Validation of clustering solutions for clinical data through biologically meaningful simulations and mixed-distance dissimilarity methods

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Caitlin E. Coombes

Graduate Program in Public Health

The Ohio State University

2020

Thesis Committee

Guy Brock, PhD, Advisor

Courtney Hebert, MD, Committee Member

Chi Song, Committee Member

Abstract

Unsupervised clustering poses unique challenges in clinical data due to heterogeneous size and mixed type. We hypothesize that these limitations can be overcome by calculating dissimilarity by combining multiple distance methods. A review of the literature suggests that solutions for mixed, clinical data are sparse and lack rigor. In an initial experiment on real clinical data, we find limitations in a common approach: converting a mixed data set to a single data type. To rigorously test dissimilarity metrics and clustering methods, we develop 32,400 simulations of realistic, mixed-type clinical data and test 3 clustering algorithms (hierarchical clustering, Partitioning Around Medoids, and self-organizing maps) on 5 single distance metrics (Jaccard Index, Sokal & Michener distance, Gower coefficient, Manhattan distance, Euclidean distance) and 3 multiple distance methods of calculating dissimilarity (DAISY, Supersom, and Mercator, a method of our own devising). We apply the superior solution for a data mixture predominated by binary features, DAISY with Ward's hierarchical clustering, to the data set from our initial experiment, and recover important prognostic features. These experiments raise future questions for clustering problems in clinical data, including identifying minimum size for successful clustering (relevant when clustering clinical trials) and addressing concerns for validation of sometimes variable outcome.

Dedication

To my grandfather, Anthony Abruzzo, who showed that it is never too late to go back to graduate school, and who always believed I could do anything I put my mind to. We miss you.

Acknowledgments

I am very grateful to a number of people whose support, encouragement, and scientific example have been instrumental in seeing this thesis to completion. Some deserve special mention:

I would like to thank my supervisor, Guy Brock, for his infinite patience for long-winded emails, and my committee members, Courtney Hebert and Chi 'Chuck' Song.

My parents, Kevin R. Coombes and Lynne V. Abruzzo, without whom being here would be impossible – genetically or scientifically. It took me 30 years, but I couldn't be prouder to follow in your footsteps.

My fiancé, Alex Rasmussen, for helping me be a good scientist and a better person.

My grandmother, Leonore M. Abruzzo. Please keep moving the goal posts.

To all of Coombes Lab, because I never could have learned to much without such a warm, talented, challenging, and collaborative group of co-trainees, but especially to Zachary Abrams, for starting out a great teacher and becoming a great friend. I hope we never stop collaborating.

To all the friends who were instrumental in helping me stay sane, keep perspective, recharge, and have fun, but especially Jessica Pollock, Matt Phillips, Lindsay Mlynarek, Sally Trout, Katie Balaskas, and Eileen Hu

Vita

MD Candidate, The Ohio State University College of Medicine, Class of 2022

BA, Rice University, 2009

Publications

Coombes CE, Gregory ME. The Current and Future Use of Telemedicine in Infectious Diseases Practice. Curr Infect Dis Rep. 2019 Oct 19;21(11):41.

Fields of Study

Major Field: Public Health

Biomedical Informatics

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

With improvements in data mining of the electronic medical record and the expansion of large clinical databases, the scale of data available for clinical knowledge discovery is increasing dramatically. This expanding data size and complexity demands new analytical approaches.[1, 2] Some analytical techniques refined in bioinformatics for the analysis of high-throughput data may provide translational methods to analyze large-scale clinical data, if they can be properly transformed. Pattern discovery with unsupervised machine learning (ML), a common approach for multi-omics data,[3, 4] has the potential to revolutionize our understanding of patient phenotypes and clinical outcomes.[5] However, clinical data, which are characterized by considerably greater heterogeneity than common high-throughput datasets, pose unique problems for unsupervised ML applications.[1]

For two decades, clustering has been a common tool for pattern discovery in bioinformatics.[3, 4] Unsupervised analysis of high-throughput omics experiments have uncovered new patterns to annotate the genome, elucidate chromatin structure,[6] reveal molecular subtypes in cancer from gene expression,[7] and functionally segment the human genome by understanding histone modification.[8] In recent years, unsupervised ML began to be applied to clinical data. Clustering analyses have found

applications from heart disease,[9] chronic obstructive pulmonary disease,[10, 11] critical care,[12] and sepsis [13] to health services applications.[14]

Clinical data pose unique challenges for clustering analyses. Unsupervised ML in bioinformatics commonly involves the uniform application of a mathematical distance metric to a matrix of continuous or binary data that are homogeneous in type. Unlike omics data, clinical data are heterogeneous, characterized by a mixture of data types, which can impede easy application of unsupervised ML in clinical informatics. Heterogeneity of data type raises new challenges in feature selection, choosing a distance metric that captures biological meaning, and visualizing clinical data.

A Brief Survey of Clustering Algorithms

Unsupervised ML broadly encompasses algorithms that aim to uncover hidden structure in data from input features alone. Clustering analyses, a subcategory of unsupervised ML, attempt to partition these input data into distinct groups based on calculated similarities between observations.[15] Clustering produces a taxonomy of subjects, and has useful applications at an early stage in an area of scientific research.[16] Clustering analyses require two fundamental steps, which may be implemented separately or simultaneously: the calculation of similarity, dissimilarity, or distance between features or subjects, followed by the application of an algorithm to uncover latent clusters.

Many general classes of clustering algorithms exist. Older approaches – and still the most popular methods [4] – include hierarchical and partitional algorithms. More recent methods include neural network-based clustering and kernel-based learning, arising from support vector machines.[3, 17] Hierarchical methods constitute the oldest

(and still some of the most popular) clustering algorithms.[16] Hierarchical algorithms construct clusters based on a similarity matrix, calculated *a priori,* and a linkage criterion to determine the distance between pairwise set of observations.[17] Hierarchical clusters may be partitioned as if growing from a single cluster (divisive hierarchical clustering) or formed from merging of singleton clusters (agglomerative hierarchical clustering).[3, 4] Hierarchical clusters are visualized as a dendrogram, where leaves (data objects) branch from a single root note (representing the entire data set), such that the distance in height between pairs or clusters increases with their dissimilarity.[3] In 1963, Ward published a general agglomerative procedure for cluster formation that merges optimally similar subjects based on an "objective function" that "reflects the criterion chosen by the investigator."[18] Ward's method is common use today, typically implemented with the error sum of squares.[14, 19-22]

Partitional methods include k-means and related algorithms and fuzzy clustering.[17] In k-means clustering, the algorithms initializes a number of points in space equal to a researcher-provided desired number of clusters. These cluster "centroids" describe the mean of the coordinates of the data points in the cluster. The algorithm then iteratively optimizes the membership of each remaining data point to one of these clusters, recalculating the positions of the centroids to reach fixed (assumed to be optimal) centroid positions and final cluster memberships.[23] K-means is considered a "staple" of clustering algorithms: historical, well-known, and influential. K-means produces compact clusters with less computational intensity than hierarchical methods, and is therefore suited to large data sets.[3, 16] K-modes clustering is an extension of k-

means clustering for nominal data: a non-parametric method that iteratively reduces a loss function representing the count of mismatches between points,[24] implementing the Hamming distance. In k-medoids and the related algorithm PAM (Partitioning Around Medoids), a representative data point (a "medoid") is chosen for each cluster center. Thus, distances do not need to be calculated at each iteration, as each distance can be obtained from a distance matrix, calculated *a priori*.[23, 25] In fuzzy clustering, a data point can be partitioned such that it belongs to more than one cluster, with its membership to multiple cluster described as a set of membership probabilities.[3, 17] The fuzzy c-means algorithm is the counterpart to k-means with fuzzy partitions.[3]

Statistical finite mixture models, which "regard the observations to be clustered as a random sample from a finite mixture of distributions," converge on conditional probabilities of group membership by maximum likelihood. The flexibility of statistical mixtures allows applications to mixed-type data. Early implementations included latent class analysis which calculates similarity between categorical variables based on observed and expected counts.[16, 26] Latent class methods were expanded in the 1970's using maximum likelihood estimation from the EM (expectation, maximization) algorithm, an iterative method to maximize expected log-likelihood from given parameters.[27, 28] Due to their flexibility, finite mixture models are applied to mixed-type data today, and the literature describing their modern application is extensive.[16] However, mixture models could be viewed as not true clustering analysis, better suited to a "mature" stage in research development in a field, requiring some a priori understanding of the nature of the data at hand.[16] Finite mixture model approaches to

uncovering latent subclasses present unique problems that differ from concerns in other clustering analyses.[16] Thus, to keep this manuscript focused in scope, we will not discuss mixture models further.

Less common clustering approaches include neural networks, particularly self-organizing maps (SOM), and density-based clustering. Neural networks generate clusters through a competitive learning scheme, where input data are patterned into units through competition and weighting to produce the strongest output pattern.[3] SOM are shallow, unsupervised neural networks that create topographical, ordered, one- and two-dimensional mappings in a lattice from similarities in high-dimensional data, from which cluster identities can be recovered.[3, 29] Traditional implementations of SOM can cluster continuous features but cannot handle categorical data. Instead, these techniques transform categorical features to binary and handle them as continuous features.[17] Density-based clustering clusters data points above a distance threshold and frequency threshold by computing a local density criterion as triangular similarity between points, evaluating the frequency of points within a certain distance as a density and cluster membership. Non-dense points are removed, allowing the algorithm to remove "noise."[4, 30] Some instances of density-based clustering applied for the clustering of clinical data will be discussed later in this chapter.

Clustering Approaches for Mixed-Type Data

The general algorithms discussed in the previous section can be applied to cluster single-type or mixed-type data. In the handling of mixed-type data, the method of expressing similarity or distance is a primary way in which clustering techniques differ.[16] The

problem of calculating similarity or dissimiliarity in mixed-type data is as old as clustering algorithms themselves. In 1966, Goodall [31] proposed a method of calculating similarity by ordering calculated similarity between pairs and expressing their similarity as "the complement of the probability that a random sample of two will have a similarity equal to, or greater than, the pair in question." Variations on the method allowed its application to nominal, ordinal, binary, and continuous attributes, with the total probabilities calculated by a final ordering step.

Modern approaches implement several methods of handling mixed data. One approach is to convert features to a single data type. The researcher may convert all categorical features to continuous.[17, 32] Conversely, continuous features can be transformed to categorical.[11, 32] However, data conversion risks information loss, so we consider mixed-data solutions that handle mixed data directly, without transformation, to be the most desirable approaches.

The approach proposed by Goodall over 50 years ago – constructing a measure of dissimilarity for variables of each data type and combining them, possibly with a method of differential weighting, into a single coefficient – remains a dominant approach today.[16] In 1990, Chiodi [33] proposed an approach to iteratively partition data objects based on distinct measures of distance for continuous, ordinal, and nominal data. In 2002, Li and Biswas [34] implemented the Goodall similarity for agglomerative hierarchical clustering of simulated and real mixed-type data. More commonly, mixed-type approaches implement two methods of calculating similarity: one for continuous and one for categorical data. Often, the Hamming distance is used for categorical features with the

Minkowski distance (a generalization of the Euclidean and Manhattan distances),[35] the Euclidean distance,[36], or the Gower coefficient (which implements the Manhattan distance for continuous data)[33, 37, 38] for continuous features. Modha and colleagues [39] implement the Euclidean distance for continuous features with the cosine distance for categorical features. Huang's k-Prototypes algorithm implements k-means for numeric data and k-modes for categorical features.[40, 41] Some approaches introduce cost and weighting functions between feature types to prevent imbalance between data types.[42] For example, Ahmad and Dey proposed an extension of the k-Prototypes algorithm where features are weighted by significance, with similarity and significance calculated as a function of co-occurrence of categorical and (discretized) continuous features.[43, 44] Regardless of the chosen solution, a common, core theme remains the same: clustering of mixed-type data implements a well-established algorithm (or some variation thereof) based on a mixed dissimilarity metric calculated to standardize a mixed data set into a homogeneous set of distances.

Clinical Applications of Clustering Mixed Data

Unsupervised clustering analyses have been used to uncover subgroups within clinical data since the 1960s.[19] Then and now, hierarchical methods have been a dominant approach for the clustering of clinical data.[14, 19-22] Recently, a few studies have applied k-means and k-medoids algorithms to cluster clinical data.[12, 45, 46] Increasingly, studies have emerged comparing traditional hierarchical clustering approaches to k-means, k-medoids, and density-based algorithms.[10, 11, 47, 48]

Studies clustering clinical data apply several approaches to integrating heterogeneous data. These include restricting data sets to only one data type, such as an experiment on Z-normalized continuous data in the critical care setting,[12] normalizing on frequency,[46] or transforming mixed-type data to categorical.[11] However, often no evidence or description of mixed data handling is reported.[45, 47] In many cases, analyses are applied to low-dimensional feature spaces as small as six or seven features.[10, 45] Within the recent studies assessed here, the most common distance metric employed was Euclidean distance, a metric more suited for continuous data, regardless of the data type being clustered.[11, 47] Often, no distance metric was reported.[12, 45, 46]

We identified 4 recent studies, published in the past 3 years, comparing the performance of unsupervised algorithms on mixed-type, clinical data. Yan and colleagues [47] compared Ward's agglomerative hierarchical clustering, k-medoids, and density-based clustering with the OPTICS algorithm to identify subgroups of high-cost patients among a population burdened by comorbidities to inform care management strategies. Data were preprocessed by removal of variables with <1% variance, highly correlated features (Pearson correlation > 0.85) and data imputation. An estimate for the number of *k* clusters was chosen by visualization with t-Distributed Stochastic Neighbor Embedding (t-SNE). All three algorithms were applied using the Euclidean distance, with the winning solution (and value of *k*) chosen by maximizing average silhouette width. Algorithm performance was compared with ridge regression models and inter-cluster variance. The three algorithms returned markedly distinct

results, especially density-based clustering. Hierarchical and k-medoids algorithms produced clusters of relatively uniform size. By ridge regression, k-medoids cluster identities appeared to be driven by a small number of dominant features, while density-based clustering had a larger range for the next 11 variables than either hierarchical clustering or k-medoids, suggesting that a larger number of variables drive subgroup differentiation in the density-based algorithm. However, OPTICS returned clusters of an excessively broad range of sizes (from 3686 to 56 patients in a cluster) with 382 outliers unclustered, suggesting the possibility of poor fit or distortion. The strength of the study lies in its large sample size: 6154 patients with 161 features remaining after pre-preprocessing. However, although the data in the study were mixed-type, Euclidean distance is the only metric used, and there is no evidence reported on ways mixed-type data were handled.

Bose and colleagues [49] compared hierarchical clustering, k-means, and k-medoids on a data set of 557 patients with heart failure utilizing home telehealth services. The Waikato Environment for Knowledge Analysis package was used to reduce 300 first search features to a small feature space of 7 variables for clustering. Clusters were validated by the Dunn Index, silhouette width, and connectivity. Although the study team reports clusters of roughly uniform size (153-233) that differ significantly by chi-square test or one-way analysis of variance, inspection of reported identifying features was discouraging. In a table of characterizing features of each cluster, most of these features are present in less than 60% of members within a cluster, suggesting weakness in the clustering output. The strength of this study is in study

design for mixed data: the team implemented the Gower dissimilarity, which has provisions for the continuous and nominal features present in the data set. However, small feature space limits the ability to truly assess the merits of the methods in comparison.

Two teams have undertaken a comparison of hierarchical and partitioning algorithms in the setting of chronic obstructive pulmonary disease (COPD). Pikoula and colleagues [11] compared hierarchical and k-means clustering on a mixed data set of 30,961 patients with COPD and 15 clinical features. Data type was predominated by binary data and contained continuous and categorical variables. Mixed data were handled by converting continuous variables to categorical, and Euclidean distance was implemented. The number of k clusters was selected by maximizing mean silhouette width. Cluster stability was assessed with iterative resampling of 30% of the training data set. Evaluation was tested against a non-parametric decision tree classifier: an unusual choice for a standard. In resampling, k-means demonstrated higher stability than hierarchical clustering. In an analysis of cluster identities in a small feature space, 89% of variance resulted from 2 features, with low (often <50%) defining features in a cluster. Visualization did not produce discernably separate clusters. The means of mixed data handling in this study is concerning. First, discretization of continuous features to categorical results in information loss. Furthermore, as Chapter 4 will show, the Euclidean distance, although frequently unquestioningly viewed as "default" distance, is inappropriate for categorical features.

Across 10 cohorts of 17,146 patients with COPD, Castaldi and colleagues [10] demonstrated similar performance between k-medoids clustering and general hierarchical clustering. The features space consisted of 7 features: 6 continuous and 1 categorical. Using unsupervised random forests for feature selection and similarity matrix calculation, the team recovered best performance with hierarchical clustering with removal of "poorly classifiable subjects," but in some experiments this resulted in removal of up to 86% of subjects from a clustering approach, which suggests the potential for loss of biological meaning. Three spirometric measurements dominated cluster formation. No clear methods for handling mixed data were provided.

These comparative studies may raise suggestions for the merits of one approach over another, but methodological problems in each raises concerns with these results. First, although many have used large sample sizes from patient cohorts, small feature spaces do not bring with them rich and nuanced information for knowledge discovery. Questionable or excessive methods for feature reduction, such as the implementation by Bose and colleagues [49] of the Waikato Environment for Knowledge Analysis, which reduced 300 features to 7, has real potential to introduce knowledge loss or bias. Similar attitudes towards outlier removal, such as those seen in Castaldi and colleagues [10] raise similar concerns for knowledge loss and bias, while increasing the difficulty of judging the true performance of an algorithm. The problem is compounded by poor handling of mixed data from using inappropriate distance metrics for a given data type [11] to ignoring it entirely [10, 47] Perhaps it is unsurprising that clustering outcomes of these experiments, regardless of type, often produced clusters with limited

coherence. When the percentage frequency of a defining feature in a cluster was reported in a study, the most common features defining a cluster had low frequencies, sometimes less than 50%. This indicates that the recovered clusters lacked strong identities and had reduced potential for clinical discovery.[11, 12, 45, 49]

Outline of the Master's Thesis

Successes in bioinformatics show the promise of unsupervised ML in big data. As available data from cohorts and the electronic health record grow, clinical data – and therefore clinical research and patient care – can benefit from this promise. However, unlike omics data, mixed data types in clinical setting introduce an important problem. Many algorithms, both old standards (e,g., hierarchical clustering, k-medoids) and newer approaches (e.g., SOM) are readily available for clinical applications. These can be accessed through mixed-data centered approaches to calculating similarity and distance, without the need for the development of novel algorithms. The problem of calculating dissimilarity for mixed data remains unsolved, but an encouraging legacy is built on the idea of calculating distance within data types and combining these into a whole. However, these approaches have not been fully and fruitfully applied in the clinical realm. Often, clinical clustering in its current state is plague by poor handling of mixed data, producing poor results.

This thesis probes challenges within methods for unsupervised machine learning of clinical data resulting from heterogeneous size and, particularly, mixed data type. In it, we take 5 major steps to test the hypothesis that clustering methods that calculate dissimilarity on mixed-type clinical data from algorithmic combinations of multiple

distance metrics geared towards individual data subtypes outperform single-distance dissimilarity metrics for clustering knowledge discovery. Here, in Chapter 1, we began this process by establishing context with a survey of relevant algorithms and methods of calculating dissimilarity and their application to date in the clinical literature.

In Chapter 2, we perform an initial experiment on a real clinical data set consisting of clinical features and biomarkers collected on 247 patients with chronic lymphocytic leukemia (CLL). Real clinical data sets lack a validation measure. However, we chose this disease as our case study because prognosis and risk factors of CLL are well understood, providing "biological validation" for our discoveries. In this experiment, we discretize all data types to binary, an approach raised in this chapter, and cluster mixed data as a single type. We find some success, but also elucidate limitations in the method.

In response, we undertake a series of steps to explore best methods for clustering mixed type data. These tests require a gold standard of known cluster identities, which cannot be obtained from real data. Therefore, in Chapter 3 we generate a large series of simulations of mixed-type clinical data with known cluster identities. In Chapter 4, we test 5 distance metrics and 3 common clustering algorithms on these simulations and compare them to reveal the best of these methods for single and mixed data types. In Chapter 5 we return to our CLL data set from Chapter 2 with the best method we uncovered. This final experiment demonstrates improvement but also raises further questions. In summary, we address concluding results in Chapter 6.

Chapter 2. Clustering by Binary Transformation

Calculating appropriate measures of distance or dissimilarity is complicated in mixed-type data sets by the need for different data handling for each data type. The simplest or "most straightforward" approach to mixed-type data handling involves transforming all variables to the same data type.[50] For example, continuous variables can be converted to categorical variables by discretization into intervals.[50] In a clinical data setting, this approach was taken by Pikoula and colleagues,[11] who transformed continuous variables to categorical features for unsupervised machine learning in a mixed-type dataset of electronic health record data on patients with Chronic Obstructive Pulmonary Disorder.

This chapter outlines a preliminary informatic experiment in data transformation before calculating dissimilarity. We present methods for a clustering approach based on rigorous, clinically-grounded decision-making to capture known markers of prognosis and outcome with high fidelity. In this experiment, we hypothesized that unsupervised ML, when applied to clinical data, could discover biologically significant clusters of patients with different prognoses. We chose Chronic Lymphocytic Leukemia (CLL), a disease with well-understood prognosis and outcomes, to "biologically validate" the discoveries generated by our methodologic approach. Using CLL as a case-study, we applied k-medoids clustering to a set of clinical features by transforming them to binary

vectors, exploring 10 metrics for calculating a distance matrix and two common methods of visualization. Although hierarchical algorithms are more commonly implemented in the clinical literature, we used k-medoids clustering in this chapter, as they are more novel and possibly more stable approaches to clinical data clustering. The two primary challenges for either of these approaches (e.g. hierarchical clustering or k-means or k-medoids) are solutions for mixed-type data and associated concerns for selection of an appropriate distance metric. This paper transforms a mixed data set to binary features to eliminate conflict from multiple data types and assesses 10 distance metrics to make a judicious choice to maximize biological meaning recovery.

Materials and Methods

Samples and Clinical Findings

This study uses deidentified data that were previously published. Originally, peripheral blood (PB) samples were obtained from 247 treatment-naïve CLL patients after obtaining informed consent at the University of Texas M.D. Anderson Cancer (MDACC).[51-53] The studies were approved by the Institutional Review Board. Clinical and routine laboratory data were obtained by review of the medical records. These data and sample testing included 23 markers with known and unknown prognostic significance. Key characteristics of the sample are summarize in Table 2.1. The somatic mutation status of immunoglobulin heavy chain variable region (*IGHV*) genes, and ZAP70 expression, measured by either flow cytometry or immunohistochemistry, were assessed on blood or bone marrow samples according to established protocols.[54-56] Mutated *IGHV* status and negative ZAP70 expression are associated with better prognosis. Common CLL-

Table 2.1 Clinical characteristics of Chronic Lymphocytic Leukemia (CLL) patients.

	Patients n (%)
Total	247
Sex	
Male	173 (70.0%)
Female	74 (30.0%)
Race	
Asian	1 (0.4%)
Black	11 (4.5%)
Hispanic	7 (2.8%)
White	228 (92.3%)
Rai Stage	
Low (0-2)	196 (79.4%)
High (3-4)	51 (26.0%)
Döhner Classification	
del13q	90 (36.4%)
+12	37 (15.0%)
FISH normal	73 (29.6%)
del11q	34 (13.8%)
del17p	13 (5.3%)
***IGHV* Mutation Status**	
Mutated	106 (43.1%)
Unmutated	140 (56.9%)
Treatment Status	
Never treated	20 (8.1%)
Treated with FCR	227 (91.9%)
Age at Diagnosis	**Years**
Minimum	26.74
Median	55.87
Maximum	82.41

associated cytogenetic abnormalities were assessed by array-based SNP genotyping.[52, 56] Cases were grouped according to the Döhner hierarchy, which ranks survival from longest to shortest in the following order: del(13)(q14.3); trisomy 12; FISH normal karyotype; (del(11)(q22.3); del(17)(p13.1).[57, 58] Our analysis included seven measures of outcome collected over 15 years of follow-up: overall survival (OS), time

from diagnosis to treatment (TTT), time from sample collection to treatment, event-free survival (EFS), progression-free survival (PFS), time-to-progression (TTP), and survival after treatment (TxOS).

Clinical data transformation

The clinical data are heterogeneous, and include binary, nominal, ordinal, and categorical features. Because our unsupervised ML approach requires homogenized data, we transformed all clinical features to a binary matrix. Reclassifying categorical and continuous data as binary required decision-making steps that inherently result in information loss. So, we compared two distinct approaches to the transformation, which we refer to as "Data transformation A" and "Data transformation B".

Both transformations included several binary features, which can be subclassified into two types. For symmetric binary features, such as sex, both values are about equally likely and there is no reason to prefer coding either value as 0 or 1 in a vector of zeros and ones. In our data set, both the *IGHV* somatic mutation status and ZAP70 expression were symmetric and either presence or absence is relevant to predict clinical outcome. For asymmetric binary features, one of the values tends to be much rarer than the other and is usually coded as a "1". This value is viewed as more informative since people who share the attribute have more in common that people who lack it. For example, clinical features such as anemia, splenomegaly, and hypogammaglobulinemia are asymmetric binary features of our data. For symmetric binary features in both data transformations, we retained two binary vectors – one vector for presence and one vector for absence of a

feature. For asymmetric binary features, we retained one vector capturing a positive result, or presence of a feature.[25]

Discretization of continuous data into categories carries certain pitfalls, including information loss and the criteria by which intervals were chosen. Different approaches were taken based on the variable in question. For example, we binned age along decade lines (e.g., 40-49, 50-59) following common clinical conventions. For prolymphocyte count, which does not have interval conventions in clinical use, we demarcated bin size by plotting a histogram and selecting intervals that were both clinically sensible (e.g. all patients with a count of "0" were placed in a single bin and patients with a prolymphocyte count greater than 10 were placed in a bin, following the cutoff in clinical use) and contained similarly large patient populations. In data transformation A, we preserved categorical and continuous data in more detailed form than in data transformation B.

For categorical data, we transformed each category into binary dummy variables and retained a set of vectors for each category. Thus, for the Döhner classification we retained 5 binary vectors, each corresponding to one cytogenetic abnormality. We binned categorical data along clinically interpretable lines. We binned age by decade and prolymphocyte count by percentage into 6 categories each. We converted these two sets of dummy variables using the same approach that we applied to categorical data. The greatest number of dummy variables for any given category was 6.

In data transformation B, we converted all categorical and continuous features into two clinically meaningful binary categories. Each of these features was divided

along a meaningful clinical cutoff and retained as two symmetric binary vectors. For example, the continuous variable “age” was split into two vectors corresponding to age greater or less than 65 years, and prolymphocyte count was split into two vectors at a cutoff of 10%. Although this method of transformation was smoothly applied to continuous data, Döhner classification, an ordinal variable, could not be collapsed into two meaningful binary categories. Thus, we retained only the Döhner classification as a non-binary set of dummy variables, with a total of 5 vectors.

Unsupervised machine learning

We applied an identical ML workflow to both data transformations. We began with principal component analysis and clustering using the Thresher R package.[59, 60] Using the Mercator R package, we assessed 10 binary distance metrics representing meaningful groupings of 76 distance metrics.[61] Mercator provides streamlined tools for principal component analysis of dissimilarity matrices from different distance metrics, by application of the Thresher algorithm and several types of visualization. To select an appropriate distance metric, we recovered clusters at a range of k values, calculated the categorical distance between clusters, and visualized the similarity between distance metrics with hierarchical clustering. For analysis of both data transformations, we selected the Sokal and Michener distance for representativeness of trends among recovered clusters. Developed as a taxonomic tool, the Sokal and Michener distance $d_{SOKALMICHENER}$ tolerates symmetric binary variables and categorical data.[25, 62] This metric also benefits from ease of interpretability by calculating dissimilarity as a ratio of positive or negative concordant matches to all pairs:[61]

$$d_{SOKALMICHENER} = \frac{N_{00} + N_{11}}{N_{00} + N_{01} + N_{10} + N_{11}}$$

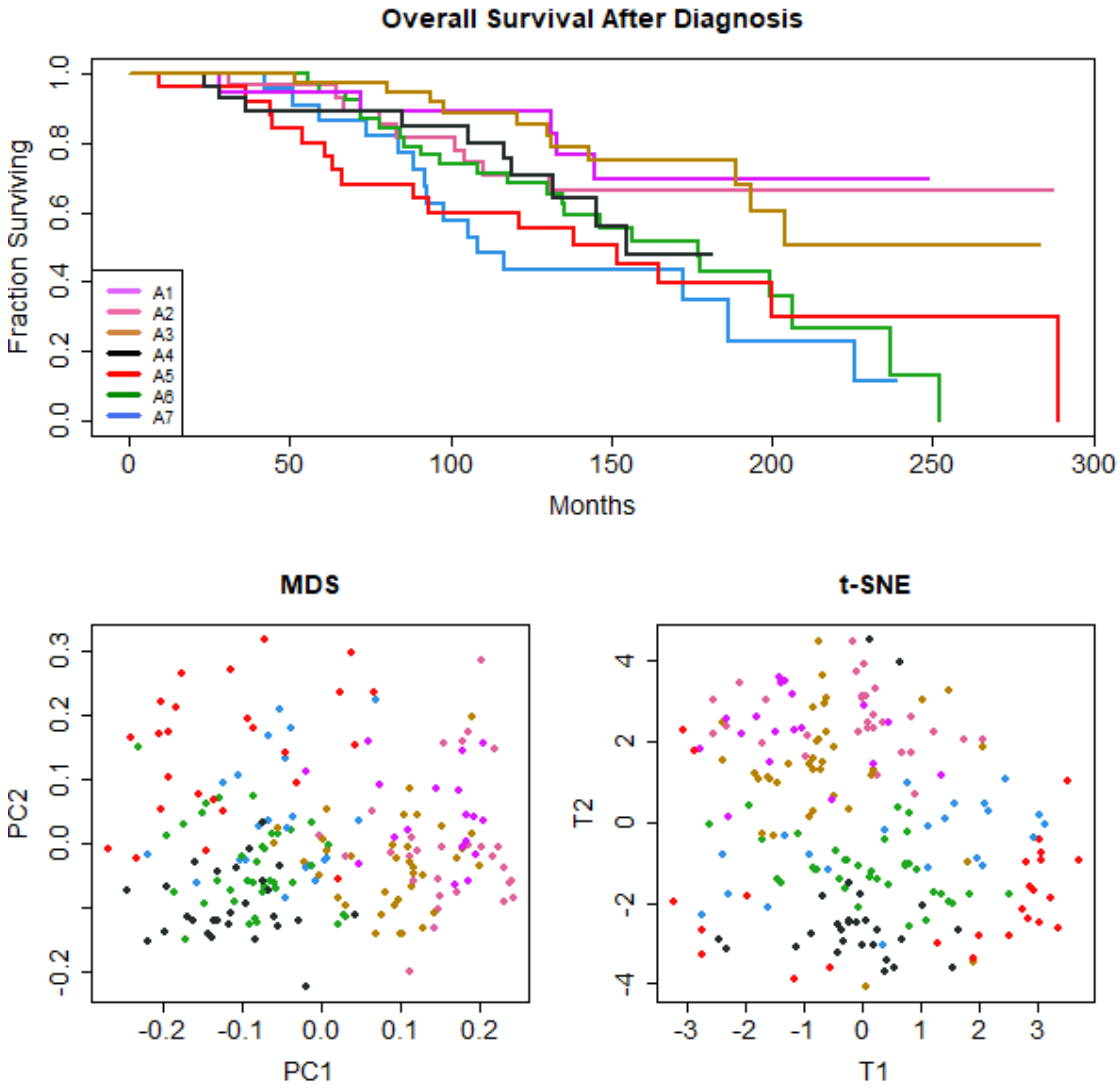
We recovered clusters using the Partitioning Around Medoids (PAM) algorithm.[25] The goodness-of-fit for each cluster was determined from the silhouette width, which quantitatively and visually represents the tightness of clustering as a function of the dissimilarity within versus between groups.[63] The number of clusters was determined by maximizing the average silhouette width. For each cluster, we defined "salient" clinical features as those that characterize greater than 75% of patients within a given cluster. We visualized clusters with both linear (multi-dimensional scaling, MDS) and nonlinear dimension reduction methods (t-stochastic neighbor embedding, t-SNE).[64] To assess prognostic utility and evaluate our methodology, we performed survival analysis using a Cox proportional hazard model, evaluated with the log-rank test and visualized by Kaplan-Meier curves.

Results

Data transformation A

After data transformation A, which preserved categorical features, our dataset contained 39 binary vectors. PAM clustering on a dissimilarity matrix constructed on the Sokal and Michener distance returned k = 7 clusters (average silhouette width = 0.10). Survival analysis with a Cox proportional hazard model revealed a statistically significant association between seven clusters and OS from the time of diagnosis (log-rank test, p = 0.0164; Figure 2.1.A) Visualization by MDS and t-SNE (Figure 2.1.B&C) demonstrated loose clusters arrayed along a gradient in the first dimension that mirror the OS order seen in the Kaplan-Meier curves. Visualizations of the other 9 tested distance metrics

Figure 2.1 Data transformation A: Kaplan-Meier survival curve, MDS plot, and t-SNE plot for seven unsupervised clusters of CLL patients. Unsupervised machine learning, using k-means clustering with Partitioning Around Medoids (PAM) and the Sokal-Michener distance yields seven clinical phenotypes with significant differences in overall survival (OS) (p = 0.0164). Clusters separated by MDS along the first dimension reflect OS outcomes.



using MDS and t-SNE mirrored similar patterns that reflected groupings of distance metric results described by Choi and colleagues.[61](Supplemental Figure A.2, A.3)

Informative features that varied between clusters with differing survival outcomes included *IGHV* somatic mutation status, sex, ZAP70 expression, immunoglobulin light

chain subtype, hypogammaglobulinemia, anemia, and Döhner classification. A subset of informative features are presented in Table 2.2.A, with complete results as Supplemental Table A.1. The three clusters with the longest survival (A1, A2, A3) were associated with mutated *IGHV* status and lack of ZAP70 expression. The cluster with second-longest survival was the only cluster associated with female sex. The two clusters with shortest survival (A6, A7) were associated with unmutated *IGHV* status and ZAP70-positivity, regardless of sex. The clusters with second- and third-longest survival were associated with del(13q), the only Döhner classification abnormality identified by the analysis. Only two clusters were associated with dummy categorical features, specifically del(13q), all other clusters are based on informative binary features. Light chain subtypes lambda (A1) and kappa (A4, A6) were identified as salient features.

Some common features characterized a majority of patients in many of the recovered clusters. In all clusters, 75% or more patients were diagnosed at low Rai stage (Rai stage < III). All clusters featured low CD38 except for cluster A4, which had high CD38. Some clusters were notable only for the absence of a common feature. All clusters had low beta-2 microglobulin except cluster A7. All clusters had low white blood cell counts at diagnosis except cluster A6. All clusters had typical Matutes except clusters A4 and *A7.*

Data transformation B

Collapsing categorical values to binary classifiers reduced the dataset for transformation B to 39 features. Using the Sokal and Michener distance, PAM clustering recovered k = 6 clusters (average silhouette width = 0.17). Transforming categorical

Table 2.2 Informative, identifying features for clustered data transformations A and B. Presented in order of overall survival. Clusters are ordered by predicted survival outcome, from longest survival (A1 or B1) to shortest (A7 or B6). Characteristic features of each cluster, defined as a feature present in at least 75% of members of a given cluster, include known indicators of superior prognosis (*IGHV*-mutated status and female sex) and poor prognosis (ZAP70 positivity). Döhner classification, known to be one of the best predictors of prognosis in CLL, failed to be captured by the analysis for most clusters. For complete results and percentages, see Supplementary Table A.1.

Data Transformation A.

Cluster	Sex	*IGHV* Status	ZAP70	Döhner	CD38	Light Chain	Other
A1	Male	Mutated	Negative		Low	Lambda	Hypogammaglobulinemia
A2	Female	Mutated	Negative	del13q	Low		
A3	Male	Mutated	Negative	del13q	Low		
A4	Male	Unmutated			High	Kappa	
A5	Male	Unmutated	Negative		Low		Anemia
A6		Unmutated	Positive		Low	Kappa	
A7		Unmutated	Positive		Low		Anemia

Data Transformation B.

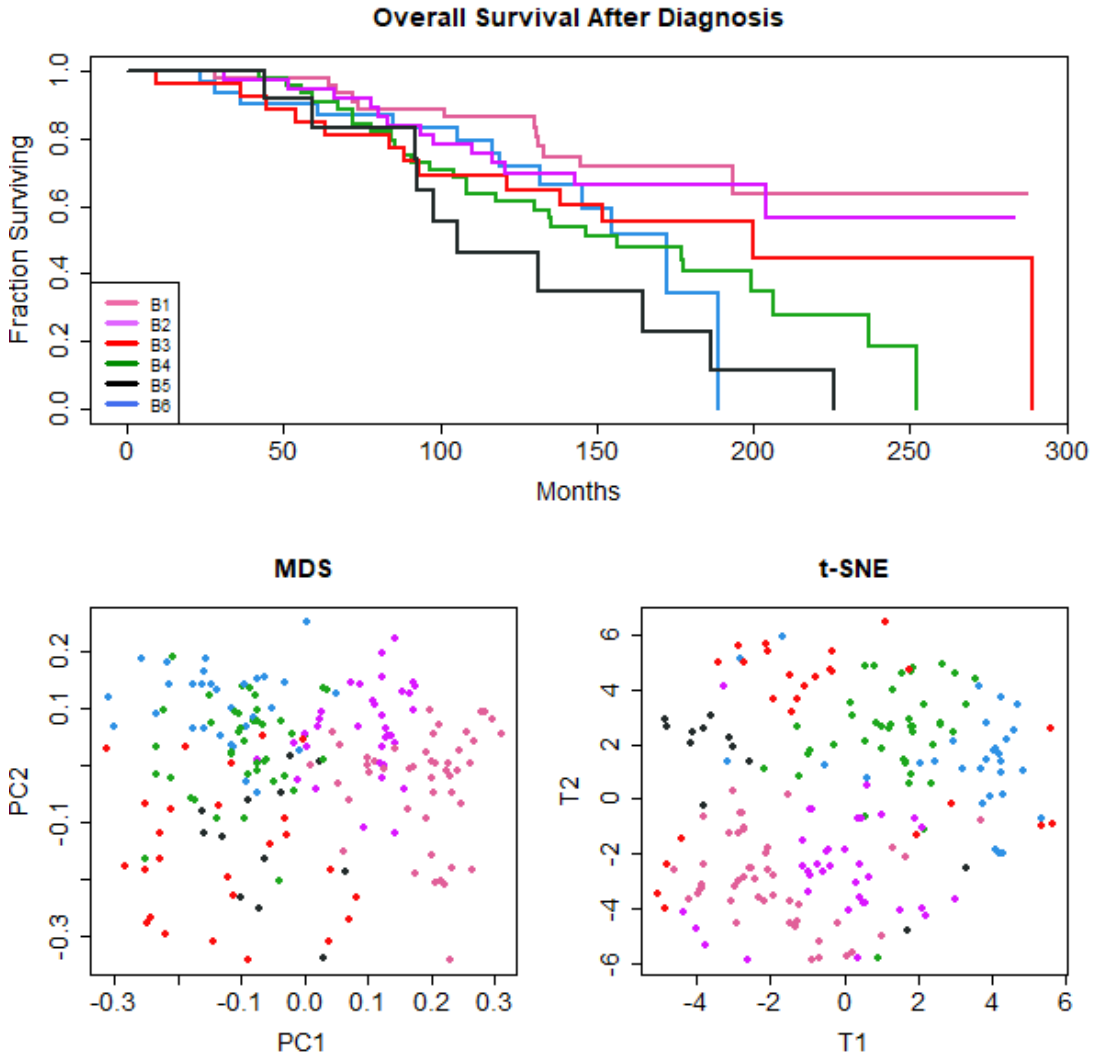
	Sex	*IGHV* Status	ZAP70	CD38	Age	Prolymphocytes	Light Chain	Anemia
B1		Mutated	Negative	Low	Under 65	Under 10	Lambda	
B2		Mutated	Negative	Low	Under 65	Under 10	Kappa	
B3	Male	Unmutated			Under 65	Under 10		Anemia
B4		Unmutated	Positive	Low	Under 65	Under 10	Kappa	
B5	Male	Unmutated	Positive	Low	Over 65		Lambda	Anemia
B6	Male	Unmutated	Positive	High	Under 65	Under 10	Kappa	

features to binary form improved silhouette widths. Survival analysis by Cox proportional hazard on several outcome measures revealed statistically significant associations between six recovered clusters and TTP (log-rank test, $p = 0.0451$; Figure A.1) and the related metric, time from diagnosis to treatment ($p = 0.0039$; Figure A.1). Visualization by MDS and tSNE displayed loose clusters along a gradient that mirrored OS, but not other outcome measures, such as TTP (Figure 2.2.B&C). However, even though MDS separates clusters on the first dimension along the order of OS, the association between clusters and OS was not statistically significant ($p = 0.391$).

Informative features that defined 75% of the patients in a given cluster, ordered by separation on MDS first dimension and OS, are presented in Table 2.2B, with complete results in supplemental Table A.1. Clusters with improved OS had mutated *IGHV* status and ZAP70-negativity. Clusters with shorter OS had unmutated *IGHV* status and ZAP70 positivity. The cluster with second-shortest survival was associated with older age (greater than 65 years) at the time of diagnosis. The cluster with shortest survival was associated with CD38 positivity. As in data transformation A, lambda (B1, B5) and kappa (B2, B4, B6) were alternately identified as salient features.

As in data transformation A, some common features were represented in many clusters. For example, all clusters were associated with low Rai stage, except omissions in clusters B3 and B5. All clusters had typical immunophenotypes by Matutes score, except clusters B5 and B6, which had fewer than 75% of members with typical immunophenotypes, but no cluster was characterized by atypical immunophenotypes by Matutes.

Figure 2.2 Data transformation B. Kaplan-Meier survival curve, MDS plot, and t-SNE plot for six unsupervised clusters of CLL patients. Unsupervised machine learning, using k-means clustering with Partitioning Around Medoids (PAM) and the Sokal-Michener distance yields seven clinical phenotypes with significant differences in time-to-progression (TTP) (p = 0.0451).(Supplemental Figure 1) Clusters separated by MDS along the first dimension reflect order of overall survival (OS) outcomes.



Discussion

Applying methods common in bioinformatics to clinical data entails potential problems and pitfalls. The difficulties of these approaches are rooted in the nature of clinical data itself. The most important hurdle to overcome is clinical data's heterogeneity. Our analysis captured symmetric, binary classifiers with high fidelity. Three of the best-understood prognostic features in CLL are sex, *IGHV* mutation status, and ZAP70 expression. Both data transformations identified these three features as salient and informative. Some common features proved uninformative because they characterized a majority of patients in most or all of the recovered clusters. Binary features for which one of the two categories predominated within the dataset, such as Rai stage or white blood cell count, were sufficiently common as to be identified as salient features in each cluster by our 75% cutoff. Such features are meaningless due to their frequency across the data as a whole.

In data transformation A, a high proportion of categorical and binned continuous data led to loose clusters and low silhouette widths. Salient clinical features identified by our workflow failed to capture meaningful categorical data, including age, a well-understood prognostic indicator. These limitations led us to explore data transformation B. Collapsing categorical data to binary form improved silhouette width and led to the inclusion of two important classifiers, age and prolymphocyte count, in cluster definitions.

Both data transformations captured light chain subtype (lambda or kappa) as a salient feature for the majority of clusters. Light chain subtype is a commonly recorded

variable with known function in physiological B-cell development and differentiation. However, its role as a prognostic indicator is poorly understood. Furthermore, although light chain subtype helps distinguish cluster identities, the alternating pattern of light chain subtype along the survival spectrum in transformation B suggests that association with overall survival is unlikely. Our analysis captured several pairs of clusters differentiated by few features other than light chain subtype. In data transformation B, clusters B1 and B2 are differentiated only by light chain subtype. Applications of unsupervised ML to clinical data hold the potential to explore other poorly understood clinical features and their role in predicting treatment response or survival outcomes. Future work remains to refine methodologies to elucidate these clinical features.

Critically, neither transformation succeeded in capturing what is perhaps the most important, best understood, prognostic indicator in CLL: the Döhner classification. Data transformation A identified a Döhner abnormality in two clusters only, and data transformation B failed to identify Döhner classification for any cluster at all. The Döhner classification was the only ordinal feature that could not be meaningfully collapsed into a binary form, which may explain why it was not well captured by either model. Another possible explanation arises because the Döhner classification is strongly associated with both *IGHV* mutation status and ZAP70 expression. Cases which only have del(13q) tend to be *IGHV*-mutated and negative for ZAP70 and have good prognosis. Cases with del(17p) or del(11q) tend to be *IGHV*-unmutated and positive for ZAP70 and have poor prognosis. Nevertheless, this finding suggests that important categorical factors with many levels may have less influence on clustering than associate

symmetric binary factors. It also reflects the fact that (by definition) unsupervised ML methods are inherently less powerful than supervised methods at finding factors relevant to a particular clinical outcome; we know that the Döhner classification is prognostic because it was found during supervised analyses of clinical data from CLL cohorts.

A simple binary transformation and subsequent application of dissimilarity metrics uniformly across a clinical data set is clearly insufficient to capture all medically important facets within the data. Collapsing age to a binary classifier of greater or less than 65 years successfully led to its inclusion as a salient feature in data transformation B. However, patient ages in the data set ranged from less than 40 to over 80 years old at diagnosis. Clearly, rich and important clinical information was lost with at least some binary transformations. Ideally, the power of an application of unsupervised ML to clinical data would be in capturing details of clinical significance not previously identified. Collapsing continuous data to a binary classifier may prevent the realization of this important potential.

Clinical data are inherently complex. Our dataset, though small, is representative of this complexity. These data contain features that are symmetric, balanced, and binary (e.g., sex); symmetric and binary, but strongly unbalanced (e.g., Rai stage); binary but asymmetric (e.g., anemia); nominal (e.g., Döhner classification); continuous on an interval scale (e.g., age); and continuous on a ratio scale (e.g., prolymphocyte count). Any unsupervised ML approach must capture and leverage this complexity.

A primary methodological concern of our analysis and future directions is fitting an appropriate distance metric to a given problem. Here, we selected the Sokal and

Michener distance for appropriateness of data type, representativeness of other distance metrics, and representativeness of trends within our data. First, the Sokal and Michener distance, although originally developed for small, categorical data,[62] is appropriate for use in symmetric binary data,[25] such as the important features in our data that are prognostically meaningful both when absent or present. Second, the Sokal and Michener distance produces results highly correlated with other well-understood measures of binary distance, including Manhattan, Minkowski, and Gower distances, so we can view the Sokal and Michener distance as representative of other approaches to calculating dissimilarity. Finally, when visualizing our data across 10 distance metrics, the Sokal and Michener distance qualitatively reproduced plotting trends across multiple methods of calculating dissimilarity.

Although clustering is a common approach in bioinformatics, both current bioinformatics and future clinical informatics applications can benefit from careful attention to this problem. Within the realm of bioinformatics, homogeneous datasets can easily be subjected to a single dissimilarity metric. However, analysts typically resort to software defaults, such as the Euclidean distance for continuous metrics, as opposed to selecting the metric that best fits the particular experiment. Failure to disclose distance metrics in the construction of a dissimilarity matrix or linkage metrics in hierarchical clustering is an impediment to reproducibility. In 1980, in response to publication of cluster experiments characterized by insufficient methodological reporting to allow reproducibility, Blashfield [19] called for reporting of the chosen similarity metric in all published clustering analyses. Forty years later, this recommendation and reporting need

still stands. Although using a default measure is convenient, thoughtlessly applying an artificial, mathematically-constructed distance metric may rest on faulty assumptions that an arbitrary metric can correspond to meaningful biological reality.

Many distance metrics currently used in bioinformatics have their roots in taxonomic and speciation problems of the early- to mid-twentieth century.[61, 62] Clustering remains, in many ways, a taxonomic problem. Creating meaningful biological classifications requires thoughtful assignment of a distance metric to a particular set of data. Any solution for clustering clinical data must capture relationships between data types without information loss. Although the heterogeneity of clinical data stresses the most complex aspects of this problem, we argue that exploring multiple distance metrics to select the best fitting calculation of dissimilarity for a given data set should be an integral step in any unsupervised ML workflow. To tackle the heterogeneous data problem, Kaufman and Rousseuw [25] suggest clustering a dissimilarity matrix, as opposed to raw data. They sub-compartmentalize distinct data types, each requiring different solutions and metrics in the construction of a dissimilarity matrix, including symmetric or asymmetric binary data, ordinal, nominal, interval-scale continuous, and ratio-scale continuous. Each data type is subjected separately to targeted distance calculations, then recombined in a dissimilarity matrix for clustering. This methodology and more elegant solutions merit further exploration. The remainder of this thesis undertakes a series of experiments to probe the hypothesis generated in this preliminary experiment further.

Chapter 3. Simulations of Realistic Clinical Data

Any evaluation of clustering methodologies in clinical data demands a known validation standard for comparison. The experiment described here in Chapter 2 relied on "biological validation": testing a clustering methodology on a disease with well-understood relationships between clinical features and patient outcomes. However, there are important underlying assumptions beneath biological validation as an approach. First, one of the primary benefits of understanding clinical data through clustering analyses is improved understanding of the interactions between features, both correlation and exclusion. In many clinical cases, interaction effects are far less understood than the individual effects of features. When interactions between features are identified in a clustering experiment, biological validation leaves us unable to assess the quality of our emergent findings. Secondly, we must assume that most or all information about the clustered features, their relationships to each other, and their relationships to the disease are known. This leaves us unequipped to deal with surprises: either new discovery, or the discovery that something previously believed is revealed to be wrong. The goal of any clustering experiment should be to uncover novel findings. However, biological validation leaves us unable to assess the validity of any novel findings that emerge through the approach.

The best solution to the shortcomings of biological validation is the identification of a gold standard for analysis in which all cluster assignments of each feature are known. If it were possible to identify these with full certainty from clinical data, no further experiments in this thesis would be needed. However, artificial clinical data, simulated with known cluster identities, can serve to test and validate unsupervised ML algorithms.

Here, when we refer to "clinical data," we refer to a data set characterized by heterogeneity and measurement in a clinical setting. Clinical data sets vary widely in scale, from early-stage clinical trials with fewer than 100 patients to prospective cohorts following 10,000 patients to large-scale mining of electronic health records. They consist of data collected in the clinical setting, including demographic information, laboratory values, results of physical exam, disease and symptom histories, dates of visits or hospital length-of-stay, pharmacologic medications and dosing, and procedures performed, possibly with associated ICD-9 or -10 codes. The most salient, identifying feature of clinical data is that it is of mixed-type, consisting of continuous, categorical, and binary data.

Simulating realistic clinical data faces may challenges. The wide range in feature spaces and sample sizes demands simulation solutions that vary by orders of magnitude. Rather than simply simulating data of a single type, simulated clinical data must be of mixed-type, and must reflect the variation in distribution of these types found in clinical scenarios: Frequently, clinical data sets with mixed types of data are predominated by one type over the others. In addition, in order to conclusively test algorithms for use in clinical contexts, simulations of clinical data must replicate the noisiness of these data

that results from variation in human and technological features in measurement and the biological variation between individuals.

Although simulating realistic clinical data poses challenges, a real need exists, both for the experiments conducted in this thesis and for researchers at large, for these tools. As we demonstrate later in this chapter, there are a paucity of publicly available solutions. Clinically meaningful simulations are vital for testing and developing superior machine learning techniques for new clinical data challenges. This chapter offers a tool to meet this need.

In this chapter, we briefly review existing approaches to simulate mixed-type data. Finding the need for new development in this area, we identify an existing R-package, Umpire, which was developed to simulate clusters of continuous, gene expression data. Although a package able to directly simulate mixed-type data would facilitate our process, mixed-type data can be constructed from judiciously binning continuous data to create categorical and binary features. We describe a process to adapt the Umpire gene expression simulations to the characteristics and behavior of clinical data, creating appropriately noisy, mixed-type data. Finally, we describe a cohort of simulations that could represent the range of clustering problems in a clinical context and apply these methods to generate many simulations. In Chapter 4, we describe the use of these simulations to test dissimilarity and clustering methods for mixed-type data.

Table 3.1. Parameters for simulations of clinical data.

Patients	200, 800, 3200
Features	9, 27, 81, 243
Clusters	2, 6, 16
Data types and mixtures	Single data type: continuous, binary, nominal, ordinal, mixed categorical* Mixtures: balanced, continuous unbalanced, binary unbalanced, categorical unbalanced

Representative Clinical Simulations for Testing Mixed-Type Clustering Methods

Our goal was to create a set of simulations that represent a spectrum of important problems that could be encountered in clinical data contexts. These simulations should meaningfully reflect sample sizes corresponding to multiple common study designs and mixtures of data types reflecting that seen in the published literature. In addition, the data we simulate must represent heterogeneity from biological variation and methods of data capture characteristic of clinical data. By generating representative structure and noisiness, we will be able to generalize the results of our tests algorithms to real data situations.

The central purpose of these simulations is to generate mixed-type data. However, our review of the literature encountered many data sets were composed of unbalanced mixtures predominated by a single data type.[10, 11, 65] For this reason, we chose to simulate 9 different mixtures of data types. As controls, we generated noisy but unbinned continuous data and control data sets of binary or categorical data alone. Because nominal and ordinal data require different clustering solutions [25] and because categorical data pose unique challenges in unsupervised ML, [38] we generate three types of categorical controls: data sets that are exclusively nominal, exclusively ordinal, and those containing a mixture of nominal and ordinal features. Mixed type data were

generated in four mixtures: three mixtures that were dominated by continuous, binary, or categorical data, respectively, and a mixture that equally balanced all three data types. Balanced mixtures were 1/3 composed of each category. Unbalanced mixtures were 7/9 composed of one dominant data type, with 1/9 dedicated to the remaining two types. As such, parameters for number of features in a simulation were selected as powers of 9.

We generated a comprehensive set of simulations designed to capture the breadth of clinical data applications seen in our review of the literature conducted in Chapter 1.(Table 3.1) Clustering analyses have been implemented on data sets ranging from clinical trials with less than 200 patients and 10 or fewer features [10, 49] up to large, longitudinal cohorts with following 200 or more features on 1,000 or more patients.[12, 47, 65] We represent this spectrum with three simulated patient populations, such that 200 simulated patients represent a clinical trial; 3,200 simulated patients represent a large cohort; and 800 simulated patients represent a study of moderate size between these. Although many studies uncover small numbers of clusters (e.g. 4 or fewer), we created simulations open to the possibility of clinical data sets with many clusters. We create simulations of the minimum possible number of clusters (e.g. 2). Simulations with 6 clusters represent a moderate number of clusters, as reflected in the literature. Simulations of 16 clusters represent a very large number of clusters that is hypothetically possible, but also a larger number of clusters than was seen in our review of the literature, thereby allowing us to capture a full spectrum of cluster possibilities. Finally, we generated 100 repeats of each simulation profile for a total of 32,400 simulations.

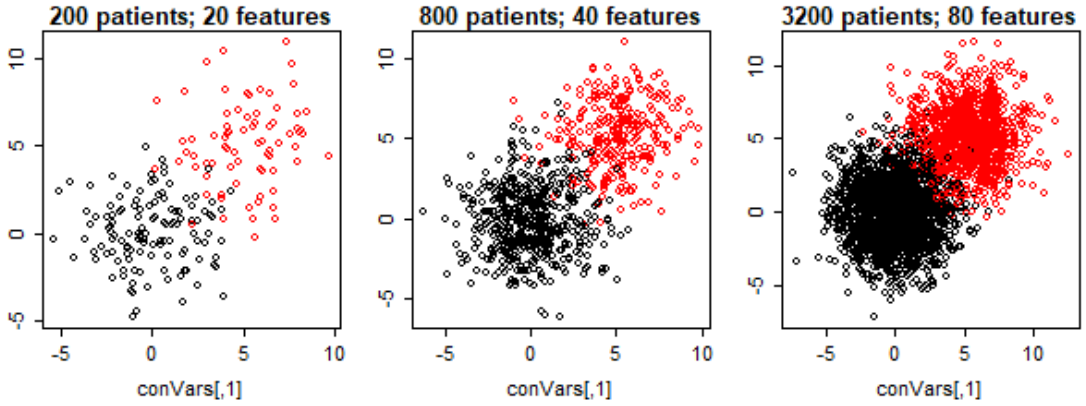
Existing Tools to Simulate Mixed Data

We began this process with a search for existing R-packages to simulate mixed-type data. From this search, we uncovered one extant tool: KAMILA (*k*-means for mixed large data). KAMILA is an R-package implementing an algorithm for k-means clustering of mixed continuous and categorical data and a model for generating mixed-type simulated data.[66] KAMILA simulates mixed-type data sets, allowing the user to control cluster separation, parameterized as an overlap of cluster densities; to control the proportion of number of categorical and continuous variables; to assign error to a certain number of these variables; and to select the relative prevalence of each cluster as a vector of probabilities. KAMILA can be easily implemented to produce clusters that suggest clinically meaningful structure.(Figure 3.1) However, KAMILA fails to provide functionality crucial to our analyses: KAMILA can only be used to generate 2 clusters. Knowing that clinical data often contain more than two clusters, potentially many more, we cannot meaningfully test algorithms for a wider variety of mixed-type data problems without simulating data sets with a wider variety of numbers of clusters. Therefore, we proceeded to explore other solutions.

Methods to Simulate Mixed-Type Clinical Data

Unable to find an available tool to simulate complex, clinically realistic, mixed-type data, we chose to build our own as an extension of an existing tool for simulating gene expression data. The Ultimate Microarray Prediction, Inference, and Reality Engine (Umpire) was published in 2009 to “simulate complex, realistic microarray data” with a known, “ground truth” underlying structure.[67] Although Umpire was developed to

Figure 3.1. Plots of simulated mixed-type clusters generated using the KAMILA R-package. KAMILA (k-means for mixed large data) can be implemented to simulate mixed-type data reflecting a variety of patient populations and feature spaces. However, KAMILA can only be used to simulate data with a maximum of 2 clusters.



simulate gene expression data, parallels between microarray data and clinical data realities allow for translation of the method to simulate data that captures the behavior of clinical data. Instead of simulating cancer subtypes from many expressed genes measured across a sample, we re-frame our thinking to simulate clusters from clinical features measured across a sample of patients and adapt our models accordingly. Following the Umpire model, we ascribe cluster membership to the perturbation of correlated blocks of features, representing functional biological groups, by latent "hits" resulting from underlying etiological processes. To these simulations, we apply noise following a clinically representative *Noise Model*. We segment these data into binary or categorical features. Finally, we combine binary, categorical, and continuous features into mixtures to create mixed-type data sets.

UMPIRE: A Base for Mixed-Type Simulations

The Umpire R-Package can be used to simulate complex, heterogeneous data with known cluster identities and survival outcomes. Originally developed to simulate microarray

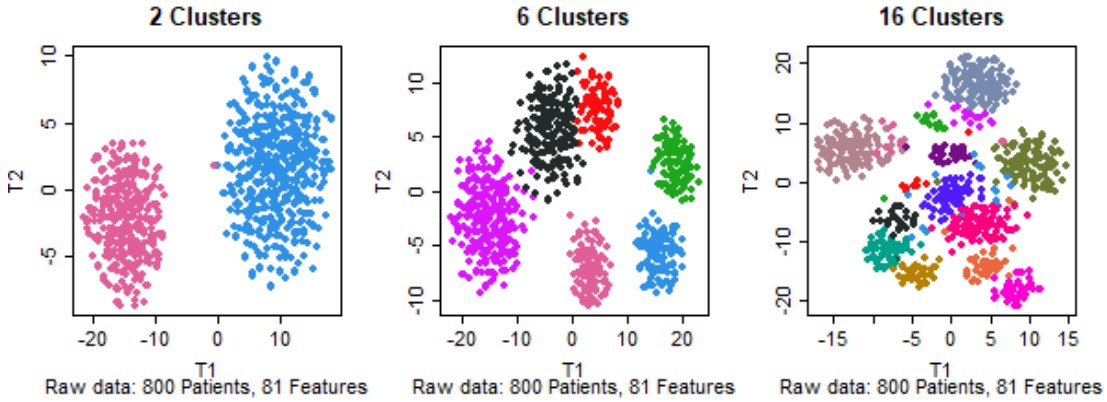
data, Umpire simulates continuous gene expression data. Genes are not simulated as independent entities, but rather as correlated blocks of fixed or variable size developed to simulate the functioning of genes in complex biological networks and pathways. Umpire simulates data with known clusters ("cancer subtypes") of pre-defined size ("prevalence") based on a multi-hit model of cancer. Under this model, cluster identity is defined by a number of informative, latent variables ("hits"). Each subject receives a combination of multiple "hits," simulating population heterogeneity. The user defines correlation and cluster behavior for an unrestricted number of clusters. Each cluster is simulated with paired survival data, in the form of a binary outcome and length of follow up. To these simulated data, Umpire can be used to apply additional noise, mimicking biological variation and experimental random error. The final results of Umpire simulations are large, continuous data sets with "realistic" noise reflecting gene expression experiments with known cluster identities and survival outcomes.[67]

Simulating Complex Data Structure

Meaningful simulations of clinical data must represent the complexity, biological variation, and measurement error that accompany patient data. Because we know that clusters of equal size are unrealistic, we generate clusters of both equal and unequal sizes to increase the complexity in the structure of simulated clusters. We sample a vector r of sizes of k clusters of unequal prevalence from the Dirichlet distribution $r \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$, choosing a set of parameters α that generate a wide variation in cluster size while ensuring that all clusters have patient members. For small numbers of clusters (e.g., $k = 2,6$), we set all $\alpha = 10$. For larger numbers of clusters (e.g., $k = 16$), we

set one quarter each of parameters α to 1, 2, 4, and 8, respectively, accepting only a vector of cluster sizes r in which every cluster has at least 1% of patients as members. (Figure 3.2)

Figure 3.2. Simulated clusters with heterogeneous cluster sizes. Clusters are simulated with variation in size of membership, ensuring that, in simulations with large numbers of clusters, each cluster has at least 1% membership. Heterogeneity in cluster size is present across a range of number of clusters in a simulation.



Simulating Clinically Meaningful Noise

Clinical data are frequently victim to complex noisiness. Marlin and colleagues [68] argue that all clinical data "must be treated as fundamentally uncertain" due to human error in measurement and manual recording, variability in sampling frequencies, and variation within automatic monitoring equipment. Clinical experience teaches us that noise in clinical data arises from many sources. Here, I illustrate this with the example of the measurement of blood pressure, one of the most common clinical measurements taken. Some error arises from frequent changes in the person performing a measurement (e.g., nurses coming off and on shift, who may take a blood pressure with the same technique but achieve slightly different measurements). Random error also arises from changes in a measurement device (e.g., two separate clinics or two nurses in the same

clinic that take a blood pressure measurement differently – one with an automatic cuff and one by hand). Measurement error can arise from variation in the patients' posture at the time of measurement (elevated with crossed legs), from the patient's behavior an hour before arriving for the measurement (a cup of coffee or a cigarette), or from the time of day (lower in the morning). Noise in clinical data may also arise from individual biological variation that causes deviation in some measurements without association with the disease process of interest. However, because clinical measurements are integral to the provision of patient care, demanding high accuracy and reliability, we also assume that many clinical variables have low measurement error. For example, we would expect almost no error in the measurement of height, where deviations from time to time of half an inch on an adult 65 inches tall would be insignificant.

Umpire simulates additive and multiplicative noise on top of a simulated continuous data set. The true biological signal S_{li}, distorted by additive noise E_{li} and multiplicative noise H_{li} results in the observed signal Y_{li}:

$$Y_{li} = \exp(H_{li})\, S_{li} + E_{li}$$

The noise terms H_{li} and E_{li} are normally distributed, and the additive noise includes a bias term (ν). In the gene expression context, this bias term represents a global elevation in feature measurements unrelated to disease outcomes, from sources such as a low level of cross-hybridization across an array, contributing some level of signal at all genes. The multiplicative noise term *H* represents the experimental factors in gene expression data commonly related to normalization problems.[67]

When we apply Umpire preset values for additive and multiplicative noise, we generate noise that is unrealistic both in conceptual model and quantity.(Figure 3.3) An

excessively high quantity of noise produces so much noise that simulated clusters lose coherence, with no identifiable clusters on visual inspection with t-SNE (Euclidean distance). This reflects a lack of fitness of the existing Umpire noise model for clinical data, with invalid key assumptions:

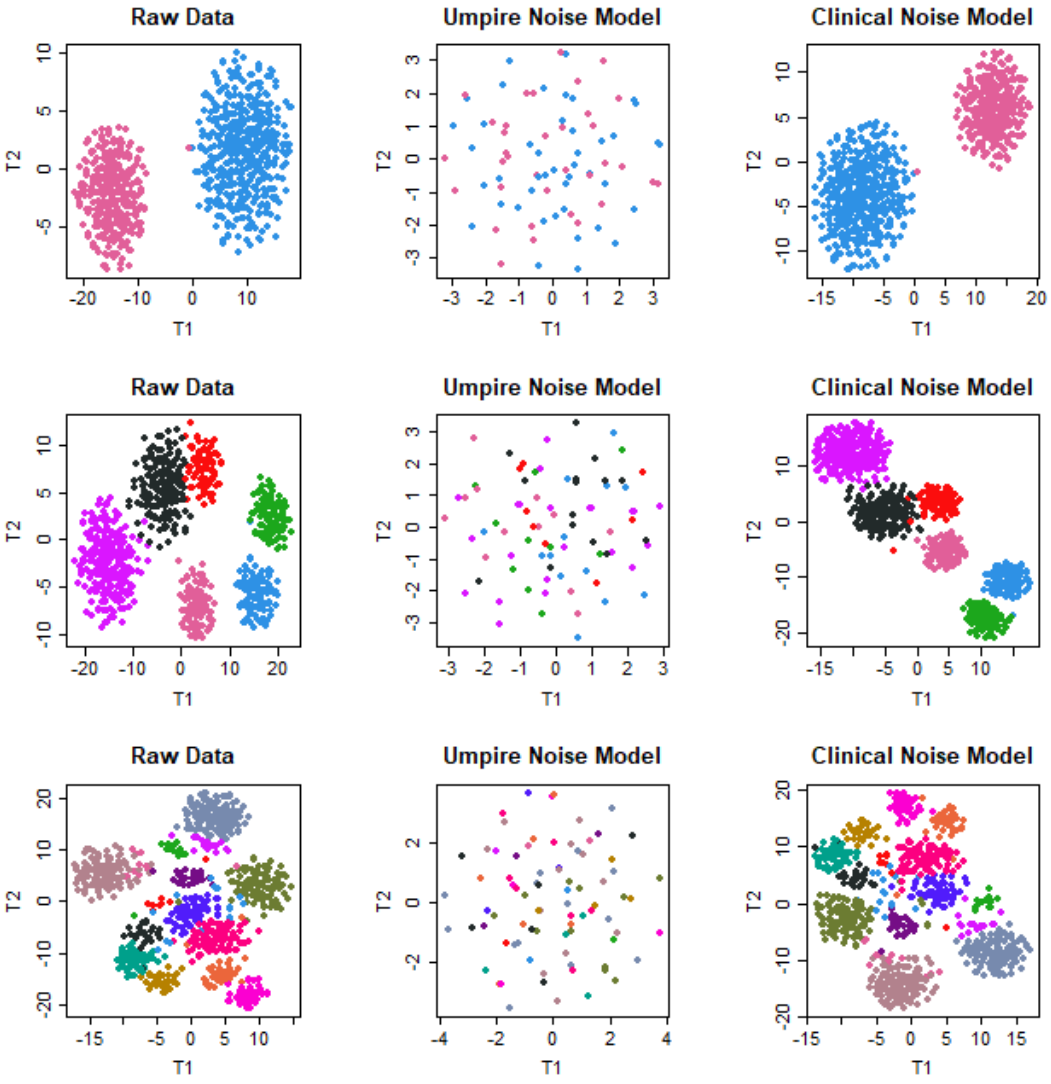
1. Gene expression data have much larger values than clinical data (e.g., 0-65,000; mean 1000 or range 0-16 and median 3-4 on a log2 scale). Thus, Umpire's parameter defaults are improperly scaled.
2. The additive noise *E* bias term ν, simulating global elevation in mean additive noise, represents cross-hybridization across an array: an experimental situation without clinical analogue.
3. The multiplicative noise term *H* represents the experimental factors in gene expression data commonly related to normalization problems, for which there are no clinical correlates.

We assume that the origin of most noise in clinical data from machine measurement noise. Within our simulations, for a given feature *f* measured on patient *i*, we model the observed signal *Y* from additive measurement noise *E* applied to the true biological signal *S* (a raw, simulated data set).

$$Y_{fi} = S_{fi} + E_{fi}$$

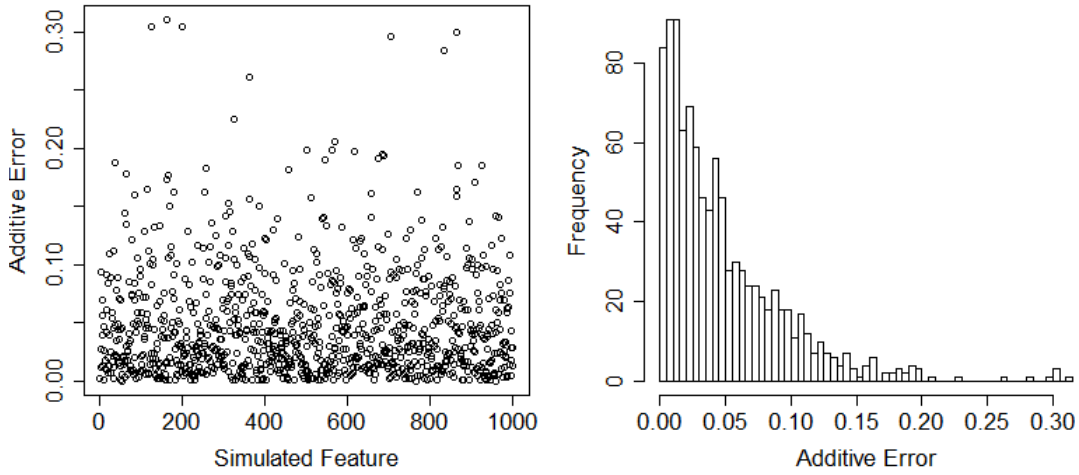
We model the additive noise following the normal distribution $E \sim N(0, \tau)$ with mean 0 and standard deviation τ.[67]

Figure 3.3. Two models for simulating experimental noise on clinical data. Using a simulated patient population of 800 patients with 81 features to represent a clinical experiment of moderate size, we simulate a range of cluster numbers and visualize raw data with 2 comparative noise models with t-Stochastic Neighbor Embedding (t-SNE) plots. Raw data without noise (left) are compared to competing models for experimental noise simulated by Umpire defaults for gene expression data (center) and by our clinical noise model (right). Compared to the clinical model, the gene expression model generates excessive noise that elides cluster identities. Application of a clinical noise model to raw, continuous, simulated data qualitatively results in clusters with mild diffusion without obscuring coherent cluster identities.



In a clinical context, we assume that many features have very low noisiness (such as height or calibrated, automated lab values) and a small number have high noisiness (e.g. blood pressure), we model τ following the gamma distribution $\tau \sim Gamma(c, b)$ such that the mean standard deviation of the additive noise $bc = 0.05$. Thus, we create a distribution in which most features have very low noise while some are victim to very high noisiness.(Figure 3.4)

Figure 3.4. A gamma distribution defines the standard deviation of additive noise in the clinical noise model. The noise is modeled with variability in the standard deviation of the noise, such that many features have low noise, but some have high noise (large standard deviations), as visualized in a scatterplot (left) and histogram of frequencies (right).



Simulating Binary Features from Continuous Data

We convert a continuous feature into a binary vector by selecting a cutoff and assigning values on one side of this demarcation to “zero” and the others to “one.” Ideally, we wish to divide a continuous feature at a meaningful level to create a binary feature in which the two components differ statistically. For each continuous feature we wish to convert to binary, we begin by calculating a “bimodality index” for the vector.[69] The bimodality

index assumes that a vector with bimodal expression can be described as a mixture of two normal distributions, calculated from the fraction of members in one distribution π or the other standardized distance between the means of the two populations $\delta = (\mu_1 - \mu_2)/\sigma$:

$$BI = [\pi(1-\pi)]^{1/2}\delta > 1.1$$

We bisect a feature with bimodal distribution, defined as a bimodal index of 1.1.

Although the bimodality index allows a statistically meaningful partition of a continuous feature into a binary vector, not all continuous features obey this distribution. For continuous features without bimodal distribution, we partition them to binary features by selecting an arbitrary cutoff between 5% to 35%. Although arbitrariness feels uncomfortable in an informatics sphere, we believe that this approach reflects a fundamental arbitrariness in many clinical definitions. For example, immunohistochemistry describes a histological technique in which a specimen slide is stained for expression of a biomarker of interest. This marker is evaluated visually by a pathologist by microscopy, who counts the frequency of this biomarker in cells of interest. A specimen is considered positive for the biomarker if some percentage of cells of interest, such as 20%, express the biomarker. Cells or persons with 19% expression probably show little difference from persons with 21% expression, and even making this distinction would be difficult. In binary clinical definitions resulting from physical exam, such as the assessment of hepatomegaly, or enlargement of the liver, physician judgement based on palpation is used to make the assessment. Like immunohistochemistry, there is a certain subjectivity in the measurement, even though it is obtained by a highly trained physician, and there is no precise demarcation where the separation lies. This

arbitrariness of cutoff pertains also to quantitative values, such as lab assessments. For example, an adult female with a hemoglobin of 12.0 is said to be anemic, even though the clinical presentation and symptoms of a woman with a hemoglobin of 11.9 probably do not differ from those of a woman with a hemoglobin of 12.1. The choice of an arbitrary cutoff reflects these clinical decision-making processes: along a spectrum of phenotype, a value is chosen based on experience to define the edge of the syndrome. By choosing an arbitrary cutoff, we replicate this process.

To reduce bias that could result if all low values were assigned "0" and all larger values were assigned "1," we randomly choose for values above or below the cutoff to be assigned 0. We mark binary features in which 10% or fewer values fall into one category as asymmetric and mark the remainder as symmetric binary features.

Simulating Categorical Features from Continuous Data

To simulate a categorical feature, we rank a continuous feature from low to high and bin its components into categories, which we label numerically (i.e. 1, 2, 3, 4, 5). From our knowledge of clinical data, we know that common categorical features have relatively few categories, so, for each feature, we sample a number of categories between 3 and 9.

There are several approaches for drawing the cut points to bin a continuous feature. Distributing an equal number of observations into each bin does not reflect the realities we see in our data, so we eliminate this option. Dividing a continuous feature by values (e.g., dividing a feature of 500 observations between 1 and 100 into units of 1-10, 11-20, etc.) could lead to overly disparate distributions of observations into categories.

Here, we risk very large categories at intermediate values and sparse tails. For *c* categories, we model a vector of *R* sizes along the Dirichlet distribution,

$$R_c \sim Dirichlet(\alpha_1, \ldots, \alpha_c)$$

$$\alpha_1 = \cdots = \alpha_c = 20$$

such that we create categories of unequal membership without overly sparse tails.

To generate an ordinal categorical feature, we bin a continuous feature and number its bins sequentially by value of observations (e.g., 1, 2, 3, 4, 5). To generate a nominal categorical feature, we number these bins in random order (e.g., 4, 2, 5, 1, 3).

Evaluation of Mixed-Type Simulated Data

Methods to Evaluate the Quality of Mixed-Type Simulated Data

How can we identify successfully simulated clusters? Here, we define "good" clusters as having two characteristics:

1. *Identity*. The cluster is defined (identified) by a core group of salient features that are strongly associated with that cluster assignment. Among real data, these clusters could be assessed to have strong, high-frequency features that distinguish one cluster from another.
2. *Distinctness*. Distinct clusters are tightly grouped with separation from other clusters with minimal overlap. Separation between one group and another can be easily discerned.

Because we wish to simulate realistic data, however, we seek to simulate clusters that possess both these properties "in moderation," to mimic the noisiness of real data. We undertake tests of both of these properties iteratively in the development of our

simulation methods, first confirming the quality of continuous simulations with realistic additive noise, then confirming the quality of simulations of a single binned type, and finally assessing data mixtures. In all cases, the clusters we assess in the evaluation of these simulations are the "known" simulated cluster identities, without testing an additional clustering algorithm.

We assess the property of *identity* quantitatively by looking at the number and strength of features defining a cluster relationship for each generated type of binned data. When we generate raw, continuous data from a Cancer Model, a subset of features within our simulation are perturbed in such a way that generates cluster identities. We assume, therefore, that some, but not all, features are significantly associated with the identity of the cluster to which they belong. In small feature spaces (e.g., 9 features), we expect percent of significant features to be as high as 100%. For large feature spaces, where many features are simulated as not belonging to a cluster for greater realism, we expect percent significant features to fall as low as 20%. We consider a mean percent significant features between 30% to 50% across all feature spaces to represent a clinically realistic scenario: where the researcher has clustered a data set of many features, each of unknown significance, and only a portion contribute to the formation of strong cluster identities.

The process of adding clinically meaningful noise and binning continuous data into binary and categorical features by its nature increases the noisiness of these data. However, we assume that, allowing for some variation introduced by our method of allocating hits and extra blocks, noisy and binned data should retain a similar number and distribution of significant features as were present in the raw data. To assess this in noise-

transformed continuous data, we construct a multivariate linear model, which we assess by ANOVA to obtain a p-value describing the significance of each feature to its cluster membership. In binary and categorical data, we repeat this process using the chi-squared test. We compared the percentage of significant features between data types with bean plots qualitatively and ANOVA or the Kruskal-Wallis test (assessing for the need for a non-parametric test with Bartlett's test of equal variances) quantitatively. We analyzed the relationship between percent of significant features and mean silhouette with using linear regression.

We assess *distinctness* by 1 qualitative assessment (t-SNE plots) and 1 quantitative assessment (silhouette width). Although we assess the "true" cluster assignments, both of these tests require the imposition of a distance metric. Thus, any data type that is ill-served by available methods for calculating distance will demonstrate poor performance on both measures. Therefore, we assume that a simulated data type that performs well under the test for *identity* but poorly on a test of *distinctness* demonstrates the shortcomings of available methods to calculate distance but not unfitness of simulated clusters. Limitations and appropriateness of distances for different data types will be explored more fully in Chapter 4.

Initially, we assessed simulation *distinctness* qualitatively by visual inspection of T-distributed Stochastic Neighbor Embedding (t-SNE) plots, which visualize high-dimensional data in low-dimensional space without imposition of a clustering algorithm.[64] As such, these plots allow the visual assessment of latent data structure without testing a specific clustering method. By default, t-SNE plots are often generated

using the Euclidean distance. Where appropriate and with support from the literature, we implemented distance metrics that were more fitting to the data type in question. For example, we visualized categorical data using the Gower coefficient and the Manhattan distance. The history and rationale for the choice of these distance metrics is described in Chapter 4. This initial filtering provided a subjective, working glimpse of the quality of our algorithm. However, more rigorous, objective evaluation was required to confirm the validity of our methods.

Quantitatively, we assessed the *distinctness* of simulated data clusters using two approaches. First, we assessed intrinsic properties of generated clusters and their behavior using silhouette width. Silhouette width is a measure of cluster tightness and separation. For a given object being clustered, a silhouette width of 1 represents maximum tightness and separation between clusters (impossibly optimal clustering) and a value of -1 represents maximally sub-optimal clustering (clear misclassification). For this individual, a silhouette width of 0 describes an intermediate case where goodness of assignment to one cluster or another is approximately equal, and an assignment to either is unclear. Although individual silhouette widths can describe the state of each object being clustered, average silhouette width across a data set can describe a general state of clustering quality resulting from an algorithm.[63]

We evaluated *identity* and *distinctness* of our simulated mixed-type data qualitatively and quantitatively, building over progressive rounds. First, we visualized and assessed the noise model. Next, we assessed the quality of our procedures for binning binary, nominal, and ordinal data. We conceptualize categorical data as presenting three

data problems, corresponding to nominal data, ordinal data, and a mixture of these two. As such, we assessed these three types of categorical data. Finally, we inspected our four mixtures of data types: a balanced mixture of continuous, categorical, and binary data and three unbalanced mixtures, each predominated by continuous, categorical, or binary data.

Results of Evaluations of Clinical Simulations

The first phase in simulating these data was the application of a representative clinical noise model. Henceforth, we will refer to the initial, raw, continuous simulated output prior to the application of the noise model as "raw" data and refer to continuous data produced as the result of the clinical noise model applied to raw data as "continuous" data. For initial comparison, we inspected 972 plots, representing 3 repeats of all desired combinations of parameters. (Table 3.1 in a later section of this chapter discusses chosen parameters and their rationale in detail.) "Noisy" continuous data qualitatively presented with a mild increase in cluster diffusion without obscuring cluster formation in sample sizes reasonable for clinical data.(Figure 3.3) Comparison of average silhouette width between raw and continuous by visual assessment of silhouette width, a measure of cluster tightness and separation, revealed similar high concordance with mild deviation from the raw standard. A paired, two-sample t-test found no difference in average silhouette between raw simulations and those with the clinical noise model applied.($p < 0.0001$)

Four data types were simulated as single-type data sets to act as controls for evaluation and assessment: continuous, binary, nominal, and ordinal. At three repetitions of our desired parameters, we generated 432 test simulations. We assessed the quality of

noisy and binned data through silhouette width and percentage of significant features. Binning continuous data to binary and categorical types increased variability between data sets. An increase in variation and noisiness is also represented in decreased value and increased variability of silhouette width, as clusters expand and their separation softens. Although adding noise to raw data to generate continuous data introduces only mild variability to simulations, binning binary and categorical data reduces average silhouette widths and increases their variability.(Figure 3.5)

Figure 3.5. Paired scatter plots comparing average silhouette widths of raw simulated data, continuous data simulated with a clinical noise model, and mixed data types simulated with a clinical noise model. Binned data of three types were tested: binary, a mixture of nominal and ordinal data representing categorical data, and a balanced mixture of binary, categorical, and continuous data representing mixed data. Observation reveals concordance between raw and noisy data, with greater variation and depression of silhouette widths in binned data.

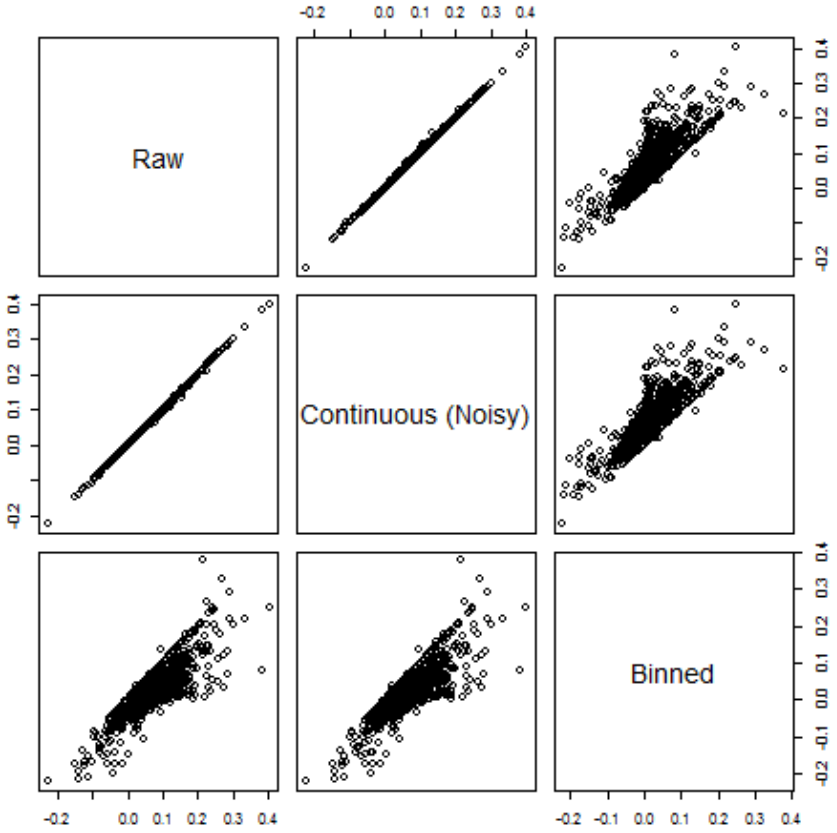


Table 3.2 Mean average silhouette width and percent significant features per simulations vary disparately by data type.

Data Type	Silhouette Width	Significant Features (%)
Continuous	0.072	33.4
Ordinal	0.035	47.0
Nominal	-0.012	46.2
Binary	0.036	46.1

The mean average silhouette width for continuous, binary, nominal, and ordinal data types differ.($p < 0.0001$ by Kruskal-Wallis rank sum test) (Table 3.2) Silhouette widths were highest among continuous data. Mean average silhouette widths for binary and ordinal data were higher than those of nominal data, which had the lowest mean value. Conversely, although variation was present in the percentage of significant features per simulation ($p < 0.0001$ by Kruskal-Wallis rank sum test), continuous data had the lowest percentage of significant features while binary, nominal, and ordinal data had elevated values.(Figure 3.6) Silhouette width was only weakly correlated with the percentage of significant features in a sample (linear refression; $R^2 = 0.085$). Rather, silhouette widths cluster in a narrow range with wide variation in the percent of significant features in a simulation. (Figure 3.7)

Figure 3.6. Bean plots of mean silhouette width (top) and percent significant features (bottom) of four primary simulated data types. Continuous, binary, nominal, and ordinal are simulated from raw data and a clinically representative noise model. Significant association was defined as $p < 0.01$ by chi-squared test (binary and categorical data) or ANOVA (continuous data). Continuous data present with the highest average silhouette width and lower percent significant features per simulation than other data types.

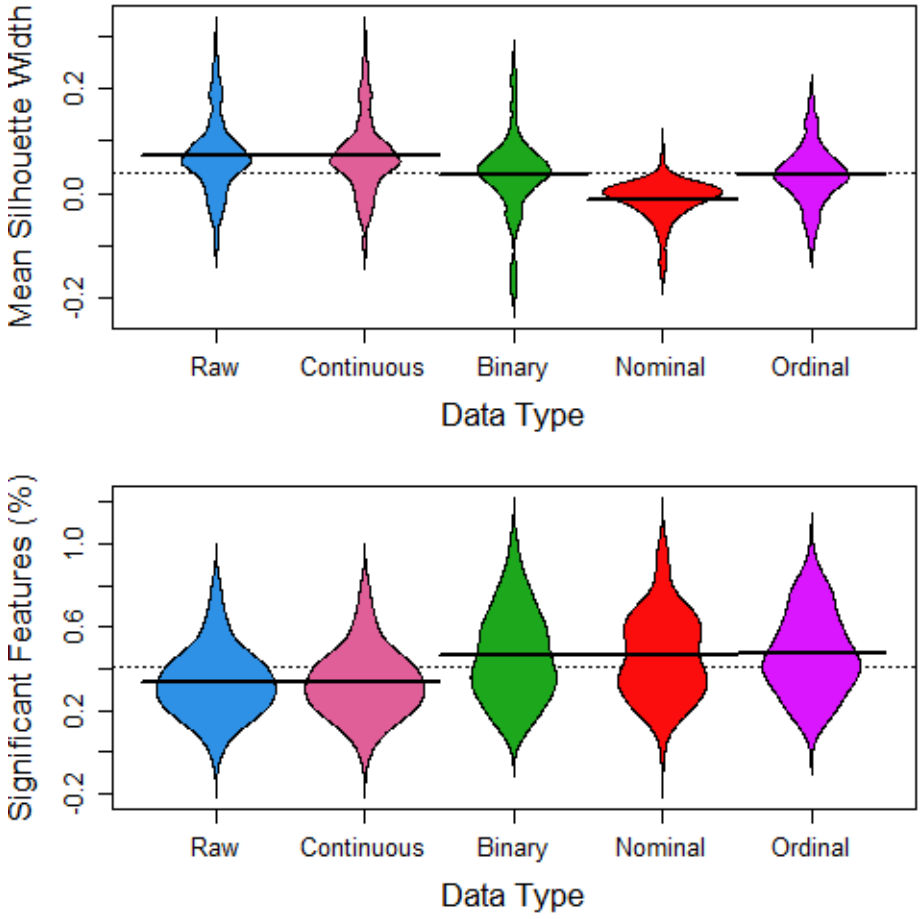
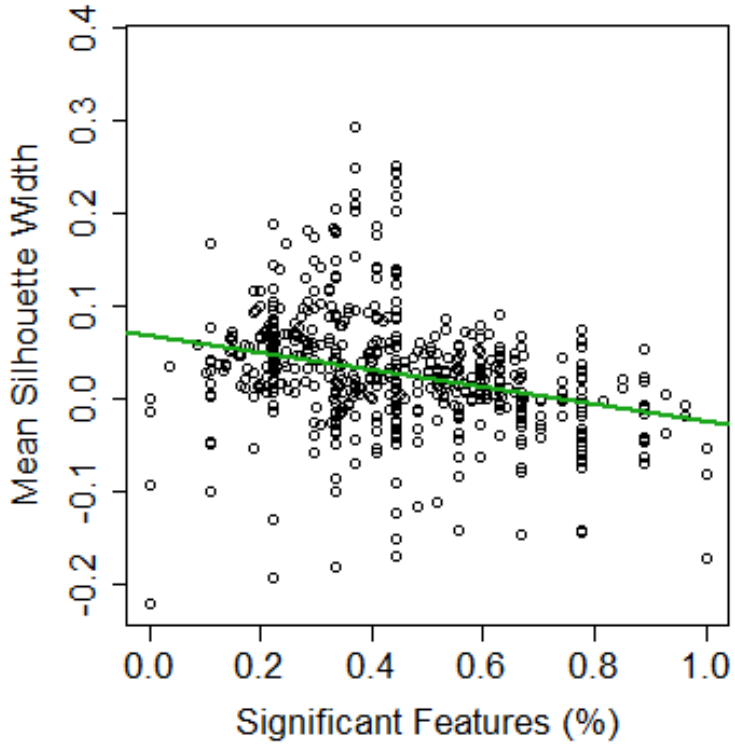


Figure 3.7. Scatter plot of average silhouette width and percent of significant features for 423 simulations of four simulated data types (continuous, binary, nominal, ordinal). Silhouette widths cluster in a range of roughly -0.5 to 0.1 with only weak correlation to the percentage of significant features in a sample (linear refression; $R^2 = 0.085$).



Simulated binary data were visualized forming coherent clusters at 3 different distance metrics appropriate for binary data.(Figure 3.8)[61]

We assessed 108 simulations each of ordinal, nominal, and mixed categorical data. Although percentage of significant features did not vary (mean SF = 46.3%; $p = 0.92$, ANOVA), mean silhouette width differed between the three data types ($p < 0.0001$, Kruskal-Wallis rank sum test). Mean silhouette width was highest among ordinal data (mean SW = 0.0346), lowest among nominal data (mean SW = -0.0123), and intermediate among mixed categorical data (mean SW = 0.0100).(Figure 3.9) With t-SNE, clusters were most easily visualized among ordinal data, forming distinct

groupings. Nominal and mixed categorical data could be visualized using the Gower coefficient, but failed to coalesce using other distance metrics.(Figure 3.10)

Figure 3.8. Representative visualizations of simulated, binary data with 3 distance metrics. Using a simulated patient population of 800 patients with 81 features to represent a clinical experiment of moderate size, we simulate a range of cluster numbers and visualize t-Stochastic Neighbor Embedding (t-SNE) plots. Coherent clusters can be visualized using 3 different distance metrics. Cluster distinctness and separation decreases as the number of clusters increases.

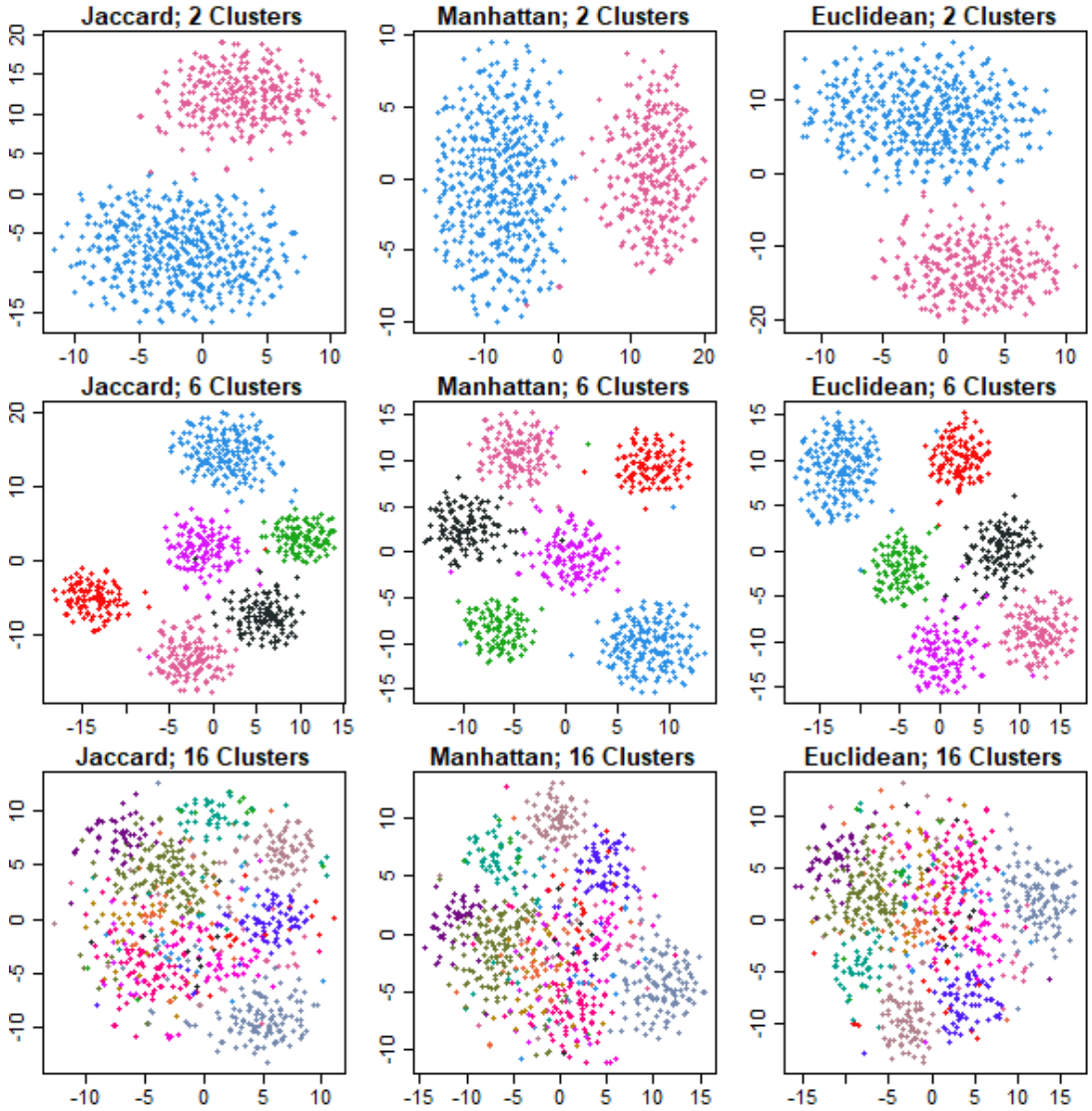


Figure 3.9. Bean plots of mean silhouette width (top) and percent significant features (bottom) of 3 types of simulated, categorical data. Significant association was defined as $p < 0.01$ by chi-squared test. Percent of significant features is the same among the three data types. Mean silhouette width is highest among ordinal data, lowest among nominal data, and intermediate in mixed categorical data.

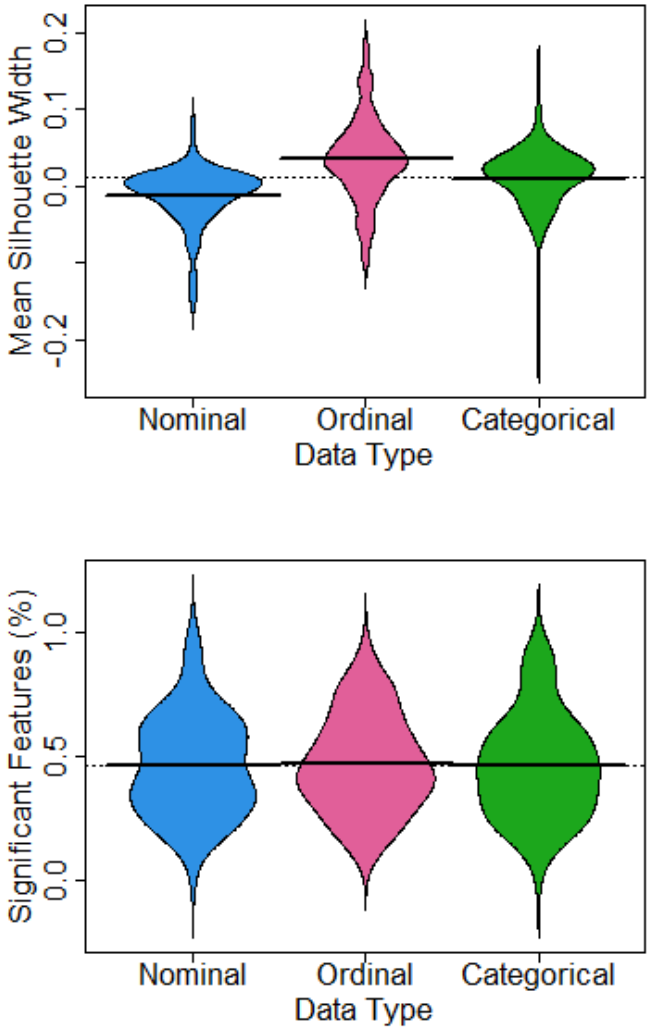
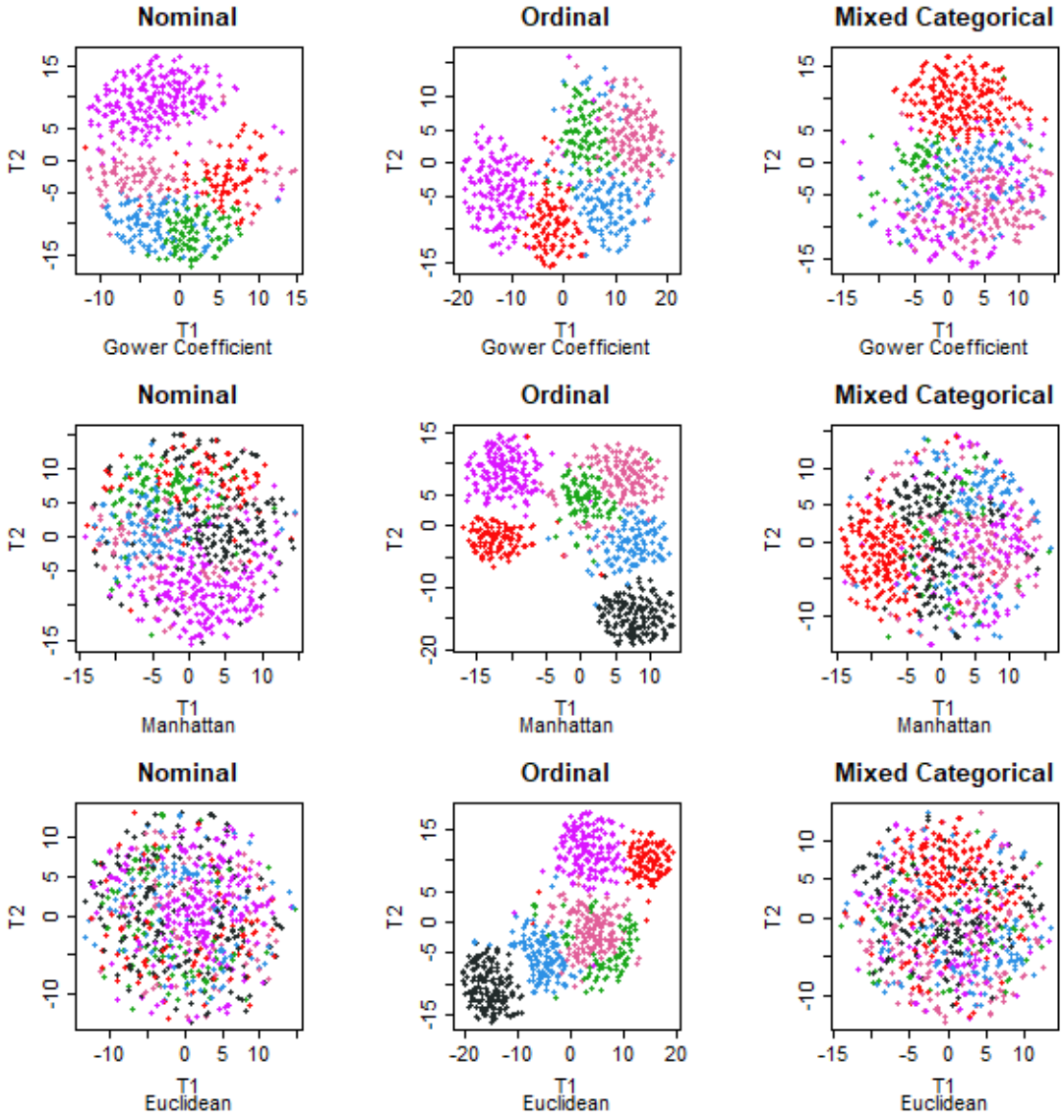


Figure 3.10. Representative visualizations of simulated, categorical data with 3 distance metrics. Using a simulated patient population of 800 patients, 81 features, and 6 clusters to represent a clinical experiment of moderate size, we simulate categorical data types (nominal, ordinal, and a mixture of these) and visualize with t-Stochastic Neighbor Embedding (t-SNE) plots. Ordinal data form distinguishable clusters with all 3 distance metrics. Distinct clusters are visualized in nominal and mixed categorical data with the application of the Gower coefficient, but diffuse with the application of the Manhattan and Euclidean distance.



We assessed 432 simulations of mixed-type data: 108 simulations each of a balanced mixture and three unbalanced mixtures predominated by continuous,

categorical, or binary data (here referred to as “unbalanced continuous mixture”, “unbalanced categorical mixture”, and “unbalanced binary mixture”). Mean silhouette widths varied among the 4 mixtures (p = 0.0018, Kruskal-Wallis rank sum test). Unbalanced continuous mixtures had the highest mean silhouette width (0.037). The 3 other mixtures had similar, lower mean silhouette widths (balanced = 0.012, unbalanced binary = 0.010, unbalanced categorical = 0.012). (Figure 3.11) We visualized using t-SNE with a single distance metric, recovering distinct clusters. (Figure 3.12)

Figure 3.11. Bean plots of mean silhouette width for four simulated data mixtures. Mean silhouette width is highest among an unbalanced mixture dominated by continuous features, with similar mean silhouette width amoung the remaining 3 data mixtures (unbalanced binary mixture, unbalanced categorical mixture, and balanced mixture.)

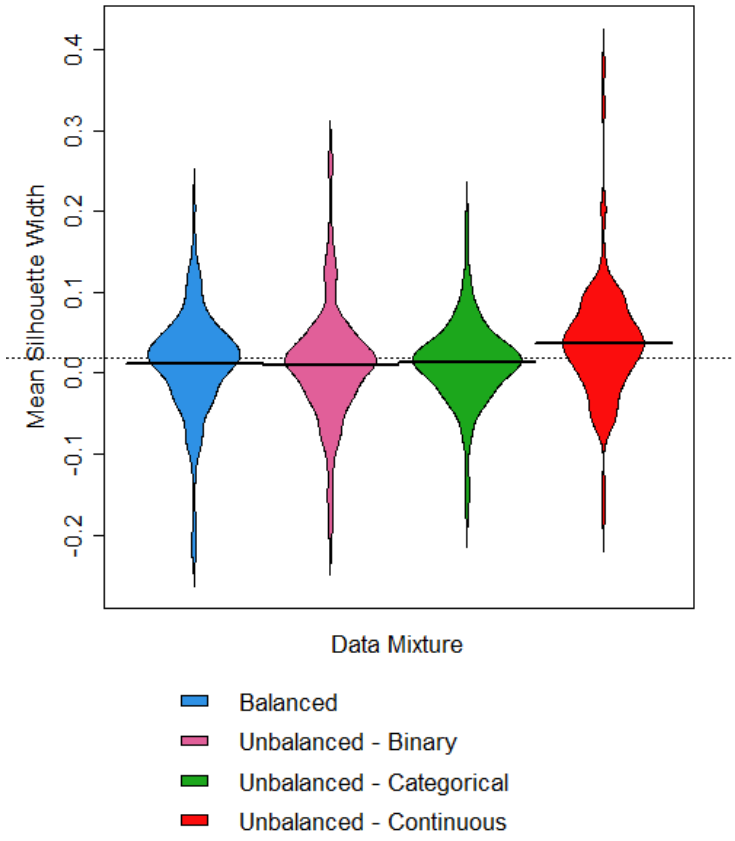
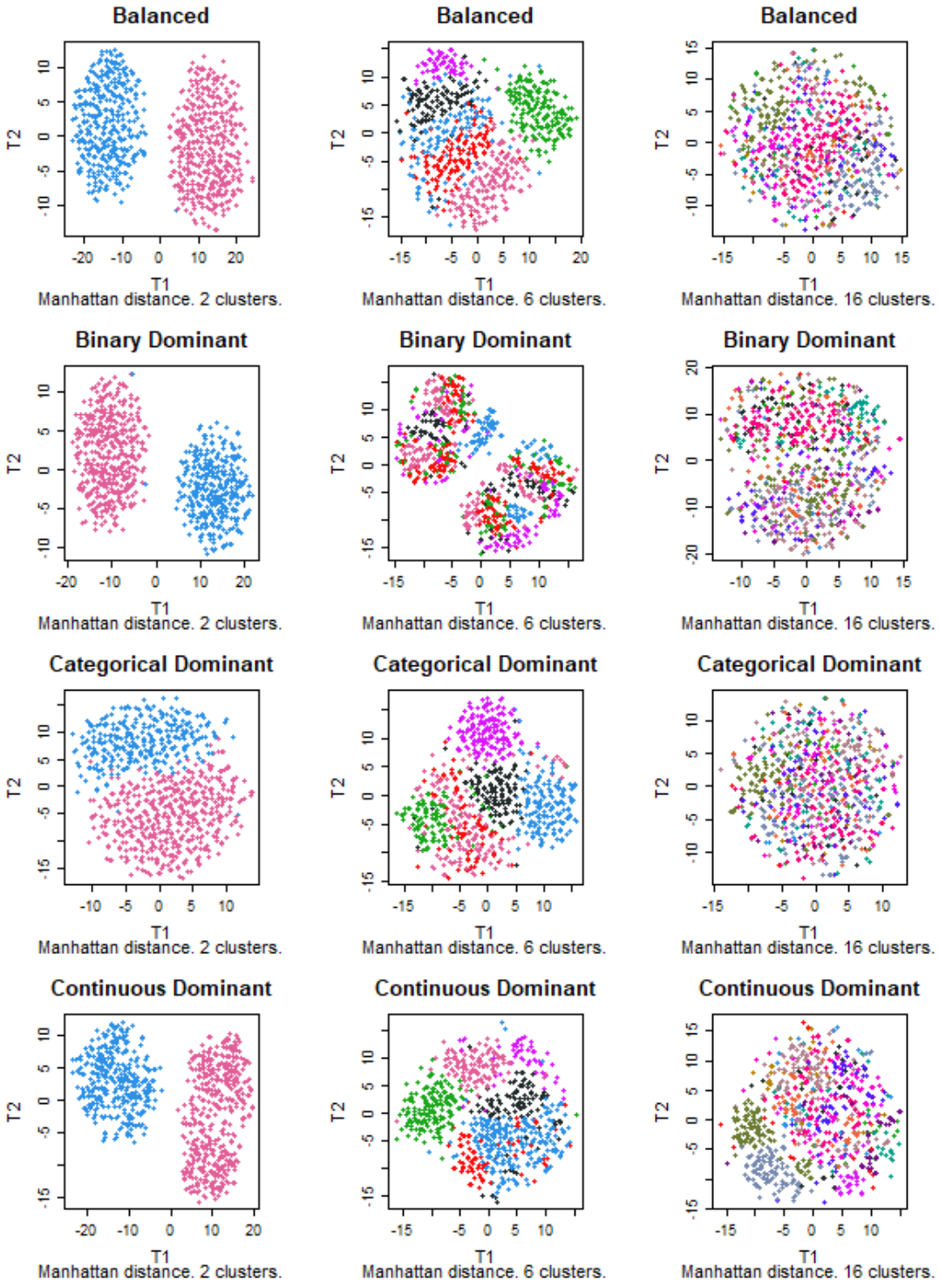


Figure 3.12. Representative visualizations of simulated, categorical data in 4 data type mixtures.



Discussion

In this chapter, we outlined our process for constructing simulated data sets that can realistically mimic clinical data corresponding to clinical trials and cohort studies with known cluster identities, for downstream use testing clustering approaches. The foundation of our simulations is the Umpire R-package. Although this tool was originally developed to simulate continuous gene expression data, we chose to proceed with Umpire because of its ability to generate heterogeneous data with complex intra-cluster relationships. Having made clinically relevant adjustments to the noise model implemented in the package, we outlined our steps for binning and assessing binary, nominal, and ordinal data and combining these into mixtures.

In assessing binned data, we rely heavily on qualitative assessment (e.g. visual inspection of plots). Often, we chose to proceed with plots that looked fuzzy or imperfect. We argue that messy or imperfect plots are what we see in real clinical data, and that challenging simulations are needed to demonstrate relevant outcomes that can be reproduced on real data. Furthermore, some noise in plot results from the need for a dissimilarity matrix calculated on a single distance to be chosen to generate a given plot: the unsolved question that underlies this thesis. Therefore, we expect that, in generating plots of certain data types where we know clustering algorithms are less successful, such as categorical or mixed data types (Chapter 4), we may need to accept a certain level of imperfection to accommodate for the state of the art. These methods are inherently subjective, even when we supplement them with inspection of behavior of significant features or silhouette widths. We argue that, as there are no perfect data, there are no

perfect simulations. Our goal is simulations that approximate an experimental reality that is noisy and challenging. We assume that if we successfully generate simulations of this type, that we will fulfill our needs for algorithm testing in Chapter 4.

The clinical noise model applied low levels of noise to the simulated data, as represented by low levels of increased diffuseness of simulated clusters seen in low rates of change of average silhouette width. Although it could be argued that low levels of noise insufficiently perturb the data, rendering it unable to represent clinical reality, we argue that our clinical noise model, with some features having large standard deviations and others having little variability, is a biologically meaningful starting point. We also must acknowledge that, before the addition of noise, the hit function and correlated block method of cluster generation results in heterogeneous clusters. Furthermore, the binning process introduces significant additional noise, as seen in reduced value and increased variation in silhouette width in binned data. If our goal were to study noisy, continuous data alone, these simulations would require the addition of noise beyond that applied here. However, generation of excessive noise at an early simulation stage would be amplified by the binning process, resulting in indistinguishable clusters.

The most notable conclusion of this chapter results from the relationship between our chosen measures of *identity* and *distinctness*, which displayed only weak correlation. Simulated silhouette width, describing cluster tightness and simulation, were low (i.e. near 0) for many data types. Our experience with clinical data, both seen in the literature, from our analytic experience, and represented in the preliminary experiment presented in Chapter 2, suggests that low silhouette widths and fuzzy distinctions between clusters are

common, if not characteristic, of real, clinical data sets. We are able to view low silhouette widths as a faithful simulation of this characteristic of clinical data due to our second measure of assessment: the percentage of features significantly associated with cluster identities. Our method of simulating clusters, with a limited number of hits perturbing correlated blocks with some or many blocks uninvolved, dictates that driving, significant features will compose only a portion of the features in a given simulation. The presence of moderately high percentages of significant features demonstrates that we have created clusters that possess fuzzy boundaries around coherent identities demarcated by driving, meaningful features. In this sense, low silhouette widths confirm that we are simulating clinically representative data, while high percentages of significant, driving features indicate that these fuzzy clusters have been successfully simulated with coherent, feature-driven identities. Notably, that clusters can have coherent *identities* without *distinctiveness* that is captured by available means suggests room for growth in clustering methods, beginning with more careful distance selection. In this sense, low silhouette width may provide more telling information about means of calculating distance than the quality of simulated clusters. These questions are a staging ground for the tests in Chapter 4.

It is our belief that these simulations capture the characteristics of study design and behavior and error of clinical data in such a way that they can be meaningfully used to test clustering approaches for clinical contexts downstream (Chapter 4).

Chapter 4: Evaluation of 18 Methods for Clustering Mixed-Type Data

Introduction

In this project, we set out with the goal of evaluating best practices for calculating a dissimilarity matrix and clustering mixed-type, clinical data. In this chapter, we first describe our development of a tool, the Mercator R-package, a pipeline to facilitate the calculation of multiple methods of dissimilarity on high-dimensional data and visualize the results with multiple methods. Mercator was applied extensively in the experiments in Chapter 2. Next, we implement the 32,400 simulations generated in Chapter 3 in a series of tests of 18 methods to undertake a clinical clustering problem. The methods we implement take the form of pairs of common clustering algorithms with methods for clustering dissimilarity. We implement 13 "algorithm-dissimilarity pairs" with hierarchical clustering, k-medoids, and self-organizing map algorithms that implement a single, common dissimilarity metric for binary, categorical, and continuous data on simulations of single-type and mixed-type data. Then, we test 5 mixed-metric dissimilarity methods on our 3 algorithms of choice on 4 types of mixed-type data simulations. These mixed-metric methods of calculating dissimilarity include two existing approaches, including the DAISY algorithm proposed in 1990[70] and the Supersom extension of self-organizing maps proposed by Wehrens and Kruisselbrink,[71] and a novel method of our own devising: extending Mercator for a mixture of multiple distances. This allows us to make statements about the performance of our varying methods of calculating dissimilarity specific to each single and mixed data type; to make general statements about the performance of the three algorithms of choice; and to suggest best practices for clustering each data type simulated and tested here.

The Mercator R-Package

Tools for calculating measures of dissimilarity and visualizing high-dimensional, big data in biomedical research are scattered. The limited distance metrics in common use, as seen in Chapter 1, and the potentially inappropriate analytic choices for mixed data sets that may follow, , may have roots in inaccessibility of high-quality tools for dissimilarity calculation, comparison, and visualization. Based on the type of data, an appropriate distance metric must be chosen to quantify a separation between data objects, based on the best fit for specific data types and experimental criteria. Thus, there is not one distance metric seen as superior to all others. Mercator is an R-package that provides a pipeline to calculate multiple single distance metrics and visualize them.[72]

First, Mercator takes the user input in the form of a data matrix of binary or continuous variable data. Second, some initial data filtering is performed using Thresher,[60] an R package that performs clustering using a combination of outlier detection, principal components analysis, and von Mises Fisher mixture models. By identifying significant features, Thresher performs feature reduction through the identification and removal of noninformative features and the nonbiased calculation of the number of groups (K) for downstream use. Third, the user calculates the appropriate distance metric based on data type and biological meaning. Mercator supports 10 distance metrics representing core subgroups defined by Choi and colleagues [61]: Jaccard, Sokal & Michener, Hamming, Russell-Rao, Pearson, Goodman & Kruskal, Manhattan, Canberra, Binary and Euclidean. Finally, Mercator offers 5 visualization methods, including both standard techniques (hierarchical clustering, heat maps) and large-scale multi-dimensional visualizations (multidimensional scaling (MDS),[73] T-distributed Stochastic Neighbor Embedding (t-SNE)[64], and iGraph[74].) Users may easily mix and match distance metrics and visualization techniques to gain a better understanding of patterns in their data. Mercator

streamlines appropriate distance metric selection by facilitating visualization of clusters with multiple distances.

Methods

Clustering Algorithms

For our clustering methods, we choose 3 algorithms with common use, representativeness of trends, and historical significance within the field. First, we use agglomerative hierarchical clustering with Ward's criterion (HC), which we saw as the most commonly implemented algorithm on clinical data in chapter 1. Several related hierarchical clustering algorithms have been in use for over 50 years.[18] As such, they represent an important standard for comparison in any survey of clustering algorithms. Second, we represent partitioning algorithms, an important class of clustering algorithms in common use on clinical data sets, with Partitioning Around Medoids (PAM). PAM is a *k*-medoids clustering algorithm that is related to *k*-means. However, PAM resolves some problems in the *k*-means algorithm including greater robustness to outliers [47] and ability to implement a variety of distance metrics.[25] Third, we represent neural-network based clustering algorithms with self-organizing maps (SOM). The computational methods, advantages, and disadvantages are outlined in Table 4.1.[3, 4, 47]

Table 4.1. Features of 3 implemented clustering algorithms.

Algorithm	**Year**[1]	**Class**	**Computational Method**	**Advantages**	**Disadvantages**
Agglomerative hierarchical clustering with Ward's method	1963	Connectivity-based.	Sequential, bottom-up merging of objects into clusters to increase within-cluster error sum of squares	1. Does not require *a priori* designation of number of clusters 2. Can be implemented with a variety of distance metrics and linkage methods	1. Geometric interpretation assumes objects are in Euclidean space 2. Tends to result in hyperspherical clusters of similar size 3. Not robust to outliers 4. High computational cost with high-dimensional data 5. Requires designation of a level to cut the hierarchy to obtain a final cluster solution 6. Every outlier observation is forced into a cluster
Partitioning Around Mediods (PAM) (*k*-medoids)	1990	Partitioning	Iteratively defines a central observation within a cluster (medoid) and assigns each object to the nearest medoid	1. Robust to outliers 2. Can be implemented with a variety of distance metrics and linkage methods 3. Low computational cost	1. Requires *a priori* designation of number of clusters. 2. Tends to result in hyperspherical clusters of similar size 3. Every outlier observation is forced into a cluster

[1]Year of 1st commonly available published citation

Table 4.1 continued.

Self-organizing maps (SOM)	1973-1982	Neural-network based	High-dimensional data are projecting into a 1-D or 2-D lattice of neurons, preserving the proximity relationships of original data as a topological map	1. Low computational intensity; very fast. 2. Can be implemented with a variety of distance metrics and linkage methods	1. Classically considered a method of visualization, not a clustering approach 2. Every outlier observation is forced into a cluster 3. Requires *a priori* designation of number of clusters.

Distance Metrics

In Chapter 1, we briefly outlined common approaches to calculating dissimilarity for mixed data. Many methods involved the combined use of two approaches for calculating distance: one for continuous data and another for categorical data. Several astute data scientists have suggested mixed data approaches that segregate data types more finely. In the 1971 publication of his mixed-distance coefficient, Gower defines three data types: qualitative (nominal), quantitative (continuous), and dichotomous (binary).[38] In *Finding Groups in Data*, Kaufman and Rousseeuw define 6 types of variables - symmetric binary, asymmetric binary, nominal, ordinal, interval, and ratio – each with different concerns with calculating distance.[25]

Stanley Smith Stevens classically separates continuous data into ratio or interval data based on the presence or absence (respectively) of a true zero. Because ratio data contains a meaningful, true zero, ratios of data points can be compared, not only their differences.[75] Kaufman and Rousseeuw paradoxically define the difference between interval and ratio continuous variables logarithmically. By their definition, interval-scaled variables are positive or negative real numbers on a linear scale.[25] Examples of clinical interval-scaled variables include systolic blood pressure and temperature. Conversely, ratio-scaled variables are always positive and frequently describe

quantities that follow exponential growth or decay curves in time, such as bacterial growth or radioactive decay. They propose, that ratio variables, by this definition, can be treated as though they were interval-scaled, which can introduce distortion; they can be logarithmically transformed and converted to interval-scaled variables; or their ranks can be used in place of their values, treated as ordinal data.[25] The simulated data here thus represent the problems proposed by continuous data more generally or logarithmically transformed data by Kaufman and Rousseeuw's definition. Dissimilarity between continuous variables can be calculated as distances in space. The most common distance metric for continuous data is the Euclidean distance. Also commonly implemented is the Manhattan or City-Block distance. Table 4.2 outlines features of these and other distance metrics implemented in experiments in this chapter. The Euclidean and Manhattan distances are generalized as the Minkowski distance:[25]

$$d(i,j) = \left(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q\right)^{1/q}$$

Although binary data are frequently typed as a special case of nominal data and considered as a unit with the categorical data problem, as we saw in Chapter 1, both Gower and Kaufman and Rousseeuw give binary features special treatment. Kaufman and Rousseeuw separate dichotomous variables into 2 types. For symmetric variables, both possible states (i.e. 0 or 1) carry equal value and weight. A classic example is the binary variable "sex." Conversely, in asymmetric binary variables, the outcomes are not equally important, such that the presence of a relatively rare attribute is more valuable than its more common absence.[25, 61] Choi and colleagues clustered the behavior of 76 binary similarity and distance metrics on a random binary data set into 6 groups, which we used to inform the selection of the distance metrics implemented here.[61] Many dissimilarity metrics for binary data computed from a 2-by-2 contingency table of co-occurring 0 or 1 values:

Table 4.2. Comparison of distance metrics

Distance	Data type	Mathematical Expression	Method
Euclidean	Continuous	$d(i,j) = \left(\lvert x_{i1} - x_{j1}\rvert^2 + \lvert x_{i2} - x_{j2}\rvert^2 + \cdots + \lvert x_{ip} - x_{jp}\rvert^2\right)^{1/2}$ $d_{BINARY} = \sqrt{(b+c)^2}$	Distance in real space
Manhattan	Continuous	$d(i,j) = \lvert x_{i1} - x_{j1}\rvert + \lvert x_{i2} - x_{j2}\rvert + \cdots + \lvert x_{ip} - x_{jp}\rvert$ $d_{BINARY} = b + c$	Distance in real space
Jaccard Index	Asymmetric binary	$d = \frac{a}{a+b+c}$	Negative match exclusive
Sokal & Michener	Symmetric binary	$d = \frac{a+d}{a+b+c+d}$	Hamming-like
Gower	Nominal, ordinal; binary, continuous[1]	$s(i,j) = \sum_{k=1}^{n} s_{ijk} \Big/ \sum_{k=1}^{n} \delta_{ijk}$ $s_{ijk;\,BINARY} = \frac{a}{a+b+c}$ $s_{ijk;NOMINAL} = 1 \; if \; x_{ik} = x_{jk};$ $s_{ijk;NOMINAL} = 0 \; if x_{ik} \neq x_{jk}$ $s_{ijk;QUANTITATIVE}{}^{2} = 1 - \lvert x_{ik} - x_{jk}\rvert / r_k$	Simple matching

[1]Although the Gower coefficient can be implemented for multiple data types, in this study it is implemented for only nominal and ordinal data.
[2] "Quantitative" = ordinal or continuous

Table 4.3. Contingency table for the calculation of binary dissimilarity metrics.

		Object j	
		1	0
Object i	1	*a*	*b*
	0	*c*	*d*

For asymmetric binary data, we chose the Jaccard distance (1908), a negative match exclusive distance with easy interpretability.[25, 61]

For symmetric binary data, we implement two related distances. In Chapter 1, we saw the most commonly implemented distance for binary and categorical data when two metrics were used in a mixed-type study was the Hamming distance:

$$d_{HAMMING} = b + c$$

In Chapter 2, we clustered symmetric binary data using the Sokal & Michener distance:

$$d_{SOKAL\&MICHENER} = \frac{a+d}{a+b+c+d}$$

Choi clusters these distances together in the "Hamming-like" distances.[61] Furthermore, we can see that, given the same number of $a+b+c+d$ patients in a sample (as occurs in our simulations), the Hamming distance and Sokal & Michener distance become equivalent:

$$1 - \frac{a+d}{a+b+c+d} = \frac{b+c}{a+b+c+d}$$

Implementing self-organizing maps with the Kohonen R-package, we are provided with a limited set of distance metrics. For binary data, Kohonen implements a metric called, under "incorrect naming" (for "backwards compatibility") a distance that "returns (for two binary vectors of length n) the fraction of cases in which the two vectors disagree." This is calculated as "basically the Hamming distance divided by n," which we see is analogous to the Sokal & Michener distance.[71] Because the Sokal & Michener distance is easily interpretable; is analogous to the common Hamming distance for our purposes; is analogous to the Kohonen "Tanimoto" distance for our purposes; and maintains continuity with the experiments performed in Chapter 2, we here implement the Sokal & Michener distance for symmetric binary data.

The Gower coefficient of similarity [38] is a historic mixed-distance metric for mixed-type data, which, as we saw in Chapter 1, is still in use today. In its simplest implementation, the data types are unweighted. Gower's coefficient provides solutions for three data types: binary, nominal ("qualitative"), and continuous ("quantitative"). The similarity S_{ij} between features i and j is calculated from their similarity at each feature k and a marker δ_{ijk} of whether a comparison at this point is possible (1) or not possible (0):

$$S_{ij} = \sum_{k=1}^{n} s_{ijk} \Big/ \sum_{k=1}^{n} \delta_{ijk}$$

Similarity between binary features can be described by simple matching,[61] where a binary feature may be present, which Gower represents as + and is commonly represented as 1, or absent (-; 0):

Table 4.4. Table for the calculation of binary Gower dissimilarity.

		Values of k			
Individual	i	+	+	-	-
	j	+	-	+	-
s_{ijk}		1	0	0	0
δ_{ijk}		1	1	1	0
Concordance table equivalency		a	b	c	d

We can see that, in the case that all features in a data set are binary, the Gower coefficient is equivalent to the Jaccard index. Dissimilarity between nominal features is calculated from simple matching with $s_{ijk} = 1$ if $x_{ik} = x_{jk}$ and $s_{ijk} = 0$ if $x_{ik} \neq x_{jk}$. Dissimilarity is calculated from the ratio of difference between objects to the range of values for the k^{th} variable, r_k. (Table 4.2) Here, we implement the Gower coefficient through the daisy function of the cluster R-package. In this implementation, the Gower coefficient can be implemented to calculate categorical distances but not continuous distance. For this and the binary equivalency with the Jaccard, we treat the Gower coefficient in this thesis a measure of categorical distance, and not as a mixed-distance metric per se.

The 5 measures of distance outlined in Table 4.2 have various availabilities based on the restrictions of the R-packages we implemented. The cluster R-package, used for HC through the hclust function,[76] and Mercator, used to implement PAM,[72] can be used with a wider variety of user-defined distance functions. The kohonen R-package,[71] used to implement SOM, provides a narrower range of package-restricted defaults. Table 4.5 describes the single distance measures used for each implemented algorithm in this study.

Table 4.5. Single distance metrics implemented with 3 clustering algorithms.

Algorithm	**Distance**	**Data Type**
Agglomerative hierarchical clustering with Ward's method	Jaccard	Binary (asymmetric)
	Sokal-Michener	Binary (symmetric)
	Gower	Nominal; categorical
	Manhattan	Ordinal; continuous; binary
	Euclidean	Continuous; binary
Partitioning Around Medoids (PAM) (*k*-medoids)	Jaccard	Binary (asymmetric)
	Sokal-Michener	Binary (symmetric)
	Gower	Nominal; categorical
	Manhattan	Ordinal; continuous; binary
	Euclidean	Continuous; binary
Self-organizing maps	Tanimoto	Binary
	Manhattan	Ordinal; continuous; binary
	Euclidean	Continuous; binary

For these 3 algorithms, we implemented 2 methods of single distance and 3 methods of calculated dissimilarity from the combination of multiple distance metrics, as allowed with package restrictions.(Table 4.6) Because the Manhattan distance and Euclidean distance can be applied to a variety of data types, albeit with varying efficacy, we applied these two distance with all 3 algorithms as single distance controls. First among our multiple-distance methods, we implemented the DAISY algorithm as currently available through the cluster R-package.[76] In this iteration, DAISY implements a strategy we saw in Chapter 1: the Gower coefficient for categorical and binary data paired with the Euclidean distance for continuous data. Because the Gower coefficient is implemented as a component of DAISY, we did not apply it separately to the data mixtures. Our remaining two solutions allowed the used to select a distance metric for a given data type. We used a combination of literature knowledge and results from our study of single-distance metrics to inform these choices. The "best" performing single distance metric for this purpose was defined as highest mean Adjusted Rand Index.(See following section) Our final choices are shown in Table 4.6 and our reasoning further explained in the Results section of this

chapter. SOM were implemented within package guidelines. We developed our third solution using the Mercator package, calculating distinct measures of distance following Kaufman and Rousseeuw's guidelines [25] for the 5 data types present in these simulations. We combined these separate measures of distance as an unweighted sum of squares. In the implementation of DAISY and Mercator, distance cannot be calculated in the case of a data type containing only 1 feature. This feature must be excluded from analysis.

Table 4.6. Mixed distance metrics to calculate dissimilarity from mixed data

Dissimilarity Method	Clustering Algorithm	Distance Metric	Data Type
Manhattan distance	Hierarchical clustering PAM SOM	Manhattan distance (single)	Binary, ordinal, continuous
Euclidean distance	Hierarchical clustering PAM SOM	Euclidean distance (single)	Binary, continuous
DAISY	Hierarchical clustering PAM	Gower coefficient	Nominal, ordinal, binary
		Euclidean	Interval-scaled continuous
Mercator	Hierarchical clustering PAM	Jaccard	Binary (asymmetric)
		Sokal-Michener	Binary (symmetric)
		Gower coefficient	Nominal
		Manhattan	Ordinal
		Euclidean	Continuous
Supersom	SOM	Manhattan	Categorical, binary
		Euclidean	Continuous

Clustering method validation

Andreopoulos and colleagues [4] outline 7 "desirable features" to evaluate the "suitability" of a clustering method for a biomedical problem: scalability to high-dimensional data within reasonable computational limits, robustness to outliers, insensitivity to ordering of input objects, minimum user-specified input (including the need to specify the number of input lusters), ability to find arbitrary-shaped clusters, point proportion admissibility (such that adding or removing data redundancy does not change results), and ability to handle mixed-type data. Some of these qualities are properties of a clustering algorithm, not a method of

calculating dissimilarity, and therefore fall beyond the scope of this project.(Table 4.1) In these tests, we are able to assess two of Andreopoulos' features: handling of mixed-type data and, secondarily, scalability within reasonable computational limits.

Our central assessment of the 18 algorithm-dissimilarity pairs is their ability to accurately cluster single- and mixed-type data. There are three classes of methods for validation of a clustering algorithm: external criteria, internal criteria, and relative criteria. External criteria validate a clustering assignment against previous knowledge about the data in the form of a "gold standard" of known cluster identities. The cluster assignment is compared for consensus against this "ground truth."[77, 78] Internal criteria validate the clustering assignment based exclusively on information intrinsic to the data.[78] These measures assess compactness, connectedness, separation, stability, predictive power, and/or correlation of clusters.[77] Relative criteria evaluate a clustering structure by comparison with other clustering schemes.[78, 79]

In this study, our previous simulations have provided us with the ability to externally validate the clustering methods tested here because we have simulated "ground truth" cluster identities.(Chapter 3) There exists many statistics to score external validity against a ground truth, including the Jaccard coefficient (also a measure of binary distance discussed above), the Minkowski Score, the F-measure, the Fowles and Mallows Index, and the Rand Index.[77, 79, 80] Based on matching of pairs of elements into the same or separate cluster assignments, the Rand Index [81] R calculates assignment concordance from a contingency table:

$$R = \frac{a + b}{a + b + c + d}$$

The Adjusted Rand Index (ARI) [82] corrects the Rand Index for chance assignment into concordant clusters. Possible values range from 0 to 1, where 1 is perfect concordance. ARI is in common use and has been considered an important external validity measure for over 30

years.[80] For this reason, we implement it here. Many options also exist to assess internal validity. Silhouette width is a well-known internal measure that computes a score to assess both intra-cluster homogeneity or compactness and inter-cluster separation.[63, 77, 78] Silhouette width takes values from -1 (worst) to 1 (best) as measures of suitability of cluster assignment. An object with a silhouette width of 1 is located in an ideally well-fitting cluster. An object with a silhouette width of -1 falls in a maximally poorly-fitted cluster. An object with a silhouette width of 0 sits on the edge of belonging to 2 clusters. Average silhouette width (SW) describes the fitness of clustering assignments across the structure, and also takes values from -1 to 1.[63] It is in common use and reflects the experiment we performed in Chapter 2, and we continue with this measure in this chapter. In this study, we assess the quality of a clustering assignment with ARI or SW in two ways: quantitatively, by comparison of means, and qualitatively, by comparison of the distribution of these statistics across the test set. We visualize these patterns with beanplots.[83] Finally, because we have a known "ground truth" from our simulations, relative measures of validity have reduced use. We do not implement a relative measure here.

Andreopoulos [4] identifies scalability within reasonable computation limits as an important criterion of a clustering method for a biological problem. To reflect the computational realities of many biomedical projects, these experiments were run on a desktop personal computer. For DAISY and Mercator, we documented the CPU time charged for the execution of the calculation of the distance metric and for each clustering algorithm. SOM implements the calculation of distance and the clustering process in a single step, for which we documented CPU time. Computational time was compared by mean and standard deviation. When identified, slower runtimes within an algorithm were compared by simulation characteristics (number of

patients, features, or clusters and type of data mixture) by mean, standard deviation, ANOVA where applicable, and visualized.

These three assessments – external criterion, internal criterion, and computational scalability – were applied to repetitions of the simulations described in the previous chapter. The 13 single distance metrics were tested on 100 simulation repeats for 34,200 test simulations. For reasons of computation intensity (discussed in a later section of the results of this chapter), mixed distance metrics were tested on a subset of 30 repeats of simulated data mixtures (4,320). Of these simulations, we identified that certain simulated combinations of features, patient populations, and clusters were implausible, given our knowledge of the literature:

- 9 or 27 features simulating greater than 2 clusters
- 200 or 800 patients simulating 16 or more clusters

To produce the most realistic tests of these simulated clinical data mixtures, we removed 1,560 simulations meeting the above criteria from analysis. Thus, we assessed the 5 mixed-distance methods of calculated dissimilarity on 2,760 unique simulations.

Results

Single-Distance Methods

Clustering performance for each data type varied by both clustering algorithm and distance metric.(Table 4.7) On noisy simulations across each data type and distance metric, HC had higher ARI than PAM. HC had higher silhouette widths than PAM on continuous, ordinal, and mixed categorical data. The two algorithms had similar silhouette width performance for nominal and binary data. SOM had highest ARI and SW for all data types and distance metrics, except nominal data.

Continuous data had higher ARI and SW across distance metrics compared to other data types. SOM with Euclidean distance produced the highest mean ARI (0.611 ± 0.336) and highest

mean SW (0.093 ± 0.051). Visualization with bean plots can show the consistency of high-quality solutions from a given method. All distance methods and algorithms produced a range of ARI from very poor (near 0) to nearly perfect (near 1).(Figure 4.1) While PAM produces a bolus of clustering solutions with low ARI (0.1-0.5), HC and SOM produce a bolus of solutions with very high ARI. SW does not vary strongly across algorithms.

Binary data had second highest ARI and SW across distance metrics compared to other distance types. SOM with Euclidean distance resulted in the highest mean ARI (0.516 ± 0.357) followed by SOM with Manhattan distance (0.513 ± 0.359). SOM with Manhattan distance also produced the highest mean SW (0.177 ± 0.120). Across HC or PAM, performance of the 4 distance metrics in question (Jaccard, Sokal & Michener, Manhattan, and Euclidean) produced similar results. By visualization with bean plots, all distance methods and algorithms produced solutions spanning a range of ARI from 0 to 1.(Figure 4.2) PAM ARI's were heavily weighted towards inaccurate solutions (ARI between 0 and 0.4). HC and SOM produced bipolar results, with ARI clustered either near 1 or near 2. The bolus of solutions near 1 was larger for SOM than HC. The strongest bipolar distribution of ARI resulted from the Tanimoto distance. SW were uniformly lower than for continuous data. PAM produced some solutions with low SW with a group of simulations with higher SW between 0.2-0.4. SOM with the Manhattan distance produced many solutions with lower silhouette widths, but resulted in a group of simulations with higher SW than other solutions, including PAM. The Tanimoto distance presented with the lowest range of SW, with a tail of many values less than 0.

Nominal, ordinal, and categorical data had lowest ARI and SW across distance metrics, compared to continuous and binary data. Clustering solutions for nominal data produced the lowest ARI or any data type. Among nominal data, the HC with Gower distance produced the

solution with both highest mean ARI and largest ARI standard deviation (0.283 ± 0.298). The highest silhouette width was produced by SOM with the Manhattan distance (0.052 ± 0.046). PAM produced solutions with lower mean ARI but with the smallest standard deviations. (Table 4.7) By visualization of ARI with bean plots, all methods produced a range of values with most solutions clustered near 0 with no evidence of the bipolar distribution seen in binary and continuous data. SW also clustered near 0, with PAM and SOM producing a fraction of solutions with elevated SW.(Figure 4.3)

Clustering solutions of ordinal data produced intermediate ARI and SW. SOM with the Manhattan distance produced the solutions with highest mean ARI (0.405 ± 0.368) and SW (0.081 ± 0.044). The Gower distance had lower ARI and SW performance by quantitative measures and bean plot visualization than the Manhattan or Euclidean distance.(Figure 4.4) HC, PAM, and SOM all visualized with a range of ARI from 0 to 1. PAM solutions weighted towards 0. SOM solutions displayed a bipolar distribution, with solutions clustered either near 0 or a bolus of solutions near 1. All implementations of the Manhattan and Euclidean distance resulted in range of SW weighted between 0 and 0.2.

Mixed categorical data resulted in low ARI and SW, like nominal data. SOM with the Manhattan distance produced the highest mean ARI (0.301 ± 0.342) and SW (0.066 ± 0.044). Like nominal data, visualization of mixed categorical data resulted in a range of solution ARI with a heavy distribution near 0.(Figure 4.5) SOM produced a small fraction of solutions near 1. The Euclidean and Manhattan distances with all 3 algorithms produced a range of SW between 0 and 1, with PAM producing many low solutions and a portion of solutions with elevated SW.

Table 4.7 Results of single-distance methods for simulations of single data types

		Data Type									
		Binary		**Nominal**		**Ordinal**		**Categorical**[1]		**Continuous**	
Algorithm	**Distance**	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]
HC[2]	Jaccard	0.430 ± 0.342	0.129 ± 0.108	-	-	-	-	-	-	-	-
	Sokal & Michener	0.433 ± 0.341	0.147 ± 0.106	-	-	-	-	-	-	-	-
	Gower	-	-	**0.283 ± 0.298**	0.024 ± 0.020	0.271 ± 0.293	0.023 ± 0.019	0.276 ± 0.295	0.023 ± 0.020	-	-
	Manhattan	0.434 ± 0.341	0.148 ± 0.107	0.141 ± 0.219	0.035 ± 0.038	0.376 ± 0.336	0.064 ± 0.042	0.280 ± 0.305	0.049 ± 0.039	0.561 ± 0.341	0.06 ± 0.050
	Euclidean	0.426 ± 0.343	0.101 ± 0.106	0.047 ± 0.100	0.036 ± 0.044	0.335 ± 0.327	0.059 ± 0.042	0.211 ± 0.272	0.046 ± 0.043	0.602 ± 0.345	0.085 ± 0.053
PAM[3]	Jaccard	0.314 ± 0.269	0.133 ± 0.131	-	-	-	-	-	-	-	-
	Sokal & Michener	0.331 ± 0.276	0.156 ± 0.138	-	-	-	-	-	-	-	-
	Gower	-	-	0.103 ± 0.105	0.020 ± 0.017	0.096 ± 0.097	0.019 ± 0.017	0.099 ± 0.100	0.019 ± 0.017	-	-
	Manhattan	0.331 ± 0.276	0.156 ± 0.138	0.044 ± 0.057	0.035 ± 0.038	0.207 ± 0.195	0.056 ± 0.040	0.121 ± 0.137	0.043 ± 0.039	0.373 ± 0.277	0.063 ± 0.051
	Euclidean	0.329 ± 0.277	0.109 ± 0.125	0.018 ± 0.025	0.038 ± 0.045	0.189 ± 0.192	0.055 ± 0.044	0.094 ± 0.123	0.044 ± 0.045	0.446 ± 0.302	0.076 ± 0.056
SOM[4]	Tanimoto	0.437 ± 0.361	0.086 ± 0.114	-	-	-	-	-	-	-	-
	Manhattan	0.513 ± 0.359	**0.177 ± 0.120**	0.131 ± 0.224	**0.052 ± 0.046**	**0.405 ± 0.368**	**0.081 ± 0.044**	**0.301 ± 0.342**	**0.066 ± 0.044**	0.568 ± 0.344	0.083 ± 0.049
	Euclidean	**0.516 ± 0.357**	0.117 ± 0.112	0.092 ± 0.188	0.052 ± 0.049	0.380 ± 0.368	0.075 ± 0.044	0.274 ± 0.337	0.062 ± 0.046	**0.611 ± 0.336**	**0.093 ± 0.051**

[1] A mixture of nominal and ordinal features.
[2] Agglomerative hierarchical clustering with Ward's criterion
[3] Partitioning Around Medoids
[4] Kohonen self-organizing maps
[5] Adjusted Rand Index; mean ± standard deviation
[6] Average silhouette width; mean ± standard deviation

Figure 4.1 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated continuous data.

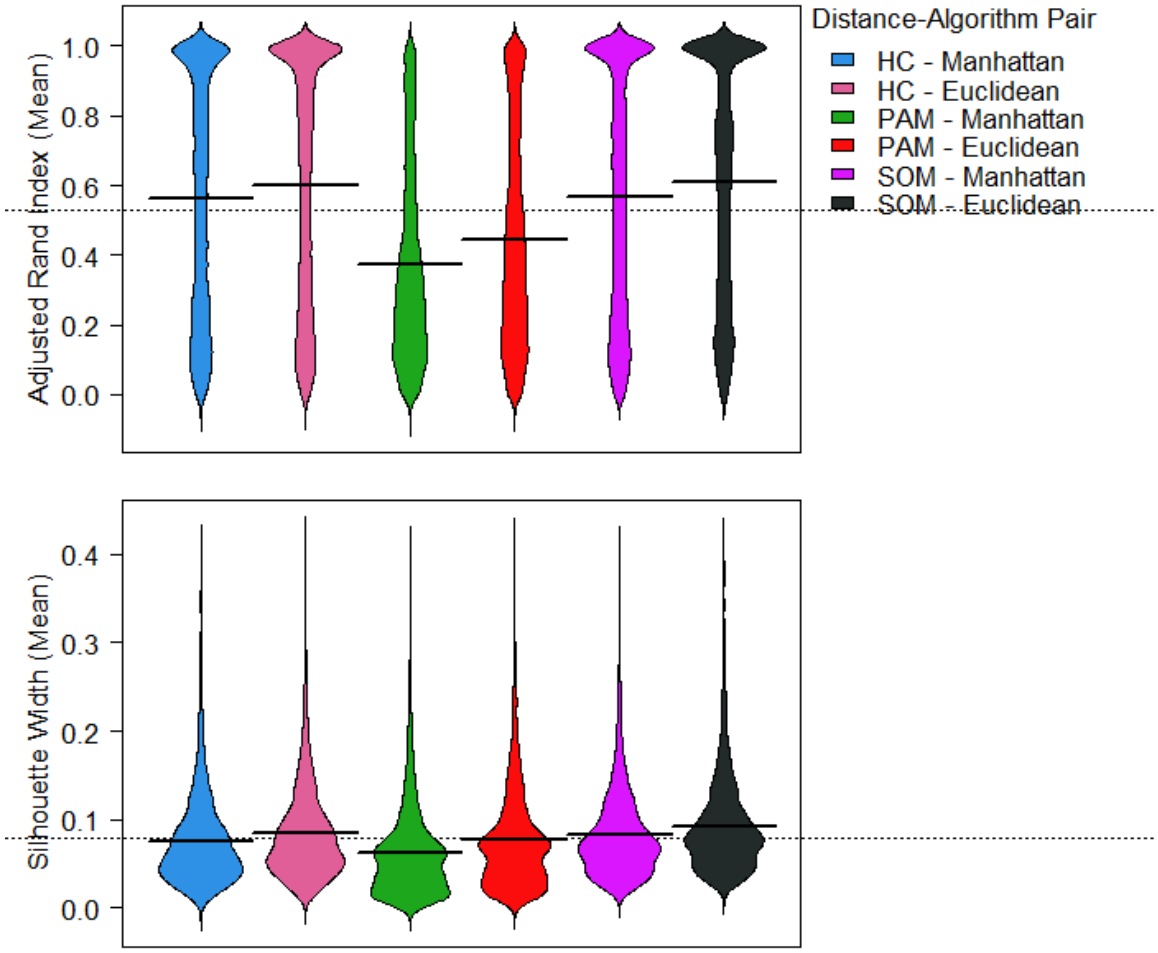


Figure 4.2 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated binary data.

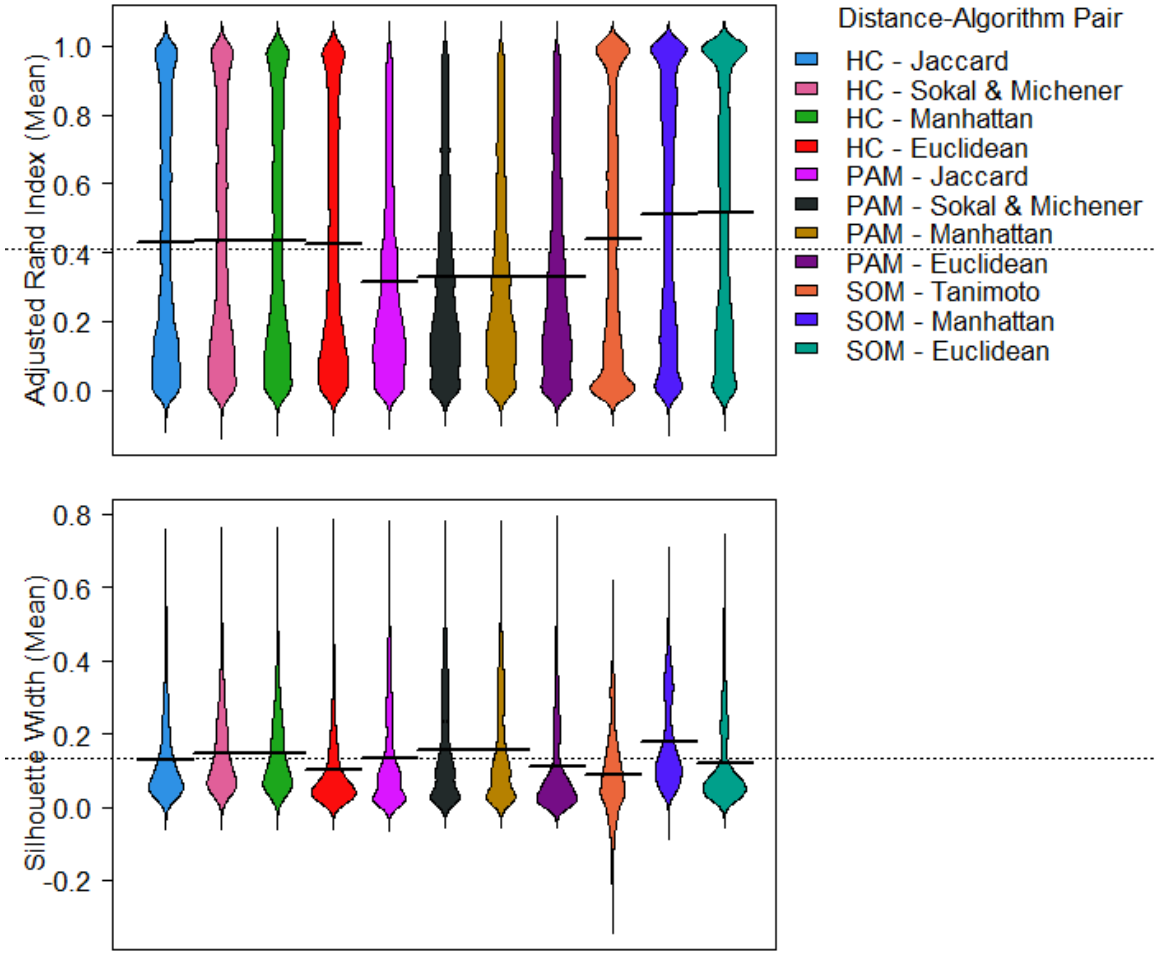


Figure 4.3 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated nominal data.

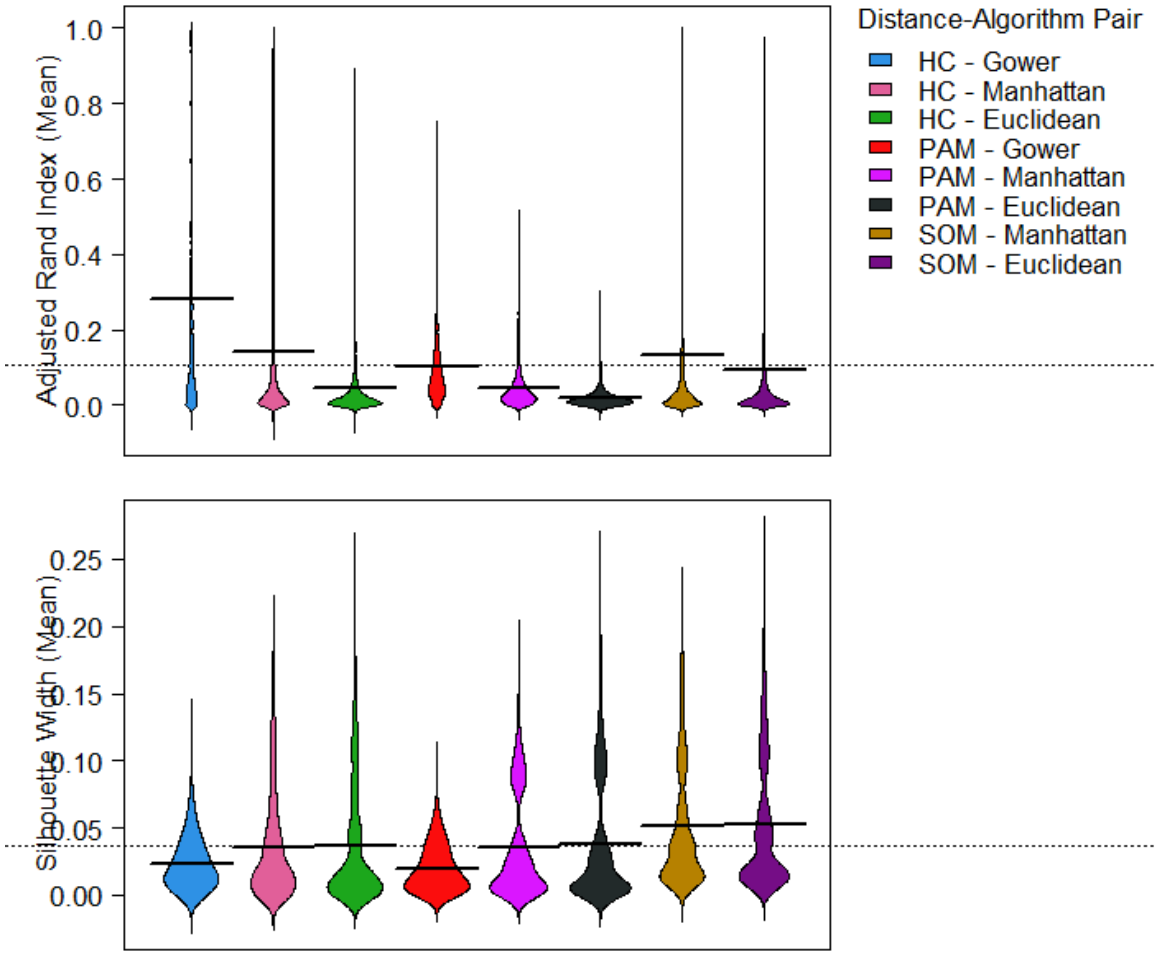


Figure 4.4 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated ordinal data.

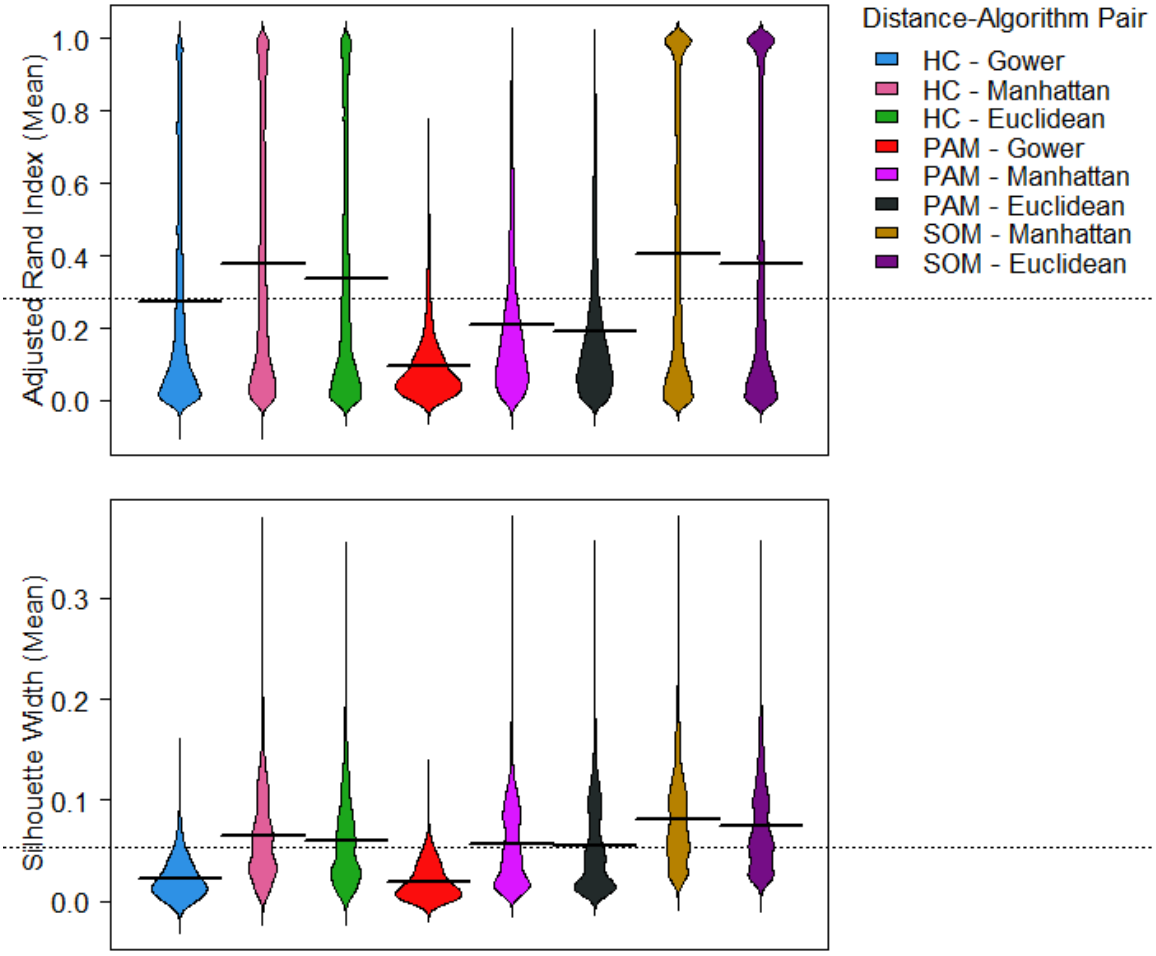
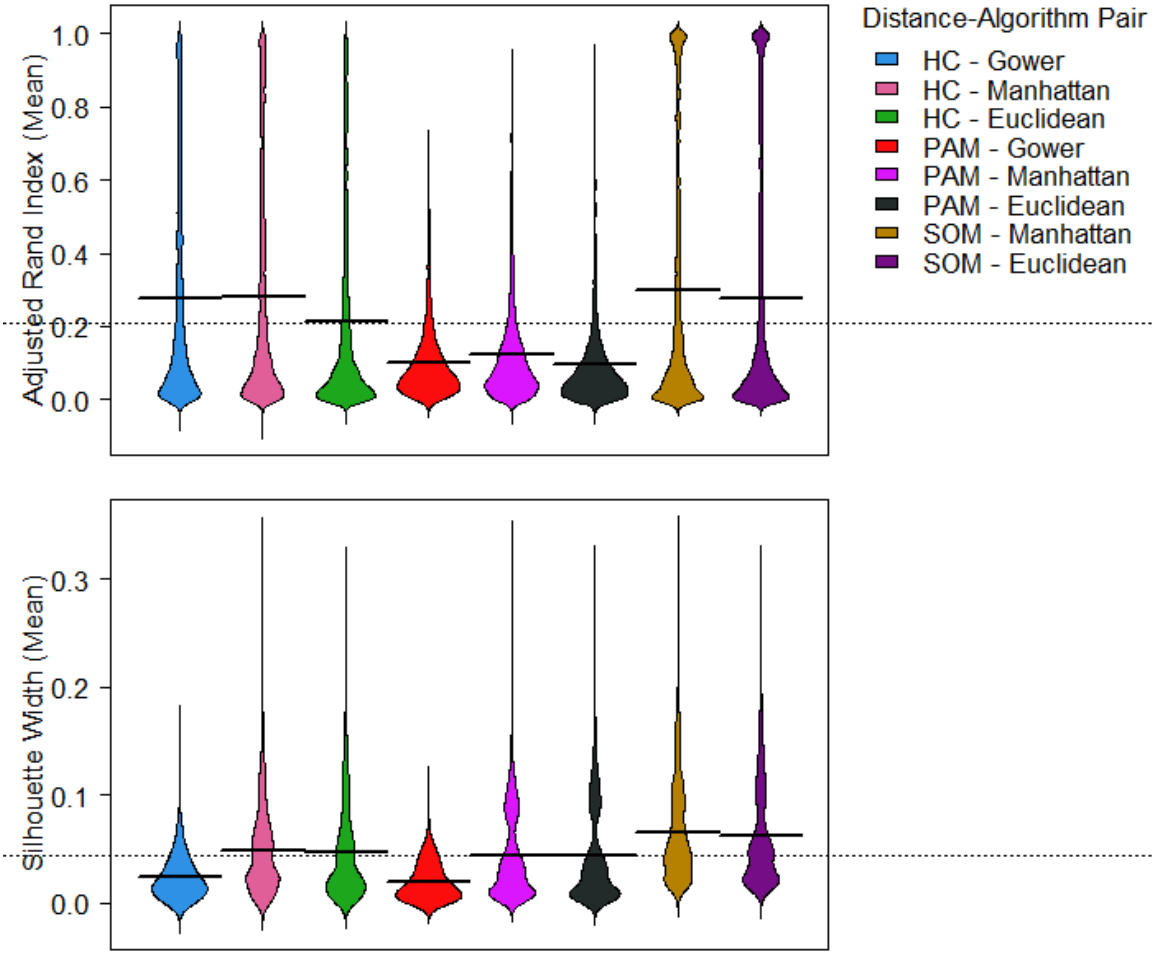


Figure 4.5 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated mixed categorical data.



Multiple-Distance Methods

Here we arrive at the core question of this thesis: for a mixture of data types, does a dissimilarity matrix calculated from a mixture of data types result in superior clustering outcomes, as measured by extrinsic and intrinsic criteria? ARI and SW (mean ± standard deviation) for 2,760 plausible clinical simulations of data mixtures clustered with 3 algorithms on 2 single distance controls and 3 mixed-distance dissimilarity metrics are displayed in Table 4.8. Because of the poor performance of the Tanimoto distance in

SOM on binary data, we supplemented the Manhattan distance for the treatment of binary data in Supersom.

For balanced, unbalanced binary, and unbalanced categorical mixtures, the DAISY algorithm with HC outperformed all tested algorithm-distance pairs by mean ARI. DAISY with HC resulted in the highest SW in balanced data, as well (0.099 ± 0.085). However, for the other three data types, DAISY with PAM or HC resulted in low SW when compared with other methods. Supersom produced the highest SW for all 3 unbalanced data types and the second highest SW for balanced data (0.098 ± 0.080), but produced low mean ARI compared to other methods.

For balanced mixtures, the superior solution was produced by DAISY with HC (mean ARI = 0.474 ± 0.352; mean SW = 0.099 ± 0.085), closely followed by Mercator with HC (mean ARI = 0.467 ± 0.366; mean SW = 0.093 ± 0.069). By visualization with bean plots, all algorithms tested produce solutions with a range of ARI between 0 and 1. (Figure 4.6) HC with Manhattan distance, DAISY, or Mercator results with a distribution with bipolar weighting. DAISY with HC and Mercator with HC are weighted heavily towards 1. SOM with the Manhattan distance presents with a strong bipolar distribution. ARI with PAM, regardless of distance metric, produces a distribution of solutions weighted towards 0.

For unbalanced binary mixtures, the superior solution by mean ARI was produced by DAISY with HC (0.574 ± 0.324) with highest mean SW produced by Supersom (0.193 ± 0.137). By visualization with bean plots, DAISY and Mercator result in a range of ARI distributed across the range of 0 to 1, with DAISY with HC weighted towards 1.

(Figure 4.7) Implementations of the Euclidean distance and Supersom result in solutions near 0. The Manhattan distance implemented with HC or SOM results in a bipolar distribution of ARI. Higher SW are produced by single distance measures, SOM, and Supersom, with DAISY and Mercator producing SW below the overall mean.

For unbalanced categorical mixtures, the highest mean ARI was produced by DAISY with HC (0.341 ± 0.311) with highest mean SW produced by Supersom (0.071 ± 0.049). By visualization with bean plots, HC with DAISY or the Manhattan distance produces solutiosn with a range of ARI between 0 and 1.(Figure 4.8) SOM and Supersom produce bipolar distributions of ARI. SW are low, with single distance metrics, SOM, and Supersom outperforming DAISY and Mercator.

For unbalanced continuous mixtures, the highest mean ARI solutions were produced by SOM with the single Manhattan distance (0.564 ± 0.392) with the highest mean SW produced by Supersom (0.174 ± 0.123). By visualization with bean plots, SOM, HC with single distances, and HC with DAISY produce bipolar distributions of ARI, with solutions with PAM, Mercator, and Supersom weighted towards 0.(Figure 4.9) DAISY and Mercator result in low SW, below the overall mean, compared to single distance metrics, SOM, or Supersom.

Table 4.8 Results of single- and mixed-distance methods for plausible, simulated mixed data types

		Data Mixture Type							
		Balanced		**Binary Unbalanced**		**Categorical Unbalanced**		**Continuous Unbalanced**	
Distance	**Algorithm**	***ARI***[1]	***SW***[2]	***ARI***[1]	***SW***[2]	***ARI***[1]	***SW***[2]	***ARI***[1]	***SW***[2]
Manhattan	HC[3]	0.430 ± 0.357	0.081 ± 0.068	0.349 ± 0.366	0.142 ± 0.123	0.267 ± 0.303	0.055 ± 0.044	0.472 ± 0.385	0.105 ± 0.085
	PAM[4]	0.203 ± 0.210	0.068 ± 0.072	0.204 ± 0.228	0.118 ± 0.127	0.121 ± 0.135	0.049 ± 0.045	0.271 ± 0.258	0.080 ± 0.087
	SOM[5]	0.460 ± 0.278	0.098 ± 0.070	0.402 ± 0.417	0.153 ± 0.117	0.288 ± 0.338	0.071 ± 0.049	**0.564 ± 0.392**	0.110 ± 0.080
Euclidean	HC[3]	0.232 ± 0.299	0.079 ± 0.083	0.075 ± 0.175	0.156 ± 0.141	0.195 ± 0.263	0.053 ± 0.050	0.335 ± 0.359	0.119 ± 0.105
	PAM[4]	0.115 ± 0.157	0.077 ± 0.089	0.073 ± 0.134	0.140 ± 0.144	0.088 ± 0.114	0.052 ± 0.053	0.219 ± 0.240	0.101 ± 0.109
	SOM[5]	0.278 ± 0.353	0.097 ± 0.085	0.083 ± 0.204	0.160 ± 0.136	0.248 ± 0.325	0.069 ± 0.053	0.353 ± 0.385	0.123 ± 0.100
DAISY	HC[3]	**0.474 ± 0.352**	**0.099 ± 0.085**	**0.574 ± 0.324**	0.091 ± 0.053	**0.341 ± 0.311**	0.034 ± 0.026	0.393 ± 0.359	0.060 ± 0.043
	PAM[4]	0.279 ± 0.248	0.084 ± 0.093	0.387 ± 0.279	0.077 ± 0.060	0.146 ± 0.139	0.025 ± 0.020	0.205 ± 0.197	0.041 ± 0.034
Mercator	HC[3]	0.467 ± 0.366	0.093 ± 0.069	0.327 ± 0.165	0.089 ± 0.064	0.183 ± 0.253	0.054 ± 0.045	0.127 ± 0.219	0.085 ± 0.068
	PAM[4]	0.274 ± 0.248	0.074 ± 0.071	0.165 ± 0.187	0.069 ± 0.064	0.136 ± 0.163	0.030 ± 0.032	0.101 ± 0.135	0.065 ± 0.061
Supersom	SOM[5]	0.243 ± 0.312	0.098 ± 0.080	0.061 ± 0.158	**0.193 ± 0.137**	0.270 ± 0.326	**0.071 ± 0.049**	0.079 ± 0.190	**0.174 ± 0.123**

[1] Adjusted Rand Index; mean ± standard deviation

[2] Average silhouette width; mean ± standard deviation

[3] Agglomerative hierarchical clustering with Ward's criterion

[4] Partitioning Around Medoids

[5] Kohonen self-organizing maps

Figure 4.6 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated balanced data mixtures.

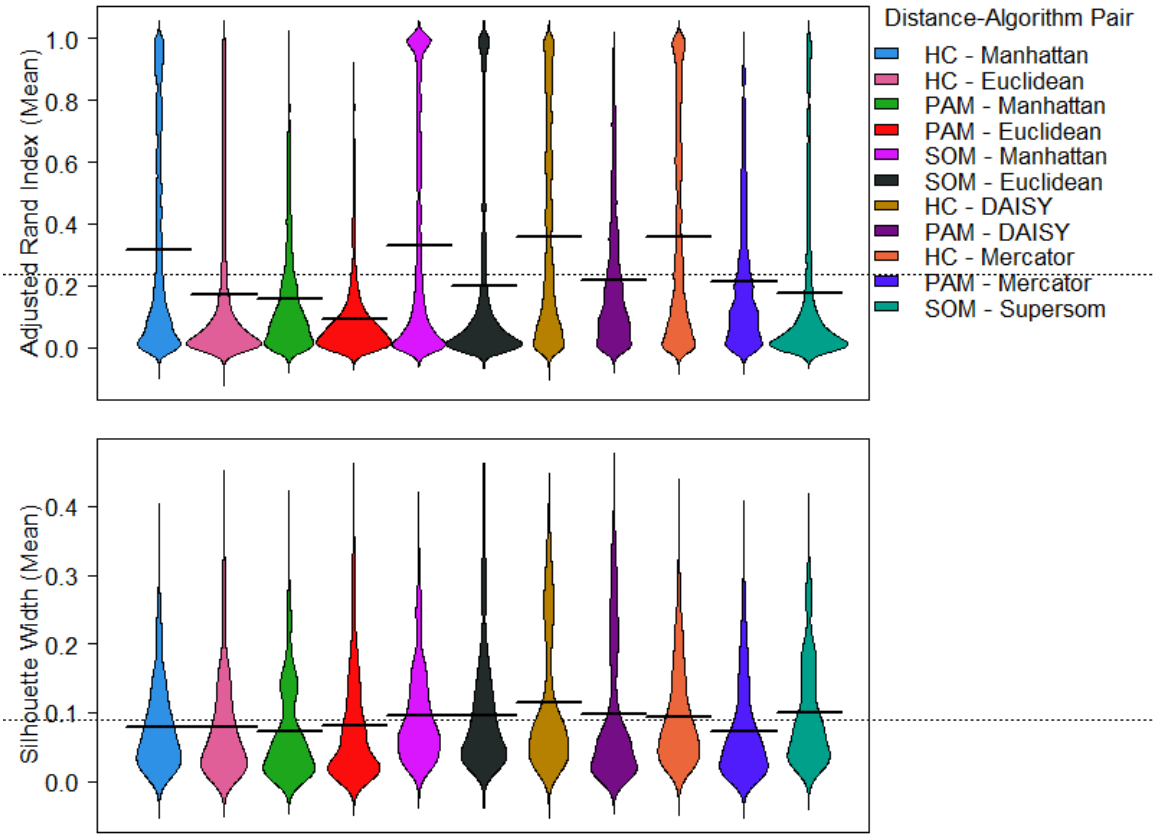


Figure 4.7 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated unbalanced, binary-dominant data mixtures.

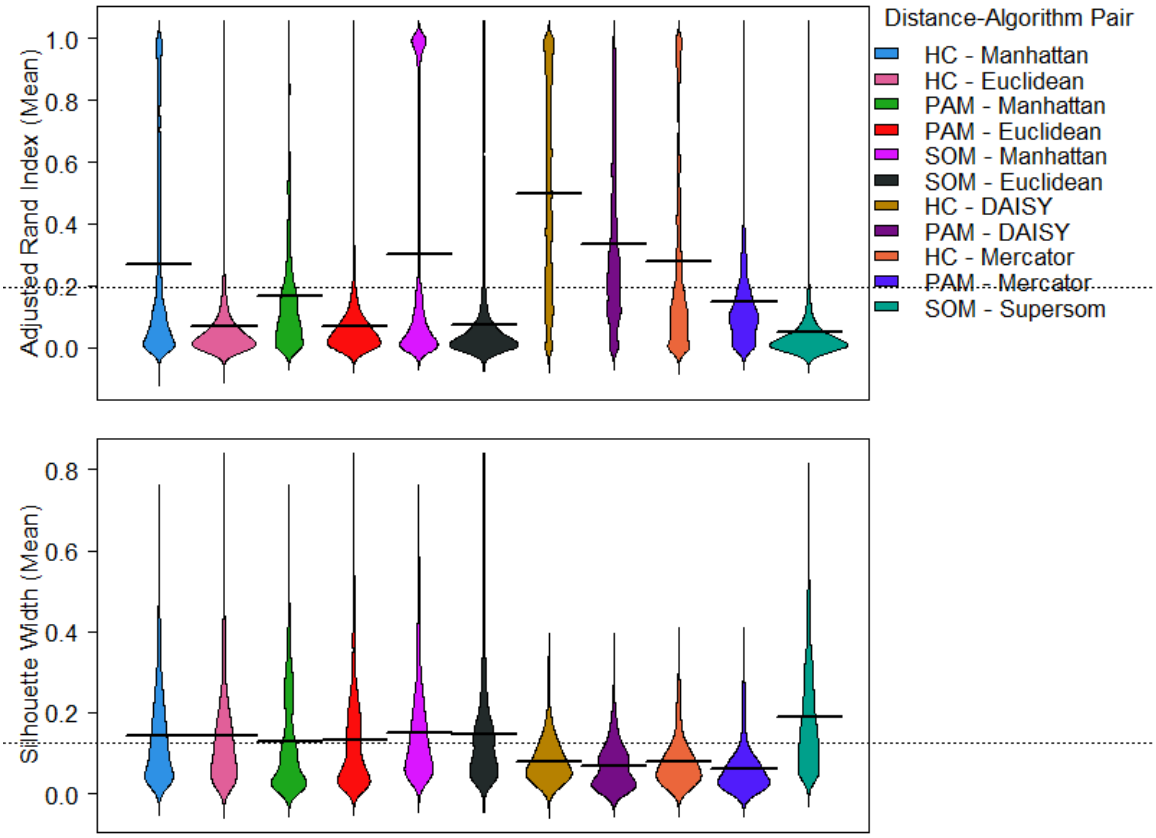


Figure 4.8 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated unbalanced, categorical-dominant data mixtures.

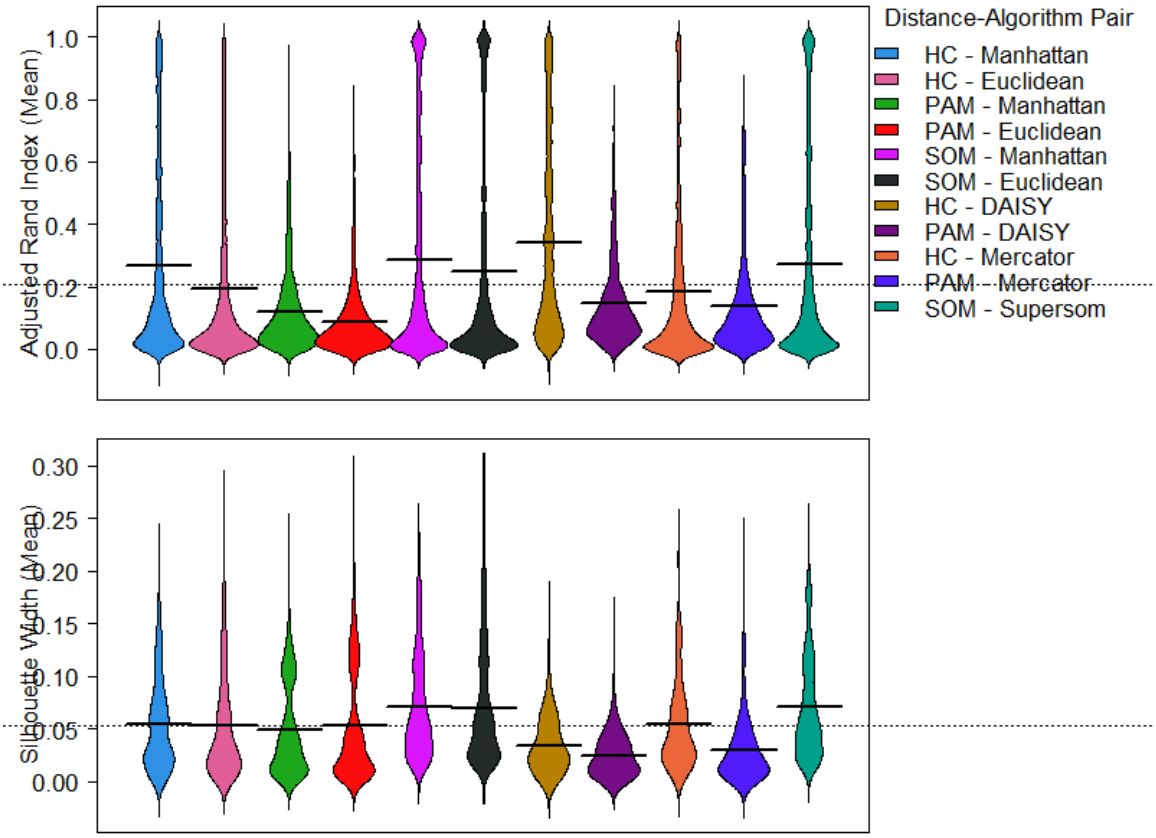
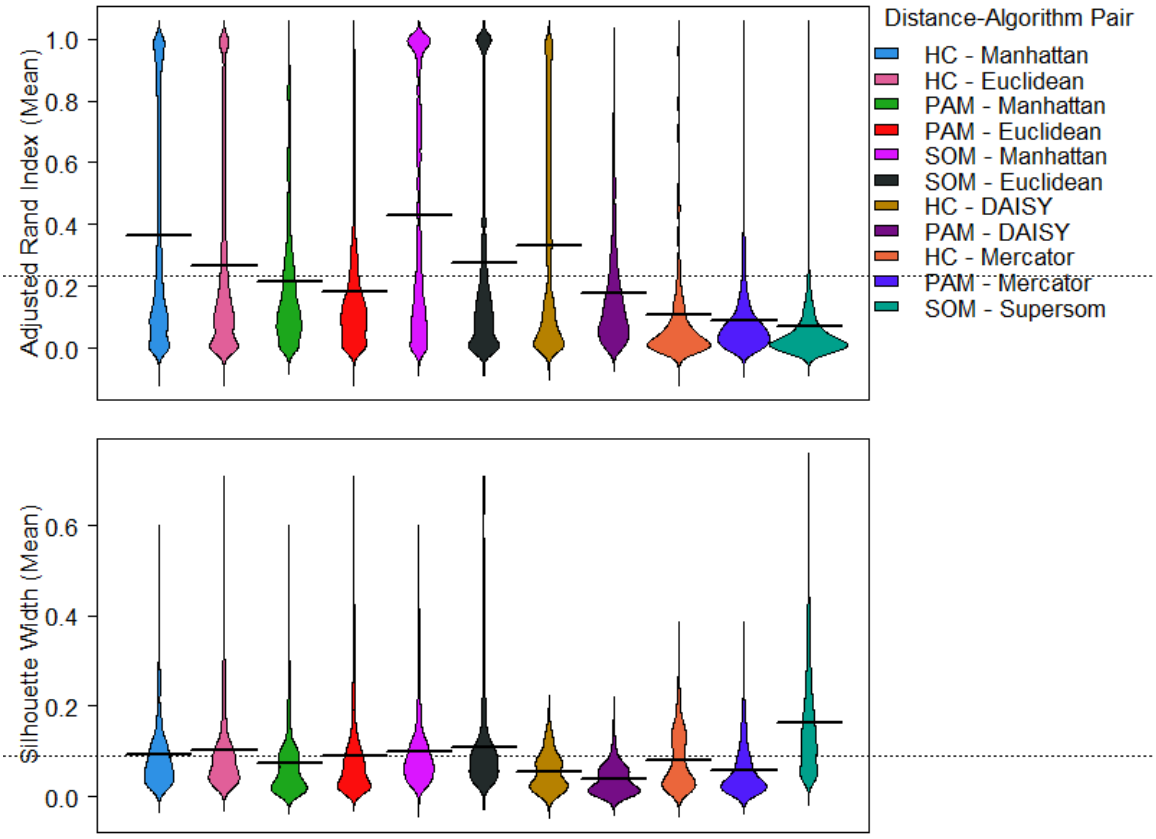


Figure 4.9 Bean plots of adjusted rand index (top) and silhouette width (below) for simulated unbalanced, continuous-dominant data mixtures.



Variability of Adjusted Rand Index Across Simulation Parameters

Violin plots of clustering solutions displayed variability of ARI in the form of broad spectra and bipolar distributions. Here, we employ lattice plots to break down these results across number of simulated patients and features. For single data types, (Figures 4.10-4.13) we plot using the Euclidean distance, calculated for all 4 single data types, and three algorithms (HC, PAM, and SOM). For binary data, we also plot the Tanimoto distance, which showed the strongest bipolar distribution in previous plots. For mixed data types, (Figures 4.14-4.15) we employ lattice plots to observe the behavior of the DAISY and Mercator multiple distance algorithms across simulation parameters.

Across lattice visualizations of single distances, a common trend emerged: ARI varied strongly by number of features, but not by number of patients. ARI was lowest among simulations with 9 features and highest among simulations with 243 features. Intermediate features spaces displayed higher degrees of variability, represented by broad spectra of ARI across many simulations. Categorical simulations displayed poorer performance, even at larger feature spaces. Even at simulations with 243 features, ordinal simulations presented broad, variable spectra. Nominal data, characterized by poor performance at 81 or fewer features, presented with improved, though variable, performance at 243 features. This pattern of poor performance at low feature numbers and improved performance at higher feature spaces was also present with the DAISY and Mercator algorithms across data mixtures.

Figure 4.10 Lattice violin plot of ARI of continuous simulations by number of features and patients with 3 algorithms and Euclidean distance.

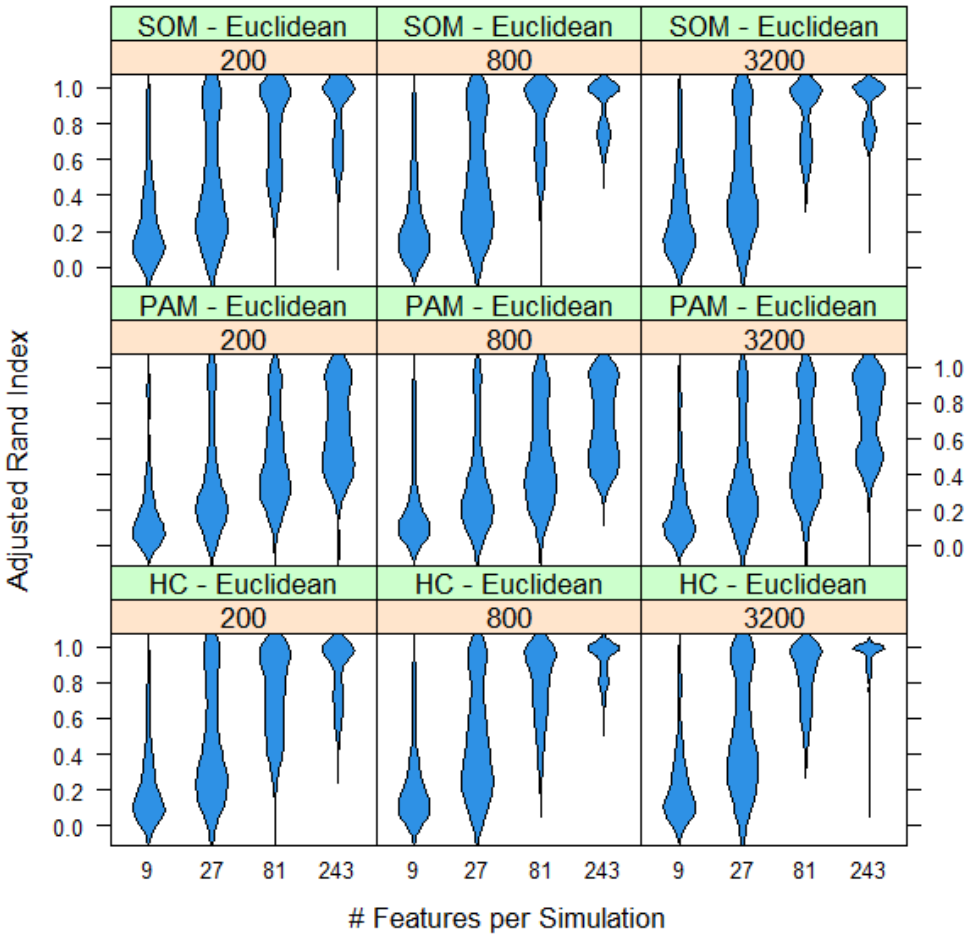


Figure 4.11 Lattice violin plot of ARI of binary simulations by number of features and patients with 3 algorithms with Euclidean and Tanimoto distance.

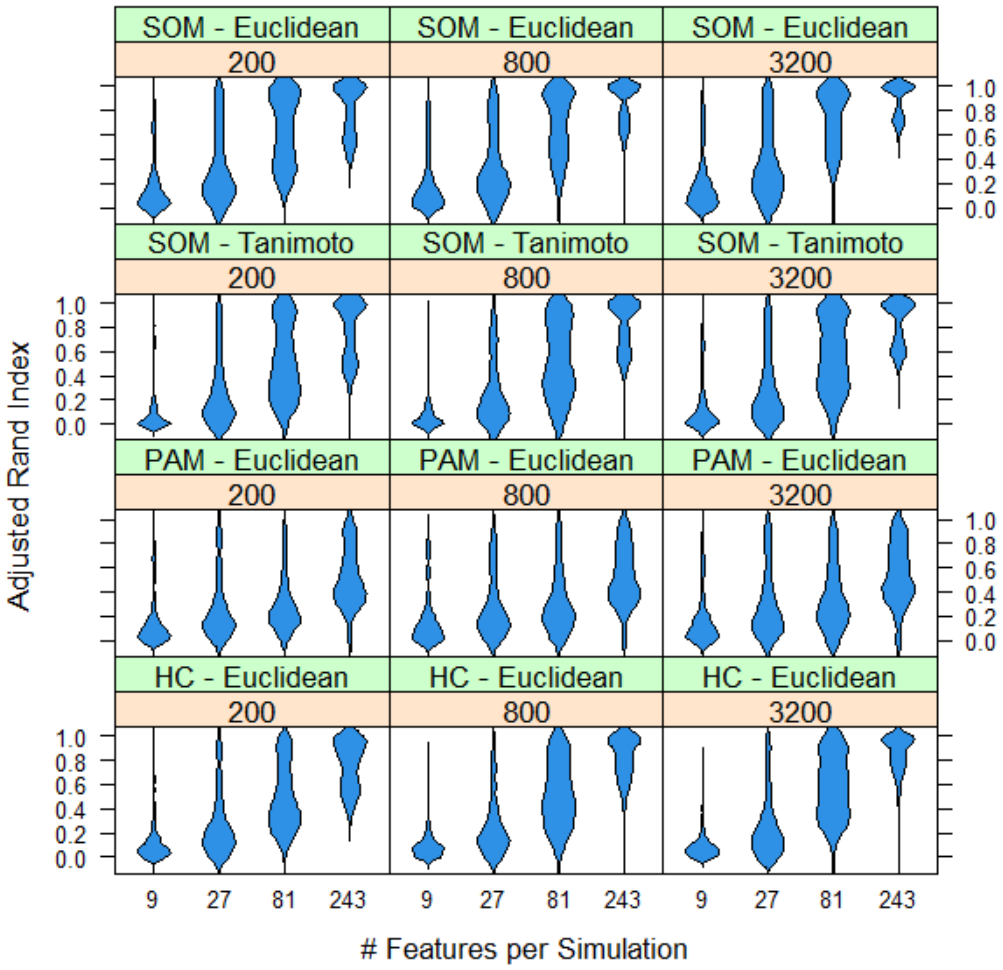


Figure 4.12 Lattice violin plot of ARI of nominal simulations by number of features and patients with 3 algorithms and Euclidean distance.

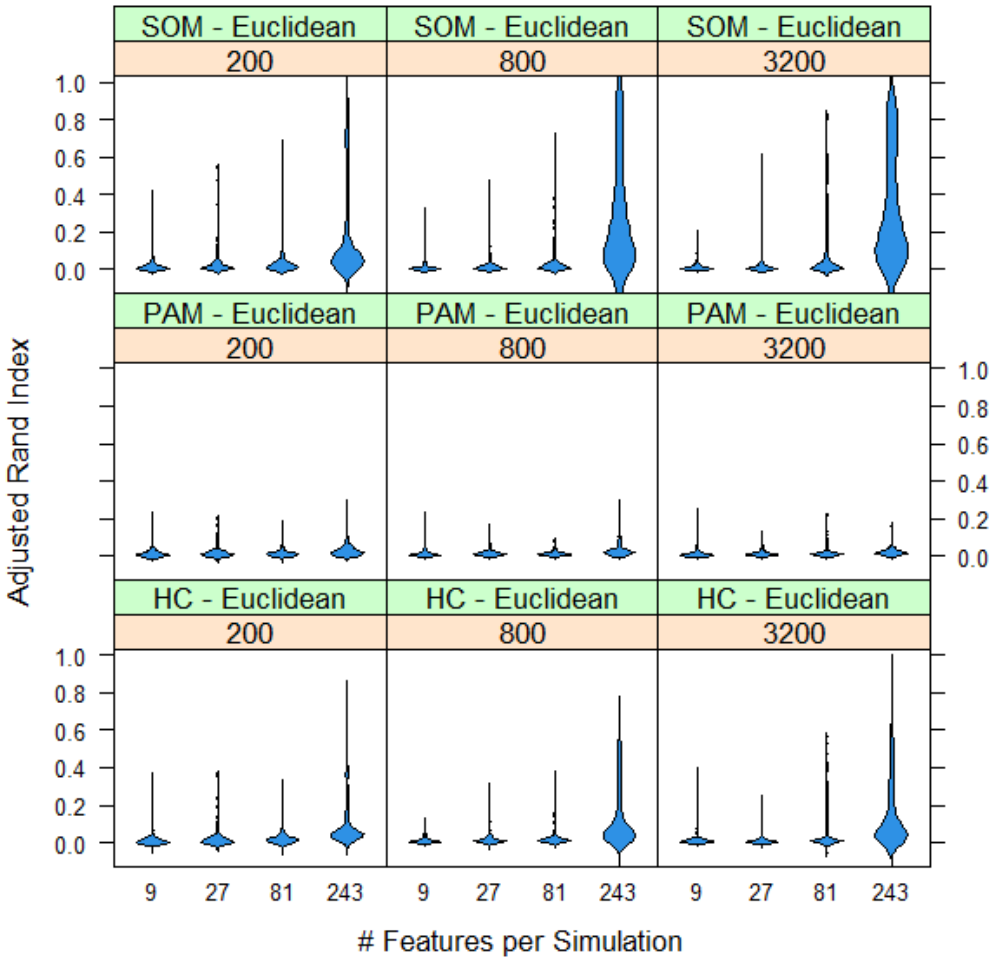


Figure 4.13 Lattice violin plot of ARI of ordinal simulations by number of features and patients with 3 algorithms and Euclidean distance.

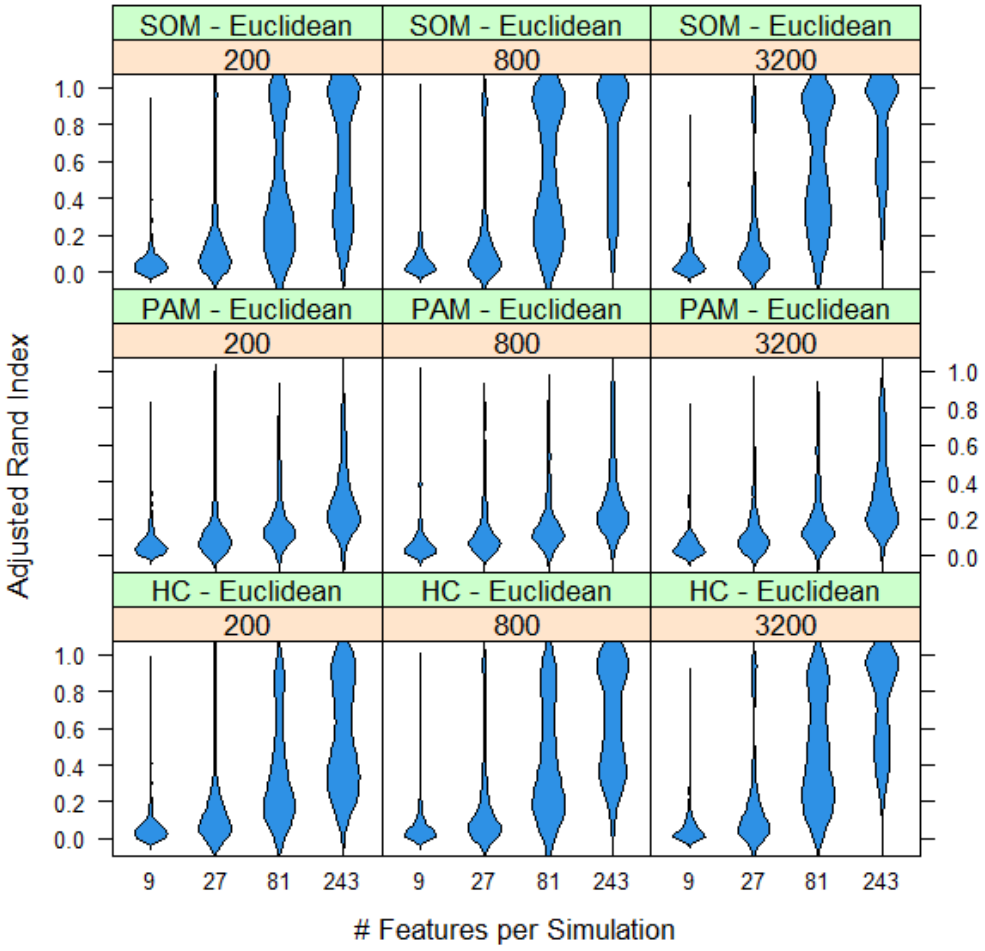


Figure 4.14 Lattice violin plot of ARI of 4 dat mixtures with the DAISY distance algorithm and hierarchical clustering.

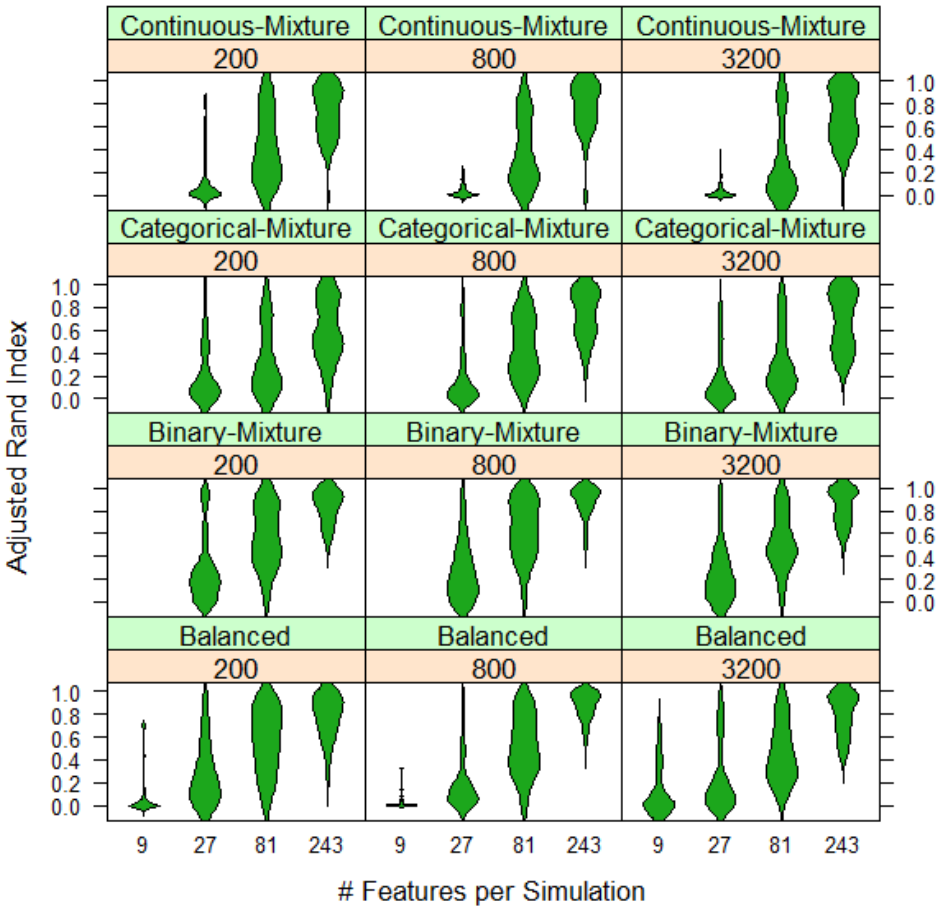
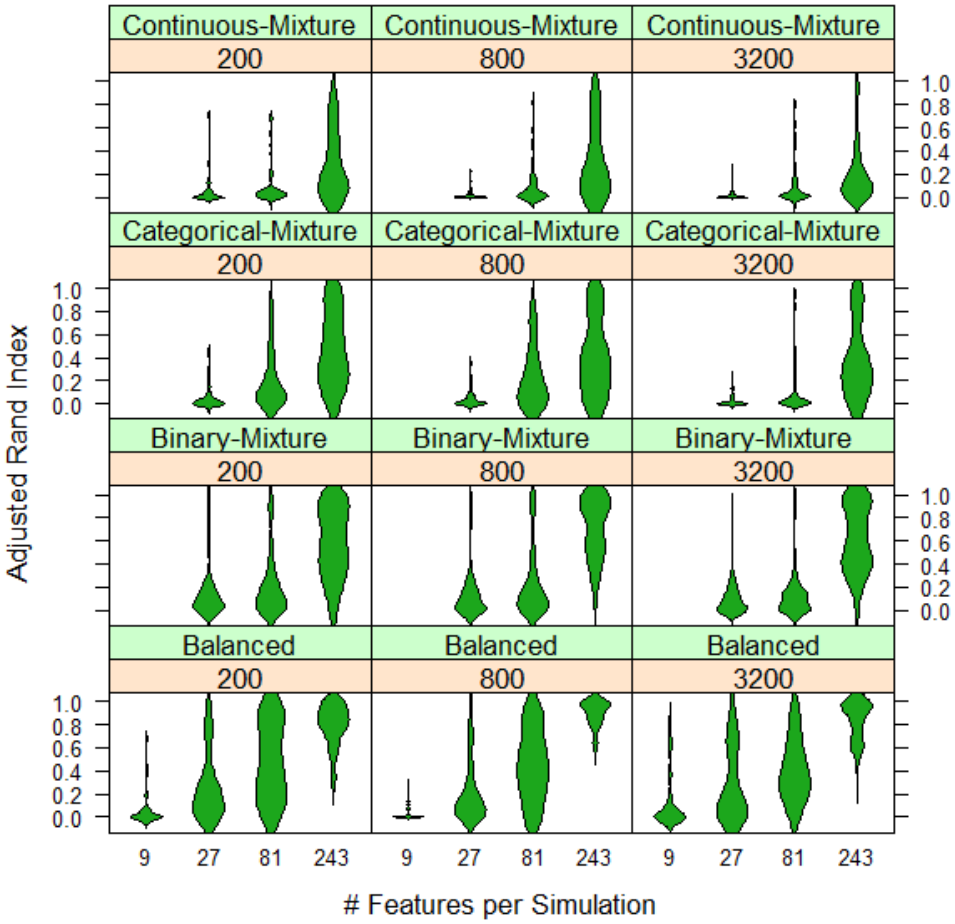


Figure 4.15 Lattice violin plot of ARI of 4 dat mixtures with the Mercator distance algorithm and hierarchical clustering.



Computational Scalability of Mixed-Distance Methods

CPU time to calculate a mixed-distance dissimilarity matrix varied by algorithm. Time costs predominantly resulted from time to calculate dissimilarity, not from time to execute a clustering algorithm.(Table 4.9) SOM, which calculates dissimilarity and clusters within a single process, had fastest overall execution for any simulation size (mean 0.533s), while the DAISY algorithm had the slowest CPU time averaged over all simulations sizes (mean 0.372s).

Table 4.9 Computational (CPU) time (s) for 3 algorithms to calculate mixed-distance dissimilarity

Dissimilarity Algorithm	Dissimilarity Time (s)	Clustering Time (s)	
		HC[1]	PAM[2]
DAISY	372.461 ± 983.599	0.105 ± 0.142	1.623 ± 3.611
Mercator	99.859 ± 139.127	0.097 ± 0.13	1.598 ± 3.46
Supersom[3]	0.533 ± 0.794	-	-

[1] Agglomerative hierarchical clustering with Ward's criterion
[2] Partitioning Around Medoids
[3] Kohonen self-organizing maps and their related mixed-distance implementation calculate dissimilarity and cluster in a single process.

Time to calculate dissimilarity with DAISY varied (min = 0.3s; max = 3869.72s = 1hr4.5m). Mean CPU time for DAISY increased with increasing numbers of features and patients.(Figure 4.16) Mean time to cluster a data set with 200 patients was 12.88s, while mean time to cluster a data set of 3,200 patients was 1039.32s. Mean time to cluster a data set with 9 features was 2.42s, while mean time to cluster a data set of 243 features was 1111.907s. By interaction, calculating the DAISY dissimilarity was slowest in simulations with both large numbers of features and large numbers of patients.(Figure 4.17) Comparatively, variation between data types was statistically significant ($p < 0.0001$) but limited in range with balanced data requiring shortest mean time over all mixed-type simulations (292.89s) and unbalanced binary simulations requiring longest CPU time (489.05s). There was no significant difference in time to calculate dissimilarity by number of clusters in a simulation ($p=0.99$).

Figure 4.16 Mean CPU time to calculate DAISY dissimilarity for 4 simulation parameters

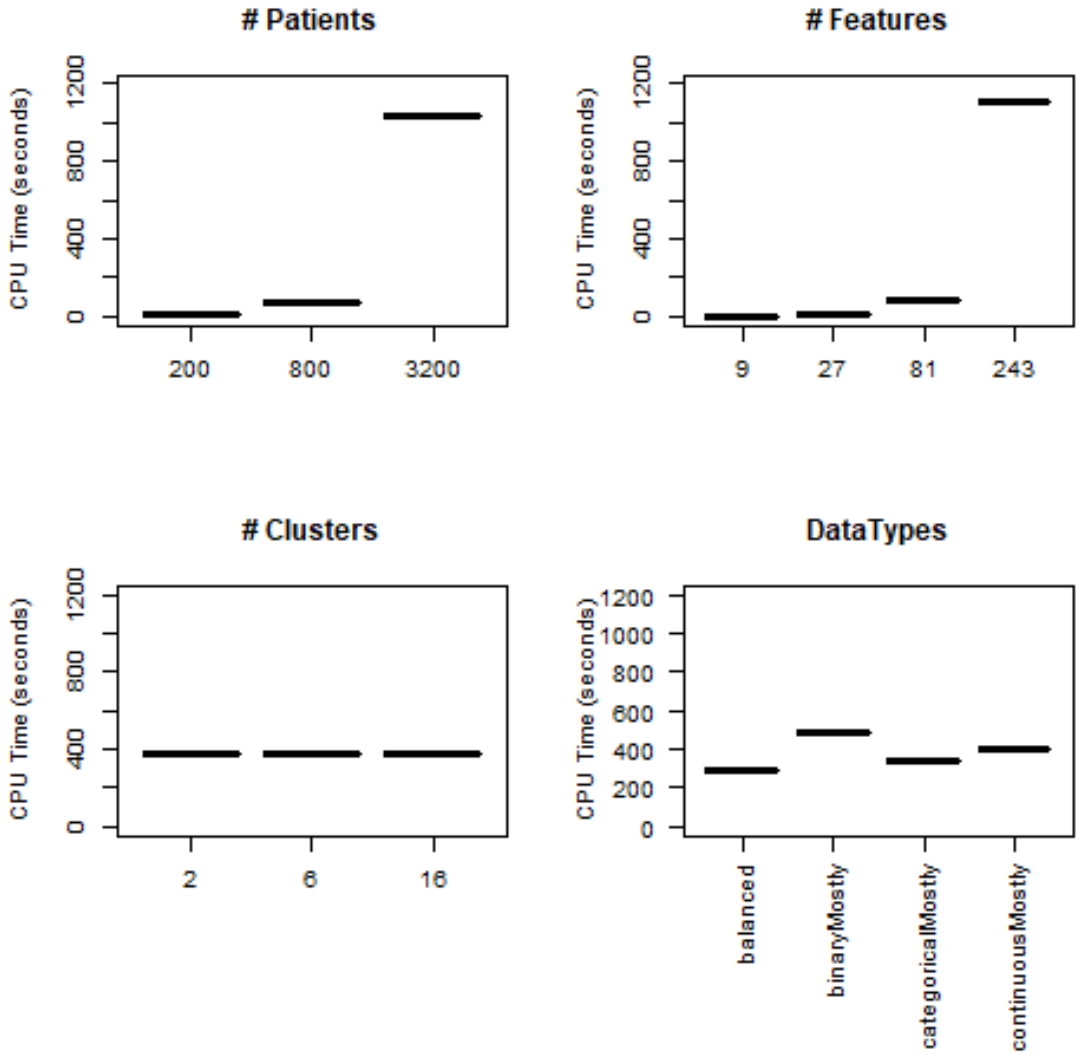
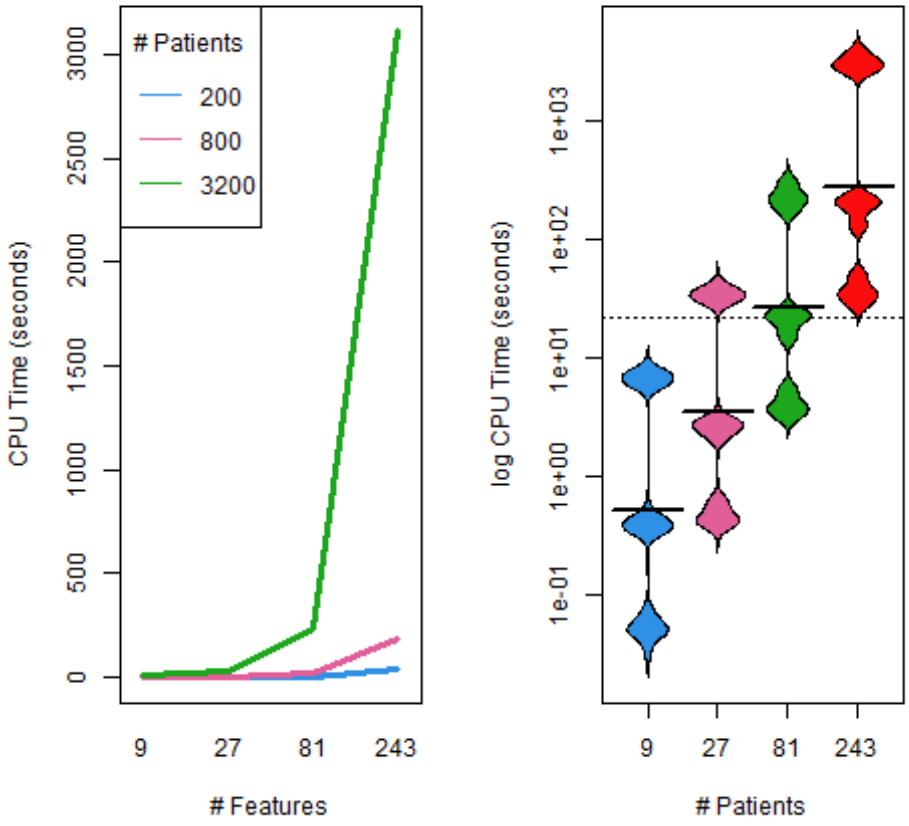


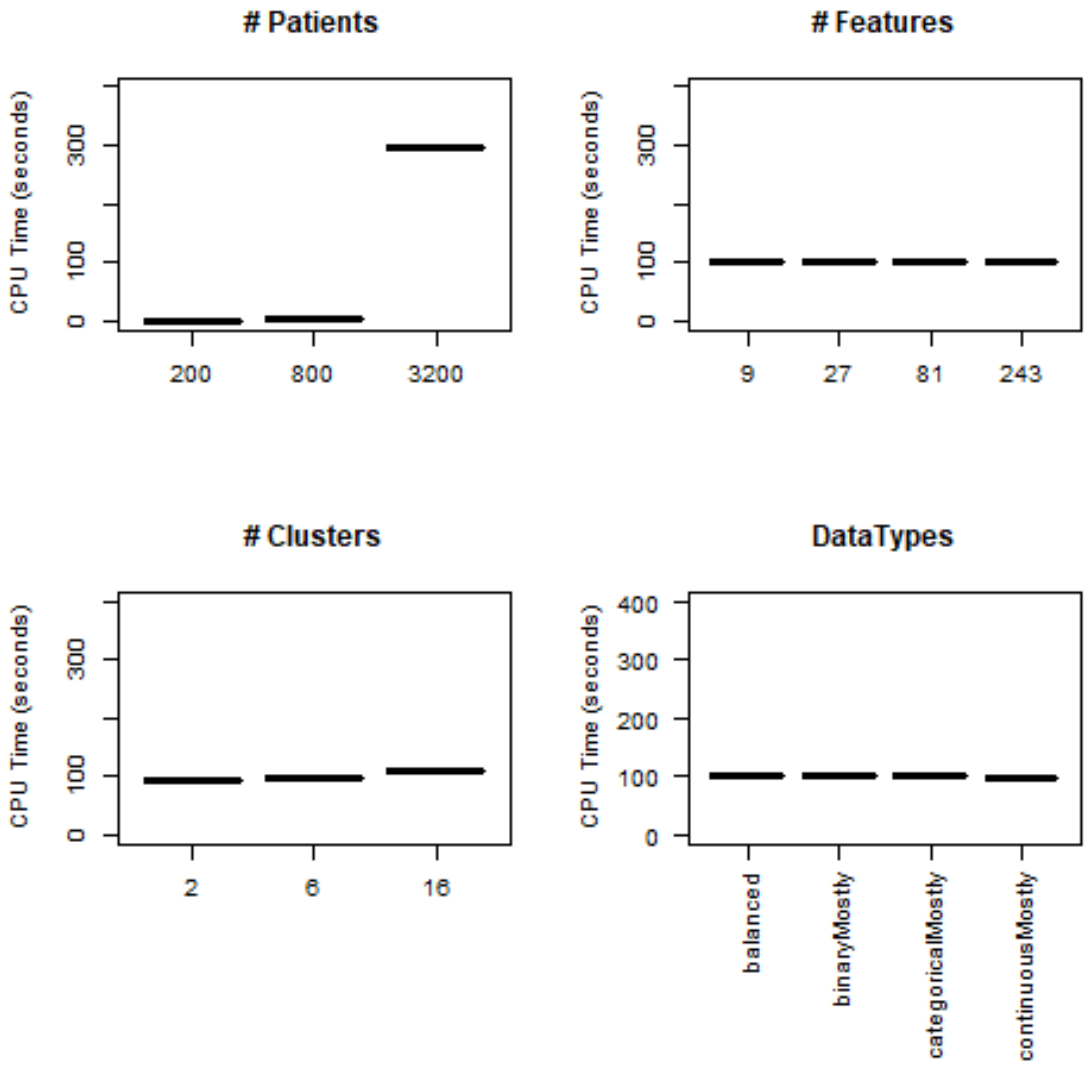
Figure 4.17. CPU time to calculate DAISY dissimilarity from the interaction of number of patients and number of features in a simulation.



Time to calculate dissimilarity with Mercator also varied (min = 0.01s; max = 405.64s). However, both mean and max CPU time for the set of simulations was shorter than that of DAISY. Mean CPU time for Mercator increased with increasing numbers of patients.(Figure 4.18) Mean time to cluster a data set with 200 patients was 0.080s, while mean time to cluster a data set of 3,200 patients was 295.66s. Variation by number of clusters in a simulation was statistically significant (p=0.005) but comparatively limited in range, with shortest time for 2 clusters (93.80s) and longest time for 16 clusters

(109.00s). There was no significant difference in time to calculate dissimilarity by number of features (p=0.83) in a simulation or between different data types (p=0.98).

Figure 4.18. CPU time to calculate Mercator dissimilarity for 4 simulation parameters



Discussion

In Chapter 1, we described 4 recently published studies that compared methods for clustering clinical data.[10, 11, 47, 49] The studies undertaken here are more extensive in

number, testing a larger number of methods than these studies, which appraised 2-3 approaches each. More importantly, this chapter applied these tests to simulated data with a known ground truth for comparison, allowing a more rigorous and nuanced validation of methods than possible on real clinical data. Chapter 5 in this manuscript follows in their footsteps with an application of the best methods described here to real, clinical data.

In our analysis of mixed data, we compared three measures of mixed distance. DAISY represents an old standard: an algorithm developed in the late 1980's implementing, primarily, a distance coefficient developed in the early 1970's. Supersom is a recent algorithm proposed as an extension to an existing neural net technology. Mercator is a proprietary solution, based on a pipeline we developed, to implement a wider variety of distance metrics. Surprisingly, Supersom were markedly outperformed by SOM using single distance alone, but these are troubled by a bipolar distribution of outcomes that suggests variability in results.

DAISY produced the highest mean ARI for mixed data types for all mixtures except unbalanced mixtures dominated by continuous data. It is notable that an accurate solution, as measured by a high ARI, may present with a poor silhouette width. This observation should be of note and concern for researchers using silhouette widths to drive the selection of a particular algorithm or solution over another. DAISY produced broad ranges of solutions, suggesting that some have a high degree of accuracy and others lower. This suggestion of variability is of concern. DAISY also presented with an unexpected impediment to usability. Initial tests revealed extensive computational time

needed to calculate the DAISY dissimilarity. It is for this reason that the tests on mixed-type data are performed on only a limited number of unique simulation repeats (30 instead of 100). A researcher attempting to implement DAISY on a personal computer on very large data may find the time cost prohibitive.

Mercator performed poorly compared to DAISY on all mixed types except for balanced data. However, it can be calculated, without optimization, much faster than DAISY, improving usability. Because Mercator is, at the moment, an unweighted combination of distance metrics, good performance on balanced data and mediocre performance on unbalanced mixtures is unsurprising. Future directions for a potential mixed-distance extension include pursuing weighting measures to improve application to unbalanced distance measures, potentially providing an option with comparable accuracy but reduced computational intensity. Furthermore, Mercator could be used to overcome the limitation in choice of distance metrics present in other packages, such as DAISY or Supersom, which confine the user to limited implementations. Offering the user greater choice in distance metric could allow improved customization for data mixtures.

An important limitation arises in small data sets for both DAISY and Mercator: distance cannot be calculated within a given type if only one feature of that type is present. In this case, a single feature of a type is lost to analysis, removing important information. While this scenario is unlikely in large data sets, it is important in data sets with small features spaces, which featured prominently in our review in Chapter 1. The documentation of Supersom makes no note about handling single-type features, so it is

unclear if the package resolves the issue internally or if a lone feature is also lost to analysis.

Although we set out to study mixed-type data, this study revealed important conclusions about the analysis of single data types. First, we improved our understanding of the Sokal & Michener distance, and, by extension, the Hamming distance. Although the Hamming distance is commonly implemented for the handling of binary data in mixed-type studies, as we saw in chapter 3, this study shows little improvement in performance over the other distances assessed, including both distances commonly used for binary (Jaccard index) and continuous (Manhattan, Euclidean) data.

Perhaps more importantly, although we implemented distance metrics in common use for all single data types tested, we noted a strong disparity in ARI and SW between data types. Specifically, performance was good for continuous and binary data, intermediate for ordinal data, and poor for nominal and mixed categorical data. While high-quality solutions for binary and continuous data exist, and these can be implemented with overlapping distance metrics (i.e. Manhattan or Euclidean), we were unable to identify a strong solution for categorical data in this study. Given the frequency of categorical features in mixed clinical data, the absence of quality methods for this data type is concerning for analyzing mixed data problems. Future work analyzing mixed data may need to turn its head to solutions to generate high-quality clusters on their own, and then return to questions of combining distance metrics described here.

These studies provided unexpected insight, not only into the distance measures implemented, but into the algorithms chosen. PAM represents a newer algorithm than our

implementation of HC with Ward's criterion. Unlike HC, PAM was developed with the intention of larger data sets analyzed on computers. It is in common use. SOM, implementing a neural net, also represent a more progressive technology. We included HC as a common standard, expecting that it would be outperformed by both techniques. What we found, however, was inconsistent performance of SOM (represented by the bipolar distribution of ARI). We also found almost universal performance of HC over PAM by mean ARI and mean SW. However, although PAM presented with lower ARI, it produced more stable solutions, as indicated by narrower standard deviation. Variability in a solution, seen in HC and SOM, can be as important a problem as inaccuracy. All 3 algorithms carry benefits and risks, and the selection of one over another should be undertaken with grounding in the literature and attention to the data and the researcher's goals, not by going along with a default.

In addition to quantitative measures (mean, range, etc.) to describe ARI and SW, we also described these methods qualitatively through the inspection of bean plots. The use of bean plots allowed us to describe the distribution of these values with greater nuance than mean and range alone. Importantly, they showed that a mean ARI often reflects the presentation of a wide range of values, either in the form of a spectrum (such as seen with DAISY) or a two-headed distribution (as seen with SOM). What is revealed is algorithms producing variable results: some excellent and some problematically poor. This is especially relevant in the bipolar distribution of SOM solutions, which almost exclusively produced excellent or poor results, with little middle ground. Further inspection revealed that the source of variability resulted from variation in feature size,

with small feature spaces resulting in solutions with low accuracy and large feature spaces resulting in more reliable solutions. Conversely, the number of patients in a simulation did not have a strong effect on the ARI of a solution. This suggests that there is some minimum threshold of feature number (approximately 200) that is necessary to undertake a clustering analysis in clinical data where a solution becomes more reliable.

Not every clinical study will have at its disposal 200 or more features. In some cases, this may suggest that no clustering analysis should be performed. However, at intermediate ranges (approximately 100 features), solutions can be unreliable, but some solutions are of high quality. Even at large feature spaces, variability is present in solutions, particularly among categorical data types. In the absence of ground truth (which is obviously unavailable or there would be no point in undertaking clustering resource), an important challenge remains in identifying if the solution before the researcher is of good quality. In the absence of this test, certain techniques with high levels of variability, even if they often produce excellent solutions, may be too risky for research implementation. Important future work could seek to correlate other measures of validation that could be applied to real data with ARI, in order to identify a test that would indicate more accurate solutions. Given the numerous techniques available, including both external assessors of randomness and internal measures of consistency [79], identifying such a measure may provide fruitful. Because the simulations produced in Chapter 3 provide an opportunity for ground truth studies mimicking clinical data, such a study is feasible.

Chapter 5. Concluding Experiment: Clustering Mixed-Type Data

As a final experiment and test of concept, we return to our experiment from Chapter 2 with best methods gleaned from our tests of algorithm-distance pairs in Chapter 4.

Methods

Here, we reanalyze data collected on 247 patients with chronic lymphocytic leukemia (CLL). These data include clinical signs and symptoms, physical exam results, laboratory value, immunophenotyping, and genetic and cytogenetic markers. Cytogenetic abnormalities are stored in two forms. The Döhner classification identifies a hierarchy of abnormalities to classify a cytogenetic phenotype. Although patients have a single Döhner classification, more than 1 cytogenetic abnormality may be present. Both the Döhner classification and the 4 individual markers for cytogenetic abnormality are included here.

The data consists of 21 features: 4 continuous features, 2 nominal features, and 15 binary features, 10 of which are symmetric binary and 5 of which are asymmetric binary. As this dataset contains 71.4% binary features, we classify it as an unbalanced, binary-dominant mixture. Following the best performer for unbalanced, binary-dominant mixtures from Chapter 4, we calculate distance using the DAISY algorithm and implement hierarchical clustering using Ward's agglomerative method. To select the number of clusters *k*, we plotted and selected peak average silhouette width at a range of

k from 2 to 12 clusters. Salient categorical features were uncovered as the most common category within a feature occurring in a given cluster. Salient binary features were defined as positive (present in >75% of patients in a given cluster) or negative (present in <25% of patients in a given cluster). Continuous features were described by mean and standard deviation, with differences in these values between clusters tested by ANOVA. We visualized using t-Stochastic Neighbor Embedding (t-SNE). Patient overall survival was modeled by Cox proportional hazard and visualized with Kaplan Meier curves.

Results

Based on peak silhouette width, we recovered 5 clusters (mean silhouette width = 0.26). Visualized with t-SNE, patients formed coherent, well-separated clusters.(Figure 5.1)

Cluster A, the largest cluster, contained 93 (37.7%) patients. Cluster E, the smallest cluster, contained 10 (4.0%) patients. Table 5.1 describes salient features and frequencies defining each of the 5 clusters. Salient differences characterizing cluster A included cytogenetic abnormality del13q, low CD38, and low beta-2-microglobulin. Cluster B was characterized by normal karyotype. Cluster C was characterized by del11q cytogenetic abnormality, unmutated *IGHV* status, positive ZAP70 expression, male sex, absent hypogammaglobulinemia, and the lowest prolymphocyte count of any of the 5 clusters. Cluster D was characterized by trisomy 12 karyotype, male sex, and low beta-2-microglobulin. Cluster E was associated with the del17p cytogenetic abnormality, unmutated *IGHV* status, hypogammaglobulinemia, and the highest prolymphocyte count of the 5 clusters. In all clusters, the dominant race was white, white blood cell count was

low, and massive splenomegaly was absent. Although prolymphocyte count varied, there were no statistically significant differences in continuous features between clusters.

The 5 clusters recovered were significantly associated with overall survival (Cox proportional hazard, p = 0.0108).(Figure 5.2) Patients in cluster A had longest survival. Patients in clusters B, C, and D had intermediate survival. Patients in cluster E had shortest survival.

Figure 5.1. Visualization of 5 clusters of 247 patients with chronic lymphocytic leukemia with t-Stochastic Neighbor Embedding (t-SNE). T-SNE recovers 4 coherent clusters (A-D). A final cluster captures a small group of outliers (E).

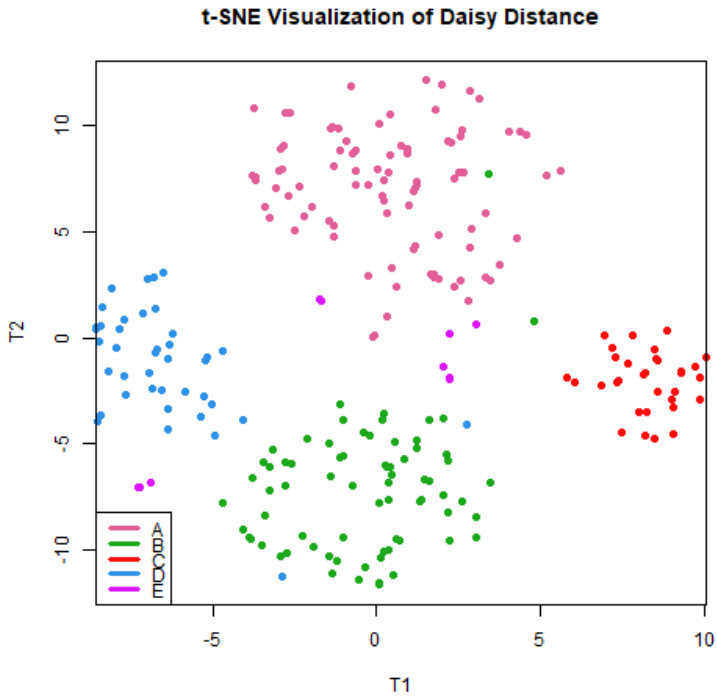


Table 5.1. Salient features for 5 clusters of CLL patients recovered with hierarchical clustering and the DAISY dissimilarity metric.

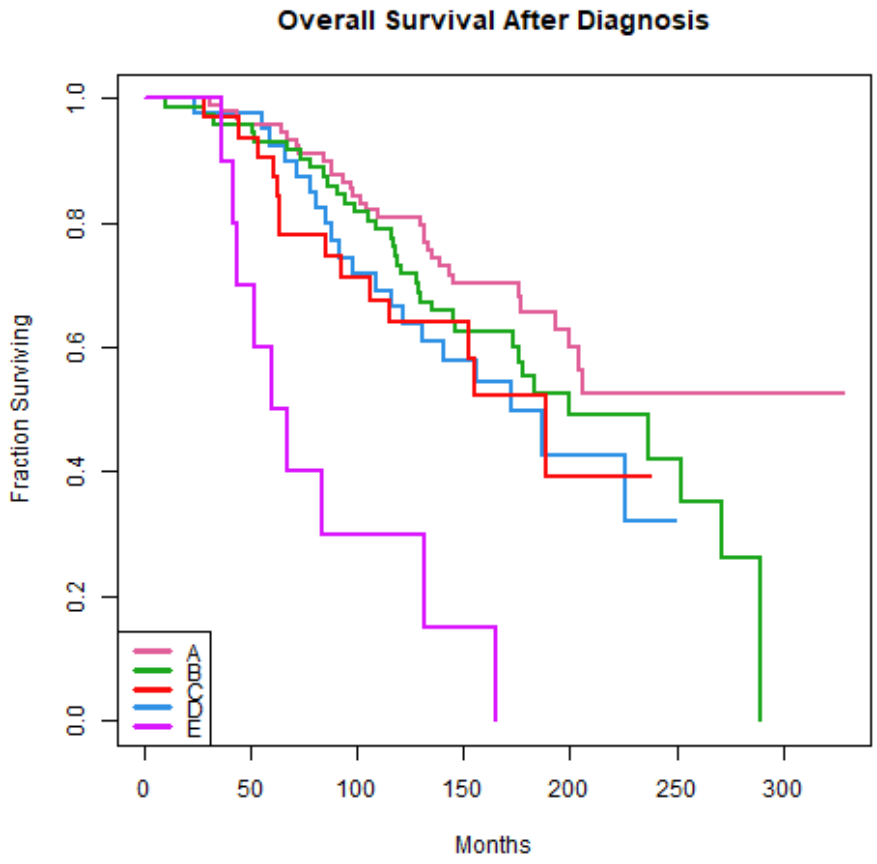
Cluster	**A**	**B**	**C**	**D**	**E**
Patients (n)	93	72	32	40	10
Genetics					
Döhner Classification	**Del13q 95.7%**	**Normal 97.2%**	**Del11q 100%**	**Trisomy 12 92.5%**	**Del17p 100%**
13q	***Abnormal 100%***	*Normal 97.2%*		*Normal 82.5%*	
12	*Normal 100%*	*Normal 100%*	*Normal 96.9%*	***Abnormal 65.0%***	*Normal 100%*
11q	*Normal 97.8%*	*Normal 98.6%*	***Abnormal 100%***	*Normal 100%*	*Normal 100%*
17p	*Normal 97.8%*	*Normal 98.6%*	*Normal 100%*	*Normal 100%*	***Abnormal 100%***
IGHV mutation status			**Unmutated 93.8%**		**Unmutated 100%**
Demographics					
Race	White 95.7%	White 94.4%	White 90.6%	White 85%	White 80%
Sex			**Male 84.4%**	**Male 75%**	
Age at Diagnosis	55.7 ± 10.9	54.8 ± 11.1	58.4 ± 10.9	58.7 ± 10.1	60.0 ± 10.5
Clinical signs					
Rai Stage	Low 76.3%	Low 75.0%	Low 93.8%	Low 85%	
Massive splenomegaly	Absent 92.4%	Absent 90.3%	Absent 93.8%	Absent 92.5%	Absent 100%
Laboratory values and immunophenotype					
ZAP70 expression			**Positive 75.0%**		
Beta-2 microglobulin	**Low 83.7%**			**Low 75%**	
WBC*	Low 90.2%	Low 79.2%	Low 87.5%	Low 87.5%	Low 80%
CD38	**Low 88.4%**				
Light chain subtype		Lambda 75.7%	Lambda 84.4%		Lambda 90.0%

Table 5.1 continued

Matutes score	Typical 88.2%	Typical 86.8%	Typical 87.5%		
Hypogamma-globulinemia			**Absent 75.9%**		**Present 77.8%**
Hemoglobin (mean ± sd)	12.9 ± 1.8	12.7 ± 1.7	13.6 ± 1.5	12.7 ± 1.6	11.7 ± 2.7
Platelets (mean ± sd)	173.6 ± 63.7	176.8 ± 82.5	202.3 ± 74.6	187.4 ± 78.2	193.1 ± 109
Prolympho-cytes (mean ± sd)	3.9 ± 4.1	4.4 ± 3.2	**3.5 ± 4.5**	7.2 ± 5.9	**9.9 ± 7.0**

*WBC = White blood cell count

Figure 5.2. Kaplan Meier curves of overall survival after diagnosis for 5 clusters of 247 patients with chronic lymphocytic leukemia. The dominant cytogenetic abnormality by Döhner classification for each cluster is: A = Del13q, B = normal karyotype, C = Del11q, D = Tri12, E = 17p.



Discussion

In this concluding experiment, we demonstrate that we can successfully recover the most important prognostic marker in CLL, Döhner classification, by unsupervised clustering on a clinical data set. We returned to our data set from chapter 2 with methods informed by knowledge gained in the production of this thesis. Identifying our data as an unbalanced, binary-dominant mixture, we applied the DAISY dissimilarity metric with hierarchical clustering: the best performing method for this data type identified by our experiments in chapter 4. Applications of these methods returned superior clusters compared to the initial experiment in chapter 2. Quantitatively, this concluding experiment uncovered clusters with a higher average silhouette width (initial average silhouette width = 0.17, concluding = 0.26). Qualitatively, visualized clusters in the concluding experiment had better tightness and separation.

The recovered clusters were characterized by high frequencies of prognostic markers, in which the categorical Döhner classification, understood to be the most important prognostic marker in CLL, was a defining feature. This approach also recovered important binary prognostic markers, including sex, ZAP70 expression, and *IGHV* mutation status. Notably, this clustering approach failed to reveal statistically significant variation in continuous features. Although this could represent a limitation of the DAISY dissimilarity algorithm, other methodological concerns may merit further study. In some cases, continuous features may be more meaningfully represented as binary or ordinal features. For example, hemoglobin is relevant in clinical use as a marker of anemia (low hemoglobin) or polycythemia (high hemoglobin): The meaningfulness of

this continuous measure is as a categorical construct. In the same way, platelet count is meaningful in that it represents low, normal, or high values, but small, numeric variations are not clinically interpreted to have substantial meaning. The clinical interpretation of continuous features may need to drive their data type.

In other cases, population homogeneity could also reduce the distinctiveness of a feature's contribution to cluster formation. For instance, all 5 clusters recovered an age at diagnosis between 55 and 60 with a standard deviation of approximately 10. CLL is a disease of older adults. Older age greater than 65 years is understood to imply poor prognosis. However, in a small sample, age may not be a dominate driver of subclassification. This raises the point that clustering and subclassification is only one problem solving tool in understanding prognosis, risk, and outcome. Some important risk factors may not drive clustering, and additional approaches, including regression and supervised methods, must be brought to bear to fully understand these processes.
These recovered clusters were strongly associated with overall survival. Survival is concordant with survival predicted by the Döhner classification, which predicts longest survival in patients with Del13q (Cluster A) and shortest survival in patients with Del17p (Cluster E). Among intermediate phenotypes, the Döhner classification predicts intermediate survival among patients with Tri12 and normal karyotype with poorer survival among patients with Del11q. In these data, the cluster characterized by normal karytotype (Cluster B) presents with intermediate survival. The survival curves of patients in clusters characterized by Tri12 (Cluster D) and Del11q (Cluster C) are more similar than predicted. Although this distortion in expected survival outcomes could

represent limitations in this clustering method, the size of this data set (247 patients) could be small enough that sample variability is the root of this result.

Some limitations of this approach did arise. In this data set, key prognostic information from cytogenetic tests was stored in 5 closely related fields comprised of a categorical field for Döhner classification and binary fields marking presence or absence of these 4, sometimes co-occurring, abnormalities. The clusters recovered were dominated by high and well-separated frequencies of these entities. This indicates that, especially in small data sets, a subgroup of features can drive the clustering process through relatedness of concept and collinearity. It merits future exploration if relatedness among different data types, such as the combination of categorical and binary fields, may more strongly drive a mixed-type clustering process than a single type. These limitations are in greater focus due to the small size of the data set: Small numbers of features increase the ability for a small subgroup to dominate clustering. In addition, we have seen in Chapter 4 that when some mixed-distance dissimilarities are calculated (i.e. with Mercator or DAISY) data types represented in only one feature may be lost to analysis. Although we described in Chapter 1 that clustering of data sets with small population sizes and small features spaces is common in unsupervised analysis of clinical data, we may need to consider that there is some minimum threshold below which clustering is inappropriate.

Chapter 6. Conclusion

This thesis set out to tackle core problems in unsupervised clustering of clinical data: clinical data are of heterogeneous size and mixed type. We hypothesized that these limitations could be overcome by implementing methods of dissimilarity using a mixture of formulae to calculate distance and undertook 5 steps towards a potential solution.

In Chapter 1, we began with a review of the current state of the literature for common clustering approaches, common applications of distance metrics for mixed-type data, and existing studies comparing methods for mixed-type, clinical data. We found that, although there were many clustering algorithms at our disposal, solutions for mixed-type data, particularly within the clinical literature, were sparse and lacked rigor. Common approaches contained limited applications of distance-method mixtures, which encouraged us in our approach.

In Chapter 2, we applied a solution suggested in the literature for handling mixed-type data: converting all features to a single data type. We chose a real data set on a disease with well-understood prognosis: chronic lymphocytic leukemia. A disease with well-understood prognosis does not allow us to make grand contributions to new understanding in the literature. However, perhaps more importantly for assessment of a method, a well-understood disease provides “biological validation,” allowing us to test the success of our clustering method against a standard. We transformed mixed data,

containing categorical, continuous, and binary features, to a binary feature space and clustered. Our success was mixed: important binary features were salient, but a crucial categorical feature – the Döhner cytogenetic classification, known to be one of the best prognostic markers for the disease – was not uncovered by the analysis. With an understanding of the limitations of a transformation to a single data type, including information loss and distortion in an unweighted transformation, we undertook further steps to explore our core problems.

A fundamental problem in the assessment of clustering algorithms is the inability to test a method against a known "ground truth." In Chapter 3, we tackled this problem by developing realistic simulations (i.e. noisy, complex, and heterogeneous) of mixed-type clinical data. We validated these simulations in three ways. First, we qualitatively inspected T-distributed Stochastic Neighbor Embedding (t-SNE) plots for realistic cluster shape. Second, we quantitatively assessed cluster form, compactness, and separation by silhouette width. Third, we assessed the statistical association of generated features with their cluster identity, taking this to signify clusters with sound foundations. Confirming that our simulations were biologically representative and of high quality, we generated 32,400 simulations with parameters representing clinical data ranging from clinical trials to large cohorts.

With a "gold standard" available, in Chapter 4 we undertook a series of 18 tests of dissimilarity-algorithm pairs on single- and mixed-type data. These pairs consisted of the application of 5 related single distance metrics for binary, categorical, or continuous data (Jaccard index, Sokal & Michener or "Tanimoto" distance, Gower coefficient, Manhattan

distance, and Euclidean distance) and 3 methods for calculating dissimilarity from multiple distance metrics (DAISY, Supersoms, and our own proposed solution, Mercator). We applied these metrics with 3 algorithms: agglomerative hierarchical clustering with Ward's criterion, Partitioning Around Medoids (PAM), and self-organizing maps (SOM). Our analysis suggested the superiority of some solutions, particularly DAISY, for mixed type data. However, visualization with bean plots, which qualitatively captures a frequency distribution of measurements across a range, raised important concerns for reliability and reproducibility of clustering solutions across all simulated data types tested.

As a return to biological validation of a method, we returned to our data set from Chapter 2 using the best method suggested by our tests in Chapter 4: Ward's method of hierarchical clustering with the DAISY dissimilarity method. This time, we captured 5 clusters reflecting Döhner classification. However, our results raised concerns for bias that can arise in small data sets and small feature spaces.

The results of this thesis raise important concerns and avenues for future directions in clustering clinical data – or data of any type. Two key concerns arose that raise avenues for future exploration. First, evidence from our simulations, our tests in Chapter 4, and our concluding experiment in Chapter 5 suggest that some minimum threshold of patient population and, particularly, feature space is necessary for an optimal clustering solution. These experiments suggest that the impact of a number of patients greater than 200 may be small, but that large number of features (e.g., 200 or greater) may be required to obtain accurate and reliable clustering solutions. Future experiments

with simulations and/or random subsets of varying and progressive size from a large data set are important steps to define this standard. Finally, evidence of variability of solutions generated by clustering algorithms is of great concern. When an algorithm returns either an excellent solution or one that is very wrong, any answer it provides must be held as suspect. When the appropriate distance metric is employed, reliability in some large sample sizes and feature spaces is encouraging. However, the researcher may only have a feature space of intermediate size available for study, and issues with variability remain, particularly in categorical data types. Although it could be argued that clustering analyses of insufficiently large feature spaces should not be undertaken, important discoveries may arise from data sets with increased risk of variability. Future work testing methods of validation not requiring a gold standard (such as measures of entropy or intrinsic properties) for correlation with measures of ground truth (such as adjusted rand index) may be important steps to validate a solution on these data sets. In the absence of these measures, some clustering solutions with smaller feature spaces must be viewed with suspicion. Clustering holds promise for important subclassification problems in chronic and acute clinical medicine, but work remains to ensure that solutions are rigorous, valid, and reproducible.

In Chapter 3, we demonstrated that we can realistically simulated mixed-type, clinical data. These simulation methods may provide the tools needed to hone clustering approaches for mixed-type and clinical data. Because of easy manipulations of feature number and sample number, these simulations could be used to explore appropriate solutions for varying data set size. Such simulated data could establish boundaries for an

appropriate sample size and feature space to obtain a meaningful clustering solution or to explore new methods appropriate for data sizes that prove challenging. Because the simulations produced in this modification of Umpire generate a known, "ground truth" clustering assignment, they can be used to address variability of clustering results. By testing correlation between methods of cluster validation that do not assess ground truth (e.g., tests of entropy or intrinsic cluster properties) and gold standard validation (e.g., Adjusted Rand Index, Adjusted Mutual Information), these simulations could be used to identify a measure of validation for real data that better indicates a "true" solution. Finally, realistic, mixed-type, clinical simulations can be used to develop more advanced methods for clustering noisy, mixed-type data. Thus, although Chapter 4 identified important problems in clustering clinical data, Chapter 3 provides future directions to solve them.

Clustering analysis in clinical contexts hold promise to improve the understanding of patient phenotype and disease course. Better subclassification of disease opens avenues for targeted treatments, precision medicine, and improved patient outcomes. Diseases with high morbidity and mortality but subtle presentation, such as sepsis and delirium, can benefit from subclassification to identify treatment response groups, prognosis, and underlying etiology. As an extension, rigorous subclassification could be used to inform clinical decision support and improve practical treatment outcomes. We hope this work opens the door for improvements in methods for unsupervised ML that can make these important advances a reality.

References

1. Raghupathi, W. and V. Raghupathi, *Big data analytics in healthcare: promise and potential.* Health Inf Sci Syst, 2014. **2**: p. 3.
2. Cook, J.A. and G.S. Collins, *The rise of big clinical databases.* British Journal of Surgery, 2015. **102**(2): p. e93-e101.
3. Xu, R. and D.C. Wunsch, 2nd, *Clustering algorithms in biomedical research: a review.* IEEE Rev Biomed Eng, 2010. **3**: p. 120-54.
4. Andreopoulos, B., et al., *A roadmap of clustering algorithms: finding a match for a biomedical application.* Brief Bioinform, 2009. **10**(3): p. 297-314.
5. Basile, A.O. and M.D. Ritchie, *Informatics and machine learning to define the phenotype.* Expert Rev Mol Diagn, 2018. **18**(3): p. 219-226.
6. Libbrecht, M.W. and W.S. Noble, *Machine learning applications in genetics and genomics.* Nat Rev Genet, 2015. **16**(6): p. 321-32.
7. Sørlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proceedings of the National Academy of Sciences, 2001. **98**(19): p. 10869-10874.
8. Greene, C.S., et al., *Big data bioinformatics.* Journal of cellular physiology, 2014. **229**(12): p. 1896-1900.

9. Inohara, T., et al., *A Cluster Analysis of the Japanese Multicenter Outpatient Registry of Patients With Atrial Fibrillation.* Am J Cardiol, 2019. **124**(6): p. 871-878.

10. Castaldi, P.J., et al., *Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts.* Thorax, 2017. **72**(11): p. 998-1006.

11. Pikoula, M., et al., *Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records.* BMC Med Inform Decis Mak, 2019. **19**(1): p. 86.

12. Williams, J.B., D. Ghosh, and R.C. Wetzel, *Applying Machine Learning to Pediatric Critical Care Data.* Pediatr Crit Care Med, 2018. **19**(7): p. 599-608.

13. Mayhew, M.B., et al., *Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models.* J Biomed Inform, 2018. **78**: p. 33-42.

14. Fareed, N., et al., *Inpatient portal clusters: identifying user groups based on portal features.* J Am Med Inform Assoc, 2019. **26**(1): p. 28-36.

15. Bastanlar, Y. and M. Ozuysal, *Introduction to machine learning.* Methods Mol Biol, 2014. **1107**: p. 105-28.

16. Hunt, L. and M. Jorgensen, *Clustering mixed data.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011. **1**(4): p. 352-361.

17. Ahmad, A. and S.S. Khan, *Survey of State-of-the-Art Mixed Data Clustering Algorithms.* IEEE Access, 2019. **7**: p. 31883-31902.

18. Ward, J.H., *Hierarchical Grouping to Optimize an Objective Function.* Journal of the American Statistical Association, 1963. **58**(301): p. 236-244.

19. Blashfield, R.K., *Propositions regarding the use of cluster analysis in clinical research.* J Consult Clin Psychol, 1980. **48**(4): p. 456-9.

20. Burgel, P.R., et al., *Clinical COPD phenotypes: a novel approach using principal component and cluster analyses.* Eur Respir J, 2010. **36**(3): p. 531-9.

21. Inohara, T., et al., *Association of of Atrial Fibrillation Clinical Phenotypes With Treatment Patterns and Outcomes: A Multicenter Registry Study.* JAMA Cardiol, 2018. **3**(1): p. 54-63.

22. Egan, B.M., et al., *A cluster-based approach for integrating clinical management of Medicare beneficiaries with multiple chronic conditions.* PLoS One, 2019. **14**(6): p. e0217696.

23. Reynolds, A.P., et al., *Clustering rules: a comparison of partitioning and hierarchical clustering algorithms.* Journal of Mathematical Modelling and Algorithms, 2006. **5**(4): p. 475-504.

24. Chaturvedi, A., P.E. Green, and J.D. Caroll, *K-modes clustering.* Journal of classification, 2001. **18**(1): p. 35-55.

25. Rousseeuw, P.J. and L. Kaufman, *Finding groups in data.* Hoboken: Wiley Online Library, 1990.

26. Lazarsfeld, P.F. and N.W. Henry, *Latent structure analysis*. 1968: Houghton Mifflin Co.

27. Goodman, L.A., *Exploratory latent structure analysis using both identifiable and unidentifiable models.* Biometrika, 1974. **61**(2): p. 215-231.

28. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society: Series B (Methodological), 1977. **39**(1): p. 1-22.

29. Kohonen, T., *Self-organized formation of topologically correct feature maps.* Biological cybernetics, 1982. **43**(1): p. 59-69.

30. Kriegel, H.P., et al., *Density-based clustering.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011. **1**(3): p. 231-240.

31. Goodall, D.W., *A new similarity index based on probability.* Biometrics, 1966: p. 882-907.

32. Balaji, K. and K. Lavanya, *Clustering algorithms for mixed datasets: A review.* International Journal of Pure and Applied Mathematics, 2018. **18**(7): p. 547-556.

33. Chiodi, M., *A partition type method for clustering mixed data.* Rivista di statistica applicata, 1990. **2**: p. 135-147.

34. Li, C. and G. Biswas, *Unsupervised learning with mixed numeric and nominal data.* IEEE Transactions on Knowledge and Data Engineering, 2002. **14**(4): p. 673-690.

35. Sangam, R.S. and H. Om, *An equi-biased k-prototypes algorithm for clustering mixed-type data.* Sādhanā, 2018. **43**(3): p. 37.

36. Ren, M., et al. *An improved mixed-type data based kernel clustering algorithm*. in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. 2016. IEEE.

37. Philip, G. and B. Ottaway, *Mixed data cluster analysis: an illustration using Cypriot hooked-tang weapons.* Archaeometry, 1983. **25**(2): p. 119-133.

38. Gower, J.C., *A general coefficient of similarity and some of its properties.* Biometrics, 1971: p. 857-871.

39. Modha, D.S. and W.S. Spangler, *Feature weighting in k-means clustering.* Machine learning, 2003. **52**(3): p. 217-237.

40. Huang, Z. *Clustering large data sets with mixed numeric and categorical values*. in *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. 1997. Singapore.

41. Huang, Z., *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.* Data Mining and Knowledge Discovery, 1998. **2**(3): p. 283-304.

42. Chae, S.-S., J.-M. Kim, and W.-Y. Yang, *Cluster analysis with balancing weight on mixed-type data.* Communications for Statistical Applications and Methods, 2006. **13**(3): p. 719-732.

43. Ahmad, T., et al., *Clinical Implications of Cluster Analysis-Based Classification of Acute Decompensated Heart Failure and Correlation with Bedside Hemodynamic Profiles.* PLoS One, 2016. **11**(2): p. e0145881.

44. Ahmad, A. and L. Dey, *A k-mean clustering algorithm for mixed numeric and categorical data.* Data & Knowledge Engineering, 2007. **63**(2): p. 503-527.

45. Lee, J.H., et al., *Identification of subtypes in subjects with mild-to-moderate airflow limitation and its clinical and socioeconomic implications.* Int J Chron Obstruct Pulmon Dis, 2017. **12**: p. 1135-1144.

46. Ta, C.N. and C. Weng, *Detecting Systemic Data Quality Issues in Electronic Health Records.* Stud Health Technol Inform, 2019. **264**: p. 383-387.

47. Yan, J., et al., *Applying Machine Learning Algorithms to Segment High-Cost Patient Populations.* J Gen Intern Med, 2019. **34**(2): p. 211-217.

48. Yan, S., et al., *Identifying heterogeneous health profiles of primary care utilizers and their differential healthcare utilization and mortality - a retrospective cohort study.* BMC Fam Pract, 2019. **20**(1): p. 54.

49. Bose, E. and K. Radhakrishnan, *Using Unsupervised Machine Learning to Identify Subgroups Among Home Health Patients With Heart Failure Using Telehealth.* Comput Inform Nurs, 2018. **36**(5): p. 242-248.

50. van de Velden, M., A. Iodice D'Enza, and A. Markos, *Distance-based clustering of mixed data.* Wiley Interdisciplinary Reviews: Computational Statistics, 2019. **11**(3): p. e1456.

51. McCarthy, H., et al., *High expression of activation-induced cytidine deaminase (AID) and splice variants is a distinctive feature of poor-prognosis chronic lymphocytic leukemia.* Blood, 2003. **101**(12): p. 4903-4908.

52. Duzkale, H., et al., *LDOC1 mRNA is differentially expressed in chronic lymphocytic leukemia and predicts overall survival in untreated patients.* Blood, 2011. **117**(15): p. 4076-4084.

53. Schweighofer, C.D., et al., *The B cell antigen receptor in atypical chronic lymphocytic leukemia with t (14; 19)(q32; q13) demonstrates remarkable stereotypy.* International journal of cancer, 2011. **128**(11): p. 2759-2764.

54. Rassenti, L.Z., et al., *ZAP-70 compared with immunoglobulin heavy-chain gene mutation status as a predictor of disease progression in chronic lymphocytic leukemia.* New England Journal of Medicine, 2004. **351**(9): p. 893-901.

55. Admirand, J.H., et al., *Immunohistochemical detection of ZAP70 in chronic lymphocytic leukemia predicts immunoglobulin heavy chain gene mutation status and time to progression.* Modern Pathology, 2010. **23**(11): p. 1518.

56. Schweighofer, C.D., et al., *Genomic variation by whole-genome SNP mapping arrays predicts time-to-event outcome in patients with chronic lymphocytic leukemia: a comparison of CLL and HapMap genotypes.* The Journal of Molecular Diagnostics, 2013. **15**(2): p. 196-209.

57. Dohner, H., et al., *Genomic aberrations and survival in chronic lymphocytic leukemia.* N Engl J Med, 2000. **343**(26): p. 1910-6.

58. Zenz, T., H. Döhner, and S. Stilgenbauer, *Genetics and risk-stratified approach to therapy in chronic lymphocytic leukemia.* Best practice & research clinical haematology, 2007. **20**(3): p. 439-453.

59. Auer, P. and D. Gervini, *Choosing principal components: a new graphical method based on Bayesian model selection.* Communications in Statistics—Simulation and Computation®, 2008. **37**(5): p. 962-977.

60. Wang, M., et al., *Thresher: determining the number of clusters while removing outliers.* BMC bioinformatics, 2018. **19**(1): p. 9.

61. Choi, S.-S., S.-H. Cha, and C.C. Tappert, *A survey of binary similarity and distance measures.* Journal of Systemics, Cybernetics and Informatics, 2010. **8**(1): p. 43-48.

62. Michener, C.D. and R.R. Sokal, *A quantitative approach to a problem in classification.* Evolution, 1957. **11**(2): p. 130-162.

63. Rousseeuw, P.J., *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of computational and applied mathematics, 1987. **20**: p. 53-65.

64. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.

65. Powers, B.W., et al., *Subgroups of High-Cost Medicare Advantage Patients: an Observational Study.* J Gen Intern Med, 2019. **34**(2): p. 218-225.

66. Foss, A.H. and M. Markatou, *kamila: clustering mixed-type data in R and Hadoop.* Journal of Statistical Software, 2018. **83**(1): p. 1-44.

67. Zhang, J. and K.R. Coombes, *Sources of variation in false discovery rate estimation include sample size, correlation, and inherent differences between groups.* BMC Bioinformatics, 2012. **13**(13): p. S1.

68. Marlin, B.M., et al. *Unsupervised pattern discovery in electronic health care data using probabilistic clustering models*. in *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*. 2012. ACM.

69. Wang, J., et al., *The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data.* Cancer Inform, 2009. **7**: p. 199-216.

70. Kaufman, L. and P.J. Rousseeuw, *Partitioning around medoids (program pam).* Finding groups in data: an introduction to cluster analysis, 1990: p. 68-125.

71. Wehrens, R., *kohonen: Supervised and Unsupervised Self-Organising Maps*. 2019.

72. Coombes, K.R.a.C., Caitlin E., *Mercator: Clustering and Visualizing Distance Matrices*. 2019.

73. Cox, T.F. and M.A. Cox, *Multidimensional scaling*. 2000: Chapman and hall/CRC.

74. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, Complex Systems, 2006. **1695**(5): p. 1-9.

75. Stevens, S.S., *On the theory of scales of measurement.* 1946.

76. Maechler, M., Rousseeuw, Peter, Struyf, Anja, Hubert, Mia and Hornik, Kurt, *cluster: Cluster Analysis Basics and Extensions*. 2019.

77. Handl, J., J. Knowles, and D.B. Kell, *Computational cluster validation in post-genomic data analysis.* Bioinformatics, 2005. **21**(15): p. 3201-3212.

78. Rendón, E., et al., *Internal versus external cluster validation indexes.* International Journal of computers and communications, 2011. **5**(1): p. 27-34.

79. Brock, G., et al., *clValid, an R package for cluster validation.* Journal of Statistical Software (Brock et al., March 2008), 2011.

80. Milligan, G.W. and M.C. Cooper, *Methodology review: Clustering methods.* Applied psychological measurement, 1987. **11**(4): p. 329-354.

81. Rand, W.M., *Objective Criteria for the Evaluation of Clustering Methods.* Journal of the American Statistical Association, 1971. **66**(336): p. 846-850.

82. Hubert, L. and P. Arabie, *Comparing partitions.* Journal of classification, 1985. **2**(1): p. 193-218.

83. Kampstra, P., *Beanplot: A boxplot alternative for visual comparison of distributions.* Journal of statistical software, 2008. **28**(1): p. 1-9.

Appendix A. Supplemental Tables and Figures to Chapter 2

Figure A.1. Data transformation B: Kaplan-Meier survival curves for time-to-progression and time from diagnosis to treatment. Unsupervised machine learning, using k-means clustering with Partitioning Around Medoids (PAM) and the Sokal-Michener distance yields seven clinical phenotypes with significant differences in time-to-progression (TTP) ($p = 0.0451$) and the related metric time from diagnosis to treatment ($p = 0.0039$).

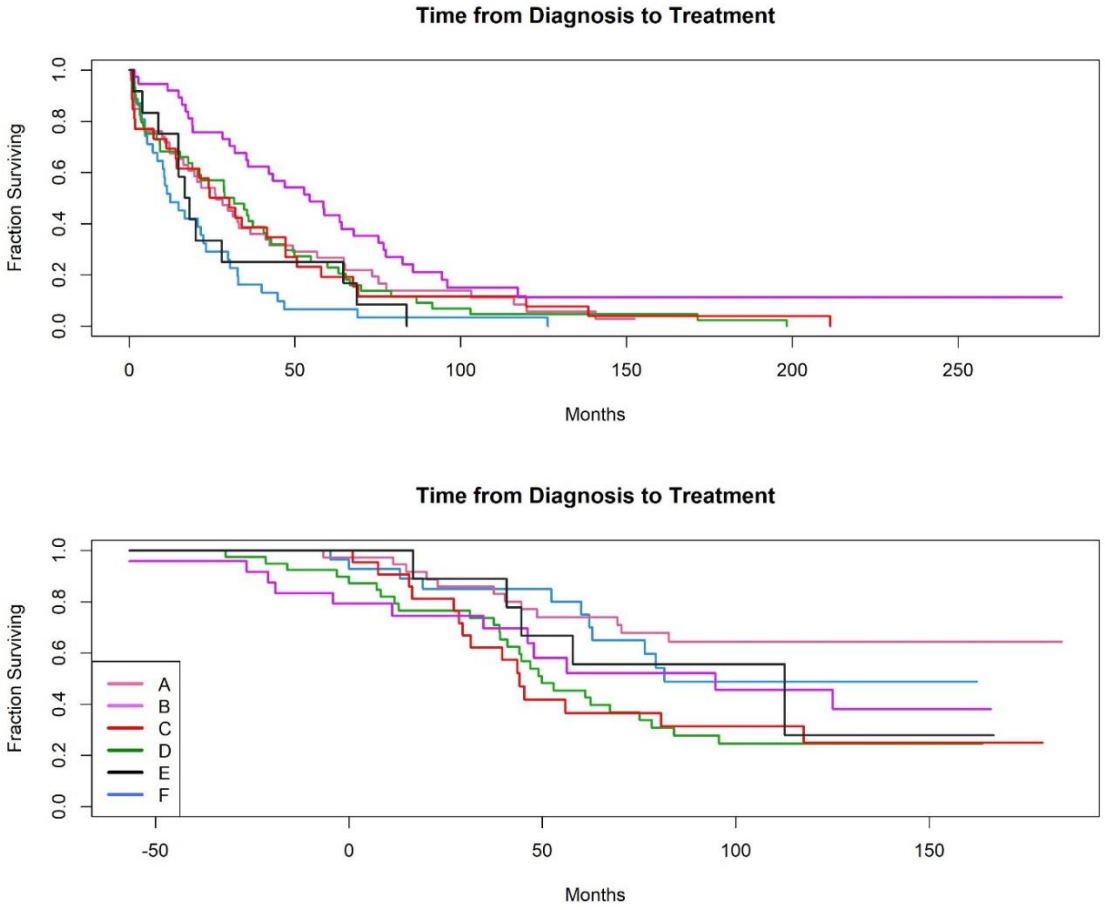


Figure A.2. Multi-dimensional scaling (MDS) plots constructed from 9 different distance metrics in data transformation A. In selecting a distance metric for applying unsupervised machine learning using k-medoids clustering with Partitioning Around Medoids (PAM), we assessed 10 distance metrics representing meaningful groupings of 76 distance metrics for binary data. Similarity and difference between outputs of distance metrics grouped qualitatively. Although the Sokal & Michener distance was chosen for ease of interpretability, similar results were obtained from the Manhattan, Hamming, and Goodman & Kruskal distances.

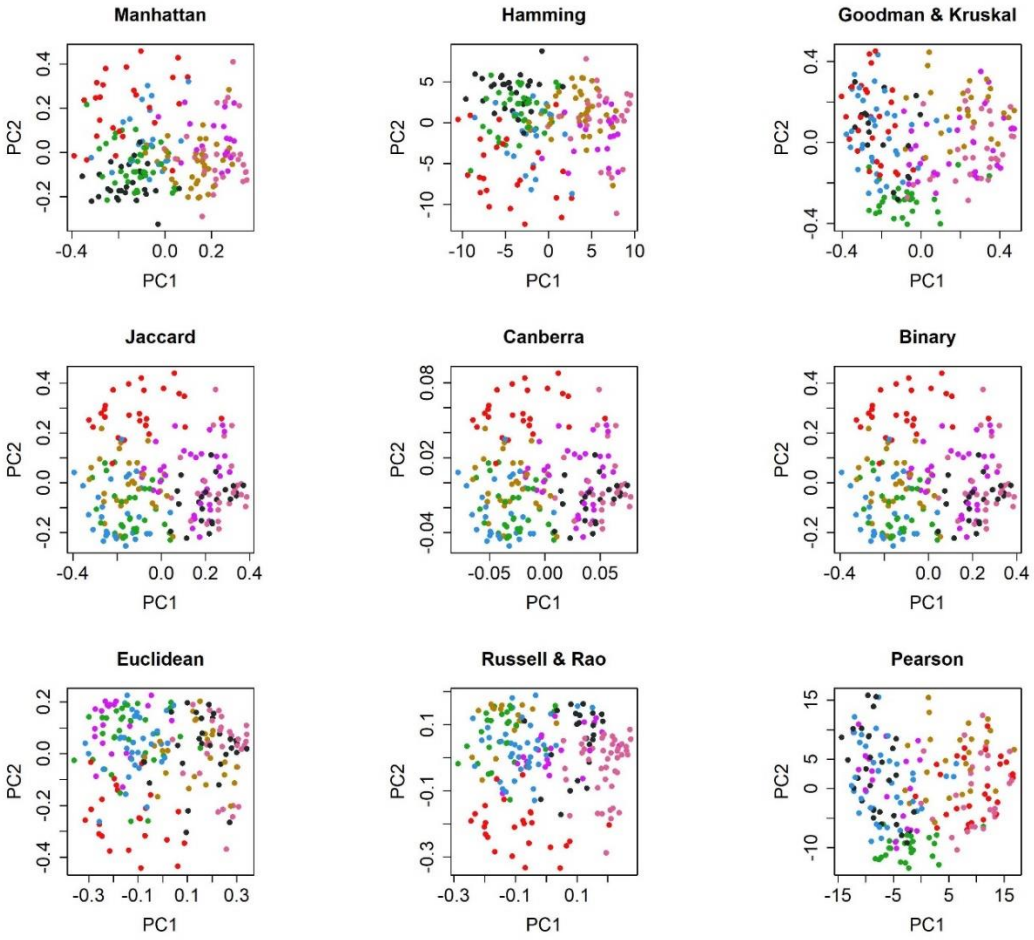


Figure A.3. t-Stochastic Neighbor Embedding (t-SNE) plots constructed from 9 different distance metrics in data transformation A. In selecting a distance metric for applying unsupervised machine learning using k-medoids clustering with Partitioning Around Medoids (PAM), we assessed 10 distance metrics representing meaningful groupings of 76 distance metrics for binary data. Similarity and difference between outputs of distance metrics grouped qualitatively. Although the Sokal & Michener distance was chosen for ease of interpretability, similar results were obtained from the Manhattan, Hamming, and Goodman & Kruskal distances.

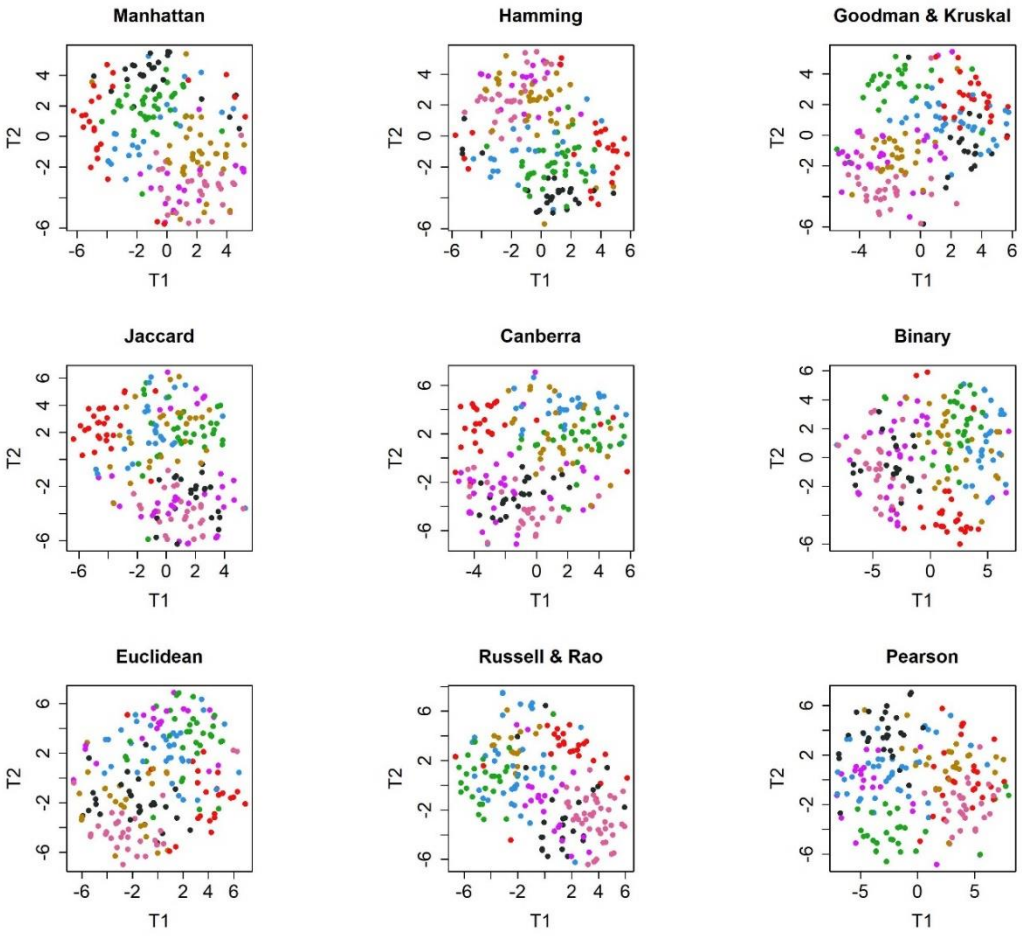


Table A.1. Comprehensive identifying features of clusters for data transformations A and B, in order of overall survival. Clusters are ordered by predicted survival outcome, from longest survival (A1 or B1) to shortest (A7 or B6). Percentages represent frequency of a feature within a given cluster. Characteristic features of each cluster, defined as a feature present in at least 75% of members of a given cluster, include known indicators of superior prognosis (*IGHV*-mutated status and female sex) and poor prognosis (ZAP70 positivity). Döhner classification, known to be one of the best predictors of prognosis in CLL, failed to be captured by the analysis for most clusters.

Data Transformation A.

Cluster	Sex[1]	*IGHV* Status[2]	ZAP70	Döhner	CD38	Light Chain	Cytopenia	RAI Stage	WBC Count	B2M[4]	Matutes
A1	M 89.5%	M 89.5%	- 84.2%		Low 100%	λ 84.2%	HGG[3] 84.2%	Low 78.9%	Low 94.7%	Low 89.5%	Typical 78.9%
A2	F 79.3%	M 89.7%	- 86.2%	del13q 75.9%	Low 93.1%			Low 86.2%	Low 96.6%	Low 82.8%	Typical 75.9%
A3	M 88.9%	M 80.6%	- 86.1%	del13q 80.6%	Low 83.3%			Low 88.9%	Low 83.3%	Low 91.7%	Typical 80.6%
A4	M 85.2%	U 96.3%			High 88.9%	κ 88.9%		Low 100%	Low 81.5%	Low 85.2%	
A5	M 84%	U 84%	-		Low		Anemia 80%		Low 80%	High 96%	Typical 80%
A6		U 94.7%	+ 89.5%		Low 94.7%	κ 92.1%		Low 92.1%		Low 86.8%	Typical 86.8%
A7		U 86.4%	+ 90.9%		Low 86.4%		Anemia 81.8%	Low 81.8%	Low 81.8%		

[1]M = Male, F = Female
[2]M = Mutated, U = Unmitated
[3]Hypogammaglobulinemia
[4]Beta-2 microglobulin
[5]Prolymphocytes

Table A.1 continued

Data Transformation B.

Cluster	Sex[1]	*IGHV* Status[2]	ZAP70	CD38	Age	PL[5]	Light Chain	Anemia	RAI Stage	WBC Count	B2M[4]	Matutes
B1		M 91.3%	- 82.6%	Low 93.5%	< 65 82.6%	< 10 82.6%	λ 100%		Low 84.8%	Low 95.7%	Low 91.3%	Typical 78.3%
B2		M 78.4%	- 97.3%	Low 86.5%	< 65 75.7%	< 10 83.8%	κ 100%		Low 89.2%	Low 81.1%	Low 86.5%	Typical 78.4%
B3	M 76.9%	U 76.9%			< 65 76.9%	< 10 80.8%		Anemia 84.6%		Low 84.6%	High 88.5%	Typical 84.6%
B4		U 95.5%	+ 93.2%	Low 95.5%	< 65 93.2%	< 10 86.4%	κ 86.4%		Low 90.9%		Low 86.4%	Typical 86.4%
B5	M 83.3%	U 83.3%	+ 83.3%	Low 91.7%	> 65 91.7%		λ 83.3%	Anemia 83.3%		Low 83.3%		
B6	M 87.1%	U 96.8%	+ 80.6%	High 80.6%	< 65 83.9%	< 10 87.1%	κ 83.9%		Low 100%	Low 80.6%	Low 87.1%	

[1]M = Male, F = Female
[2]M = Mutated, U = Unmitated
[3]Hypogammaglobulinemia
[4]Beta-2 microglobulin
[5]Prolymphocytes