EVALUATION OF ARCHETYPAL ANALYSIS AND MANIFOLD LEARNING FOR PHENOTYPING OF ACUTE KIDNEY INJURY

by

Dylan Michael Rodriquez

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer Science

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Ananth Grama, Chair

Department of Computer Science

Dr. Muhammad Adibuzzaman

Department of Computer Science

Dr. Petros Drineas

Department of Computer Science

Approved by:

Dr. Kihong Park

To my family and Chris.

ACKNOWLEDGMENTS

I am grateful and appreciative of my committee for their continuing support. Additionally, I would like to extend my gratitude to Professor Xavier Tricoche in allowing me the freedom of exploring other scientific domains from the perspective of computer science and Dr. David T. Rubin for allowing me to explore the possibilities of computer science in the domain of medicine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$^\circ$	degrees
$\|\| \cdot \|\|$	2-norm
$\|\| \cdot \|\|_F$	Frobenius norm
ρ	Pearson correlation coefficient
X	Matrix
x	Vector

ABBREVIATIONS

AKI	Accute Kidney Injury
CHF	Cerner Health Facts
GFR	Glomerular filtration rate
WBC	White Blood Cell
mg	Milligram
g	Gram
mm	Millimeter
bpm	Beats per minute
br	Breaths
min	Minute
μL	Microliter
dL	Deciliter
L	Liter
mmol	millimol

NOMENCLATURE

AKI	Accute Kidney Injury
CHF	Cerner Health Facts
GFR	Glomerular filtration rate

GLOSSARY

Creatinine	A nitrogenous compound present in muscles responsible for quickly replenishing energy stores
Glomerular filtration rate	The estimation of how much blood passes through the kidneys

ABSTRACT

Disease subtyping has been a critical aim of precision and personalized medicine. With the potential to improve patient outcomes, unsupervised and semi-supervised methods for determining phenotypes of subtypes have emerged with a recent focus on matrix and tensor factorization. However, interpretability of proposed models is debatable. Principal component analysis (PCA), a traditional method of dimensionality reduction, does not impose non-negativity constraints. Thus coefficients of the principal components are, in cases, difficult to translate to real physical units. Non-negative matrix factorization (NMF) constrains the factorization to positive numbers such that representative types resulting from the factorization are additive. Archetypal analysis (AA) extends this idea and seeks to identify pure types, archetypes, at the extremes of the data from which all other data can be expressed as a convex combination, or by proportion, of the archetypes. Using AA, this study sought to evaluate the sufficiency of AKI staging criteria through unsupervised subtyping. Archetype analysis failed to find a direct 1:1 mapping of archetypes to physician staging and also did not provide additional insight into patient outcomes. Several factors of the analysis such as quality of the data source and the difficulty in selecting features contributed to the outcome. Additionally, after performing feature selection with lasso across data subsets, it was determined that current staging criteria is sufficient to determine patient phenotype with serum creatinine at time of diagnosis to be a necessary factor.

1. INTRODUCTION

The prevalence of electronic medical records (*EMR*) has enabled the exploration of vast amounts of biomedical data in the pursuit of precision medicine. Structured data present in EMR such as demographics, diagnostic codes, and laboratory values can be collected in real time and warehoused such that it may be leveraged for research purposes [1]. Warehoused data may also include information not directly contained within a patient's medical record. Patient metadata, such as payer information, hospital setting, and referral data expands the utility of EMR and provides additional context for patient encounters [2].

A specific endeavor of precision medicine is disease subtyping, where a particular disease is stratified or further classified into distinct types [3], [4]. Traditionally considered a byproduct of clinical experience and expertise, advances in computation and processing of *omics* data; such as genomics, transcriptomics, or proteomics, have advanced this effort. Additional advances in curation and processing of EHR data and metatdata have allowed *clinarrays*, aggregated laboratory and clinical information, to enable quantitative methods previously utilized to analyze genomic data to be applied to clinical data [5]. As a result, it is possible to construct disease phenotypes with unsupervised learning techniques to personalize treatment according to patient phenotype [6].

Several techniques, such as deep learning, clustering, and manifold learning are currently used in patient subtyping, most of which are unsupervised [7], [8] or semi-supervised [9]–[11]. Notably, variants of non-negative matrix (NMF) and tensor factorization have become prevalent in preserving interpretability of phenotypes [7], [12]. Archetypal analysis (AA) follows suit, as it seeks to increase interpretability though the use of additive pure types,*archetypes*, while incorporating nonnegativity constraints and convexity constraints such that all data may be expressed as a convex combination of archetypes [13]. While factorization methods such as PCA express variability, the coefficients associated with the principal components may be negative, making it more difficult to translate directly to practical or real world physiological parameters in clinical applications. Additive coefficients in NMF seek to mitigate this difficulty, where each pure type contributes to the data; however, AA extends this idea to proportionality as each archetype constitutes a proportion to a particular datum.

Using separability of archetypes as a measure of disease subtype uniqueness, archetyal analysis finds potential utility in discovering subtypes and providing descriptors of phenotypes in diseases considered to be heterogenous mixtures.

1.1 Acute Kidney Injury

Acute kidney injury (AKI) is such a disease considered to be a heterogenous group of conditions. AKI is characterized by a sudden decrease in glomerular filtration rate (GFR) resulting in increased serum creatinine. It is traditionally discretized into three stages based on severity and etiology enumerated from least to most severe [14]. Stages are determined from the ability of the patient to metabolize creatinine with respect to an initial measurement. Other criteria, such as urinary clearance of creatinine or supportive treatments, contribute to staging [15]. Etiology is equivalent in importance to staging, as determination of etiology can guide treatment.

The etiology of AKI can be grouped into categories. These groups are defined by underlying pathophysiology and include decreased bloodflow through the kidneys, blockage of the urinary tract, kidney diseases, and damage to tubule cells in the kidneys. Although etiologies are grouped, it is not uncommon that AKI originates from more than one cause.

Although affecting approximately 20% of hospitalized patients, determining the true incidence of AKI is also difficult. Clinical signs that would indicate AKI may be difficult to observe or absent, resulting in it being harder to identify. It is most prevalent in those greater than 65 years of age and various morbidities including those with diabetes, chronic kidney disease, heart failure, or anaemia. AKI can also be contributed to external factors [16]. Independent of these factors, severity of AKI is still highly associated with poor outcomes.

Although standard criteria is available and has been updated [17][18], it is emphasized that AKI is more often a continuum of disease and is constituted by subtypes rather than a series of progressive stages describing one disease [19]. However this has not yet been demonstrated quantitatively [20].

1.2 Aims

The primary aim of this study is to assess the utility of archetypal analysis in determining two disease constructs. The first is disease staging with the second being disease subtypes. While disease stages are well defined by clinical criteria, it is of additional clinical utility to determine if these stages are sufficient in predicting patient outcomes or if additional disease subtypes of AKI may contribute to the predictive utility of staging. The following hypothesis is such that AA is able to extract or approximate disease stages as archetypes. In the case that it is unable to, then additional insight into phenotypes of patient subtypes may be obtained. The secondary aim of the study is to establish a mapping of archetypes to stages and patient outcomes, as defined by a reduction in AKI stage.

2. METHODS

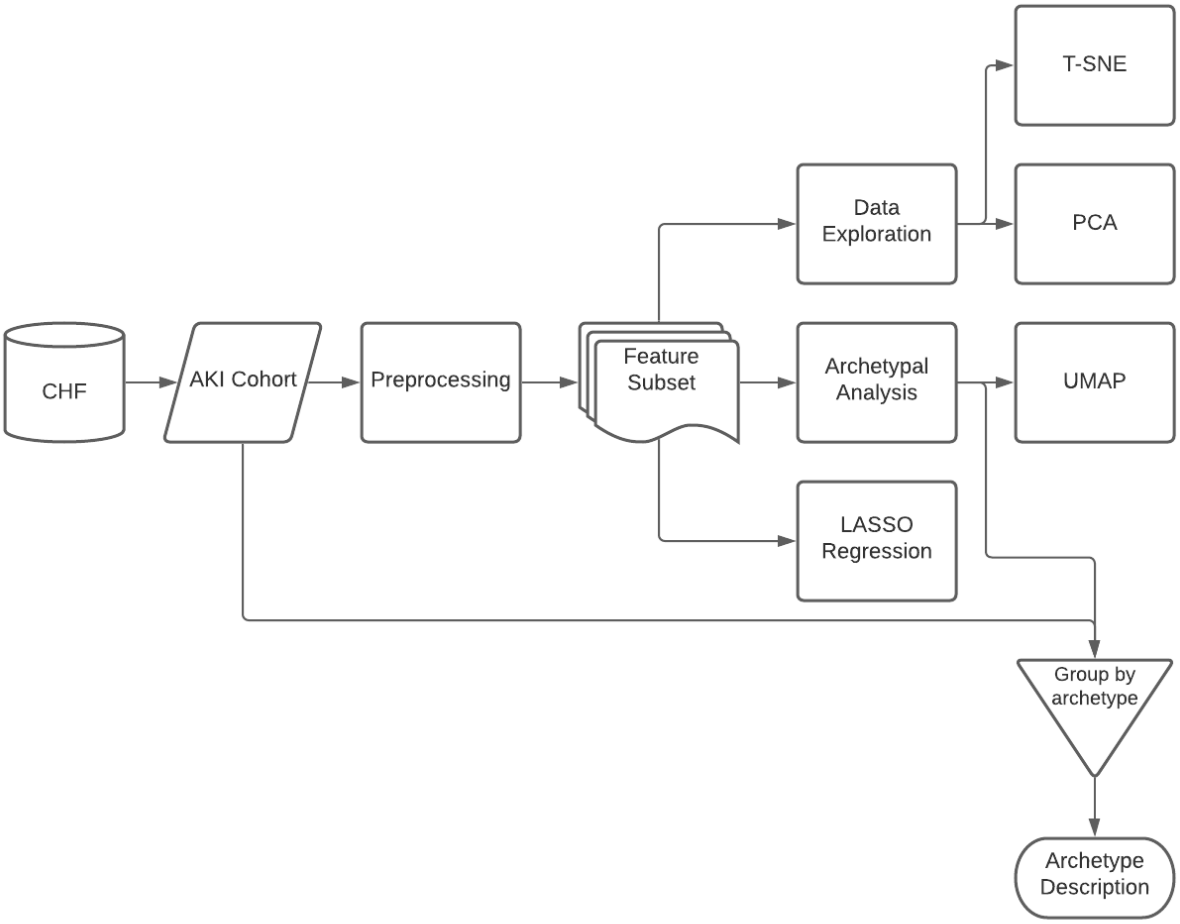


Figure 2.1. Overview of methods

2.1 Data Source and Cohort Selection

Data were collected from the Cerner Health Facts (CHF) database. CHF utilizes an automated electronic medical record system to capture patient and encounter information, laboratory results, surgical encounters, and medication information. 750 facilities contribute to these data, including 388 inpatient facilities.

All hospitalized patients with cirrhosis were queried from CHF (n = 117,991). Of this cohort we continued to filter patients whom did not meet KDIGO AKI criteria (table 2.1)

Table 2.1. AKI definition

AKI diagnostic criteria
Increase in serum creatinine by $\geq$ 0.3 ml/dL within 48 hours
OR
Increase in serum creatinine to $\geq$ 1.5 times baseline measured within 7 days prior
OR
Urine volume $<$ 0.5 ml/kg/h for at least 6 hours

[15] (n = 12,201). Patients whom had undergone hemodialysis, surgical cases, ICU cases were excluded (n = 12,025). Patients without serum creatinine values and albumin or crystalloid administration data 7 days after AKI diagnosis were excluded. The final AKI cohort consisted of 4,338 patients and 217 variables (a rank deficient 4338×217 matrix). Missing data were imputed using K-nearest neighbor imputation.

Table 2.2. AKI staging criteria

Stage	Serum Creatinine	Urine Output
1	1.5 - 1.9 times baseline	$<$ 0.5 ml/kg/h for 6 - 12 hours
	OR	
	$\geq$ 0.3 mg/dL increase	
2	2.0 - 2.9 times baseline	$<$ 0.5ml/kg/h for $\geq$ 6 - 12 hours
3	3.0 times baseline	$<$ 0.3 ml/kg/h for $\geq$ 24 hours
	OR	OR
	Increase in serum creatinine to $\geq$ 4.0 ml/dL	Anuria for at least 12 hours
	OR	
	initiation of renal replacement therapy	

2.2 Outcomes

AKI stages were defined using KDIGO Criteria (table 2.2). The primary outcome was defined as a reduction in AKI stage [15].

2.3 Archetypal Analysis

Archetypal analysis (AA) aims to represent data with respect to extreme types found on the corners of the data, as estimated by the convex hull. AA will be used to determine disease subtypes with each archetype mapping to a disease subtype. The representation of the data are convex combinations of the extreme types which can be calculated in an alternating least squares problem [13]. Given a matrix of data, $X \in \mathbf{R}^{m \times n}$, and each record, $[x_1, \ldots, x_n] \in \mathbf{R}^m$, AA seeks to find vectors of Z, $[z_1, \ldots, z_p] \in \mathbf{R}^p$, that characterize the p pure types in the data as mixtures of X_i. Each archetype, Z_i is a convex combination of the data such that

$$Z_k = \sum_\mathrm{j} \beta_{k\mathrm{j}} x_\mathrm{j} \qquad\qquad k = 1, \ldots, p$$

with non negativity and convexity constraints:

$$\text{(i) } \beta_{k\mathrm{i}} \geq 0 \qquad\qquad \text{(ii) } \sum_\mathrm{i} \beta_{k\mathrm{i}} = 1$$

and α are the minimizers of

$$||x_\mathrm{i} - \sum_k^p a_{\mathrm{i}k} z_k||^2$$

with non negativity and convexity constraints:

$$\text{(i) } \alpha_{\mathrm{i}k} \geq 0 \qquad\qquad \text{(ii) } \sum_k \alpha_{\mathrm{i}k} = 1$$

Archetypal patterns are then defined as the mixtures of Z that minimize

$$RSS(p) = \sum_{\mathrm{i}} ||x_{\mathrm{i}} - \sum_{k}^{p} a_{\mathrm{i}k} z_k||^2$$

$$RSS(p) = ||X - \alpha Z^T||^2 \qquad\qquad Z = X^T \beta$$

Algorithm 1: Archetypal Analysis

```
Standardize data and randomly initialize $\beta$ subject to constraints ;
while RSS value is sufficiently large or maximum iterations has not been reached do
    $\underset{\alpha}{\text{minimize}} \frac{1}{2}||X - \alpha Z^T||$ subject to constraints;
    $\alpha^+ X = \tilde{Z}$;
    $\underset{\beta}{\text{minimize}} \frac{1}{2}||\tilde{Z} - X\beta||$ subject to constraints;
    $Z = X\beta$;
    $RSS = ||X - \alpha Z^T||^2$;
end
```

2.4 SVD Denoising

Given the data form a $4338 \times n$ matrix where n is the number of at most 217 features, there exist two orthogonal matrices

$$U = [u_1, \ldots, u_m] \in \mathbf{R}^{4338 \times 4338} \qquad\qquad V = [v_1, \ldots, v_n] \in \mathbf{R}^{n \times n}$$

such that:

$$U^T D V = diag(\sigma_1, \ldots, \sigma_p) \in \mathbf{R}^{4338 \times n} \qquad \sigma_1 \geq \cdots \geq \sigma_r > \sigma_r + 1 = p = 0$$

$$p = min(4338, n) \qquad n \leq 217$$

$$p = n$$

$$rank(D) = r$$

$$rank(D) \leq n$$

where the k-rank approximation is given by:

$$D = \sum_{\mathrm{i}}^{k} \sigma_{\mathrm{i}} u_{\mathrm{i}} v_{\mathrm{i}}^{T}$$

When D is rank deficient, there exist an infinite number of solutions to least squares. Given this problem, the SVD is used to derive a k-rank approximation that best minimizes the least squares solution:

$$||D_k x - b||$$

In all cases of selected features, D has fewer linearly independent columns, and therefore is rank deficient. Utilizing the k-rank approximation filters singular values, reducing noise [21]. Consider a matrix comprised of noisy data:

$$\tilde{X} = X + N$$

where $\tilde{X}$ denotes data containing noise and N contains random noise with distribution $N(0, \epsilon)$. Reconstruction of the clean data can then be evaluated as

$$\Delta = ||D - D_k||_F$$

seeking to minimize Δ [22], [23]

2.5 Feature Selection

Several methods for feature selection were explored. The first method used domain expert knowledge for selection of relevant features. Additional methods included regression with the least absolute shrinkage and selection operator (LASSO) and finding the minimum correlation vertex cover of a given feature set.

2.5.1 Variable Selection with Domain Expertise

Several column subsets were selected under the guidance of a domain expert. Features were selected based on temporal relation to diagnosis and outcomes. Features at the time of AKI diagnosis were selected for their temporal relation to AKI. Additionally, these features are used to stage the disease as well as guide treatment decisions. Discrete data, such as presence of comorbidities, are not ordinal. Matrix factorization methods assume that noise is Gaussian, wheras the Bernoulli distribution is a more appropriate descriptor of discrete noise [24]. Thus, all continuous features were selected from this set as an additional feature subset.

2.5.2 Minimum Weighted Vertex Cover

To reduce correlation between variables in a given feature selection set, a variable selection algorithm dependent on minimizing pair-wise relationships was developed. The Pearson correlation coefficient:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

was calculated for each 2-combination tuples (i, j). Coefficients were mapped to a weighted adjacency matrix, $X \in \mathbf{R}^{n \times n}$, where the Pearson correlation coefficient of the edge connecting vertices i and j is the entry x_{ij}

A minimum spanning tree (MST) was constructed using Kruskal's algorithm and a minimum weighted vertex cover was subsequently calculated using a local ratio vertex cover approximation [25].

2.6 Dimensionality reduction and Visualization

Two major generalized categories that dimensionality reduction algorithms lie are the realms of matrix factorization and manifold learning. Archetypal analysis, as discussed above, is related non-negative matrix factorization with additional constraints. Additionally, principal component analysis (PCA) is also expressed by the former category. The later category generalizes the two algorithms below which were used for visualization. Both construct a neighbor graph and find an approximately optimal embedding to a low dimensional space.

Uniform Manifold approximation and Projection (UMAP)

UMAP is a nonlinear dimensionality reduction technique that seeks to preserve local and global structure of the data with utility in high dimensional biological data [26], [27]. UMAP relies on two main assumptions:

1. Data are uniformly distributed on an existing Riemannian manifold
2. The manifold is locally connected

Assumption 1 forms the basis for UMAP, for which distances on the manifold are mapped to varying distances in euclidean space. The geodesic distance can then be approximated from any datum X_{i} by normalizing distances with respect to the distance to the k^{th} nearest neighbor of X_{i}. Distance for each X_{i} is tailored to its location on the manifold.

The manifold is then represented as a k-neighbor graph with local connectivity constraints, in accordance with assumption 2, to ensure that each X_{i} is connected to at least one other point with an edge weight of at least 1. Weights of the edges, are then mapped to a probability that such edge exists. In the euclidean space, multiple edges of differing weight may exist between points. For each edge pair connecting vertices (i, j), the probability that edge $x_{\mathrm{ij}} \vee x_{\mathrm{ji}}$ maps edges to an undirected, weighted graph.

A force directed graph layout is utilized to embed the graph into a low dimensional space. Cross entropy is used to minimize the distance inside each cluster and maximize distance between clusters of data [28]

T-SNE

T-SNE is an extension of stochastic neighbor embedding (SNE), differing in the distribution used to calculate densities in the low dimensional embedding [29]. T-SNE calculates a low dimensional embedding of the high dimensional input through two phases.

Algorithm 2: Stochastic Neighbor Embedding [30]

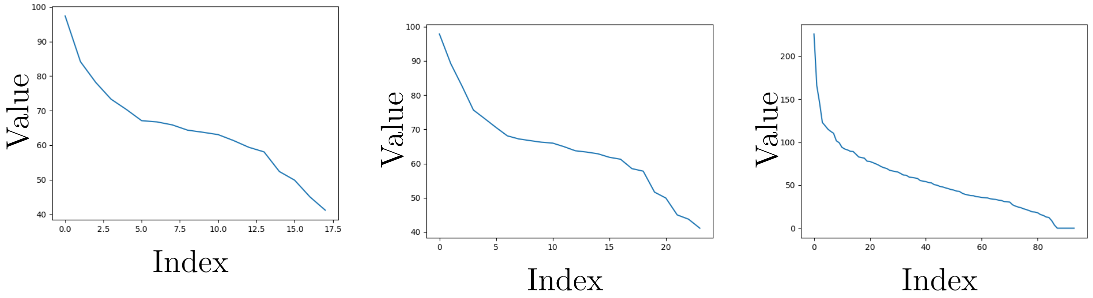
```
Data: $X \in \mathbf{R}^{m \times n}$
Parameters: iterations: $T$, learning rate: $\eta$, momentum: $\alpha(t)$
Result: Embedding $Y^T \in \mathbf{R}^{m \times k}$
for pairwise affinities $p_{\mathrm{j|i}}$ do
    $p_{\mathrm{j|i}} = \frac{\exp(-\|x_{\mathrm{i}}-x_{\mathrm{j}}\|^2/2\sigma^2)}{\sum_{k\neq \mathrm{i}} \exp(-\|x_{\mathrm{i}}-x_k\|^2/2\sigma^2)}$
end
$p_{\mathrm{ij}} = \frac{p_{\mathrm{j|i}}+p_{\mathrm{i|j}}}{2n}$
sample $Y^0$
for $t \leftarrow 1$ to $T$ do
    for pairwise affinities $q_{\mathrm{i|j}}$ do
        $q_{\mathrm{ij}} = \frac{(1+\|y_{\mathrm{i}}-y_{\mathrm{j}}\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_{\mathrm{i}}-y_l\|^2)^{-1}}$
        gradient $\frac{\delta C}{\delta y_{\mathrm{i}}} = 4\sum_{\mathrm{j}}(p_{\mathrm{ij}} - q_{\mathrm{ij}})(y_{\mathrm{i}} - y_{\mathrm{j}})(1 + \|y_{\mathrm{i}} - y_{\mathrm{j}}\|^2)^{-1}$
        $Y^{(t)} = Y^{(t-1)} + \eta\frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$
    end
end
```

The first phase computes pairwise affinities by first centering a Gaussian distribution on a point, X_{i} and calculating the conditional probability of picking a particular point, X_{j} given X_{i}. The conditional probability is proportional to the similarity of those points. After calculating these probabilities, the joint probabilities of $p_{\mathrm{i|j}}$ and $p_{\mathrm{j|i}}$ are averaged.

In the second phase, a T-distribution is centered over every point in the low dimensional space and conditional probabilities are again calculated. The difference between $q_{\mathrm{i|j}}$ and $p_{\mathrm{i|j}}$ is then iteratively minimized. The calculated gradient effectively repels dissimilar points that are represented by a small distance in the low dimensional embedding.

3. RESULTS

3.1 Data and Feature Selection



(a) Continuous Variables at AKI

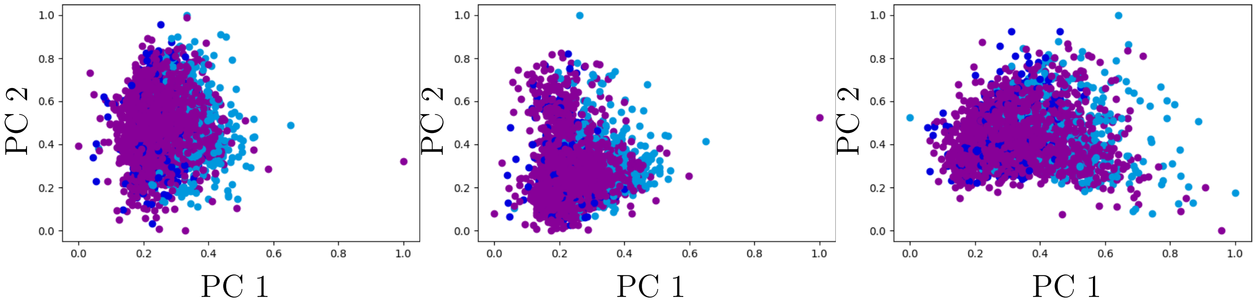
(b) All Variables at AKI

(c) All Continuous Variables

Figure 3.1. Singular Values of Data

Singular values were examined across all selected features (fig. 3.1). These values drop precipitously after the first value in all continuous diagnostic variables. There are similar declines in the set of continuous and noncontinuous diagnostic variables. The approximations used for the data subsets were 5 (fig. 3.5a), 5 (fig. 3.5b), and 8 (fig. 3.5c) rank approximations as selected in conjunction with the elbow criterion [31].

3.1.1 Dimensionality Reduction

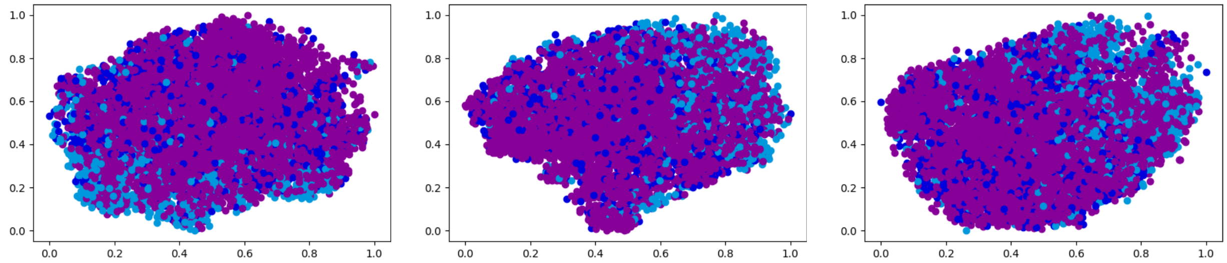


(a) Cont. Diagnostic Variables

(b) All Diagnostic Variables

(c) All Continuous Variables

Figure 3.2. Naive PCA of Feature Subsets, purple (Stage 1), blue (Stage 2), cyan (Stage 3)



(a) Cont. Diagnostic Variables (b) All Diagnostic Variables (c) All Continuous Variables

Figure 3.3. T-SNE of Feature Subsets, purple (Stage 1), blue (Stage 2), cyan (Stage 3)

The first two principal components of PCA (fig 3.2) were plotted in all variable sets. The correlation between continuous diagnostic features and principal components were measured and plotted using Pearson's correlation coefficient (table A.1 and fig. 3.4). At the time of AKI, the first principal component was inversely correlated with bilirubin and hemoglobin and moderately correlated with respiratory rate. The second principal component was highly correlated with sodium and INR. The third principal component was highly correlated to the Charlson comorbidity index. PCA plots did not show separation of AKI stages 1, 2, or 3, mapped as purple, blue, and cyan, respectively.

Additionally, T-SNE did not show separation between staging (fig 3.3), and no significant clustering or separation of data was observed. AKI stages 1, 2, and 3 were plotted as purple, blue, and cyan, respectively.

3.2 Archetypal Analysis

Three archetypes for each of the subsets of data were plotted. In each simplex plot, archetype 1 is located at the top of the polygon and archetype number numerically increases anticlockwise. The majority of points cluster near archetype 3, becoming more diffuse with the inclusion of more features (fig. 3.5). After additional physician review, continuous diagnostic criteria were utilized in further archetype analysis. As the number of archetypes increase, points within the simplex plots become more diffuse (fig. 3.7). Using 5 archetypes ($RSS = .2904$, $SSE = .8221$), the majority of points cluster near archetype 5. Increasing

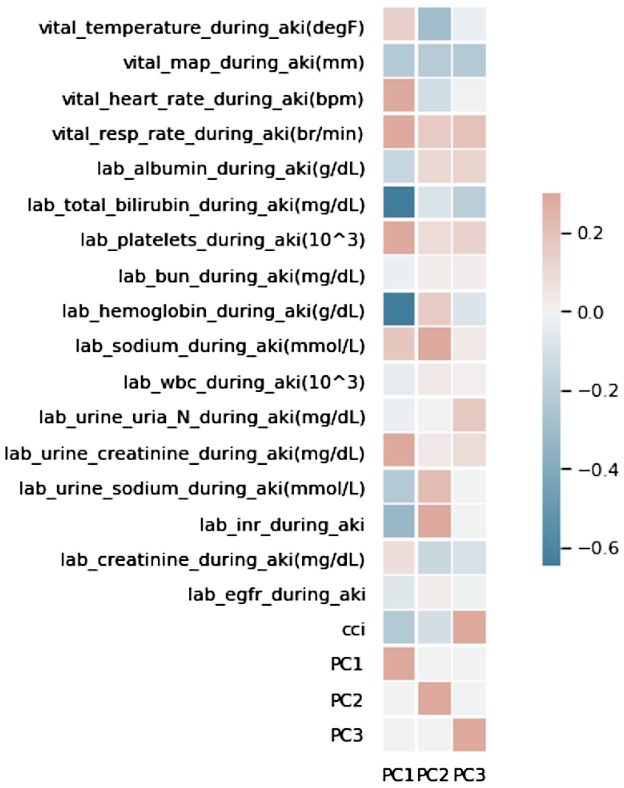
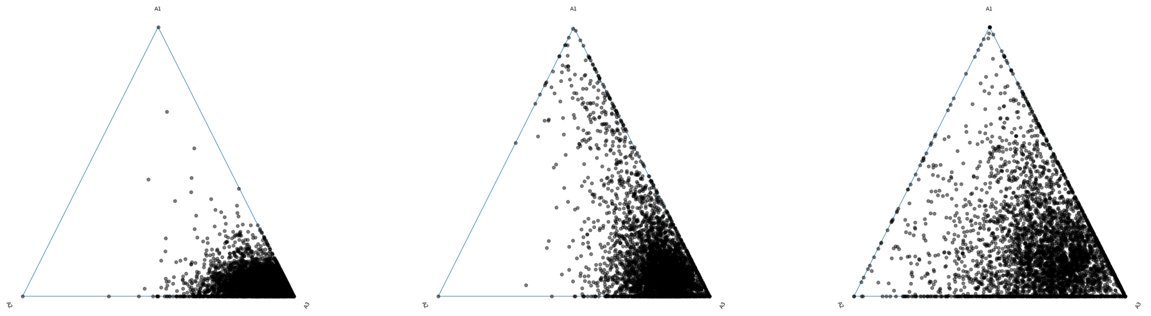


Figure 3.4. Correlation heat map of 3 principal components

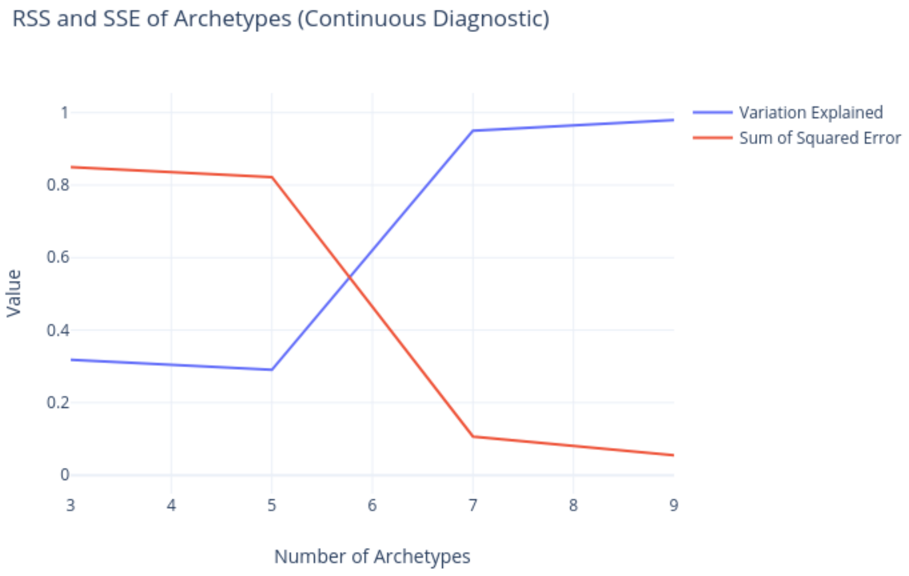
to 7 archetypes (RSS = .9503, SSE = .1067), the majority of points cluster between archetypes 3, 4, and 5. Increasing further (RSS = .9797, SSE = .0547), the majority of the points are not discernibly clustered near a particular archetype or archetypes. The residual sum of squares and sum of squared errors were plotted with respect to the number of archetypes (fig. 3.6). Variation as explained by the model increased and the squared sum of errors decreased. Phenotypes of these archetypes are presented in table A.3, and table A.2. Comorbidities across archetypes were relatively equal except for increased UTI and diabetes without complications in archetype 2. Etiologies were similar across archetypes 2 and 3 with NASH as a predominant etiology of cirrhosis, whereas alcohol was the predominant etiology in archetype 1. Complications were relatively equal across all archetypes with the exception of increased ascites and hepatic encephalopathy in archetype 1. On occasion demographics and laboratory values seem to significantly differ from other archetypes such as decreased age in archetype 1, increased heart rate in archetype 1, increased white blood cell count



(a) Cont. Diagnostic Variables (b) All Diagnostic Variables (c) All Continuous Variables

Figure 3.5. Archetypal analysis

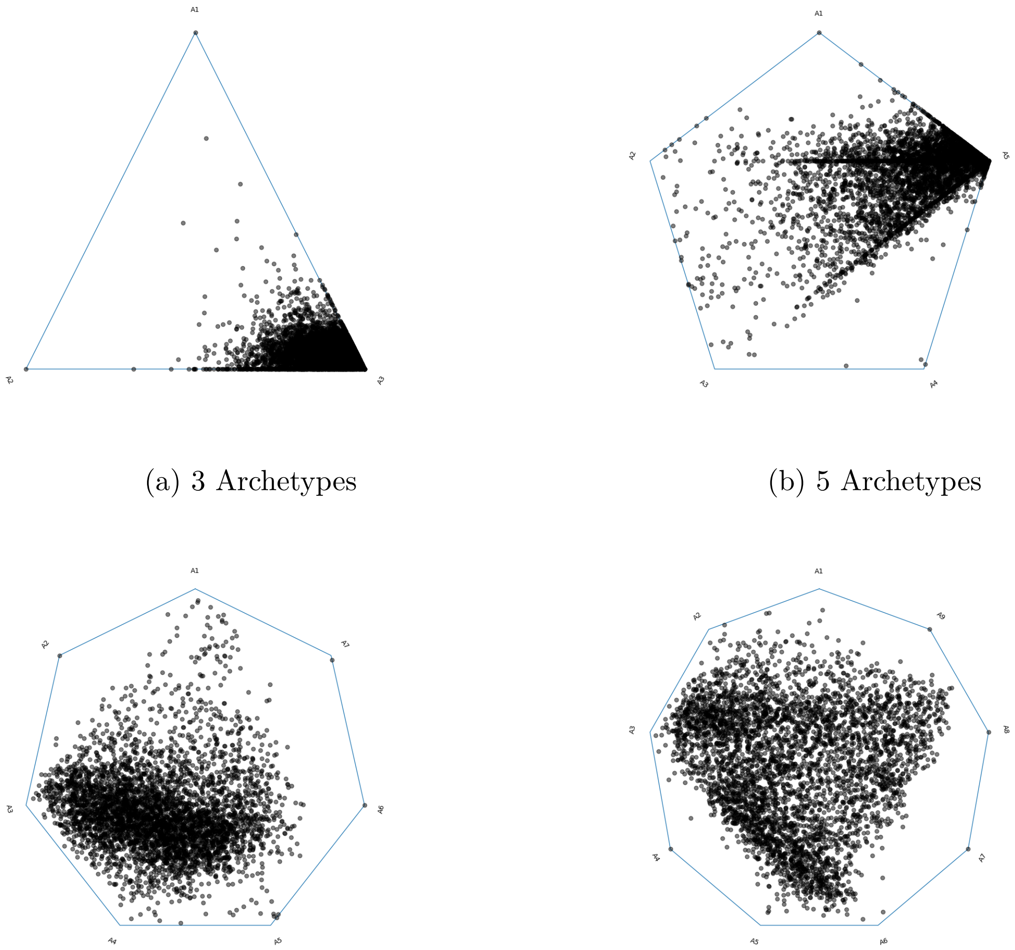
in archetype 1, increased bilirubin in archetype 1 and significantly decreased eGFR and increased serum creatinine in archetype 2. However it is difficult to describe phenotypical significance without proper clinical interpretation.



;

Figure 3.6. Model RSS and SSE

Interactive, 3-dimensional UMAP plots showed clustering of archetypes within the data, however these clusters did not show a one to one correspondence to staging nor the achieve-



(a) 3 Archetypes

(b) 5 Archetypes

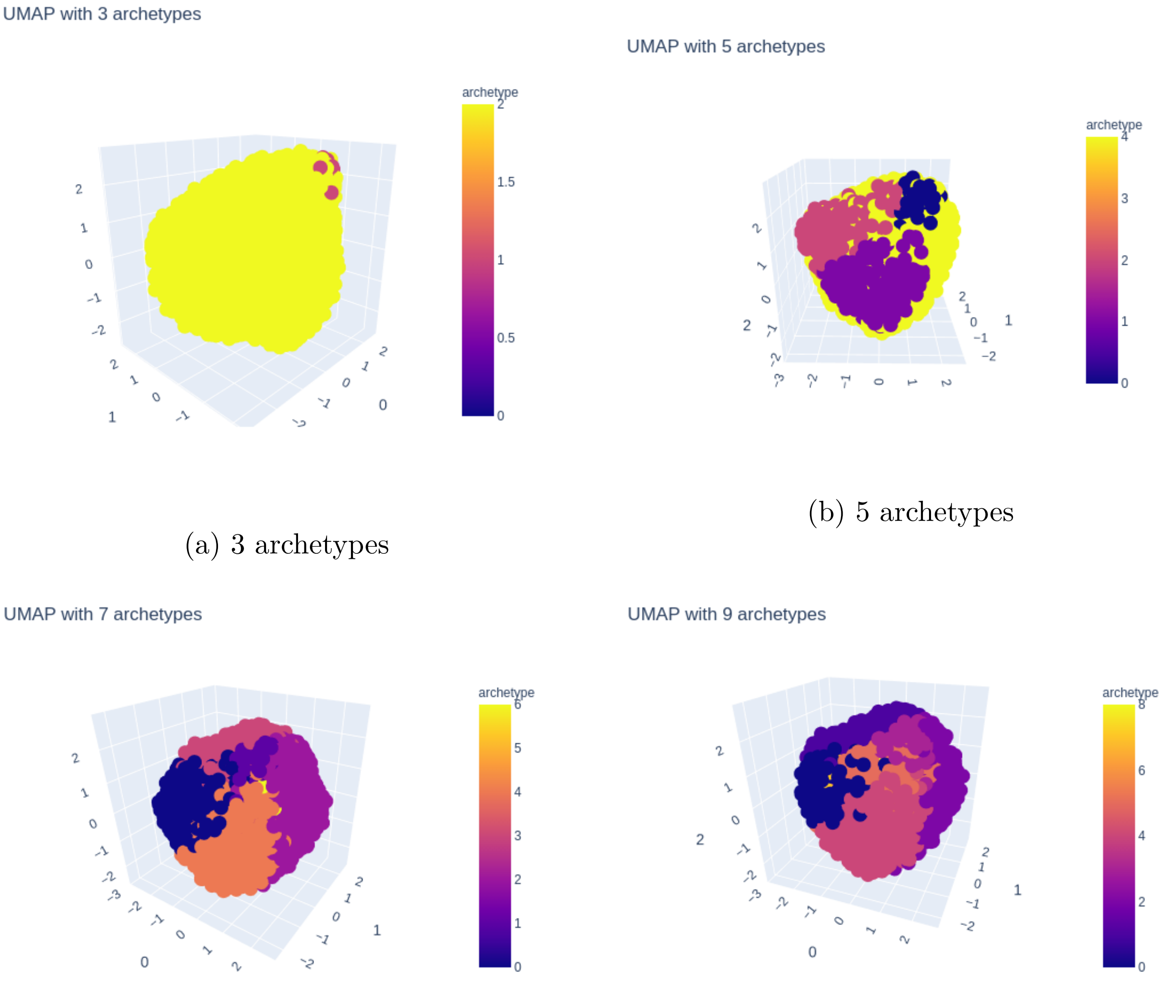
(c) 7 Archetypes

(d) 9 Archetypes

Figure 3.7. Increasing number of Archetypes with Continuous Diagnostic Variables

ment of primary outcome (table A.2). Additional clinical, etiological, and symptomatological features of the archetypes are presented in tables A.2 and A.3.

Lasso regression was performed on two feature subsets, continuous variables at time of AKI and all continuous variables. In the feature subset containing only continuous variables at time of AKI, 4 nonzero coefficients emerged for mean arterial pressure (-0.0167), total bilirubin (0.0037), hemoglobin (0.0040), and serum creatinine (1.0399) with an R^2 value of 0.15. The second feature subset had a significantly higher R^2 value (0.73) and 5 nonzero coefficients emerged for peak creatinine (0.1919), baseline creatinine (-1.1686), baseline eGFR



(a) 3 archetypes

(b) 5 archetypes

(c) 7 archetypes

(d) 9 archetypes

Figure 3.8. UMAP with labeled archetypes

(0.4283), serum creatinine during AKI (1.7414), and eGFR during AKI (-0.2032). It should be noted that the second feature subset extracted variables necessary to determine staging as well as colinear, composite score, variables.

4. DISCUSSION

This study provides a foundational step in understanding and working with the CHF database. Not only do the results exemplify several considerations that one must take when working with real world data, the utility and shortcomings of archetypal analysis as a dimensionality reduction and phenotyping technique are also explored. These shortcomings extend from several steps in the analysis.

Attempts at rectifying noise illustrates the early theoretical question posed by *Catell R.B.* as the WSF, "When shall we stop factoring", problem [32]. If a subset of variables are a linear combination of a smaller subset of variables and an additional, low-level, random background distribution is present, singular values will drop precipitously. The remaining singular values will decline in a slower fashion for the remaining factors. The elongated portion of this graph subsequent to the drop, the *scree*, corresponds to the singular values of small error terms [33], [34]. When a k-rank approximation is constructed, for which k is within the scree, the utility in minimization of reconstruction error is diminished. In the data presented, there are multiple points in which a scree may be interpreted. In these cases, it is necessary that each cutoff be subjectively assessed [35]; however, additional optimal hard cutoffs have been proposed [36].

Results from PCA exemplify its limitations as classic PCA does not differentiate between outliers due to noise or genuine variance in values [37]. These outliers tend to be exaggerated by the L_2 norm and methods utilizing the minimization of the L_1 norm are reccomended [38], [39]. Although PCA suffers from lack of robustness, T-SNE and UMAP are somewhat resistant to outliers and noise; however, only when careful initialization is performed [40], [41]. It is also important to note that with regard to results from T-SNE, there is a uniform distance between most points and a lack of clustering, indicating similarity.

Archetypal Analysis, in the setting of disease phenotyping, provides additional interpretability of dimensionality reduction techniques, giving proportionality to combinations of features presenting at the extremes of the data. While this interpretability holds potential utility, its similarity to classical PCA also begets its sensitivity and lack of robustness. Both rely on the minimization of the L_2 norm. In consequence, archetypal analysis is also sensitive to large and erratic variance in values due to measurement error, noise, or outliers. The types of errors present in a majority of the features are representative of the major sources of error and inconsistencies in electronic health records: Unit error and erroneous transcription. The recording of incorrect units has been a major source of dosing error, often causing hundred to thousand fold increases from therapeutic levels [42]–[45]. The integrity of electronic health records also increases difficulty for scientific reproducibility, as the correctness and completeness vary widely across health centers [46] and input methods [47]. In our study, some features had both a minimum and maximum value increased tenfold of what would be considered reference values and missing values were present in a majority of records, requiring imputation.

Archetypal analysis provides several advantages over NMF and PCA in interpretability. Archetypes proposed using AA presented extreme phenotypes of patients experiencing AKI. Given archetypes, it is possible to return to the original data and query demographics with respect to the calculated archetype according to majority voting. While features can be correlated with principal components, they may not necessarily be additive in contributing to the phenotype of a particular patient. While the non-negativity constraints of NMF allow for additive construction, coefficients of the factors are not normalized. Proportionality of archetype coefficients in the setting of this study allowed for their use in majority voting while classifying archetypes and ease in interpretability of extremes.

It is also notable that archetypal analysis may not produce a representative sample with respect to each archetype. As seen in plots of increasing archetypes and in the algorithm for archetype analysis, there is no requirement of a minimum quantity of data to define a particular archetype; therefore, performing statistical testing to describe the significance of features in an archetype must be done with care. Few records per archetype may not meet basic requirements for goodness of fit or ensuring robust testing with analysis of variance.

Additional results from this study regarding the use of LASSO further validate physician staging criteria in this patient cohort. The inclusion of serum creatinine in the absence of other required variables for staging implies it is necessary but the low R^2 value suggests that it, in the absence of other staging criteria, is not sufficient. However when all continuous variables were used with LASSO, the inclusion of only and all relevant creatinine variables signifies that all current staging criteria may be sufficient with serum creatinine at time of AKI bearing necessity as a clinical indicator of AKI, further validating the KDIGO model and its use in a cirrhotic setting.

Further inquiry is required in the setting of archetype analysis. First, the objective function of AA does not provide a notion of separability of archetypes, only how well the factored matrices approximate the original data. This increases difficulty in tuning hyperparameters as well in the ability of AA to be used as a measure of clustering. When coefficients of archetypes are mapped as a measured probability of a particular datum consisting of n archetypes, the probabilities may be represented in a density matrix, ρ, and purity, $\chi(\rho)$, may be defined as

$$\chi(\rho) = tr(\rho^2)$$

$$\rho = \begin{bmatrix} n_1 & 0 & 0 & 0 \\ 0 & n_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & n_n \end{bmatrix}$$

bounded by $\frac{1}{n} \leq \chi(\rho) \leq 1$[48] where the lower bound signifies an equal mix of archetypes and the upper bound implies a pure type. Therefore, multiobjective optimization minimizing RSS and maximizing average purity in archetypal analysis may be of further utility in patient phenotyping.

REFERENCES

[1] S. A. Pendergrass and D. C. Crawford, “Using electronic health records to generate phenotypes for research,” *Current protocols in human genetics*, vol. 100, no. 1, e80, 2019.

[2] V. D. Kumar and H. J. Tipney, *Biomedical literature mining*. Springer, 2014.

[3] L. Hood and S. H. Friend, “Predictive, personalized, preventive, participatory (p4) cancer medicine,” *Nature reviews Clinical oncology*, vol. 8, no. 3, pp. 184–187, 2011.

[4] S. Saria and A. Goldenberg, “Subtyping: What it is and its role in precision medicine,” *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 70–75, 2015.

[5] D. P. Chen, S. C. Weber, P. S. Constantinou, T. A. Ferris, H. J. Lowe, and A. J. Butte, “Clinical arrays of laboratory measures, or “clinarrays”, built from an electronic health record enable disease subtyping by severity,” in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2007, 2007, p. 115.

[6] J.-E. Bibault and L. Xing, “The role of big data in personalized medicine,” *Precision Medicine in Oncology*, pp. 229–247, 2020.

[7] I. Perros, E. E. Papalexakis, H. Park, R. Vuduc, X. Yan, C. Defilippi, W. F. Stewart, and J. Sun, “Sustain: Scalable unsupervised scoring for tensors and its application to phenotyping,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, ser. KDD ’18, London, United Kingdom: Association for Computing Machinery, 2018, pp. 2080–2089, ISBN: 9781450355520. DOI: 10.1145/3219819.3219999. [Online]. Available: https://doi.org/10.1145/3219819.3219999.

[8] J. K. De Freitas, K. W. Johnson, E. Golden, G. N. Nadkarni, J. T. Dudley, E. P. Bottinger, B. S. Glicksberg, and R. Miotto, “Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records,” *medRxiv*, 2020.

[9] B. K. Beaulieu-Jones, C. S. Greene, *et al.*, “Semi-supervised learning of the electronic health record for phenotype stratification,” *Journal of biomedical informatics*, vol. 64, pp. 168–178, 2016.

[10] D. C. Koestler, C. J. Marsit, B. C. Christensen, M. R. Karagas, R. Bueno, D. J. Sugarbaker, K. T. Kelsey, and E. A. Houseman, “Semi-supervised recursively partitioned mixture models for identifying cancer subtypes,” *Bioinformatics*, vol. 26, no. 20, pp. 2578–2585, 2010.

[11] T. Ma and A. Zhang, “Affinity network fusion and semi-supervised learning for cancer patient clustering,” *Methods*, vol. 145, pp. 16–24, 2018.

[12] J. Zhao, Q. Feng, P. Wu, J. L. Warner, J. C. Denny, and W.-Q. Wei, "Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of lipoprotein (a)(lpa)," *PloS one*, vol. 14, no. 2, e0212112, 2019.

[13] A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[14] A. S. Levey and M. T. James, "Acute kidney injury," *Annals of Internal Medicine*, vol. 167, no. 9, ITC66–ITC80, Nov. 7, 2017, Publisher: American College of Physicians, ISSN: 0003-4819. DOI: 10.7326/AITC201711070. [Online]. Available: https://www.acpjournals.org/doi/abs/10.7326/aitc201711070.

[15] A. Khwaja, "Kdigo clinical practice guidelines for acute kidney injury," *Nephron Clinical Practice*, vol. 120, no. 4, pp. c179–c184, 2012.

[16] Eric A. J. Hoste, John A. Kellum, Nicholas M. Selby, Alexander Zarbock, Paul M. Palevsky, Sean M. Bagshaw, Stuart L. Goldstein, Jorge Cerdá, and Lakhmir S. Chawla, "Global epidemiology and outcomes of acute kidney injury," *Nature Reviews Nephrology*, vol. 14, pages607–625, Aug. 22, 2018. [Online]. Available: https://www.nature.com/articles/s41581-018-0052-0/briefing/signup/?origin=Nature&originReferralPoint=EmailBanner.

[17] I. Acosta-Ochoa, J. Bustamante-Munguira, A. Mendiluce-Herrero, J. Bustamante-Bustamante, and A. Coca-Rojo, "Impact on outcomes across KDIGO-2012 AKI criteria according to baseline renal function," *Journal of Clinical Medicine*, vol. 8, no. 9, p. 1323, Sep. 2019, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/jcm8091323. [Online]. Available: https://www.mdpi.com/2077-0383/8/9/1323.

[18] Zaccaria Ricci, Dinna N. Cruz, and Claudio Ronco, "Classification and staging of acute kidney injury: Beyond the RIFLE and AKIN criteria | nature reviews nephrology," *Nature Reviews Nephrology*, no. 7, pp. 201–208, Mar. 1, 2011. [Online]. Available: https://www.nature.com/articles/nrneph.2011.14.

[19] K. Makris and L. Spanou, "Acute kidney injury: Diagnostic approaches and controversies," *The Clinical Biochemist Reviews*, vol. 37, no. 4, pp. 153–175, Dec. 2016, ISSN: 0159-8090. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5242479/.

[20] P. K. Moore, R. K. Hsu, and K. D. Liu, "Management of acute kidney injury: Core curriculum 2018," *American Journal of Kidney Diseases*, vol. 72, no. 1, pp. 136–148, Jul. 2018, ISSN: 02726386. DOI: 10.1053/j.ajkd.2017.11.021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0272638617311411.

[21] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Third. The Johns Hopkins University Press, 1996.

[22] S. K. Jha and R. Yadava, "Denoising by singular value decomposition and its application to electronic nose data processing," *IEEE Sensors Journal*, vol. 11, no. 1, pp. 35–44, 2010.

[23] B. P. Epps and E. M. Krivitzky, "Singular value decomposition of noisy data: Noise filtering," *Experiments in Fluids*, vol. 60, no. 8, pp. 1–23, 2019.

[24] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal components analysis to the exponential family.," in *Nips*, vol. 13, 2001, p. 23.

[25] R. Bar-Yehuda and S. Even, "A local-ratio theorm for approximating the weighted vertex cover problem," Computer Science Department, Technion, Tech. Rep., 1983.

[26] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018. DOI: 10.21105/joss.00861. [Online]. Available: https://doi.org/10.21105/joss.00861.

[27] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[28] L. McInnes, J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, 2020. arXiv: 1802.03426 `[stat.ML]`.

[29] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *NIPS*, Citeseer, vol. 15, 2002, pp. 833–840.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[31] A. Hardy, "An examination of procedures for determining the number of clusters in a data set," in *New approaches in classification and data analysis*, Springer, 1994, pp. 178–185.

[32] R. B. Cattell, "The scree test for the number of factors," *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276, 1966.

[33] B. Everett, *An introduction to latent variable models*. Springer Science & Business Media, 2013.

[34] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A practical approach to microarray data analysis*, Springer, 2003, pp. 91–109.

[35] M. O. Ulfarsson and V. Solo, "Dimension estimation in noisy pca with sure and random matrix theory," *IEEE transactions on signal processing*, vol. 56, no. 12, pp. 5804–5816, 2008.

[36] M. Gavish and D. L. Donoho, "The optimal hard threshold for singular values is 4/\sqrt {3} ," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 5040–5053, 2014.

[37] S. Bailey, "Principal component analysis with noisy and/or missing data," *Publications of the Astronomical Society of the Pacific*, vol. 124, no. 919, pp. 1015–1023, 2012, ISSN: 00046280, 15383873. [Online]. Available: http://www.jstor.org/stable/10.1086/668105.

[38] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse pca by l1-norm maximization," *Pattern Recognition*, vol. 45, no. 1, pp. 487–497, 2012.

[39] N. Kwak, "Principal component analysis based on l1-norm maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1672–1680, 2008. DOI: 10.1109/TPAMI.2008.114.

[40] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 493–501.

[41] D. Kobak and G. C. Linderman, "Initialization is critical for preserving global data structure in both t-sne and umap," *Nature Biotechnology*, vol. 39, no. 2, pp. 156–157, 2021.

[42] K. Chappell and C. Newman, "Potential tenfold drug overdoses on a neonatal unit," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 89, no. 6, F483–F484, 2004.

[43] E. Kirkendall, M. Kouril, T. Minich, and S. Spooner, "Analysis of electronic medication orders with large overdoses: Opportunities for mitigating dosing errors," *Applied clinical informatics*, vol. 5, no. 1, p. 25, 2014.

[44] M. Husch, C. Sullivan, D. Rooney, C. Barnard, M. Fotis, J. Clarke, and G. Noskin, "Insights from the sharp end of intravenous medication errors: Implications for infusion pump technology," *BMJ Quality & Safety*, vol. 14, no. 2, pp. 80–86, 2005.

[45] S. Bowman, “Impact of electronic health record systems on information integrity: Quality and safety implications,” *Perspectives in health information management*, vol. 10, no. Fall, 2013.

[46] N. G. Weiskopf and C. Weng, “Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research,” *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[47] L. A. Knake, M. Ahuja, E. L. McDonald, K. K. Ryckman, N. Weathers, T. Burstain, J. M. Dagle, J. C. Murray, and P. Nadkarni, “Quality of ehr data extractions for studies of preterm birth in a tertiary care center: Guidelines for obtaining reliable data,” *BMC pediatrics*, vol. 16, no. 1, pp. 1–8, 2016.

[48] K. Bartkiewicz, K. Lemr, and A. Miranowicz, “Direct method for measuring of purity, superfidelity, and subfidelity of photonic two-qubit mixed states,” *Physical Review A*, vol. 88, no. 5, p. 052 104, 2013.

A. TABLES

Table A.1. Correlation of features and principal components

	PC1	PC2	PC3
Temperature (°F)	0.1372	-0.2904	-0.0239
MAP (mm)	-0.2243	-0.2160	-0.2250
Heart Rate (bpm)	0.3169	-0.1147	0.0015
Resp Rate (br/min)	0.4284	0.1645	0.1906
Albumin (g/dL)	-0.1552	0.1093	0.1239
Bilirubin(mg/dL)	-0.6400	-0.0814	-0.2002
Platelets(10^3)	0.3244	0.0911	0.1272
BUN (mg/dL)	-0.0214	0.0232	0.0224
Hemoglobin (g/dL)	-0.6459	0.1512	-0.0784
Sodium (mmol/L)	0.1727	0.8071	0.0339
WBC (10^3)	-0.0349	0.0405	0.0107
Urine Urea Nitrogen (mg/dL)	-0.0182	-0.0036	0.1676
Urine Creatinine (mg/dL)	0.3528	0.0366	0.0860
Urine Sodium (mmol/L)	-0.2258	0.2186	-0.0006
INR	-0.3234	0.5044	0.0007
Creatinine (mg/dL)	0.0756	-0.1403	-0.0981
eGFR	-0.0642	0.0272	-0.0138
Charlson Comorbidity Index	-0.2280	-0.1065	0.9427

Table A.2. Comorbidities and Complications Across Archetypes.

	Archetype 1 ($n = 448$) (%)	Archetype 2 ($n = 32$) (%)	Archetype 3 ($n = 3858$) (%)
Comorbidities			
Pneumonia	46 (10.27)	3 (9.375)	357 (9.254)
UTI	15 (3.348)	4 (12.500)	184 (4.770)
Cellulitis	0 (0)	0 (0)	13 (0.3370)
Bacteremia	43 (9.598)	1 (3.125)	232 (6.014)
Sepsis	61 (13.61)	2 (6.250)	253 (6.558)
C.*diff*	13 (2.901)	2 (6.250)	68 (01.76)
Diabetes w.o. chronic complications	250 (55.80)	22 (68.750)	2071 (53.68)
Diabetes w. chronic complications	0	0 (0)	1 (0.0259)
Etiology			
Alchohol	182 (40.62)	5 (15.625)	812 (21.05)
NASH	100 (22.32)	17 (53.125)	1671 (43.31)
Hepatitis C	70 (15.62)	5 (15.62)	640 (16.58)
Other	26 (5.804)	2 (6.250)	205 (5.31)
Complication			
Ascites	95 (21.20)	5 (15.625)	641 (16.61)
Hepatic Encephalopathy	180 (40.17)	7 (21.875)	931 (24.13)
Spontaneous Bacterial Peritonitis	0 (0)	0 (0)	0 (0)
esophageal variceal hemorrhage	33 (7.366)	2 (6.250)	167 (4.329)
Hepatocellular Carcinoma	13 (2.902)	2 (6.250)	142 (3.681)
Primary Outcome Reached	232 (51.78)	13 (40.625)	2653 (68.77)

Table A.3. Archetype Lab Values

	mean	std	min	25%	50%	75%	max
age							
archetype 1	54.788	13.579	25.000	47.000	54.000	63.000	90.000
archetype 2	62.125	13.259	34.000	54.250	63.000	70.250	90.000
archetype 3	62.384	12.518	18.000	54.000	62.000	71.000	90.000
heart rate(bpm)							
archetype 1	92.849	17.437	18.000	82.000	92.000	103.000	148.000
archetype 2	83.831	15.848	55.000	70.000	85.400	95.250	112.000
archetype 3	88.908	18.514	0.000	77.000	88.000	98.375	186.000
Temperature (degF)							
archetype 1	97.569	5.569	0.000	97.520	98.000	98.420	101.000
archetype 2	95.239	10.808	37.000	96.485	97.700	98.205	98.780
archetype 3	98.006	2.304	37.000	97.600	98.060	98.420	106.700
Resp rate (br/min)							
archetype 1	20.361	8.682	-5.000	18.000	19.000	20.125	108.000
archetype 2	19.072	4.130	9.000	18.000	19.000	20.000	32.000
archetype 3	19.217	5.062	0.000	17.800	18.500	20.000	111.000
map (mm)							
archetype 1	76.116	12.194	4.000	69.000	76.000	83.592	122.670
archetype 2	87.068	19.391	58.000	75.333	82.033	96.000	135.667
archetype 3	84.230	15.062	38.000	75.000	82.683	91.000	260.000
wbc (10^3)							
archetype 1	19.183	9.882	2.700	11.675	17.450	24.350	78.200
archetype 2	13.296	7.855	3.800	8.538	11.510	16.100	42.300
archetype 3	10.018	5.290	0.000	6.300	9.149	12.500	48.600

Continued on next page

Table A.3. Archetype Lab Values

	mean	std	min	25%	50%	75%	max
Hemoglobin (g/dL)							
archetype 1	9.835	2.540	2.800	8.100	9.700	11.500	20.000
archetype 2	10.138	2.841	4.700	8.075	10.000	11.900	15.800
archetype 3	10.798	2.487	1.000	9.100	10.600	12.400	20.000
INR							
archetype 1	2.194	1.830	0.108	1.430	1.800	2.325	16.100
archetype 2	1.643	1.027	0.106	1.275	1.445	1.693	6.560
archetype 3	1.420	0.624	0.075	1.150	1.350	1.600	6.608
Platelets (10^3)							
archetype 1	172.007	112.407	8.000	96.750	146.500	212.250	892.000
archetype 2	186.834	114.206	8.500	92.250	172.000	252.500	459.000
archetype 3	159.127	96.303	1.000	94.000	141.000	201.000	1318.000
Albumin (g/dL)							
archetype 1	2.054	0.572	0.002	1.700	2.000	2.400	3.900
archetype 2	2.911	0.680	1.600	2.520	2.755	3.350	4.500
archetype 3	2.890	0.695	0.210	2.400	2.800	3.300	6.000
total bilirubin (mg/dL)							
archetype 1	14.528	12.276	0.100	3.800	11.250	23.500	70.000
archetype 2	2.032	2.521	0.100	0.500	0.800	2.075	9.070
archetype 3	3.101	3.691	0.000	0.883	1.900	3.928	43.900
BUN (mg/dL)							
archetype 1	44.757	28.139	2.000	23.000	39.000	60.000	167.000
archetype 2	135.938	50.656	8.000	103.000	137.000	160.500	264.000
archetype 3	37.828	24.763	1.000	20.000	32.000	49.000	171.000

Continued on next page

Table A.3. Archetype Lab Values

	mean	std	min	25%	50%	75%	max
eGFR							
archetype 1	35.384	21.916	3.997	17.527	32.668	47.922	179.814
archetype 2	4.913	2.624	0.408	3.226	4.101	5.571	12.929
archetype 3	41.275	26.283	2.689	22.329	36.857	54.717	262.540
Serum Creatinine (mg/dL)							
archetype 1	2.716	2.019	0.500	1.400	1.995	3.170	13.700
archetype 2	13.980	11.987	4.600	8.647	11.750	14.795	74.900
archetype 3	2.329	1.891	0.260	1.210	1.700	2.600	16.810
Sodium (mmol/L)							
archetype 1	100.350	53.208	10.600	14.075	129.500	135.000	176.000
archetype 2	120.447	41.733	13.500	128.000	135.000	140.000	154.000
archetype 3	115.065	57.495	11.300	125.000	134.000	138.000	1380.000
Urine creatinine (mg/dL)							
archetype 1	151.627	117.445	3.400	111.954	137.120	166.062	2280.000
archetype 2	140.629	77.159	52.400	103.487	123.500	147.500	403.527
archetype 3	134.709	53.045	11.000	105.280	128.502	153.764	690.100
Urine Sodium (mmol/L)							
archetype 1	33.128	22.409	5.000	23.000	31.200	39.650	213.300
archetype 2	42.388	19.876	14.700	34.350	39.450	44.525	124.100
archetype 3	39.564	28.676	5.000	27.800	35.600	44.600	960.000
urine uria (mg/dL)							
archetype 1	405.281	89.873	67.000	359.870	387.635	449.500	1045.000
archetype 2	413.572	69.750	290.400	364.300	408.000	463.025	569.100
archetype 3	431.397	90.300	0.900	375.900	426.600	467.700	1300.000

Continued on next page

Table A.3. Archetype Lab Values

	mean	std	min	25%	50%	75%	max
MELD-Na							
archetype 1	30.594	7.153	7.968	25.460	30.848	35.347	50.591
archetype 2	26.125	4.861	20.242	22.836	25.017	28.182	40.764
archetype 3	22.017	6.024	6.430	19.177	22.822	25.655	41.962
MELD							
archetype 1	28.376	8.113	7.968	22.584	27.521	33.358	54.390
archetype 2	24.137	5.021	17.900	21.202	22.707	25.074	40.764
archetype 3	19.311	5.375	6.430	16.283	19.636	22.271	43.189

Table A.4. Coefficients resulting from LASSO regression

Continuous Variables at AKI	Coefficient
Mean arterial pressure	-0.0167
Total Bilirubin	0.0037
Hemoglobin	0.0040
Serum Creatinine	1.0399
R^2	0.15

,

All Continuous Variables	Coefficient
Peak Creatinine	0.1919
Baseline Creatinine	-1.1686
Baseline eGFR	0.4283
Creatinine during AKI	1.7414
eGFR during AKI	-0.203
R^2	0.73

VITA

Education

Purdue University Master of Science, Computer Science	August 2017 - May 2021
The University of Chicago Bachelor of Arts, Biological Sciences	September 2010 - August 2014

Experience

Graduate Research Assistant Regenstrief Center for Healthcare Engineering	June 2020 - May 2020
Graduate Teaching Assistant Purdue University, CS 177	June 2018 - May 2020
UChicago Medicine:	
Clinical Research Coordinator	March 2016 - June 2017
Research Specialist	August 2014 - March 2016
Student Research Assistant	May 2011 - August 2014

Positions Held

IBD Telehealth Administrator UChicago Medicine Gastroenterology	2015 - 2017
Research Coordinator Fecal Microbiota Transplantation Comittee, UChicago Medicine	2014 - 2016
Member: Health Care Access Taskforce: Crohns and Colitis Foundation of America	2012 - 2015

Professional Service

Invited Peer Review Gastroenterology	December 2019