MULTIMORBIDITY CLUSTERS: CLUSTERING
BINARY DATA FROM MULTIMORBIDITY CLUSTERS:
CLUSTERING BINARY DATA FROM A LARGE
ADMINISTRATIVE MEDICAL DATABASE

John E. Cornell, PhD

Geriatric Research, Education, and Clinical Center
VERDICT Center of Excellence/South Texas Veterans Health Care System
University of Texas Health Science Center in San Antonio

Jacqueline A. Pugh, MD

VERDICT Center of Excellence/South Texas Veterans Health Care System
University of Texas Health Science Center in San Antonio

John W. Williams, Jr, MD, MHSc

Center for Health Services Research in Primary Care/Durham Veterans
Affairs Medical Center, Duke University Medical Center

Lewis Kazis, ScD

Austin F.S. Lee, PhD

Center for Health Quality, Outcomes, & Economic Research
Edith Nourse Rogers Memorial Veterans Hospital, Boston University

Michael L. Parchman, MD

John Zeber, Ph.D.

Thomas Pederson, MS

Kelly A. Montgomery, MPH

Polly Hitchcock Nokl, PhD

VERDICT Center of Excellence/South Texas Veterans Health Care System
University of Texas Health Science Center in San Antonio

ABSTRACT

Our purpose in this article is to describe and illustrate the application of cluster analysis to identify clinically relevant multimorbidity groups. Multimorbidity is the co-occurrence of 2 or more illnesses within a single person, which raises the question whether consistent, clinically useful multimorbidity groups exist among sets of chronic illnesses. Our purpose in this article is to describe and illustrate the application of cluster analysis to identify clinically relevant multimorbidity groups. Application of cluster analysis involves a sequence of critical methodological and analytic decisions that influence the quality and meaning of the clus-

APPLIED MULTIVARIATE RESEARCH

ters produced. We illustrate the application of cluster analysis to identify multimorbidity clusters in a set of 45 chronic illnesses in primary care patients ($N = 1,327,328$), with 2 or more chronic conditions, served by the Veterans Health Administration. Six clinically useful multimorbidity clusters were identified: a Metabolic Cluster, an Obesity Cluster, a Liver Cluster, a Neurovascular Cluster, a Stress Cluster and a Dual Diagnosis Cluster. Cluster analysis appears to be a useful technique for identifying multiple disease clusters and patterns of multimorbidity.

INTRODUCTION

Multimorbidity is defined as the presence or co-occurrence of two or more illnesses in a single individual (van den Akker, Buntinx, Metsemakers, Roos, & Knottnerus, 1998). A small but growing body of epidemiological research has focused specifically on multimorbidity, but typically only describes the average number of conditions, or the proportion of a population with a certain number of conditions (van den Akker, Buntinx, Roos, Knottnerus, 2001). A small number of studies on multimorbidity report prevalence rates ranging from 3.6% to over 50% in various populations (van den Akker et al., 1998). The reason for the widely varying rates is that most of the existing research has focused on a limited number of diseases in restricted populations such as the elderly.

Little is known about the *clustering* of diseases. Only rarely have studies reported the frequencies of specific combinations of diseases. For example, of studies reporting the frequency or co-occurrence of multiple chronic illnesses among outpatients or the general population, 2 have identified the frequency of specific, co-occurring disease pairs (Fried, Bandeen-Roche, Kasper, & Guralnik, 1999; Verbrugge, Lepkowski, & Imanaka, 1989). Although a third study provided two examples of the most common disease triads, the frequency of the triads in the study population ($N = 565$) was not stated (Grimby & Svanborg, 1997).

The paucity of research identifying the most frequently occurring clusters of diseases is perhaps not that surprising. Although it is possible to compare the observed and expected occurrence of specific combinations of diseases or the observed and expected patterns of all specific combinations possible, many calculations and large samples are needed (van den Akker et al., 2001). This is due in part to the astronomically high number of theoretically possible combinations for any set of diseases. For example, given p diseases, there are $[p \times (p-1)]/2$ possible 2-disease pairs; $[p(p-1)(p-2)]/[3(2)]$ possible disease triads; $[p(p-1)(p-2)(p-3)]/[4(3)(2)]$ possible disease quartets, and so on. For example, given 20 diseases, there would be 380 possible disease pairs, 1140 disease triads, 4845 disease quartets, and so on. What is needed is a technique to help identify potentially meaningful clusters of diseases “up front.”

The purpose of this paper is to describe our approach to the use of cluster analysis to identify clinically meaningful patterns of diagnoses in the hopes of identifying some order in the clinical “chaos” of complex patients with multimorbidity. We describe our approach to the search for multimorbid clusters within the decision-making framework described by Milligan (1996) in section 2. Section 3 describes the results of our cluster analysis, and we conclude with a general discussion of the application of clustering methods to the problem of discovering patterns among binary indicators in large datasets.

MULTIMORBIDITY CLUSTERS

2. Cluster Analysis and the Search for Multimorbidity Patterns

Cluster analysis is a statistical technique used to classify objects into coherent categories based on a set of measurements or indicator variables. A common use of cluster analysis in medicine is to categorize patients into subgroups or diagnostic categories based upon patterns of clinical signs and symptoms that empirically go together (Everitt, Landau, & Leese, 2001). Two-way clustering techniques are frequently used to organize genes into groups or clusters with similar levels of expression across relevant subgroups of patients, tissue samples or cell lines (Eisen, Spellman, Brown, & Botstein, 1988; Do, Broom, & Wen, 2003).

In practice, a cluster analysis is the end product of a series of analytical decisions. The analytic decisions made at each point in the series can significantly affect subsequent decisions, as well as the overall results of a cluster analysis (Everitt et al., 2001). This series of analytic decisions typically involve choices about what objects to cluster, what unit of measurement to use for the variables, what proximity measure to use as an index of similarity or dissimilarity among the objects, what type of clustering algorithm to use, and what criteria to use for determining the number and quality of clusters in the data.

2.1 Objects to Cluster and Unit of Measurement.

The classical approach to cluster analysis assumes that we wish to classify n objects into a set of k categories based on some index of proximity among the n objects. An index of proximity measures the closeness of one object to another computed from a set of p measurements. Closeness is defined in terms of multivariate distance for quantitative measures (e.g., Euclidean distance) or similarity coefficients (s_{ij}) for binary measures (e.g., Jaccard coefficient). Similarity coefficients range from 0.00 to 1.00 and are easily converted to a dissimilarity coefficient (d_{ij}) by taking $d_{ij} = 1 - s_{ij}$. In the classical set-up, indexes of proximity are computed across objects, yielding an $n \times n$ proximity matrix reflecting the degree of closeness among the objects. The dimensions of this $n \times n$ proximity matrix for groups of patients in large administrative databases can rapidly exceed the capacity of most systems to manipulate and analyze the data. However, we are interested in identifying multimorbidity groups based on the presence or co-occurrence of two or more chronic illnesses in a single individual. Our data are represented as a collection of binary objects $x_{ij} = \{1,0\}$ arranged in an $n \times p$ matrix with rows representing the n patients and columns representing the p indicator variables for each of the chronic illness categories. Our objective is to search for clinically meaningful clusters of chronic illness, rather than clusters of patients. This situation is similar to Q factor analysis where associations are computed across the measurements rather than across individuals (Gorsuch, 1983). Thus, we wish to produce groupings of chronic illness based on their relative similarity or dissimilarity. The problem simplifies to reducing the transposed $p \times n$ data matrix to a much smaller $p \times p$ proximity matrix among the chronic illness conditions, rather than a potentially large $n \times n$ proximity matrix.

2.2 Similarity and Proximity Measures for Binary Data.

Similarity measures (s_{ij}) for binary data are based on a 2×2 table (Figure 1.A.) where cell a records the number of positive matches and cell d records the

number of negative matches. A number of similarity coefficients for binary data have been proposed in the literature. The choice of coefficient depends on the relative weight given to positive matches (a) and negative matches (d) (Seber, 1984). The decision about how much weight to assign each type of match depends on the nature of question and the relative importance of positive and negative matches (Everitt et al., 2001). We are essentially interested in a measure of similarity based on the ratio of the number of cases contained in the intersection of the two diagnoses to the number of cases in the union of the two diagnoses. Negative matches are virtually non-informative in this case. We also have no a priori reason to weight positive matches more heavily than non-matches (cells b and c in Figure 1.A.). Thus, we chose a measure that assigns zero weight to negative matches, and equal weights to positive matches and non-match elements of the union of the two diagnoses. Jaccard's coefficient (Jaccard, 1908) satisfies these requirements (see Figure 1.C.). Thus, s_{ij} is the proportion of cases that have both condition i and condition j relative to the total number of patients who have either condition i or condition j or both. The proximity matrix consists of a set of dissimilarity coefficients easily obtained from the similarity coefficient by setting $d_{ij} = 1 - s_{ij}$.

2.3 Agglomerative Hierarchical Clustering

Clustering algorithms are broadly classified as hierarchical or nonhierarchical. Hierarchical algorithms are most appropriate for classification problems where objects are related via some underlying systematic structure (Everitt, et al., 2001; Seber, 1984). Chronic diseases can co-occur for several reasons (van den Akker, et al., 2001; Kraemer, 1995). Certain diseases share the same underlying genetic, environmental, or behavioral risk factors. For example, obesity is a risk factor for a variety of chronic illnesses, such as osteoarthritis, hypercholesterolemia, type 2 diabetes, hypertension, and sleep apnea. Individual chronic diseases may lead to multiple sequelae, as in the case of diabetes leading to complications such as neuropathy, retinopathy, nephropathy, and impotence. In addition, some chronic illnesses may simply randomly co-occur without any clinical, epidemiological, or yet defined relationship between them; this is likely to be uncommon, however (Kraemer, 1995). Thus, given our expectation that multimorbidity clusters arise from common underlying factors or as multiple sequelae of a primary condition, a hierarchical algorithm appears most appropriate for our purposes.

Hierarchical algorithms are further classified according to whether the algorithm proceeds by successively merging (agglomerative) or splitting (divisive) clusters. Agglomerative methods produce a series of partitions of data starting with the most elementary cluster composed of a single object and successively grouping objects and sets of objects together until the final grouping contains all the original objects. Divisive algorithms start with one large cluster and progressively subdivide the objects into a series of smaller clusters. With binary data, divisive algorithms use measures of association that include negative, as well as positive matches. Thus, we chose an agglomerative hierarchical method to search for multimorbidity clusters.

Algorithms for agglomerative hierarchical clustering include single linkage, (Sneath, 1957), complete linkage (Sorensen, 1948), average linkage (Sokal &

MULTIMORBIDITY CLUSTERS

Michener, 1958), median linkage (Gower, 1967), and Ward's method (Ward, 1963). Each algorithm, however, is a special case of the Lance and Williams (1967) recurrence formula for the distance between cluster k and cluster (ij) :

$$d_{k(y)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} - \gamma |d_{ki} + d_{kj}|.$$

One can reproduce the results for each agglomerative algorithm by substituting different values for the coefficients— α_i , α_j , β , and γ —into the Lance-Williams recurrence formula (see Table 4.2 in Everitt et al., 2001) for the Lance-Wallace parameter values for each agglomerative algorithms). Different choices for β provide different representations of the space defined by the proximity measures. Ideally, we want clusters that conserve the original space as much as possible.

Lance and Williams' flexible scheme provides a method to control the space-conserving and space-dilating characteristics of the solution. It imposes the following four conditions on the coefficients in the equation: $\alpha_i + \alpha_j + \beta = 1$, $\alpha_i = \alpha_j$, $\beta < 1$, and $\gamma = 0$. Under these conditions the Lance-Williams flexible recurrence formula simplifies to the following:

$$d_{k(ij)} = (d_{ki} + d_{kj}) \frac{1 - \beta}{2} \beta d_y.$$

Values of $\beta > 0$ contract the space and tend to induce “chaining” in the structure. Chaining occurs when objects are linked together by a string of intermediate, often noisy clusters, producing a long chain of clusters. Values of $\beta < 0$ dilate the space. A space-dilating system magnifies the dissimilarity among clusters. It is the opposite of chaining and often produces “nonconformist” clusters, consisting of rejects from other clusters (Seber, 1984). A desirable solution, therefore, is one that tends to conserve the space defined by the proximity measures. Values close to $\beta = 0$ conserve the space. Lance-Williams recommend that coefficient β be set to some small negative number such as $\beta = -0.25$. Setting $\beta = -0.50$, however, often provides a better solution (Scheibler & Schnieder, 1985).

2.4 Evaluative Criteria for Determining the Number and Quality of Multimorbidity Clusters

There is little consensus about how to determine the number and quality of clusters that represent a meaningful grouping of objects. Subjective criteria, based on subject expertise, are the most frequent approach used to determine the number and quality of the clusters revealed in a hierarchical cluster analysis. (Paraphrased from a quote by Baxter (1994) found in Everitt, et al., 2001). We advise that the criteria used to evaluate a cluster solution be carefully delineated prior to conducting an analysis.

For example, in the application described below, we adopted clinical criteria to evaluate the cluster solution that were based upon the informed judgments and consensus of a team of 7 health services researchers, including 3 primary care physicians (2 internists and 1 family practice physician who each had 10 or more years of clinical experience) about what comprises a theoretically or clinically meaningful set of diagnoses for the clusters. The following clinical criteria were used to identify multiple disease clusters with the most relevance for chronic disease management in primary care.

APPLIED MULTIVARIATE RESEARCH

- a) Do the groupings of the diseases within the cluster match known epidemiology?
- b) Would it be important for physicians to be aware of the frequent co-occurrence of these illnesses for co-management purposes (e.g., potential for drug-drug interactions)?
- c) Are most of, or all of, the diseases within the cluster known to respond (or have the potential to respond) to chronic disease management approaches?

3. Example: Multimorbidity in a Large Veterans Health Administration Sample

Our motivation for the application of cluster analysis to the problem of identifying multimorbid groupings of disease conditions grew out of a study designed to identify and describe the self-reported collaborative care needs among groups of veterans who are members of well defined, high frequency multimorbid groups of disease conditions. The immediate objectives of this exploratory study were: 1) Identify and describe the most prevalent chronic disease clusters among veteran primary care patients; and 2) Identify and describe differences in self-reported patient-centered needs and preferences among veterans with multiple chronic illnesses. The Veterans Health Administration is the ideal health care system in which to apply this technique to the study the multimorbidity phenomenon because it serves a large, aging population and has a well-established administrative database that contains patient encounter diagnoses.

3.1 Chronic Illness Diagnosis Groups

Based on a methodology used in a large survey of health status and outcomes among patients served by the U.S. Veterans Health Care Administration, (Perlin, Kazis, Skinner, Ren, Lee, Rogers, et al., 2000). ICD-9-CM diagnosis codes representing highly prevalent diseases for the veteran population were selected by a consensus panel of experts to create a set of 36 easily identifiable disease cohorts. Related diagnostic conditions were collapsed into a single diagnostic category. For this study we added additional ICD-9-CM codes, selected by two senior internists, to form other diagnosis groups of relevance to the primary care setting. This created a total of 45 diagnosis groups, 38 medical and 7 psychiatric disease categories, representing prevalent chronic illnesses in the VA primary care population.

3.2 Data Sources

Data were extracted from the outpatient (OPC) and inpatient (PTF) files of the National Patient Care Database (NPCD), which is an administrative database for patients served by the US Department of Veterans Affairs. (Veterans Affairs Information Resource Center, 2001). These files contain diagnostic information in the form of International Classification of Disease (ICD-9) codes. The OPC is the Veterans Health Administration's (VHA) national database for outpatient care.

MULTIMORBIDITY CLUSTERS

All outpatient encounters are stored in these datasets. The Diagnosis file can contain 10 diagnostic variables per encounter: one primary diagnosis and up to nine secondary diagnoses. The PTF is the national database for inpatient care in the VHA. All inpatient episodes of care, with the exception of admissions to Extended Care (nursing home), Observation Care, and Non-VA Care are stored in these datasets. Principal diagnosis is the admission diagnosis, while primary diagnosis is the diagnosis most responsible for the length of stay. Secondary diagnoses treated, observed, or known to have impacted the patient's length of stay are also recorded.

3.3 Study Population

The study population consisted of all patients who had primary care clinic visits at any VHA facility during a 4-year period (FY 1997-2000). Patients were considered to have had a primary care clinic visit if they had an encounter recorded for any of the following 6 "clinic stop" codes in the Visit file in the Outpatient Care System records of the NPCD: 301 (General Internal Medicine), 318 (Geriatric Clinic), 319 (Geriatric Evaluation & Management), 322 (Women's Clinic), 323 (Primary Care/Medicine), and 350 (Geriatric Primary Care). Eligible patients were required to have had at least one primary care visit per fiscal year in at least 3 of the 4 years. The three years did not have to be consecutive.

All inpatient and outpatient encounter diagnoses (including primary and secondary diagnoses), were extracted from the OPC and PTF for all patients meeting eligibility criteria. These included the Diagnosis files from the OPC and all 12 files in the PTF. For each patient, the presence of each of the 45 diagnosis groups was coded with a "1" if the patient had any of the corresponding ICD-9 codes recorded as a primary or secondary diagnosis for any outpatient or inpatient encounter over the 4 year period. The absence of ICD-9 codes corresponding to a particular diagnosis group was coded with a "0."

3.4 Statistical Procedures

A SAS PROC/IML program was written and used to reduce the transposed $p \times n$ data matrix to a $p \times p$ proximity matrix of dissimilarity coefficients among the chronic disease categories. Jaccard's coefficient was used to estimate these dissimilarity coefficients. The Lance-Williams flexible method, implemented in the agnes algorithm developed by Kaufman and Rousseeuw (1990) for agglomerative hierarchical clustering, was applied to the proximity matrix to identify the most highly associated diagnoses among the 1,327,382 patients with 2 or more chronic illnesses. We used the flexible-beta approach with β set at -0.50. The algorithm is available in the R statistical language.¹ The plot function in R takes the data object created by agnes and produces a dendrogram that depicts the successive grouping of chronic illness categories into clusters at each step in the hierarchical algorithm. The agglomerative coefficient (AC) provides a measure of the amount of structure found by the algorithm. It is based on the ratio of between-cluster dissimilarities to within-cluster dissimilarities. AC is the average ratio, and it ranges between 0 (no structure) and 1 (completely defined structure).

4. Results

1,645,314 unique patients met the eligibility criteria for primary care visits. Of these, 94.6% were male. The average age was 62.4 years, $SD = 13.6$. 76% were white, 17% were African American; 6% were Hispanic, and 1% belonged to other ethnic groups (American Indian and Asian American). The average number of diagnoses per patient was 3.49, $SD = 2.22$, and ranged from 0 to 20 diseases. Prevalence rates for the 45 chronic illnesses are displayed in Table 1. 101,814 (6.2%) had none of the 45 chronic illnesses, while 216,118 (13.1%) had only 1 of the 45 chronic illnesses. This left 1,327,382 (80.7%) patients who had 2 or more of the 45 chronic diseases. In the population with only 1 chronic disease, the frequency of individual diagnoses ranged from 46,608 (22.49 %) patients with hypertension to 113 (0.05 %) patients with Hepatitis B.

4.2 Hierarchical Cluster Analysis Results

Our expert panel of 7 health services researchers reviewed our hierarchical cluster results for the 45 chronic illness conditions relative to the three clinical criteria for identifying multiple disease clusters relevant to chronic disease management in primary care. Hypertension and hyperlipidemia had the smallest normed flexible distance and formed the first cluster to appear in the dendrogram. Ischemic heart disease was joined to this cluster at the next branch in the dendrogram. After joining substance abuse and alcohol abuse together, the algorithm added diabetes to the cluster representing the union of hypertension, hyperlipidemia and ischemic heart disease. This grouping of chronic illnesses associated with the metabolic syndrome produced the most distinct multimorbid grouping in our analysis. The normed flexible-beta distances become rapidly smaller as we move further toward the root of the dendrogram. Our hierarchical cluster analysis produced a number of single disease and predictable two disease clusters: i.e., cataract and glaucoma, congestive heart failure with irregular heart rate, and COPD with tobacco abuse, as well as two larger more heterogeneous groupings. A group of heterogeneous diseases whose common clinical link is the aging process is also present: i.e., arthritis, rheumatological conditions, hip fractures, skin cancer, and impotence. The presence of such heterogeneous diagnostic categories can substantially distort the cluster solution and creates interpretive ambiguity. Since neither of these more heterogeneous clusters nor the numerous two disease clusters met our third criterion for multimorbid clusters with relevance for chronic disease management in primary care, we trimmed these diagnostic categories from the cluster analysis. Our final cluster solution, based on 23 diagnostic categories, is displayed in Figure 3. The agglomerative coefficient for this solution is $AC = 0.52$ indicating that only a modest amount of structure is present in the data.

4.3 Clinically Meaningful Multimorbidity Clusters

Our expert panel identified six clinically meaningful multiple disease clusters that met all three a priori clinical criteria for multimorbid disease clusters with relevance to chronic disease management in primary care. These disease clusters

MULTIMORBIDITY CLUSTERS

were identified as follows: 1) “Obesity” Cluster with Osteoarthritis, Low Back Pain, Enlarged Prostate, GERD, and Obesity; 2) “Metabolic” Cluster with Diabetes, Hypertension, Hyperlipidemia, and Ischemic Heart Disease; 3) “Neurovascular” Cluster with Peripheral Vascular Disease, Stroke, TIA, Alzheimer’s Disease, and Seizures; 4) “Liver” Cluster with Hepatitis B, Hepatitis C, Chronic Liver Disease, and HIV; and 5) “Dual Diagnosis” Cluster with Substance Abuse, Alcohol Dependence, Schizophrenia, and Bipolar Disease; and 6) a “Mixed Anxiety-Depression” Cluster with Depression, PTSD, and Other Anxiety Disorders.

The cluster composed of diagnostic groups with the highest degree of association was the *Metabolic Cluster*. Of the patients who had 2 or more chronic diseases, 1,088,774 (83%) fell into one of the 15 possible subsets of the 4 diseases within this cluster: e.g., hypertension and hyperlipidemia; hypertension and diabetes; hypertension, hyperlipidemia and diabetes; etc. The prevalence for patients with all four chronic conditions in this cluster is 0.05033. The emergence of this cluster was not surprising, since it has been identified previously in the literature (e.g., the Deadly Quartet, Deedwania, 2000). The diseases within this cluster have well-established epidemiological ties; diabetes (or underlying insulin resistance) is thought to be a risk factor for the other 3 diseases.

The *Obesity Cluster* was the second most prevalent cluster; 711,272 (54%) of all patients who had two or more chronic illnesses fell into one of the 31 possible subset combinations of the 5 diseases within this cluster (obesity, osteoarthritis, low back pain, enlarged prostate, and GERD). The prevalence for patients who have all 5 chronic diseases in this cluster is 0.00038. Although the inclusion of GERD and enlarged prostate in this cluster was somewhat surprising, obesity is a known risk factor for esophagitis (Wilson, Ma, & Hirschowitz, 1999). Obesity has also been associated with benign prostatic hypertrophy (Dahle, Chokkalingham, Gao, Deng, Stanczyk, & Hsing, 2002). Recognition of this cluster may also be important considering the prevalence of these diseases and the potential for adverse events in the treatment of illnesses in this cluster. For example, NSAIDs for osteoarthritis may exacerbate GERD symptoms.

The diseases within the two clusters with the lowest prevalence rates, *Neurovascular* (186,881; 14%) and *Liver* (48,466; 4%) have strong epidemiological links in terms of risk factors or disease transmission. The prevalence rates for patients with all chronic conditions defining these clusters are 0.00025 and 0.00004 for the *Neurovascular* and *Liver* clusters, respectively. The true prevalence of subsets of the diseases within the Neurovascular cluster, however, may be higher, since studies suggest that Alzheimer’s disease and other dementias are under-recognized, especially during the early stages of the disease process.

Frequency counts indicated that 280,711 (21%) of the patients with 2 or more chronic illnesses fell into 1 of the possible subsets of diseases within the *Mixed Anxiety-Depression Cluster*, and 153,962 (12%) fell into 1 of the subsets of diseases in the *Dual Diagnosis Cluster*. The prevalence rate for patients with all 3 conditions in the *Mixed Anxiety-Depression Cluster* is 0.01077, and the prevalence rate for patients with all 4 conditions in the *Dual Diagnosis Cluster* is 0.00063. The strong associations among the psychiatric diagnoses is not surprising, given the high rates of psychiatric comorbidity that have been documented in the literature, as well as recognized problems with the reliability and validity of

APPLIED MULTIVARIATE RESEARCH

the psychiatric nosology. Interestingly, the 7 diagnosis groups within these clusters form a triad of diagnoses at lower levels in the hierarchy that parallel conventional diagnostic groupings: abuse disorders, psychotic disorders, and neuroses or mood disorders.

Many health care systems organize and deliver specialty mental health services organized along these diagnosis groupings. The large numbers of primary care patients who had recorded diagnoses across the psychiatric subgroups suggests the importance of providing integrated care for these complex patients. The *Mixed Anxiety-Depression Cluster* is consistent with the growing literature on the extensive comorbid relationship between these two disorders and the biological evidence of linkage which suggests a core disorder (Barlow, Campbell, 2000; Barbee, 1998). The presence of the *Dual Diagnosis Cluster* is particularly interesting, however, and is supported by a vast literature on the comorbid prevalence and symbiotic nature of these conditions with their complicated etiology and causal direction (Mueser, Drake, & Wallach, 1998; Dixon, Hass, Weiden, Sweeney, & Frances, 1991). The prevalence of substance misuse among patients with severe psychiatric conditions, estimated to be at least 50% (Kessler, McGonagle, Zaho, Nelson, Hughes, Eshleman, Wittchen, & Kendler, 1994; Drake & Wallach, 1989), is a matter of increasing concern to mental health workers.

Patients with comorbid substance abuse represent a complex population to treat, with a significantly higher risk of relapse, psychiatric readmission (Hunt, Bergen, & Bashir, 2002; Haywood, Kravitz, Grossman, Cavanaugh, Davis, & Lewis, 1995), and poor medication adherence (Weiss, Smith, Hull, Piper, & Huppert, 2002; Owen, Fisher, Booth, & Cuffel., 1996). Furthermore, this “double jeopardy” leads to more severe symptoms, cognitive difficulties, and greater health services utilization (Swofford, Scheller-Gilkey, Miller, Woolwine, & Mance, 2002; Curran, Sullivan, Williams, Han, Collins, Keys, & Kotrla., 2003). The cost to health care systems for treating dual diagnosis patients is more than twice that for other psychiatric patients. For example, the cost of treatment in FY 1999 for a sample N = 57,000 psychiatric patients in the Veterans Administration Health Care System found the mean average cost to be \$23,405 for dual diagnosis compared with only \$11,096 for schizophrenic patients (Zeber, personal communication). The clinical and financial significance of this multimorbid pattern of psychiatric conditions has prompted many service delivery systems to develop comprehensive, collaborative care strategies for diagnosis and management of dual diagnosis patients (Coyne, Ward, and Doran, 2003; Barry, Zeber, Blow, & Valenstein, 2003).

5. Discussion

Agglomerative hierarchical clustering was used to identify clinically meaningful multimorbid groupings among a set of 45 chronic illnesses that are highly prevalent or considered to be high priority diseases among patients served by the Veterans Health Administration. We identified 6 clinically valid clusters among 23 chronic conditions that met our *a priori* criteria for multimorbid groupings with relevance to chronic disease management: Metabolic, Obesity, Liver, Neurovascular, Mixed Anxiety-Depression, and Dual Diagnosis related chronic disease groups.

MULTIMORBIDITY CLUSTERS

Cluster analysis, however, is an exploratory data technique that involves a series of analytic decisions. Decisions made at each step in the analytic process may have a profound influence on the final results. It is critically important that researchers think through the various steps in the process and set out their criteria for making these decisions prior to conducting a cluster analysis. Following Milligan (1996), we organized our analytic work around 5 basic questions: What objects to cluster, what unit of measurement to use for the variables, what proximity measure to use as an index of similarity or dissimilarity among the objects, what type of clustering algorithm to use, and what criteria to use for determining the number and quality of clusters in the data.

5.1 Objects to Cluster and Unit of Measurement.

Since our objective was to search for multimorbid groupings among chronic diseases, we choose to identify clusters of chronic disease categories, rather than clusters of patients, as is commonly done in published cluster analyses. The natural unit of measurement in this case is a simple binary variable. Another approach that may have fit our problem would be to use principal components factor analysis to group the sets of chronic diseases, compute factor scores for each patient, and cluster the factor scores. Factor scores are continuous rather than binary variables. Hallman, et al. used this strategy to identify symptom patterns among veterans who suffer from symptoms and diseases associated with the Gulf War syndrome (Hallman, Kipen, Diefenbach, Boyd, Kand, Leventhal, & Wartenberg, 2003). Factor analysis on binary data can be problematic. Binary items may group together on a factor because they share similar distributions rather than some underlying common trait (Nunnally & Bernstein, 2003). Factor analysis uses a measure of association (phi coefficient or tetrachoric correlation) that incorporates information from negative, as well as positive matches. We regarded negative matches as non-informative and chose a similarity measure that excluded negative matches.

5.2 Proximity Measures for Binary Data.

Our decision to use Jaccard's coefficient to compute the dissimilarity measure for our proximity matrix was based on the notion that multimorbidity groups reflect the likelihood that one or more chronic conditions occur conditionally. Thus, negative matches are uninformative in forming multimorbid grouping among our 45 chronic disease categories. Jaccard's coefficient is one of several similarity measures for binary data that satisfy our conceptual model for multimorbid groupings. It differs from the other coefficients in that they assign more or less weight to the non-match cells (b and c in Figure 1); but, as we argued earlier, we had no *a priori* reason to assign different weights to the positive and non-matches.

5.3 Clustering Algorithm

In addition to our conditional model for association among chronic diseases in multimorbidity, we also postulated that each multimorbid grouping reflects a

APPLIED MULTIVARIATE RESEARCH

hierarchical structure associated with a common underlying pathological origin or process, or they are linked as sequelae of one chronic disease. These two decisions led to our adopting an agglomerative hierarchical clustering technique to explore patterns of multimorbidity among the 45 chronic disease categories. Hierarchical algorithms produce exclusive clusters that do not reflect the fact that one or more diseases in one cluster may have clinically important associations with diseases in another cluster. Primary care physicians who treat diabetic patients may be surprised to see that obesity, with its well known association with diabetes and the insulin resistance syndrome, formed its own distinct cluster rather than grouping with diabetes in our Metabolic cluster. Multimorbidity clusters are useful tools to help us organize and study the problem of multiple chronic illnesses; and, they can assist in the design of collaborative intervention strategies to efficiently and effectively manage such conditions. The exclusive appearance of the clusters is an artifact of the algorithm. Other clinically significant associations may exist between diagnoses in one cluster with those in another. A fuzzy clustering algorithm based on probability of membership (grade-of-membership) or latent-class cluster analysis may produce clusters that better reflect the complex structure of multimorbidity (Everett, et al., 2001; Vermunt and Magidson, 2002), but such overlapping structure may fail to produce a solution useful for planning health care interventions. Any clustering algorithm, however, is only one step in the process of understanding multimorbidity; clusters serve as markers and guides to give focus and direction to our research. Recognition that diseases in one cluster may have important clinical associations with diseases in another cluster is important in the design of innovative collaborative intervention strategies; but, we argue that the 6 primary clusters we identified in our hierarchical analysis provide a useful starting point for the study of multimorbid grouping of diseases.

5.4 Evaluative Criteria for Determining the Number and Quality of Multimorbidity Clusters

A wide variety of approaches are taken to choose the number of clusters and assess their quality. Most approaches rely on subject expertise and judgment to determine the number clusters and assess the consistency and utility of the cluster solution. We developed and used a set of three clinical criteria to determine the number of clusters. Others might disagree with our choices of subjective criteria or prefer more objective numeric criteria to assess the consistency and utility of a given clustering solution. Our criteria and the process we used to generate the criteria are explicit and were articulated prior to our attempt to evaluate and interpret the results.

5.5 Replicability of the Cluster Solution

In most applications of cluster analysis, we rely on relatively small samples drawn from a substantially larger population. Thus, replication of the cluster solution across multiple independent samples is essential to establish consistency and enhance our confidence in the integrity of a cluster solution. Replication is less relevant, however, when the analysis is essentially based on the entire population of interest, rather than relatively small samples. The chief advantage of

MULTIMORBIDITY CLUSTERS

working with large administrative databases is that we have access to the entire population of interest upon which to base our inferences. Replication of the hierarchical cluster analysis for the 23 chronic conditions identified in our final solution based on 10 independent random samples of either 10,000 or 100,000 was entirely consistent with our solution based on the entire population.

5.6 Summary and Conclusions

Changing our research focus to study patterns of multimorbidity, as opposed to a disease-specific focus, is important because of the enormous burdens caused by multiple chronic illnesses. Although health care systems have recognized and begun to respond to the special needs of outpatients with chronic disease through such initiatives as promoting primary care, instituting case-management programs, and implementing clinical practice guidelines, most of these initiatives have focused on single chronic illnesses. Often overlooked, however, are the issues that patients, providers, and systems face in attempting to manage multiple diseases or problems in a single individual. Although disease-specific strategies may be effective for the single illnesses that they target, these strategies can also result in fragmentation of care and are often unable to address the complex needs of patients with multimorbidity. (Starfield, Lemke, Bernhardt, Foldes, Forrest, & Weiner, 2003).

For example, case management may be successful for diabetes or heart failure, but must the patient with both go to two different case managers (or three if COPD is also present)? Most clinical practice guidelines deal with a single disease entity and are derived from efficacy studies that often exclude patients with comorbidities. Yet the exclusion of patients with comorbidity restricts the generalizability of results. Providers often cite this as a barrier to guideline implementation and argue that clinical practice guidelines are simply not relevant or applicable to their “typical” patients who have multiple chronic illnesses. {Cabana, Rand, et al. 1999}

Identification of the most commonly occurring clusters of chronic diseases creates an opportunity for further research and quality improvement initiatives in two areas: 1) design and implementation of chronic disease management programs for complex patients with multimorbidity, and 2) identification of strategies to implement multiple evidence-based guidelines for a patient with multiple chronic illnesses. Recognition of these relatively high frequency clusters may be useful in allocating resources and designing chronic disease strategies. Instead of focusing on single diseases, case managers could be transformed into “multidisease” specialists. Ideally, their training would equip them to monitor, educate, and provide medication management and supportive care for patients with clusters of high-frequency diseases. Group visits could be organized not just for patients with a single chronic illness, but for patients with highly prevalent disease clusters. In addition, patient education programs could offer menus of self-management and skills training programs of most relevance to patients with high frequency clusters.

Cluster analysis is a useful tool to help organize our research efforts by identifying clinically meaningful multimorbid grouping of diseases. Although this analysis was conducted on primary care patients served by a single health care

APPLIED MULTIVARIATE RESEARCH

system, the results should be generalizable to aging patients with similar demographics (e.g., Medicare). Furthermore, this methodology could be used to inform any health care system of the patterns of multimorbidity occurring within their own unique patient populations. Health care systems are faced with the reality that a growing number of patients have more than one chronic illness (van den Akker, 2001). Those systems that develop disease management strategies across common clusters rather than for individual diseases may improve patient-centered outcomes and increase efficiency.

ACKNOWLEDGMENTS

The research reported here was supported by the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service (IIR 01-110). Dr. Nokl is the Associate Director of the VER-DICT, a VHA HSR&D Center of Excellence at the South Texas Veterans Health Care System. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

REFERENCES

- Baxter, M. J. (1994). *Exploratory multivariate analysis in archaeology*. Edinburgh: Edinburgh University Press.
- Barbee, J. G. (1998). Mixed symptoms and syndromes of anxiety and depression: Diagnostic, prognostic and etiologic issues. *Annals of Clinical Psychiatry*, 10, 15-29.
- Barlow, D. H., & Campbell, L. A. (2000). Mixed anxiety-depression and its implications for models of mood and anxiety disorders. *Comparative Psychiatry*, 41(2 Suppl 1), 55-60.
- Barry, K. L., Zeber, J. E., Blow, F. C., & Valenstein, M. (2003). Effect of strengths model versus assertive community treatment model on patient outcomes and utilization: 2-year follow up. *Psychiatr Rehabil J*, 26(3), 268-77.
- Coyne, P., Ward, D., & Doran, R. (2003). *Strategies for dual diagnosis service provision*. London: Central and North West London Health NHS Trust.
- Curran, G. M., Sullivan, G., Williams, K., Han, X., Collins, K., Keys, J., & Kotrila, K. J. (2003). Emergency department use of persons with comorbid psychiatric and substance abuse disorders. *Ann Emerg Med*, 41(5), 659-67.
- Dahle, S. E., Chokkalingham, A. P., Gao, Y. T., Deng, J., Stanczyk, F. Z., & Hsing, A. W. (2002). Body size and serum levels of insulin and leptin in relation to the risk of benign prostatic hyperplasia. *Journal of Urology*, 168, 599-604.
- Deedwania, P. C. (2000). Mechanism of the deadly quartet. *Canadian Journal of Cardiology*, 16, 17E-20E.

MULTIMORBIDITY CLUSTERS

- Dixon, L., Haas, G., Weiden, P. J., Sweeney, J., & Frances, A. J. (1991). Drug abuse in schizophrenic patients: Clinical correlates and reasons for use. *American Journal of Psychiatry*, 148, 224-30.
- Drake, R. E., & Wallach, M. A. (1989). Substance abuse among the chronic mentally ill. *Hospital and Community Psychiatry*, 40(10), 1041-6.
- Do, K.-A., Broom, B., & Wen, S. (2003). GeneClust. In G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger (Eds.), *The analysis of gene expression data* (pp. 342 - 361). NY: Springer.
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1988). Cluster analysis and display of genome wide expression patterns. *Proceedings of the National Academy of Science*, 95, 14863-14868.
- Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). NY: Oxford University Press.
- Fried, L. P., Bandeen-Roche, K., Kasper, J. D., & Guralnik, J. M. (1999). Association of comorbidity with disability in older women: the Women's Health and Aging Study. *Journal of Clinical Epidemiology*, 52, 27-37.
- Gimby, A., & Svanborg, A. (1997). Morbidity and health-related quality of life among ambulant elderly citizens. *Aging*, 9, 356-364.
- Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). London: Lawrence Erlbaum Associates, Publishers.
- Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics*, 23, 623-628.
- Hallman, W. K., Kipen, H. M., Diefenbach, M., Boyd, K., Kang, H., Leventhal, H., et al. (2003). Symptom patterns among Gulf War registry veterans. *American Journal of Public Health*, 93, 624-630.
- Haywood, T. W., Kravitz, H. M., Grossman, L. S., Cavanaugh, J. L., Davis, J. M., & Lewis, D. A. (1995). Predicting the "revolving door" phenomenon among patients with schizophrenic, schizoaffective, and affective disorders. *American Journal of Psychiatry*, 152(6), 856-61.
- Hunt, G. E., Bergen, J., & Bashir, M. (2002). Medication compliance and comorbid substance abuse in schizophrenia: impact on community survival 4 years after a relapse. *Schizophr Res*, 54(3), 253-64.
- Jaccard, P. (1908). Nouvelles recherches sur la distribution of florale. *Bulletin de la Sociti Vaudoise de Sciences Naturelles*, 44, 223-370.
- Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the National Comorbidity Survey. *Archives of General Psychiatry*, 51(1), 8-19.
- Kraemer, H. C. (WXYZ). Statistical issues in comorbidity. *Statistics in Medicine*, 14, 721-733.
- Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Computer Journal*, 9, 373-380.
- Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. J. Hubert, & G. de Soete (Eds.), *Clustering and classification* (pp. 341 - 375). Singapore: World Scientific.
- Mueser, K. T., Drake, R. E., & Wallach, M. A. (1998). Dual diagnosis: A review of etiological theories. *Addictive Behaviors*, 23, 717-734.

APPLIED MULTIVARIATE RESEARCH

- Mueser, K. T., Bond, G. R., Drake, R. E., & Resnick, S. G. (1998). Models of community care for severe mental illness: A review of research on case management. *Schizophrenia Bulletin*, 24, 37-74.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). NY: McGraw-Hill, Inc.
- Owen, R. R., Fischer, E. P., Booth, B. M., & Cuffel, B. J. (1996). Medication noncompliance and substance abuse among patients with schizophrenia. *Psychiatric Services*, 47(8), 853-8.
- Perlin, J., Kazis, L. E., Skinner, K. M., Ren, X. S., Lee, A., Rogers, W., et al. (2000). *Health status and outcomes of veterans: physical and mental component summary scores: Veterans SF-36: 1000 Large Health Survey of Veteran Enrollees: Executive report*. Washington, D.C.: Department of Veterans Affairs, Office of Quality and Performance; Bedford, MA: VHA Health Assessment Project, Center for Health Quality, Outcomes and Economic Research.
- Scheiber, D., & Schneider, W. (1985). Monte Carlo test of the accuracy of cluster analysis algorithms — a comparison of hierarchical and non-hierarchical methods. *Multivariate Behavioral Research*, 20, 283-304.
- Seber, G. A. F. (1984). *Multivariate observations*. NY: John Wiley & Sons.
- Sneath, P. H. A. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, 17, 201-226.
- Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409-1438.
- Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of vegetation on Danish commons. *Biologiske Skrifter*, 5, 1-34.
- Starfield, B., Lemke, K., Bernhardt, T., Foldes, S. S., Forrest, C. B., & Weiner, J. P. (2003). Comorbidity: Implications for the importance of primary care “case” management. *Annals of Family Medicine*, 1, 8-14.
- Swofford, C. D., Scheller-Gilkey, G., Miller, A., Woolwine, B., & Mance, R. (2000). Double jeopardy: schizophrenia and substance use. *American Journal of Drug and Alcohol Abuse*, 26(3), 343-353.
- van den Akker, M., Buntinx, F., Metsemakers, J. R., Roos, S., & Knottnerus, J. (1998). A. Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. *Journal of Clinical Epidemiology*, 51, 367-375.
- van den Akker, M., Buntinx, F., Roos, S., & Knottnerus, J. (2001). Problems in determining occurrence of multimorbidity. *Journal of Clinical Epidemiology*, 54, 675-679.
- Verbrugge, L. M., Lepkowski, J. M., Imanaka, Y. (1989). Comorbidity and its impact on disability. *Milbank Quarterly*, 67, 450-484.
- Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89-106). Cambridge: Cambridge University Press.
- Veterans Affairs Information Resource Center (VIREC). The Medical SAS(r) inpatient datasets - FY2000: a VIREC resource guide; The Medical SAS(r) outpatient datasets - FY2001: a VIREC resource guide. 2001 Sept. [2002 Nov. 4]. Available from URL: <http://www.virec.research.med.va.gov/RGFY2000.htm>.

MULTIMORBIDITY CLUSTERS

- Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.
- Weiss, K. A., Smith, T. E., Hull, J. W., Piper, A., & Huppert, J. (2002). Predictors of risk of nonadherence in outpatients with schizophrenia and other psychotic disorders. *Schizophr Bull*, 28(2), 341-9.
- Wilson, L. J., Ma, W., & Hirschowitz, B. I. (1999). Association of obesity with hiatal hernia and esophagitis. *American Journal of Gastroenterology*, 94, 2840-2844.

Footnotes

1. R is an open-source version of the S language. It is an official part of the Free Software Foundation's GNU project and is available for download from <http://www.r-project.org/>. A powerful, user friendly implementation of the S language is the commercially available S-Plus software package from Insightful Corporation, Seattle, WA. The URL for the Insightful website is <http://www.insightful.com/>.

Table 1: Prevalence rates for each of the 45 chronic disease categories

Chronic Disease Category	Total Sample		Sample with 1 chronic illness		Sample with ≥ 2 chronic illnesses	
	(n = 1,645,314)		(n = 216,118)		(n = 1,327,382)	
	Frequency	%	Frequency	%	Frequency	%
Those with none of the 45 chronic illness groups	101,814	6.19				
Hypertension	885,055	53.79	48,608	22.49	836,447	63.01
Hyperlipidemia	497,570	30.24	13,859	6.41	483,711	36.44
Diabetes	395,382	24.03	14,708	6.81	380,674	28.68
Ischemic Heart Disease (IHD)	383,040	23.28	9,086	4.20	373,954	28.17
Osteoarthritis	301,469	18.32	13,610	6.30	287,859	21.69
COPD	249,076	15.14	10,536	4.88	238,540	17.97
Enlarged Prostate	204,529	12.43	5,357	2.48	199,172	15.00
GERD	204,456	12.43	6,441	2.98	198,015	14.92
Depression	195,660	11.89	6,520	3.02	189,140	14.25
Low Back Pain	188,799	11.47	13,353	6.18	175,446	13.22
Cancers (excluding Skin Cancer)	171,814	10.44	10,304	4.77	161,510	12.17

APPLIED MULTIVARIATE RESEARCH

Table 1 continued

Cataract	164,127	9.98	3,161	1.46	160,966	12.13
Obesity	155,387	9.44	2,699	1.25	152,688	11.50
Tobacco Abuse	135,241	8.22	4,445	2.06	130,796	9.85
Congestive Heart Failure (CHF)	108,767	6.61	1,346	0.62	107,421	8.09
Anemia	105,604	6.42	2,559	1.18	103,045	7.76
PTSD	97,811	5.94	3,289	1.52	94,522	7.12
Peripheral Vascular Disease (PVD)	91,206	5.54	1,403	0.65	89,803	6.77
Other Anxiety Disorders	91,693	5.57	2,185	1.01	89,508	6.74
Glaucoma	88,143	5.36	1,662	0.77	86,481	6.52
Peptic Ulcer Disease	89,069	5.41	2,795	1.29	86,274	6.50
Irregular Heart Rate	87,939	5.34	2,187	1.01	85,752	6.46
Alcohol	83,825	5.09	2,847	1.32	80,978	6.10
Thyroid	77,909	4.74	3,266	1.51	74,643	5.62
Other Arthritis	71,586	4.35	2,724	1.26	68,862	5.19
Stroke	67,674	4.11	1,708	0.79	65,966	4.97
Impotence	64,300	3.91	1,446	0.67	62,854	4.74
Gout	56,431	3.43	1,247	0.58	55,184	4.16
Schizophrenia	51,897	3.15	3,374	1.56	48,523	3.66
Asthma	52,573	3.20	4,381	2.03	48,192	3.63
Renal Disease	41,399	2.52	725	0.34	40,674	3.06
Bipolar	37,735	2.29	1,378	0.64	36,357	2.74
Skin Cancer	32,672	1.99	797	0.37	31,875	2.40
Substance Abuse	32,836	2.00	1,037	0.48	31,799	2.40
Alzheimer's Disease & Other Dementias	29,628	1.80	1,660	0.77	27,968	2.11
Hepatitis C	28,064	1.71	1,374	0.64	26,690	2.01
Rheumatology	22,828	1.39	1,170	0.54	21,658	1.63
Chronic Liver Disease	21,541	1.31	843	0.39	20,698	1.56
Hip Fractures	19,690	1.20	751	0.35	18,939	1.43
Transient Ischemic Attacks (TIA)	15,192	0.92	311	0.14	14,881	1.12
Seizures	13,822	0.84	1,263	0.58	12,599	0.95
Gallbladder Disease	11,354	0.69	858	0.40	10,496	0.79
Bowel Disease	8,685	0.53	1,010	0.47	7,675	0.58
HIV	8,498	0.52	1,722	0.80	6,776	0.51
Hepatitis B	4,448	0.27	113	0.05	4,335	0.33

MULTIMORBIDITY CLUSTERS

Figure 1. Jaccard's coefficient as a measure of the joint occurrence of two chronic medical conditions

A.

		Hypertension	
		Yes	No
Hyperlipidemia	Yes	338,257 <i>a</i>	498,190 <i>b</i>
	No	145,454 <i>c</i>	117,873 <i>d</i>

B.

$$s_{ij} = \frac{\text{Hyperlipidemia} \cap \text{Hypertension}}{\text{Hyperlipidemia} \cup \text{Hypertension}}$$

C.

$$\begin{aligned} \text{Jaccard's Coefficient: } s_{ij} &= \frac{a}{a + b + c} \\ &= \frac{338,257}{338,257 + 498,130 + 146,464} \\ &= 0.345 \end{aligned}$$

Figure 2. Cluster dendrogram for Jaccard dissimilarity coefficients among the 23 prevalent chronic conditions observed in the VHA patient population