, d_i;\ h = 1, \ldots, g)$ are fixed in all the settings as given in Table 2. These values imply that the first component consists of respondents with a low level of multimorbidity in all six health condition groups, while the second component consists of respondents with a high level of multimorbidity in health condition groups 1 and 3. The third component consists of respondents with a high level of multimorbidity in health condition groups 2, 4, and 5.

With reference to (6) and (7), the above setting ensures that the corresponding mixture model with multivariate generalized Bernoulli components is identifiable. In the simulation, individual component membership is generated independently in which each respondent has a probability of π_h to be from the hth component $(h = 1, \ldots, g)$ on the basis of random numbers simulated from a uniform distribution $U(0, 1)$. For those respondents belonging to the hth component, q-variate categorical random variables are then generated independently from the generalized Bernoulli distribution with probabilities $\theta_{i1h}, \ldots, \theta_{id_ih}$ for $i = 1, \ldots, q$ separately $(h = 1, \ldots, g)$. Let us consider four different sets of parameter values of π_h and n. In Set 1, they are ($\pi_1 = 0.60$, $\pi_2 = 0.30$, $\pi_3 = 0.10$) and $n = 1000$. In Sets 2 and 3, I study the effect

of changing π_h to ($\pi_1 = 0.70$, $\pi_2 = 0.15$, $\pi_3 = 0.15$) and ($\pi_1 = 0.40$, $\pi_2 = 0.30$, $\pi_3 = 0.30$), respectively. In Set 4, I study the effect of changing the sample size to $n = 500$. There are 500 replications considered in each setting.

The aim of the simulation experiments is to assess the applicability of model (5) by evaluating the bias and the variability of the estimators. Simulation results are provided in Table 2, where SE_1 and SE_2 denote the average of the standard error of the estimates and the sample standard error of the estimates, respectively, over the 500 replications. From Table 2, no appreciable bias is observed in any of the simulation settings, confirming the applicability of the mixture model of multivariate generalized Bernoulli distributions for clustering multivariate categorical data in small sample situations. A comparison of SE_1 and SE_2 provides information on whether the estimated standard errors obtained using the nonparametric bootstrap method are overestimated or underestimated. In general, good agreement between SE_1 and SE_2 is observed for mixing proportions π_h. However for component-density with a large proportion π_h (such as the first component for Sets 1, 2, and 4), the estimated standard errors of θ_{ilh} appear to be slightly overestimated. Thus, caution should be exercised in interpreting the significance levels attached to these estimates when the significance of probabilities θ_{ilh} is relevant.

4. Example: Australian National Survey of Mental Health and Wellbeing 2007 Data

The 2007 SMHWB was conducted by the Australian Bureau of Statistics (ABS) from August to December 2007 [25], collecting information on the prevalence of three mental health disorders and 21 physical health conditions from 8841 Australians aged 16–85 years. Assessment of the three mental health disorders (anxiety disorder such as social phobia; affective disorder such as depression; substance-use disorder such as alcohol harmful use) was based on the World Health Organization International Classification of Diseases, Tenth Revision [33]. The survey focused on the prevalence of 12-month mental health disorders for persons with a life-time mental health disorder who experienced symptoms in the 12 months period prior to the survey interview; see [34, 35] for key findings and descriptive comorbidity of mental health disorders. The survey data in two Confidentialized Unit Record Files (CURFs) can be obtained via the Remote Access Data Laboratory (RADL) of the ABS website [36].

The clustering method described in Section 2.1 was applied to the SMHWB data in order to cluster the 24 mental and physical health conditions into non-overlapping groups. With the FDR controlled at $\alpha = 0.01$ level, it was found that 77 out of $n_p = 276$ pairs of conditions are significant and the expected number of false positive among these 77 pairs is smaller than one. Nine non-overlapping groups of health conditions were found, as presented in Table 1, using the procedures described in Section 2.1. Medical-related commentary is not provided on the structure of the non-overlapping groups as it is out of scope of this paper. With this data set, only comorbid groups 1, 3, 4, and 6 have a sufficient number of observed counts corresponding to the presence of more than one condition ($y_{ij} = 3$); see Section 2.2. That is, $d_i = 3$ for $i = 1, 3, 4, 6$ and $d_i = 2$ for $i = 2, 5, 7, 8, 9$. With reference to (6), mixture models of multivariate generalized Bernoulli distributions with 2 components or more are all identifiable.

Clustering of respondents was then performed using the mixture of multivariate generalized Bernoulli distributions (5) described in Section 2.2. Model selection using the BIC indicates that there are four clusters. The clustering results for $g = 4$ is provided in Table 3. The probabilities of the presence of one condition or more than one condition in the nine comorbid groups are presented in Figure 2. It can be seen that the largest group is Cluster 1 (51.3%) that consists of respondents with low levels of comorbidity in all health condition comorbid groups. Based on (11) and the estimates of $\boldsymbol{\theta}_h$ in Table 3, the cluster-specific multimorbidity measure for Cluster 1 is $m_1 = 0.842$. The second largest group is Cluster 2 (22.4%) which consists of respondents with high levels of comorbidity in condition groups G1 and G4, with $m_2 = 2.871$. The next is Cluster 3 (20.3%), consisting of respondents with high levels of comorbidity in groups G3, G4, and G6; where $m_3 = 2.443$. The smallest group is Cluster 4 (6.0%) which comprises of respondents with high levels of comorbidity in G3 and very high levels of comorbidity in groups G1, G4, and G6. The cluster-specific multimorbidity measure for Cluster

Statistics
in Medicine

S. K. Ng

4 is $m_4 = 5.668$.

Demographic characteristics and health-related outcomes of respondents across the four clusters are presented in Table 4, along with the cluster-specific multimorbidity measures described above. All the factors in Table 4 were significantly different among the four clusters, partly due to the relatively large sample size. From Table 4, it can be seen that both Clusters 2 and 4, which have the two highest cluster-specific multimorbidity measures, consist of more older people (mean age is 62.8 and 59.7, respectively) compared to the other two clusters. Because of this, Clusters 2 and 4 also have significantly higher proportions of people not in the labor force, as well as those being divorced or widowed (data not shown). In contrast, both Clusters 3 and 4, which have the two highest level of multimorbidity in condition group G3 (see Figure 1), contain more people with medium to high levels of psychological distress. To summarize, compared to Cluster 1, individuals in Cluster 2 are older and are more likely to be over-weight or obese, and are less physically active. Individuals in Cluster 3 are more likely to be female, Australian born, current smokers, and have medium to high levels of psychological distress. Finally, Cluster 4 contains more people who are female, older, are more likely to be obese with medium to high levels of psychological distress. Individuals in Cluster 4 are also more likely to be a current or ex-smoker, and not be physically active.

From Table 4, it can be seen that health care utilization of respondents is different across the four clusters. Both Clusters 2 and 4 have a significantly higher frequency of hospital admissions, longer length of stay, and higher frequency of GP consultations. To demonstrate the explanatory power of the multimorbidity pattern clusters and individual multimorbidity scores, regression models were adopted to relate the number of nights in hospital and the frequency of GP consultations, respectively, with the two multimorbidity measures separately, after adjusting for the demographic factors described above. With the SMHWB data, zero nights in hospital constituted in excess of 88 per cent of the observations. Both zero-inflated negative binomial (ZINB) and zero-inflated Poisson (ZIP) regression models are commonly used to model count outcome variables with "excess" zeros [37]. The ZINB procedure in Stata (Stata IC 13.1; StataCorp, College Station, TX) was used in this study because the ZINB model is more appropriate when the count variables are overdispersed. The results are presented in Table 5. From Table 5(a) (the logit model indicating the association between zero outcomes and covariates), it was identified that females and respondents who are not born in Australia were more likely to spend zero nights in hospital in the past year due to physical health problems, whereas older respondents with higher levels of psychological distress who no longer smoke were less likely to spend zero nights in hospital. From Table 5(b) (the negative binomial count model), older and under-weight respondents were more likely to stay longer in hospital, whereas over-weight respondents had fewer nights in hospital compared to normal-weight respondents. Comparing to respondents in Cluster 1, respondents in Clusters 2 and 4 (higher levels of multimorbidity) had a significantly decreased probability of spending zero nights in hospital. Respondents in Cluster 4 were more likely to stay longer in hospital. Similarly, respondents with a higher multimorbidity score had a smaller probability of spending zero nights in hospital and tended to have a longer stay if they were admitted to hospital.

For analyzing the frequency of GP consultations, four ordinal categories of frequencies (0, 1-5, 6-11, 12+) were considered. The OLOGIT procedure in Stata was adopted to estimate an ordered logistic regression model, relating the ordinal outcomes with covariates and multimorbidity clusters. The results are presented in Table 6. It was found that current smokers (relative to non-smokers) and respondents who are not born in Australia had a lower frequency of GP consultations in the past year due to physical or mental health problems. Females and older respondents with higher levels of psychological distress, who are obese and no longer smoke had an increased frequency of GP consultations. Comparing to respondents in Cluster 1, respondents in Clusters 2, 3, and 4 (higher levels of multimorbidity) tended to have more GP consultations in the past year. Similarly, respondents with a higher multimorbidity score had significantly more GP consultations in the past year.

Two additional regression analyses on the SMHWB data were conducted to compare the proposed two-way clustering approach with two existing methods. The first method uses the total number of conditions the respondent suffers as a measure of multimorbidity; see [14]. The second method adopts an indicator (binary) variable for each condition the

respondent suffers. This method assumes independence among conditions and is based on a single disease-based paradigm. The comparative results for the covariate effects on the number of nights in hospital and the frequency of GP consultations in the past year are provided in Web Appendix B. The two existing methods provided different results compared to those presented in Tables 5 and 6. Specifically, the effects of different levels of psychological distress on the number of nights in hospital and the frequency of GP consultations in the past year were both under-estimated with the two existing methods. Moreover, the capability of the proposed methods using multimorbidity clusters or scores to differentiate individuals with different multimorbidity patterns was compared with the existing method using the total number of conditions. It was found that the method using the total number of conditions cannot differentiate the predicted mean numbers of nights in hospital and the predicted probabilities for more than 12 GP consultations in the past year among respondents with different multimorbidity patterns (see Figure S1 in Web Appendix B).

5. Discussion

I have developed a new two-way clustering framework for the identification of groups of comorbid health conditions and the subsequent clustering of individuals based on their multimorbidity patterns. For health conditions that are completely independent of one another, it is expected they will coexist in a proportion that equals the product of prevalence rates of individual conditions. The use of the Somers' D statistic to adjust for this coincidental multimorbidity by chance with control for the FDR is therefore a critical first step towards the identification of clusters of most significant non-random comorbid conditions; see, for example, [26, 38]. The identification of non-random comorbid condition groups is particularly relevant because the majority of clinical practice and research to date has relied on a single disease-based paradigm. About 88% and 56% of clinical guidelines examined in Australia and the USA, respectively, fail to make specific recommendations for patients with multiple comorbid conditions [39].

Greater multimorbidity burden among individuals is always associated with poor health-related outcomes, such as quality of life, health service utilization, and mortality, among others [2, 3, 14]. It is thus important to identify the heterogeneity in multimorbidity patterns and determine the impact of multimorbidity on individual health outcomes. The findings will improve the evidence-base for research of interventions by taking account of confounding factors due to differences in case-mix, with an ultimate goal of reducing the adverse effects of multimorbidity on health and wellbeing of Australians. Aside from the clustering of individuals into different multimorbidity patterns, I have developed a clustering-based method to formulate an index value for measuring the degree of multimorbidity of individuals. Improved understanding of the relationship between multimorbidity and protective or risk factors will contribute to the development of tailored treatment plans for preventing onset or reducing severity of diseases.

Recently, latent profile analyzes have been adopted to characterize patterns of multimorbidity and reveal clusters of individuals [40]. However, this approach is only applicable when the number of conditions is small (for example in [40], $p = 5$ conditions were considered). When the number of conditions is large, problems in determinating the number of clusters g and hence in the quality of clustering results may arise. Particularly, high-dimensional problems are issues for studying multimorbidity with national survey data, where many health conditions are to be considered. To illustrate these issues, a latent profile analysis was conducted to cluster individuals based on the 24 mental and physical health conditions in the SMHWB data set, without any grouping of comorbid conditions. The results are provided in Web Appendix C, which show that determinating the value of g is not consistent when different criteria are used (due to the high-dimensional problem). In handling such high-dimensional data, this proposed two-way clustering approach is innovative in that the clustering of conditions provides not only a generic clustering-based technique for dimensional reduction in the subsequent clustering of individuals, but also a more meaningful interpretation on presenting the multimorbidity patterns for each identified clusters.

With the mixture model (5) presented in Section 2.2, it is assumed that the mixing proportions π_h $(h = 1, \ldots, g)$ do not

depend on covariates. Without covariates, the mixing proportions representing the prior probabilities of membership in each cluster are the same for all respondents. The implied clustering based on the posterior probabilities (8) thus represents a grouping of respondents entirely on the basis of their multimorbidity patterns. If the mixing proportions are formulated to relate π_h with covariates (such as respondent's demographic factors), then the prior probabilities of cluster membership are not the same for all respondents. The effects of covariates will be incorporated into the mixture model via π_h. The clusters formed will then represent an "adjusted" clustering of respondents that accounts for the adjustment of multimorbidity patterns with the respondents' characteristics. This study is an item of future research.

In this paper, I consider a cluster analysis (unsupervised classification) approach, where there is no *a priori* information regarding the underlying group structure of individuals [27]. In contrast, discriminant analysis (supervised classification) is concerned with situations where "training" data from individuals of known origin from two or more classes were available, and the aim is to construct a discriminant rule (classifier) for assigning a new unclassified individual to one of the classes [41, 42]. In the context of medical and health research, the group origin could represent the classification of prognosis (such as, metastasis within 5 years or not [43]) or outcomes of interest (such as, health service utilization). In this context, the identified heterogeneity in multimorbidity patterns among individuals, either in terms of the estimated posterior probabilities of component membership (8) or individual multimorbidity scores (11), can be included as a distinguishing variable in the feature vector of each individual in order to construct a prediction rule via the estimation of the class-conditional distributions using the training data. An optimal rule of allocation can be defined by assigning an unclassified individual to the class to which the individual has the highest posterior probability of belonging [42]. The work for discriminant analysis and the relative performance of a prediction rule that accounts for multimorbidity using the new clustering-based measures will be pursued in future research.

Acknowledgement

The author thanks an Associate Editor and two referees for their constructive comments that helped to improve this manuscript. The author is grateful to the Australian Bureau of Statistics for providing the national survey data set in CURF and to Dr Kui Wang for developing the R-coded program to implement the clustering procedures. The work was supported by a grant from the Population & Social Health Research Program, Griffith Health Institute, Griffith University, Australia.

References

1. van den Akker M, Buntinx F, Knottnerus JA. Comorbidity or multimorbidity: what's in a name? A review of literature. *European Journal of General Practice* 1996; **2**:65–70.
2. Britt HC, Harrison CM, Miller GC, Knox SA. Prevalence and patterns of multimorbidity in Australia. *Medical Journal of Australia* 2008; **189**:72–77.
3. Caughey GE, Vitry AI, Gilbert AL, Roughead EE. Prevalence of multimorbidity of chronic diseases in Australia. *BMC Public Health* 2008; **8**:221.
4. Glynn LG, Valderas JM, Healy P, Burke E, Newell J, Gillespie P, Murphy AW. The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Family Practice* 2011; **28**:516–523.
5. Feinstein AR. The pretherapeutic classification of comorbidity in chronic disease. *Journal of Chronic Disease* 1970; **23**:455–468.
6. Kraemer HC. Statistical issues in assessing comorbidity. *Statistics in Medicine* 1995; **14**:721–733.
7. Ng SK, Holden L, Sun J. Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. *Statistics in Medicine* 2012; **31**:3393–3405. DOI: 10.1002/sim.5426.
8. Batstra L, Bos EH, Neeleman J. Quantifying psychiatric comorbidity: Lessions from chronic disease epidemiology. *Social Psychiatry and Psychiatric Epidemiology* 2002; **37**:105–111. DOI: 10.1007/s001270200001.
9. Willcutt EG, Pennington BF, Olson RK, DeFries JC. Understanding multimorbidity: A twin study of reading disability and attention-deficit/hyperactivity disorder. *American Journal of Medical Genetics Part B* 2007; **144B**:709–714.

10. Yau KKW, Lee AH, Ng SK. Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics & Data Analysis* 2003; **41**:359–366.
11. Marengoni A, von Strauss E, Rizzuto D, Winblad B, Fratiglioni L. The impact of chronic multimorbidity and disability on functional decline and survival in elderly persons. A community-based, longitudinal study. *Journal of Internal Medicine* 2009; **265**:288–295. DOI: 10.1111/j.1365-2796.2008.02017.x.
12. Fortin M, Hudon C, Bayliss EA, van den Akker M. Multimorbidity's many challenges. Time to focus on the needs of this vulnerable and growing population. *British Medical Journal* 2007; **334**:1016–1017.
13. Violán C, Foguet-Boreu Q, Hermosilla-Pérez E, Valderas JM, Bolíbar B, Fàbregas-Escurriola M, Brugulat-Guiteras P, Muñoz-Pérez MÁ. Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity. *BMC Public Health* 2013; **13**:251. DOI: 10.1186/1471-2458-13-251.
14. de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity: a critical review of available methods. *Journal of Clinical Epidemiology* 2003; **56**:221–229. DOI: 10.1016/S0895-4356(02)00585-1.
15. Ng SK, McLachlan GJ, Lee AH. An incremental EM-based learning approach for on-line prediction of hospital resource utilization. *Artificial Intelligence in Medicine* 2006; **36**:257–267. DOI: 10.1016/j.artmed.2005.07.003.
16. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* 1987; **40**:373–383.
17. Holden L, Scuffham PA, Hilton MF, Muspratt A, Ng SK, Whiteford HA. Patterns of multimorbidity in working Australians. *Population Health Metrics* 2011; **9**:15.
18. Kaplan MH, Feinstein AR. The importance of classifying initial comorbidity in evaluating the outcome of diabetes mellitus. *Journal of Chronic Diseases* 1974; **27**:387–404.
19. Rochon PA, Katz JN, Morrow LA, McGlinchey-Berroth R, Ahlquist MM, Sarkarati M, Minaker KL. Comorbid illness is associated with survival and length of hospital stay in patients with chronic disability. A prospective comparison of three comorbidity indices. *Medical Care* 1996; **34**:1093–1101.
20. Li M, Zhang L. Multinomial mixture model with feature selection for text clustering. *Knowledge-Based Systems* 2008; **21**:704–708.
21. Rigouste L, Cappé O, Yvon F. Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing & Management* 2007; **43**:1260–1280.
22. Topchy A, Jain AK, Punch W. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005; **27**:1866–1881.
23. Ambroise C, Matias C. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society B* 2012; **74**:3–35.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 1995; **57**:289–300.
25. Australian Bureau of Statistics. *National Survey of Mental Health and Wellbeing: user's guide, 2007, Cat. No. 4327.0*. Australian Bureau of Statistics: Canberra, 2009. Available from: http://www.abs.gov.au.
26. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. *Journal of Clinical Epidemiology* 2014; **67**:254–266.
27. McLachlan GJ, Peel D. *Finite Mixture Models*. Wiley: New York, 2000; Chapters 1 and 6.
28. Ng SK, McLachlan GJ. Mixture models for clustering multilevel growth trajectories. *Computational Statistics & Data Analysis* 2014; **71**:43–51. DOI: 10.1016/j.csda.2012.12.007.
29. Hand DJ, Yu K. Idiot's Bayes – Not so stupid after all? *International Statistical Review* 2001; **69**:385–398.
30. Allman E, Matias C, Rhodes J. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* 2009; **37**:3099–3132.
31. Dempster AP, Laird NM, Rubin DB. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 1977; **39**:1–38.
32. Ng SK. Recent developments in expectation-maximization methods for analyzing complex data. *WIREs Computational Statistics* 2013; **5**:415–431. DOI: 10.1002/wics.1277.
33. World Health Organization. *International Classification of Diseases and Related Health Problems*, 10th Revision (ICD-10), 2007. Available from: http://www.who.int/.
34. Slade T, Johnston A, Oakley Browne MA, Andrews G, Whiteford H. 2007 National Survey of Mental Health and Wellbeing: methods and key findings. *Australian and New Zealand Journal of Psychiatry* 2009; **43**:594–605. DOI: 10.1080/00048670902970882.
35. Teesson M, Slade T, Mills K. Comorbidity in Australia: findings of the 2007 National Survey of Mental Health and Wellbeing. *Australian and New Zealand Journal of Psychiatry* 2009; **43**:606–614. DOI: 10.1080/00048670902970908.
36. Australian Bureau of Statistics. *National Survey of Mental Health and Wellbeing, Confidentialised Unit Record Files, 2007, Cat. No. 4329.0*. Australian Bureau of Statistics: Canberra, 2009. Available from: http://www.abs.gov.au.

37. Sheu ML, Hu TW, Keeler TE, Ong M, Sung HY. The effect of a major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model. *Health Economics* 2004; **13**:781–791. DOI: 10.1002/hec.849.
38. Wong A, Boshuizen HC, Schellevis FG, Kommer GJ, Polder JJ. Longitudinal administrative data can be used to examine multimorbidity, provided false discoveries are controlled for. *Journal of Clinical Epidemiology* 2011; **64**:1109–1117. DOI: 10.1016/j.jclinepi.2010.12.011.
39. Caughey GE, Roughead EE. Multimorbidity research challenges: where to go from here? *Journal of Comorbidity* 2011; **1**:8–10.
40. Tsai J, Rosenheck RA. Psychiatric multimorbidity among adults with schizophrenia: A latent class analysis. *Psychiatry Research* 2013; **210**:16–20. DOI: 10.1016/j.psychres.2013.05.013.
41. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Edition). Springer: New York, 2009.
42. McLachlan GJ. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley: New York, 1992.
43. Ng SK, McLachlan GJ, Wang K, Nagymanyoki Z, Liu S, Ng SW. Inference on differences between classes using cluster-specific contrasts of mixed effects. *Biostatistics* 2015; **16**:98–112. DOI: 10.1093/biostatistics/kxu028.

S. K. Ng

Statistics in Medicine

Table 1. Conversion of 23 clusters into 9 non-overlapping groups.

23 clusters in decreasing order of strength	Procedure to obtain non-overlapping groups (G1-G9)
stroke, heart disease, arthritis, diabetes	Form G1 (the conditions removed in other clusters)
heart disease, arthritis, diabetes, oedema	A singular cluster (oedema eventually forms G4)
heart disease, arthritis, hernias, tuberculosis	Form G2 (the conditions removed in other clusters)
cancer, heart disease, arthritis, tuberculosis	A singular cluster (cancer eventually put into G4)
anxiety disorder, affective disorder, substance use disorder	Form G3 (the conditions removed in other clusters)
heart disease, arthritis, thyroid goitre, tuberculosis	A singular cluster (thyroid goitre eventually as G7)
heart disease, arthritis, oedema, kidney problem, back/neck pain	Form G4 (the conditions removed in other clusters)
anxiety disorder, affective disorder, back/neck pain	All the conditions are removed
stroke, heart disease, arthritis, back/neck pain	All the conditions are removed
heart disease, arthritis, emphysema, hernias, back/neck pain	A singular cluster (emphysema eventually forms G5)
heart disease, arthritis, emphysema, stomach ulcer, back/neck pain	Form G5 (the conditions removed in other clusters)
cancer, heart disease, arthritis, oedema, back/neck pain	A singular cluster (cancer eventually put into G4)
heart disease, arthritis, emphysema, bronchitis, back/neck pain	A singular cluster (bronchitis eventually forms G6)
hay fever, sinusitis, bronchitis, migraine, back/neck pain	Form G6 (the conditions removed in other clusters)
hay fever, sinusitis, emphysema, bronchitis, back/neck pain	All the conditions are removed
hay fever, sinusitis, oedema, migraine, back/neck pain	All the conditions are removed
asthma, hay fever, sinusitis, emphysema, bronchitis	A singular cluster (asthma eventually put into G6)
heart disease, arthritis, thyroid goitre, back/neck pain	A singular cluster (thyroid goitre eventually as G7)
hay fever, sinusitis, anaemia, migraine, back/neck pain	A singular cluster (anaemia eventually put into G6)
arthritis, anaemia, back/neck pain	A singular cluster (anaemia eventually put into G6)
anxiety disorder, anaemia, migraine, back/neck pain	A singular cluster (anaemia eventually put into G6)
arthritis, epilepsy, back/neck pain	A singular cluster (epilepsy eventually as G8)
psoriasis	A singular cluster (psoriasis eventually as G9)
9 non-overlapping groups: G1 - stroke, heart disease, arthritis, diabetes G2 - hernias, tuberculosis G3 - anxiety disorder, affective disorder, substance use disorder G4 - oedema, kidney problem, back/neck pain, cancer G5 - emphysema, stomach ulcer G6 - hay fever, sinusitis, bronchitis, migraine, asthma, anaemia G7 - thyroid goitre G8 - epilepsy G9 - psoriasis	

Statistics
in Medicine

S. K. Ng

Table 2. Estimated biases and standard errors of the ML estimators for the mixture of multivariate generalized Bernoulli distributions (500 replications).

Parameter	True value	Average bias	SE_1	SE_2	True value	Average bias	SE_1	SE_2
	Set 1 (n=1000)				Set 2 (n=1000)			
π_1	0.60	-0.0021	0.0458	0.0567	0.70	0.0029	0.0463	0.0546
π_2	0.30	0.0020	0.0447	0.0593	0.15	-0.0027	0.0451	0.0555
θ_{111}	0.80	-0.0014	0.0600	0.0410	0.80	-0.0020	0.0687	0.0380
θ_{121}	0.15	0.0000	0.0520	0.0325	0.15	0.0012	0.0598	0.0300
θ_{211}	0.85	-0.0007	0.0428	0.0203	0.85	-0.0013	0.0483	0.0188
θ_{221}	0.10	0.0005	0.0402	0.0158	0.10	0.0005	0.0430	0.0147
θ_{311}	0.90	0.0006	0.0602	0.0337	0.90	-0.0009	0.0740	0.0278
θ_{411}	0.95	0.0003	0.0421	0.0128	0.95	-0.0004	0.0431	0.0118
θ_{511}	0.90	-0.0004	0.0412	0.0176	0.90	0.0000	0.0434	0.0158
θ_{611}	0.90	-0.0009	0.0340	0.0146	0.90	-0.0008	0.0350	0.0131
θ_{112}	0.20	0.0108	0.0628	0.0722	0.20	0.0151	0.0710	0.1202
θ_{122}	0.60	-0.0092	0.0556	0.0627	0.60	-0.0142	0.0626	0.1014
θ_{212}	0.60	0.0036	0.0465	0.0425	0.60	-0.0014	0.0520	0.0675
θ_{222}	0.20	-0.0030	0.0441	0.0308	0.20	0.0018	0.0459	0.0542
θ_{312}	0.30	0.0098	0.0649	0.0870	0.30	0.0248	0.0757	0.1528
θ_{412}	0.85	0.0016	0.0453	0.0308	0.85	-0.0005	0.0461	0.0504
θ_{512}	0.80	0.0026	0.0451	0.0348	0.80	0.0007	0.0457	0.0537
θ_{612}	0.90	0.0012	0.0381	0.0221	0.90	0.0035	0.0393	0.0408
θ_{113}	0.60	-0.0045	0.0636	0.0635	0.60	-0.0005	0.0729	0.0498
θ_{123}	0.30	0.0047	0.0556	0.0576	0.30	-0.0010	0.0637	0.0474
θ_{213}	0.20	0.0046	0.0447	0.0659	0.20	0.0018	0.0516	0.0494
θ_{223}	0.70	-0.0035	0.0427	0.0717	0.70	-0.0007	0.0463	0.0517
θ_{313}	0.80	-0.0050	0.0659	0.0548	0.80	-0.0016	0.0776	0.0438
θ_{413}	0.20	-0.0010	0.0430	0.0835	0.20	-0.0011	0.0443	0.0585
θ_{513}	0.20	0.0058	0.0435	0.0695	0.20	0.0028	0.0459	0.0498
θ_{613}	0.60	0.0003	0.0363	0.0595	0.60	0.0030	0.0387	0.0470
	Set 3 (n=1000)				Set 4 (n=500)			
π_1	0.40	-0.0062	0.0451	0.0465	0.60	-0.0008	0.0613	0.0720
π_2	0.30	-0.0203	0.0444	0.0325	0.30	-0.0003	0.0621	0.0753
θ_{111}	0.80	-0.0053	0.0586	0.0557	0.80	-0.0057	0.0870	0.0608
θ_{121}	0.15	0.0034	0.0497	0.0433	0.15	0.0054	0.0792	0.0495
θ_{211}	0.85	-0.0010	0.0381	0.0278	0.85	0.0002	0.0685	0.0295
θ_{221}	0.10	0.0004	0.0339	0.0220	0.10	-0.0008	0.0638	0.0231
θ_{311}	0.90	0.0005	0.0577	0.0455	0.90	-0.0047	0.0900	0.0467
θ_{411}	0.95	-0.0007	0.0336	0.0198	0.95	-0.0002	0.0636	0.0186
θ_{511}	0.90	-0.0007	0.0351	0.0239	0.90	-0.0005	0.0642	0.0247
θ_{611}	0.90	-0.0022	0.0294	0.0192	0.90	-0.0005	0.0552	0.0201
θ_{112}	0.20	0.0121	0.0592	0.0711	0.20	0.0122	0.0872	0.1016
θ_{122}	0.60	-0.0097	0.0504	0.0601	0.60	-0.0080	0.0795	0.0852
θ_{212}	0.60	-0.0012	0.0378	0.0420	0.60	-0.0003	0.0659	0.0614
θ_{222}	0.20	0.0001	0.0357	0.0316	0.20	0.0014	0.0613	0.0462
θ_{312}	0.30	0.0048	0.0593	0.0813	0.30	0.0209	0.0891	0.1301
θ_{412}	0.85	0.0008	0.0343	0.0329	0.85	-0.0013	0.0608	0.0460
θ_{512}	0.80	0.0014	0.0360	0.0347	0.80	0.0017	0.0632	0.0479
θ_{612}	0.90	0.0025	0.0305	0.0245	0.90	-0.0000	0.0532	0.0336
θ_{113}	0.60	-0.0013	0.0596	0.0339	0.60	-0.0028	0.0859	0.0969
θ_{123}	0.30	0.0006	0.0512	0.0315	0.30	0.0010	0.0768	0.0909
θ_{213}	0.20	-0.0007	0.0392	0.0318	0.20	0.0063	0.0654	0.0926
θ_{223}	0.70	0.0004	0.0355	0.0353	0.70	-0.0097	0.0627	0.0981
θ_{313}	0.80	-0.0008	0.0605	0.0284	0.80	0.0032	0.0866	0.0817
θ_{413}	0.20	-0.0033	0.0350	0.0372	0.20	0.0188	0.0633	0.1115
θ_{513}	0.20	0.0009	0.0360	0.0332	0.20	0.0141	0.0632	0.1032
θ_{613}	0.60	-0.0005	0.0304	0.0313	0.60	0.0038	0.0561	0.0938

Table 3. Results of fitting a 4-component mixture of multivariate generalized Bernoulli distributions to the SMHWB data.

Parameter	1st component	2nd component	3rd component	4th component
π_h	0.513 (0.045)	0.224 (0.027)	0.203 (0.033)	0.060
$\boldsymbol{\theta}_{11h}$	0.800 (0.045)	0.198 (0.051)	0.787 (0.065)	0.072 (0.032)
$\boldsymbol{\theta}_{12h}$	0.151 (0.026)	0.426 (0.029)	0.212 (0.048)	0.326 (0.044)
$\boldsymbol{\theta}_{13h}$	0.049	0.376	0.001	0.602
$\boldsymbol{\theta}_{21h}$	0.990 (0.003)	0.888 (0.023)	0.980 (0.009)	0.808 (0.033)
$\boldsymbol{\theta}_{22h}$	0.010	0.112	0.020	0.192
$\boldsymbol{\theta}_{31h}$	0.875 (0.020)	0.885 (0.034)	0.578 (0.032)	0.591 (0.055)
$\boldsymbol{\theta}_{32h}$	0.104 (0.014)	0.103 (0.027)	0.289 (0.023)	0.245 (0.029)
$\boldsymbol{\theta}_{33h}$	0.021	0.012	0.133	0.164
$\boldsymbol{\theta}_{41h}$	0.890 (0.016)	0.360 (0.095)	0.470 (0.064)	0.110 (0.038)
$\boldsymbol{\theta}_{42h}$	0.111 (0.015)	0.508 (0.064)	0.483 (0.053)	0.429 (0.053)
$\boldsymbol{\theta}_{43h}$	0.000	0.132	0.047	0.461
$\boldsymbol{\theta}_{51h}$	0.990 (0.003)	0.911 (0.022)	0.956 (0.012)	0.698 (0.055)
$\boldsymbol{\theta}_{52h}$	0.010	0.089	0.044	0.303
$\boldsymbol{\theta}_{61h}$	0.756 (0.022)	0.628 (0.056)	0.362 (0.038)	0.158 (0.081)
$\boldsymbol{\theta}_{62h}$	0.206 (0.013)	0.274 (0.025)	0.338 (0.022)	0.287 (0.038)
$\boldsymbol{\theta}_{63h}$	0.038	0.098	0.300	0.555
$\boldsymbol{\theta}_{71h}$	0.984 (0.004)	0.920 (0.022)	0.973 (0.014)	0.821 (0.034)
$\boldsymbol{\theta}_{72h}$	0.016	0.080	0.027	0.179
$\boldsymbol{\theta}_{81h}$	0.993 (0.002)	0.995 (0.004)	0.993 (0.004)	0.949 (0.014)
$\boldsymbol{\theta}_{82h}$	0.007	0.005	0.007	0.051
$\boldsymbol{\theta}_{91h}$	0.989 (0.003)	0.962 (0.011)	0.939 (0.012)	0.909 (0.019)
$\boldsymbol{\theta}_{92h}$	0.011	0.038	0.061	0.092

Standard error of parameter estimates given in brackets

Statistics
in Medicine

S. K. Ng

Table 4. Multimorbidity scores, demographic characteristics and health-related outcomes among the four clusters (SMHWB data).

Characteristic	Frequency (percentage) or Mean (standard deviation)[a] Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total sample
Cluster-specific multimorbidity	1.261 (0.29)	2.881 (0.50)	2.371 (0.37)	4.968 (0.52)	1.912 (1.01)
Gender					
Male	2609 (48.5%)	800 (46.3%)	480 (35.8%)	138 (34.9%)	4027 (45.5%)
Female	2769 (51.5%)	926 (53.7%)	862 (64.2%)	257 (65.1%)	4814 (54.5%)
Age	41.6 (18.1)	62.8 (13.9)	40.2 (15.0)	59.7 (14.6)	46.3 (19.0)
Country of birth					
Australia	3891 (72.4%)	1279 (74.1%)	1059 (78.9%)	301 (76.2%)	6530 (73.9%)
English speaking[b]	596 (11.1%)	226 (13.1%)	161 (12.0%)	49 (12.4%)	1032 (11.7%)
Other	891 (16.6%)	221 (12.8%)	122 (9.1%)	45 (11.4%)	1279 (14.5%)
Psychological distress					
Low or nil (10-15)	4202 (78.1%)	1277 (74.0%)	661 (49.3%)	141 (35.7%)	6281 (71.0%)
Medium (16-29)	1116 (20.8%)	421 (24.4%)	595 (44.3%)	199 (50.4%)	2331 (26.4%)
High (30-50)	58 (1.1%)	28 (1.6%)	86 (6.4%)	55 (13.9%)	227 (2.6%)
BMI category					
Under weight (< 18.5 kg/m^2)	174 (3.3%)	28 (1.7%)	40 (3.0%)	12 (3.1%)	254 (2.9%)
Normal weight (18.5-24.99)	2468 (46.9%)	518 (30.7%)	627 (47.5%)	96 (24.9%)	3709 (42.8%)
Over weight (24.5-29.99)	1774 (33.7%)	658 (38.9%)	403 (30.6%)	134 (34.8%)	2969 (34.3%)
Obese (≥ 30)	851 (16.2%)	486 (28.8%)	249 (18.9%)	143 (37.1%)	1729 (20.0%)
Smoking status					
Current smoker	1125 (21.0%)	266 (15.4%)	395 (29.4%)	97 (24.5%)	1883 (21.3%)
Ex-smoker	1302 (24.2%)	715 (41.4%)	359 (26.8%)	142 (35.9%)	2518 (28.5%)
Never smoked	2951 (54.9%)	745 (43.2%)	588 (43.8%)	156 (39.5%)	4440 (50.2%)
Level of exercise					
High	409 (7.6%)	51 (3.0%)	125 (9.3%)	13 (3.3%)	598 (6.8%)
Moderate	1137 (21.2%)	277 (16.0%)	286 (21.3%)	59 (14.9%)	1759 (19.9%)
Low or very low	3087 (57.5%)	1028 (59.6%)	757 (56.4%)	217 (55.0%)	5089 (57.6%)
No exercise	740 (13.8%)	369 (21.4%)	174 (13.0%)	106 (26.8%)	1389 (15.7%)
Hospital admission[c]					
0	4978 (92.6%)	1379 (79.9%)	1214 (90.5%)	291 (73.9%)	7862 (88.9%)
1-2	375 (7.0%)	307 (17.8%)	116 (8.6%)	75 (19.0%)	873 (9.9%)
3 or more	25 (0.5%)	39 (2.3%)	12 (0.9%)	28 (7.1%)	104 (1.2%)
No. of nights in hospital[c,d]	0.234 (1.06)	0.805 (2.01)	0.276 (1.06)	1.311 (2.65)	0.400 (1.43)
GP consultation[e]					
0	1192 (22.2%)	112 (6.5%)	176 (13.1%)	15 (3.8%)	1495 (16.9%)
1-5	3375 (62.8%)	888 (51.5%)	786 (58.7%)	139 (35.6%)	5188 (58.8%)
6-11	563 (10.5%)	424 (24.6%)	231 (17.2%)	104 (26.7%)	1322 (15.0%)
12 or more	244 (4.5%)	301 (17.4%)	147 (11.0%)	132 (33.8%)	824 (9.3%)

[a] Test for differences in frequencies among the four clusters using chi-square tests; Test for differences in means using ANOVA
[b] Main English speaking countries (Canada, Ireland, New Zealand, South Africa, UK, USA)
[c] Admission in the past year due to physical health problems
[d] Category "5 to 7 nights" as five and Category "8 or more nights" as eight
[e] GP consultation in the past year due to physical or mental health problems

Table 5. Zero-inflated negative binomial (ZINB) regression models on the number of nights in hospital in the past year due to physical health problems (SMHWB data).

Factor	Multimorbidity clusters Coefficient	p-value	Multimorbidity scores Coefficient	p-value
(a) Logit[a]				
Female	0.273	< 0.001	0.304	<0.001
Age	-0.015	< 0.001	-0.016	<0.001
Country of birth (English speaking)	0.238	0.039	0.244	0.034
Country of birth (Other)	0.369	0.001	0.346	0.003
Psychological distress (Medium)	-0.269	0.002	-0.184	0.033
Psychological distress (High)	-0.730	< 0.001	-0.489	0.015
Current smoker	-0.118	0.237	-0.094	0.347
Ex-smoker	-0.221	0.009	-0.211	0.013
Multimorbidity cluster 2	-0.722	< 0.001		
Multimorbidity cluster 3	-0.193	0.116		
Multimorbidity cluster 4	-0.934	< 0.001		
Intercept	2.555	< 0.001	2.914	<0.001
Multimorbidity score			-0.335	<0.001
(b) Negative binomial				
Age	0.012	< 0.001	0.012	<0.001
BMI (Under weight)	0.535	0.002	0.543	0.002
BMI (Over weight)	-0.186	0.014	-0.190	0.012
BMI (Obese)	0.004	0.965	-0.003	0.972
Multimorbidity cluster 2	0.149	0.061		
Multimorbidity cluster 3	-0.007	0.951		
Multimorbidity cluster 4	0.435	< 0.001		
Intercept	0.318	0.007	0.153	0.190
Multimorbidity score			0.105	<0.001
Log(alpha)[b]	-0.573	<0.001	-0.554	<0.001
Log-likelihood	-4821.144		-4815.875	
Likelihood ratio[c] – χ^2 $(d.f.)$	106.1 (3)		116.6 (1)	

[a] Vuong test of ZINB versus standard negative binomial gives a p-value < 0.001
[b] The dispersion parameter (alpha) is significantly different from zero
[c] Compared against the full model without multimorbidity cluster or multimorbidity score variables

Statistics
in Medicine

S. K. Ng

Table 6. Ordinal logistic regression models on the frequency of GP consultations (0, 1-5, 6-11, 12+) in the past year due to physical or mental health problems (SMHWB data).

	Multimorbidity clusters		Multimorbidity scores	
Factor	Adjusted OR	p-value	Adjusted OR	p-value
Female	1.596	< 0.001	1.566	<0.001
Age	1.018	< 0.001	1.017	<0.001
Country of birth (English speaking)	0.876	0.050	0.872	0.042
Country of birth (Other)	0.797	<0.001	0.816	0.001
Psychological distress (Medium)	1.518	<0.001	1.418	<0.001
Psychological distress (High)	3.319	< 0.001	2.666	<0.001
BMI (Under weight)	1.010	0.938	1.008	0.948
BMI (Over weight)	1.034	0.505	1.015	0.772
BMI (Obese)	1.147	0.020	1.097	0.119
Current smoker	0.857	0.007	0.839	0.002
Ex-smoker	1.126	0.021	1.120	0.028
Multimorbidity cluster 2	2.682	<0.001		
Multimorbidity cluster 3	1.755	<0.001		
Multimorbidity cluster 4	4.823	<0.001		
Multimorbidity score			1.719	<0.001
Log-likelihood	-8967.206		-8918.537	
Likelihood ratio[a] $-\chi^2$ $(d.f.)$	379.7 (3)		477.0 (1)	

[a] Compared against the full model without multimorbidity cluster or multimorbidity score variables

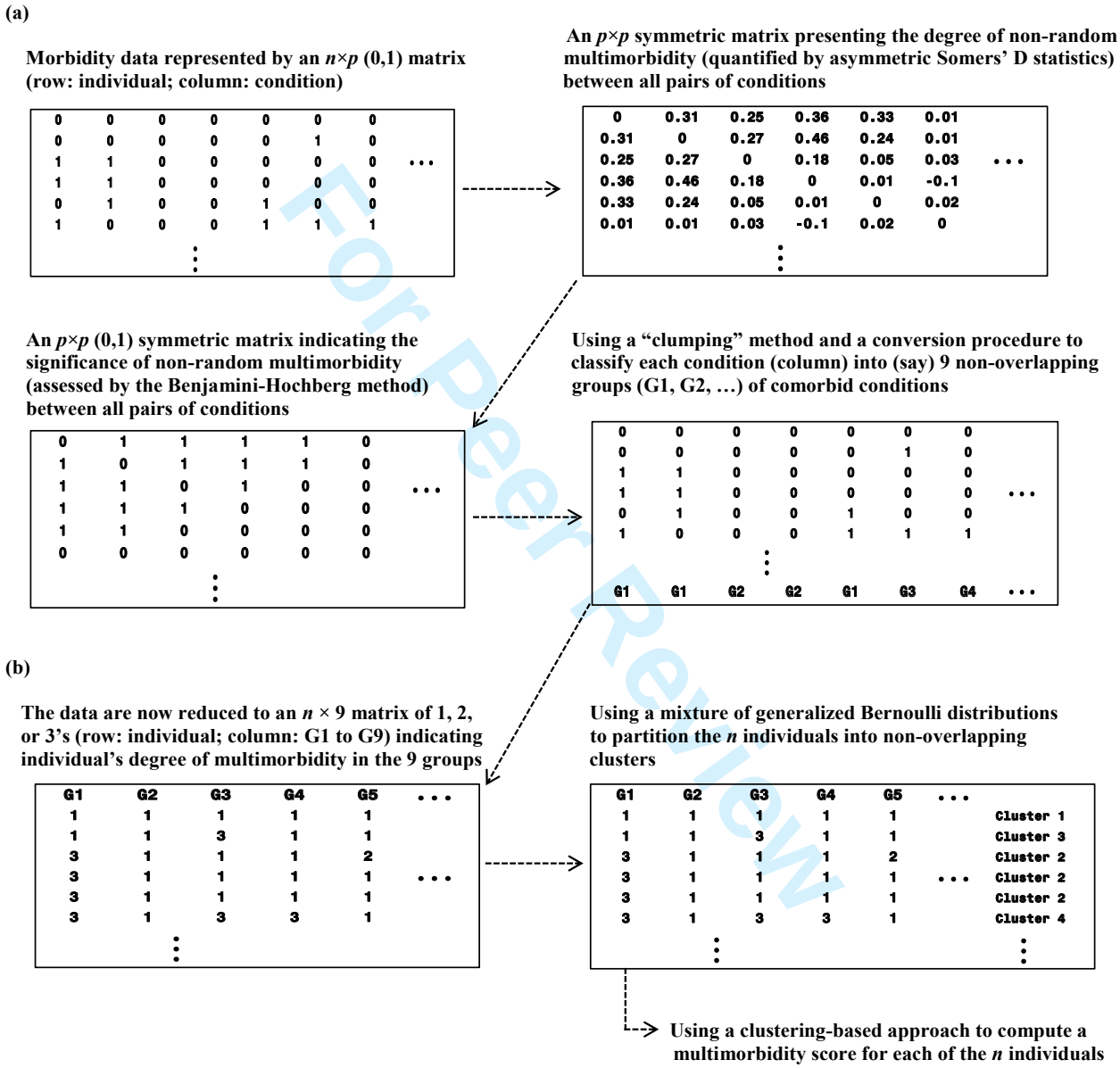

Figure 1. Rational of the proposed two-way clustering framework: (a) Clustering of conditions; (b) Clustering of individuals (Notations: n is the number of individuals; p is the number of conditions).

Statistics
in Medicine

S. K. Ng

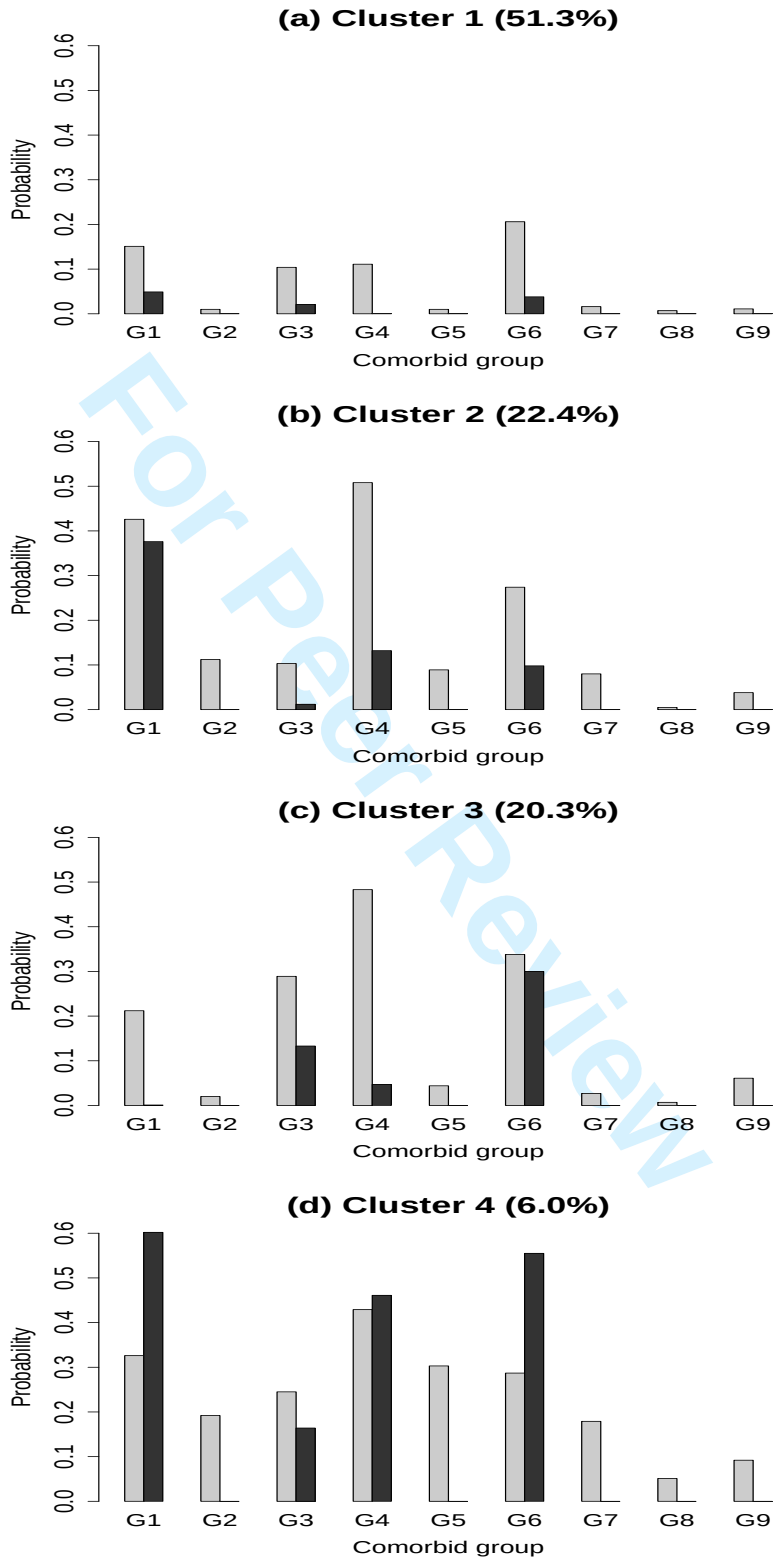

Figure 2. Probabilities of the presence of one condition (light gray bar) or more than one condition (dark gray bar) in the nine health condition comorbid groups (SMHWB data).