Interpretable Data Phenotyping for Healthcare via Unsupervised Learning

Christine Allen

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Computer Science and Systems

University of Washington

2020

Committee:

Ankur Teredesai

Juhua Hu

Vikas Kumar

Program Authorized to Offer Degree:
Computer Science and Systems

University of Washington

Abstract

Interpretable Data Phenotyping for Healthcare via Unsupervised Learning

Christine Allen

Chair of the Supervisory Committee:
Professor Ankur Teredesai
Department of Computer Science and Systems

Healthcare applications of machine learning tend toward greater requirements for model transparency than most applications. Yet the often high dimensionality of the data presents a significant impediment to meeting this requirement, particularly as it relates to the underlying relationships contributing to an individual prediction. Thus emerged the concept of "data phenotypes", clinically relevant groupings that facilitate population statistics and reduce barriers in the development of quality machine learning models. However, the results of current phenotyping methods are often difficult to interpret, and they often require clarification from an experienced clinician to be useful. This is a problem for administration-level prediction problems in particular, for example Length of Stay prediction, because those developing the models are not commonly clinicians, and because the results of these models are often desired with a fast turnaround.

With the above in mind, this thesis reviews the utility of four prominent phenotyping approaches: k-means, agglomerative clustering, non-negative matrix factorization, and non-negative tensor factorization. We propose variants of the four approaches with the goal of producing distinct feature membership. We then show that our proposals can produce easily understandable phenotypes at no detriment to prediction performance over some real healthcare tasks.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables iv

Chapter 1: Introduction 1

Chapter 2: Context and Related Work 3
2.1 Data Phenotyping and Healthcare 3
2.2 Medical Coding Systems 4
2.3 Clustering 5
2.4 Matrix Factorization 6
2.5 Tensor Factorization 7

Chapter 3: Proposed Method 9
3.1 Problem Definition 9
3.2 Phenotyping Approaches 9

Chapter 4: Experiments 12
4.1 Data 12
4.2 Setup 14
4.3 Prediction Performance 15
4.4 Feature Membership of Resultant Phenotypes 18
4.5 Feature Importance in Prediction Problems 19
4.6 Computation Time 22
4.7 Discussion 22

Chapter 5: Conclusion 25
5.1 Summary of the Thesis 25

5.2 Future Directions 25

Bibliography 27

Appendices 32

Appendix A: Distributions of ICD-9 Diagnoses 33

Appendix B: Additional Phenotype and LOS Prediction Results 37

LIST OF FIGURES

Figure Number		Page
2.1	Illustration of Marble Approach from Ho *et al.* [1]	8
3.1	NMF Feature Assignment with Constraint	11
4.1	ROC AUC with vs. without Procedure Codes	15
4.2	ROC AUC Relative to the Number of Phenotypes	16
4.3	Top Five Features Ranked by Relative Importance in True Positive Prediction	19
4.4	Top Ten ICD-9 Prediction Features Among to True Positive Results, Ranked by Mean LIME Weight and Matched to CCS Level 2 Categories	20
4.5	Top Ten CCS Level 2 Prediction Features Among to True Positive Results, Ranked by Mean LIME Weight	20
4.6	Top Ten K-Means(f) Phenotype Prediction Features Among to True Positive Results, Ranked by Mean LIME Weight	21
4.7	LIME Feature Importance for Example False Negative Prediction, K-Means(f) Phenotype-based Prediction	21
4.8	Computation Time per Number of Phenotypes	22
B.1	Feature Membership of Phenotyping Approaches	37

LIST OF TABLES

Table Number Page

2.1 Granularity of Medical Coding Systems 5

4.1 MIMIC-III Data Subsets 12
4.2 MIMIC-III Target Distributions for ICU Length of Stay 13
4.3 Length of Stay Prediction Model Metrics 17
4.4 Distribution of Feature Counts 18

A.1 Most Common Diagnoses, Neonatal subset 34
A.2 Most Common Diagnoses, 65+ Subset 35
A.3 Most Common CCS Categories, Neonatal Subset 35
A.4 Most Common CCS Categories, 65+ Subset 36

B.1 Comparison of Feature Membership, CCS vs. K-Means(f) Phenotypes 38
B.2 Comparison of Feature Importance Among LIME, SHAP, and XGBOOST Metrics 39

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to University of Washington. Thank you also to KenSci, Inc. for the use of their resources.

Chapter 1

INTRODUCTION

It is difficult to understate the volume and complexity of healthcare data generated globally every year. As of 2016 there were greater than 35 million in-patient clinical encounters annually in the United States alone.[2] That is 35 million datapoints per year considering only the event of hospital admission. When combined with outpatient events, diagnoses, laboratory values, notes, imaging, family histories, *et cetera*, the problem explodes in size.

Statistical methods, and even standard supervised machine learning (ML) models cannot capture the unknown number of intricate interactions among the variables in the data. Yet, such interactions are critical to the results of predictive models, for example in a clinical setting where they may directly affect patient care. However, a growing body of research is investigating the application of unsupervised learning to generate clinically relevant data groupings that can aid in both data analysis and prediction. These groupings have come to be known as "data phenotypes".

Many clustering and matrix factorization approaches have been applied to generate data phenotypes with demonstrable success. However, most of these applications focus either on uncovering novel health states or on achieving highly accurate detection of well-known diseases.[3][4] Moreover, the results from these applications typically require the input of an experienced healthcare provider, which can slow the model development process.

To provide understandable phenotypes, we propose alternative approaches to known clustering and factorization methods. Using data from the Medical Information Mart for Intensive Care III (MIMIC-III) database, we generated two data subsets on which to test our proposal.[5] Phenotype sets generated from these data were then compared by their feature membership and their performance in a simplified prediction problem. We demonstrate that

the proposed approaches may be used to produce phenotypes that are easy to interpret without heavy reliance on an experienced clinician.

Chapter 2

CONTEXT AND RELATED WORK

2.1 Data Phenotyping and Healthcare

Biological phenotypes are observable characteristics resulting from the interaction of an organisms genes with the environment. Flower petal color is a commonly used example. Inspired by that concept, the term "data phenotype" (a.k.a. "electronic phenotype", a.k.a. "computational phenotype") has emerged to describe groupings of electronic health records (EHR) that are associated with clinically relevant concepts. "Complications resulting from diabetes," is one example.[1] The machine learning approach used to produce these groupings depends upon the data types and the problem under evaluation. One such problem type is the prediction of administration-level outcomes such as length of in-patient hospital stay, also referred to as "LOS".

Solutions to prediction problems such as LOS can be important tools for identifying health risks, but they are significantly more useful when they can provide a clear understanding of the underlying risk factors. For example, long lengths of stay could be caused by some facet of disease progression, and so their prediction indicates the need for additional treatment. But, long lengths of stay can also be caused by other factors such as the development of a healthcare-acquired infection, or due to operational flaws, such as lax management of certain patient groups.[6][7][8] By providing assistive information to clinical staff, these predictions could lead to life saving decisions. Hence, model transparency (i.e., the linkage of individual predictions to clear rationale) is a critical component of these types of models.

However, although it is common to involve a medical specialist in the development process, the specialists developing these predictive models are not typically healthcare providers. They are machine learning specialists. In addition, the development time for these solutions

is typically short. Hence there is an exceptional incentive to automate the phenotyping process with as little need for physician validation as possible.

One promising validation solution is the use of a common phenotype dictionary, such as the Phenotype Knowledgebase (PheKB). PheKB is a tool as well as a public repository for generating phenotypes using validated rule-based algorithms.[9] Unfortunately, PheKB algorithms are distinct for each phenotype, which complicates model explanations. In addition, they may not perform well across data from different healthcare systems.[10] Moreover, the PheKB database is not comprehensive, so the phenotypes available may not be applicable to a given problem. Therefore, an alternate validation strategy is needed.

2.2 Medical Coding Systems

Among the first attempts to manage and to analyze medical data was standardization through codification. The International Classification of Disease (ICD) system has been used globally to standardize medical records since 1893, and it is now in its 11^{th} edition.[11] ICD codes are manually designed for medical reporting, with an inclination toward specificity and ease of indexing. They are also highly granular and organized by a known hierarchy of disease.

More recently, the US Healthcare Cost and Utilization Project (HCUP) developed Clinical Classification Software (CCS) with its own, manually-generated categorization system.[2] This system groups individual ICD codes into more manageable groups at multiple levels (see Table 2.1), organized by higher-level concepts such as "body systems". Level 2 CCS categories are often used in health informatics for data interpretation.

	Classification	Number of Unique Codes
Diagnosis	ICD-9 Codes	13000+
	CCS Level 1 Categories	19
	CCS Level 2 Categories	136
	CCS Level 3 Categories	367
	CCS Level 4 Categories	209
Procedure	ICD-9 Codes	3879
	CCS Level 1 Categories	16
	CCS Level 2 Categories	206
	CCS Level 3 Categories	184

Table 2.1: Granularity of Medical Coding Systems

2.3 Clustering

Clustering approaches group data based on their proximity to each other. In practice this typically means that each observation (for example, each patient or each clinical encounter) is modeled as a vector of feature values. That vector is then compared to analogous vectors for each of the other patients or encounters in the dataset. This comparison uses one of several, typically simple calculations to determine the similarity or dis-similarity between pairs of vectors. Manhattan distance is one of the most common metrics applied for this purpose.

The k-means partitional clustering algorithm is one of the most popular clustering algorithms in use.[12][13] Beginning with the centroids of a random data partition, the algorithm assigns each datum to the closest of k centroids, forming k data clusters[1]. The cluster centers (which may or may not also be previous centroids) are then re-assigned as the new set

[1]It's worth mentioning that current, prominent implementations of "k-means" use enhancements to the original algorithm, since true k-means offers no accuracy guarantees. Proposed in 2006 by David Arthur and Sergei Vassilvitskii of Stanford University, k-means++ starts from a single, uniformly-selected random centroid (rather than from a random partition), which has been shown to produce optimal clusters more frequently.[14]

of centroids for the next iteration. This re-assignment is then repeated until an optimization criterion is met. K-means is easy to implement, relatively fast, and is more capable than other clustering algorithms at handling a large number of features. Some examples of k-means applications for this purpose include studies of COPD,[4] psychiatry,[15] and multimorbidity.[16]

Another popular clustering approach is agglomerative (hierarchical) clustering. Agglomerative clustering begins with each vector as its own, single-member cluster. Then for each successive step, clusters are merged together until a stopping criterion is met (for example, if it reaches N number of clusters).[12][13] Agglomerative clustering, like k-means, is easy to implement, and it is often favored because it produces a useful dendrogram that illustrates associations in the data. Example phenotyping implementations of agglomerative clustering include analyses of multimorbidity,[17] as well as of renal disease.[18]

These cluster-based phenotyping approaches almost invariably compare data as vectors of feature values. This allows features to belong to multiple phenotypes while also producing distinct patient membership, which can be helpful in certain analyses. However, an important consequence is additional post-processing overhead due to indistinct feature membership. For example, it can easily be the case that a binary feature (such as "has leukaemia") may be present in only 50% of the patient group. This hinders interpretation, particularly for categorical data types, and requires the oversight of a trained clinician.

2.4 Matrix Factorization

Matrix factorization is the decomposition of a matrix into component matrices that are of similar or lesser complexity. It became a popular unsupervised learning tool after it was used to win the 2012 Netflix prize challenge predicting individual movie preferences.[19]

Phenotyping problems typically employ a version of matrix factorization that imposes a strict non-negativity constraint (so-called "NMF"). For example, Joshi *et al.* used NMF in combination with the *bag-of-words* technique to derive comorbidity phenotypes from clinical notes.[20] However, most phenotyping applications of NMF appear to be in the realms of

genomics or natural language processing, with only a few, very recent applications focused on categorical data types.

An issue with NMF approaches is that the results can be difficult to interpret and to explain. The association values within the resultant matrices can be of an extreme range, and they do not directly translate to real world concepts. This again leads to indistinct feature membership that requires clinician review, with a rationale that is difficult to communicate.

2.5 Tensor Factorization

Tensors are an n-dimensional mathematical objects that are often represented by multidimensional arrays. Tensors are used to transform other n-dimensional mathematical objects (e.g., other multidimensional arrays), and can be thought of as the N-dimensional equivalent of linear transformations for matrices.[21][22]

Tensor factorization is of particular interest in phenotyping because it views the problem from an inherently different perspective, and because it became considerably faster in recent years.[22] Whereas the former approaches (clustering and matrix factorization) necessarily reduce all features to a single axis on a 2-D object, tensor approaches can treat each variable type as its own axis on an N-D object.[1] This has important mathematical consequences that should result in distinctive phenotype outcomes.

Marble, proposed by Ho *et al.*, is one of a series of tensor factorization-based approaches designed for categorical health data.[1] Marble imposes the aforementioned non-negativity constraint, and it decomposes the data to a bias tensor and an interaction tensor, where the bias tensor is essentially the phenotype of the entire population. The interaction tensor itself is composed of the sum of N components, each of which is a rank-1 object (set of vectors) constituting one phenotype. For an illustration from the paper, see Figure 2.1.

An issue with Marble is that it makes the assumption that any nonzero association value (any value positively linking a given feature to a phenotype) necessarily indicates membership in the phenotype. This can lead to tenuous and confusing relationships among the resultant features, again leading to phenotypes that require the oversight of a trained clinician.

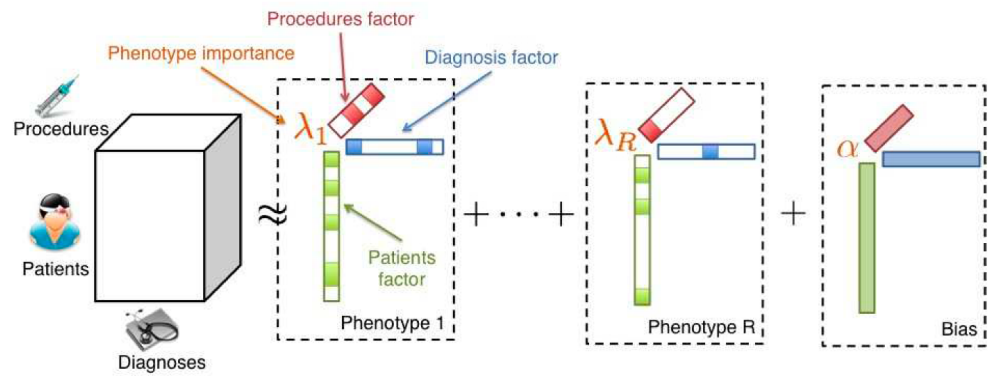


Figure 2.1: Illustration of Marble Approach from Ho *et al.* [1]

Chapter 3
PROPOSED METHOD

3.1 Problem Definition

We consider ideal phenotypes to be clinically relevant feature groupings that meet the following criteria:

1. **Distinct** – As pink petals and white petals are both distinct characteristics among groups of flowers, so to should the data phenotypes be distinct characteristics among groups of patients.

2. **Concise** – Feature membership should have minimal noise. One must be easily able to write a description of what the phenotype is expressing.

3. **Automated** – There should be as little need for physician validation as possible.

Solutions for predictive health problems such as the LOS problem employ a mix of potential data types - categorical, continuous, free-text, and otherwise unstructured. Eventually it will be useful to develop phenotypes that can encompass all variable types. However, optimal groupings of each data type are likely to require different machine learning approaches. For that reason, this investigation focused solely on categorical data types, with an eye toward eventual cohesion with other phenotyping approaches in the future.

3.2 Phenotyping Approaches

Three ideas were tested toward the goal of creating ideal phenotypes. The first idea was to transpose the feature matrix that is assessed by clustering algorithms. Rather than measuring similarity between vectors of **feature** values for each **patient**, similarity was

measured between vectors of **patient** values for each **feature**. Resultant groups were thus defined by sets of features rather than by sets of patients. K-means and agglomerative clustering were selected as representative clustering algorithms upon which to test this idea. Clusters from both approaches were generated via Python's scikit-learn package,[23] using Manhattan distance to measure the similarity between vectors. The agglomerative approach used average distance as the criterion by which to link clusters at each iteration. All features were one-hot encoded prior to processing, meaning that each diagnosis code was represented as a binary feature (e.g., "has leukaemia" vs "does not have leukaemia"). To differentiate these test versions from their standard versions, they will be referred to as "K-Means(f)" and "Agg(f)".

The second idea was to impose a strict constraint upon the results of matrix and tensor factorization such that each feature was assigned to only the one phenotype with which it held the strongest association value (see also Figure 3.1). Again, resultant phenotypes should have explicit feature membership that is mutually exclusive[1]. NMF-based phenotypes were generated using Python's scikit-learn package with default parameters, and again all features were one-hot encoded prior to processing. Marble-based phenotypes were generated using an adapted version of the epic_phenotyper package published by Robert Chen[2].[24] Results for these restricted versions of NMF and Marble are designated as "NMF(r)" and "Marble(r)".

Finally, for all approaches, patients and phenotypes were linked using an Expressivity Score, defined as one minus the ratio of the Manhattan distance to the phenotype and the maximum possible Manhattan distance to the phenotype, shown in Equation 3.1. For one-hot-encoded categories this is equivalent to the fraction of features present in the patient out of all possible features that are members of the phenotype.

$$1 - Manh(\overline{patient}, \overline{phenotype})/Manh_{max}(\overline{phenotype}) \tag{3.1}$$

[1]In the rare case of a tie the feature was assigned to all phenotypes with the maximum association value.

[2]The epic_phenotyper adaptations addressed only workflow and compatability with Python 3. No changes were made to any core components that might affect the results.

	Feature A	Feature B	Feature C	Feature D	Feature E	Feature F	Feature G
Phenotype 1	0.97	0.64	1.3	0.0074	1.23	0.75	0.23
Phenotype 2	0.6	0.0003	0	0	0.1	0.8	0.9
Phenotype 3	0.00087	0.34	1.3	0.00137	2.37	0.6	0.13
Phenotype 4	1.84	0.97	2.6	0.00274	0	0.3	0.37

In this example, Feature F and Feature G are assigned to only Phenotype 2.

Figure 3.1: NMF Feature Assignment with Constraint

Chapter 4

EXPERIMENTS

4.1 Data

MIMIC-III is a large and freely-available, de-identified database of critical care encounters at the Beth Israel Deaconess Medical Center in Boston, years 2001 through 2012.[25][5] From this database, we created two data subsets on which to test the proposed approaches. The first, "Neonatal" subset is the subset of all encounters of patients who were transferred to a neonatal Intensive Care Unit (ICU). The second, "65+" subset is that of all encounters for patients aged 65 and above. Diagnoses and procedures are encoded from the 9^{th} edition of the ICD, hence referred to as "ICD-9".[26] ICD-9 code distributions for both subsets are listed in Table 4.1. No additional feature engineering nor outlier removal were applied to the data.

Subset	Definition	Encounters	ICD-9 Diagnoses	ICD-9 Procedures
Neonatal	Neonatal ICU	8101	927	164
65+	Age $\geq$ 65	26074	4889	1581

Table 4.1: MIMIC-III Data Subsets

Prediction performance across approaches was assessed through a simple binary LOS prediction. Arbitrary cutoffs that facilitated simple analysis were selected for each dataset, as shown in Table 4.2.

Problem (Subset)	Target	Distribution
Neonatal	LOS $>$ 4 days	P(+) = 0.53
65+	LOS $\geq$ Mean	P(+) = 0.33

Table 4.2: MIMIC-III Target Distributions for ICU Length of Stay

Diagnosis features were sparse in both datasets. Only 13 of 909 (1.4%) ICD-9 diagnoses were present in at least 10% of patients in the Neonatal subset, and only 16 (0.3%) of diagnoses are at least that common in the 65+ subset. The majority, 71.3% of the Neonatal encounters involved a "Need for prophylactic vaccination and inoculation against viral hepatitis", and 68.0% were associated with "Observation for suspected infectious condition" (for more information, see A.1). Approximately 78% of Neonatal encounters involved a single liveborn patient, delivered with or without mention of cesarean section. Encounters in the 65+ subset were most commonly associated with diseases of the circulatory system, specifically "Unspecified essential hypertension" (49.03%), "Atrial fibrilation" (36.79%), "Congestive heart failure, unspecified" (33.72%), and "Coronary atherosclerosis of native coronary artery" (32.30%). See Table A.2 for more information.

Data were similarly sparse even when encoded in CCS Level 2 Categories. Only 6% of CCS codes in the Neontatal dataset and 5% in the 65+ subset were present in at least 10% of the population. Approximately 96% of the Neonatal subset is marked as "Liveborn" (Table A.3). Other common CCS diagnoses for the Neonatal subset included "Immunizations and screening for infectious disease" (71.58%), "Factors influencing health care" (71.24%), and "Other perinatal conditions" (64.88%). Among encounters in the 65+ subset, the most common CCS category was "Diseases of the heart," followed by "Hypertension," and "Diseases of the urinary sytstem" (Table A.3).

4.2 Setup

4.2.1 Phenotyping Workflow

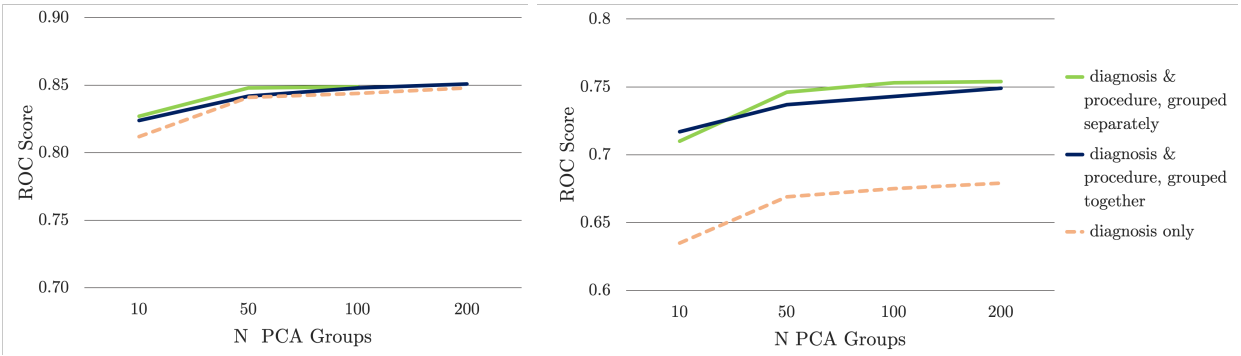
To streamline the comparison process, we developed a config-driven phenotyping pipeline tool, dubbed "ken_pheno". Ken_pheno allows the modular implementation of phenotyping approaches with standardized results. All code is written in Python using freely available, standard libraries. Note that the unsupervised learning approaches used to generate phenotypes used default parameters except where specified above.

4.2.2 Inclusion of Procedure Codes

Approaches were tested both with and without the inclusion of ICD-9 clinical procedure codes (in combination with the ICD-9 diagnosis codes). Phenotypes were generated by grouping both diagnosis and procedure together, as well as by generating separate groups ("diagnosis phenotypes" and "procedure phenotypes"). As shown in Figure 4.1a, for the Neonatal LOS problem no significant difference in predictive power was observed with the inclusion of ICD-9 procedure codes, regardless of how variables were combined. In contrast, there was a significant difference in predictive power for the 65+ LOS problem (Figure 4.1b), where the inclusion of procedure codes led to significantly higher ROC AUC. Notably, the method of combination appeared to make no significant difference. However, the trend in performance relative to the number of phenotypes was similar among all phenotype sets, regardless of the inclusion of procedure codes. Thus only diagnosis codes were used in order to simplify analysis, since results could be considered representative of the **relative** difference among the phenotyping approaches.

4.2.3 Prediction Setup

Predictions were generated via Python's XGBOOST library, again using default parameters.[27] Predictions were made purposefully naive to focus on the relative performance of each approach. No parameter tuning, feature engineering, nor outlier removal were performed, and



(a) Neonatal Results

(b) 65+ Results

Figure 4.1: ROC AUC with vs. without Procedure Codes

no other classifiers were tested. Performance metrics were averaged over 10 trials of 80-20 holdout validation. Predictions were run at the encounter level, and no constraints were imposed to prevent the same patient from appearing on both sides of the train-test split. Validation metrics included accuracy, precision, recall, and Area Under the Receiver Operating Characteristic Curve ("ROC AUC").

Feature importance was compared using weights from the Local Interpretable Model-agnostic Explanations (LIME) package.[28] Global importance was assessed as the aggregate mean of all prediction instances across all 10 trials. These mean values were then validated against SHapley Additive exPlanations (SHAP) values and mean XGBOOST weights, again averaged across the aggregate of all 10 trials.[29]

4.3 Prediction Performance

Length of Stay prediction results using the proposed approaches were compared to those using three baselines. The first was the full set of ICD-9 diagnosis codes, one-hot-encoded and designated as ICD-9(b). More will be said about the label "ICD-9" in Section 4.1. The second baseline, "CCS(b)", consisted of one-hot-encoded CCS Level 2 Categories. Finally, the one-hot-encoded ICD-9 codes were reduced to N dimensions via Principal Component

Analysis (PCA), again using scikit-learn and default settings. Prediction results using the standard NMF and Marble approaches were also generated for comparison.

For both the neonatal and the 65+ prediction problem, there was asymptotically increasing and converging performance as the data were divided into a larger number of phenotypes, demonstrated by the trend in ROC AUC (Figure 4.2).

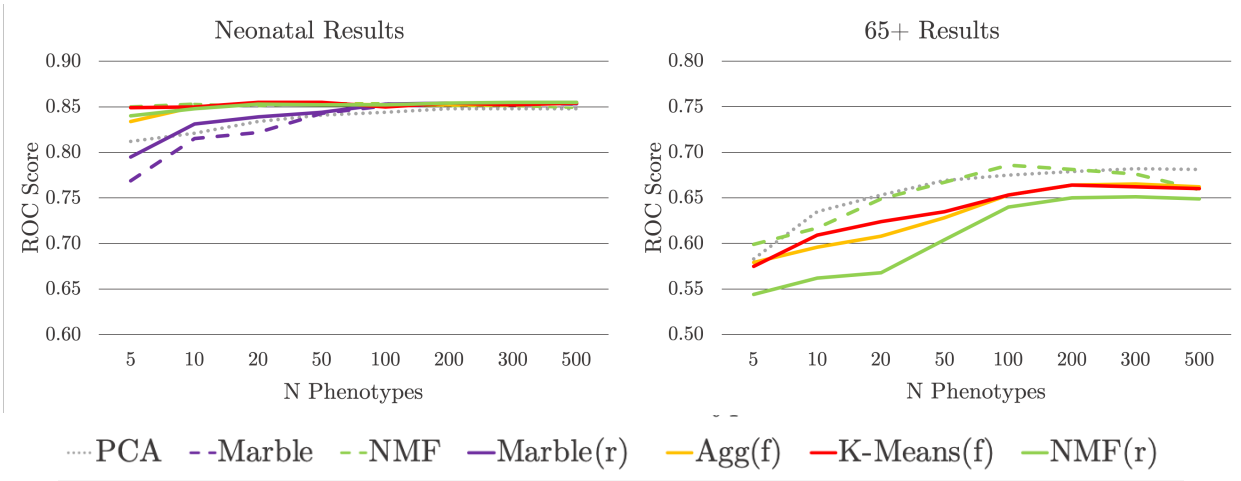


Figure 4.2: ROC AUC Relative to the Number of Phenotypes

Most approaches yielded similar performance across all four metrics, as shown in Tables 4.3a, 4.3b, and 4.3c. All approaches, including the ICD-9 and CCS baselines, had greater accuracy than would be expected from random sampling.

For the Neonatal LOS problem (Tables 4.3a and 4.3b), the CCS baseline yielded better precision than all of the other methods, yet was the worst performer by a significant margin for all other performance metrics. This appears to be true regardless of the number of phenotypes. All other approaches yielded comparable accuracy, precision, and recall - again, apparently irrespective of the number of phenotypes.

For the 65+ LOS problem (Table 4.3c), all approaches, including the CCS baseline, yielded similar accuracy. ROC AUC was also very similar across methods. And although the ICD-9 baseline and K-means phenotype predictions achieved the highest precision, precision

scores again were similar among most methods. However, there is a noteworthy spread in recall scores, the lowest of which was achieved by the ICD-9 baseline, and the highest by PCA. The standard NMF approach yielded an average recall score that was competitive with PCA, and produced the highest scores in accuracy and ROC AUC (although the spread for these metrics is small).

Prediction Scores

	ICD-9(b)	CCS(b)	PCA	K-Means(f)	Agg(f)	NMF	NMF(r)	Marble	Marble(r)
Accuracy	0.8565	0.7944	0.8380	0.8586	0.8575	0.8552	0.8565	0.8275	0.8433
Precision	0.8280	0.8555	0.8091	0.8298	0.8298	0.8247	0.8337	0.7913	0.8119
Recall	0.9198	0.7357	0.9081	0.9219	0.9193	0.9224	0.9104	0.9155	0.9161
ROC AUC	0.8526	0.7980	0.8336	0.8547	0.8537	0.8510	0.8532	0.8221	0.8388

(a) Neonatal Results for 20 Phenotypes

	ICD-9(b)	CCS(b)	PCA	K-Means(f)	Agg(f)	NMF	NMF(r)	Marble	Marble(r)
Accuracy	0.8565	0.7944	0.8529	0.8579	0.8564	0.8554	0.8572	0.8559	0.8574
Precision	0.8280	0.8555	0.8189	0.8302	0.8275	0.8222	0.8321	0.8290	0.8309
Recall	0.9198	0.7357	0.9272	0.9195	0.9203	0.9274	0.9147	0.9168	0.9172
ROC AUC	0.8526	0.7980	0.8483	0.8541	0.8524	0.8510	0.8536	0.8522	0.8537

(b) Neonatal Results for 200 Phenotypes

	ICD-9(b)	CCS(b)	PCA	K-Mans(f)	Agg(f)	NMF	NMF(r)	Marble	Marble(r)
Accuracy	0.7499	0.7490	0.7563	0.7564	0.7564	0.7640	0.7488	0.7480	0.7482
Precision	0.7543	0.6980	0.6999	0.7502	0.7487	0.7390	0.7428	0.7332	0.7349
Recall	0.3535	0.4153	0.4514	0.3871	0.3887	0.4370	0.3593	0.3654	0.3645
ROC AUC	0.6486	0.6637	0.6784	0.6620	0.6624	0.6810	0.6492	0.6502	0.6501

(c) 65+ Results for 200 Phenotypes

Table 4.3: Length of Stay Prediction Model Metrics

4.4 Feature Membership of Resultant Phenotypes

Feature membership was similar across most phenotyping approaches (shown in Figure B.1), with the expected exception of Marble. This was due in part to skewed feature distributions in the Agg(f), K-Means(f), and NMF(r) approaches, examples of which are shown in Tables 4.4a and 4.4b, and as explained in greater detail in the Discussion. Notably, agglomerative clustering generated phenotypes either with only one feature or with greater than 890 features for the Neonatal subset, and phenotypes for the 65+ subset had a similar distribution.

Distribution of Feature Counts for 20 Phenotypes

N Features	Agg(f)	K-Means(f)	NMF(r)	Marble(r)
1	19	17	15	-
2-4	-	1	4	-
5-9	-	1	1	-
10-19	-	-	-	2
20-29	-	-	-	4
30-39	-	-	-	4
40+	1	1	-	10

(a) Neonatal Results

N Features	Agg(f)	K-Means(f)	NMF(r)	Marble(r)
1	196	197	182	3
2-4	3	2	18	21
5-9	-	1	1	21
10-19	-	-	-	23
20-29	-	-	-	45
30-39	-	-	-	34
40+	1	1	-	46

(b) 65+ Results

Table 4.4: Distribution of Feature Counts

4.5 Feature Importance in Prediction Problems

Among predictions made with non-Marble phenotypes, those phenotypes with the largest absolute mean of LIME weights tended to have similar feature membership (Figure 4.3). In addition, the diagnoses that were members of these "highest weighted" phenotypes tended to be among the highest weighted diagnoses in the ICD9-based prediction problem. In contrast, the highest weighted features in the CCS baseline prediction had little overlap with the CCS categories of the highest weighted diagnoses in the ICD-9 baseline prediction (see Figures 4.4 and 4.5).

The ten highest ranked phenotypes using the mean of true positive LIME weights across ten trials is similar to the highest ranked phenotypes using the mean of true positive SHAP values (Table B.2). Both sets of ranks are similar to global ranks as determined by XGBOOST weights, with the notable exception of one phenotype with the exceptionally large feature membership.

Top Features in True Positive Predictions

Rank	ICD-9	K-Means(f)	Agg(f)	NMF(r)
1	- Neonatal jaundice associated with preterm delivery	- Neonatal jaundice associated with preterm delivery	- Neonatal jaundice associated with preterm delivery	- Neonatal jaundice associated with preterm delivery
2	- Single liveborn, born in hospital, delivered w/o mention of cesarean	- 31-32 completed weeks of gestation - Neonatal bradycardia - Primary apnea of newborn	- Single liveborn, born in hospital, delivered w/o mention of cesarean	- Neonatal bradycardia - Primary apnea of newborn
3	- Disorder of stomach function and feeding problems in newborn	- Single liveborn, born in hospital, delivered w/o mention of cesarean	- Primary apnea of newborn	- Disorder of stomach function and feeding problems in newborn
4	- Primary apnea of newborn	- Disorder of stomach function and feeding problems in newborn	- Disorder of stomach function and feeding problems in newborn	- Single liveborn, born in hospital, delivered w/o mention o cesarean
5	- Congenital pneumonia	- 33-34 completed weeks of gestation	- (Too many features to list)	- Anemia of prematurity - Chronic respiratory disease arising in the perinatal period - Other specified conditions originating in the perinatal period - Patent ductus arteriosus - Retrolental fibroplasia - Septicemia [sepsis] of newborn

Figure 4.3: Top Five Features Ranked by Relative Importance in True Positive Prediction

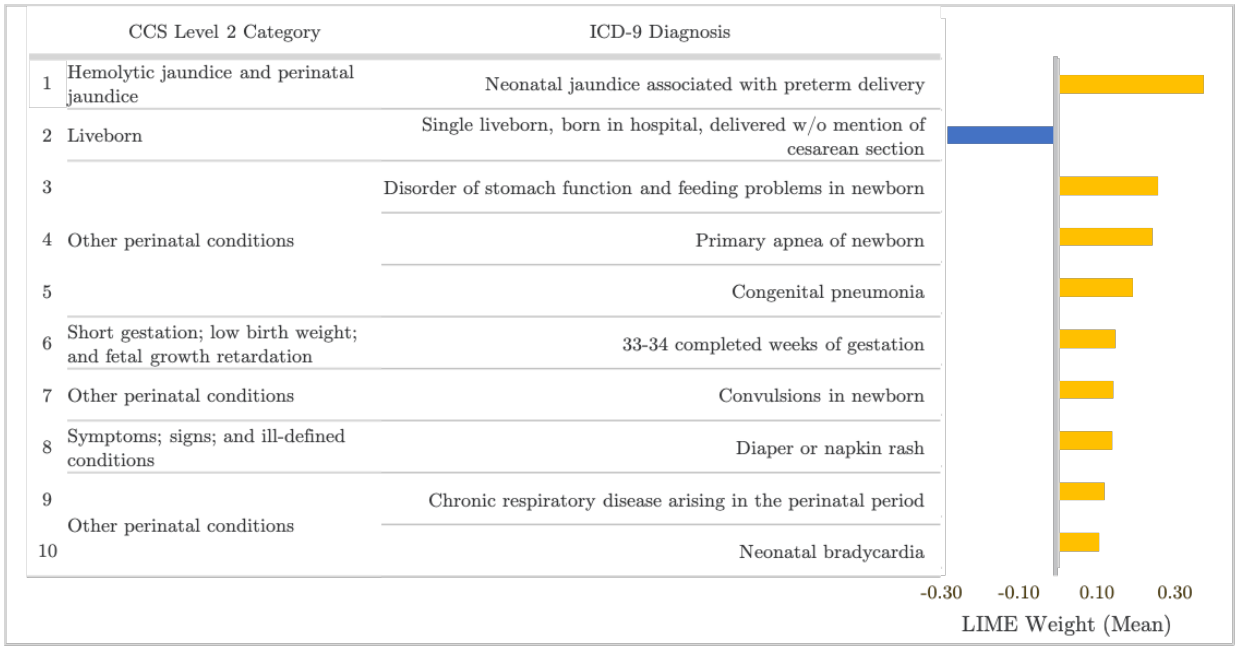


Figure 4.4: Top Ten ICD-9 Prediction Features Among to True Positive Results, Ranked by Mean LIME Weight and Matched to CCS Level 2 Categories

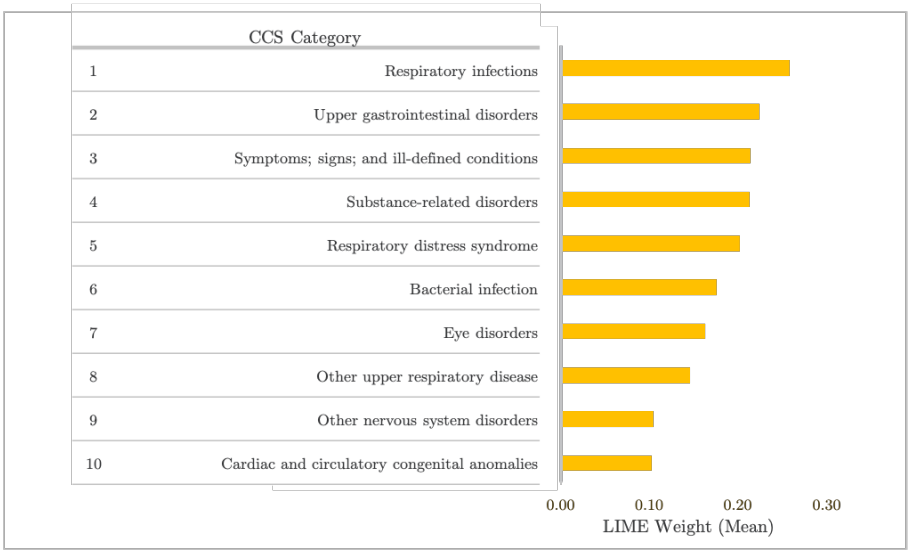


Figure 4.5: Top Ten CCS Level 2 Prediction Features Among to True Positive Results, Ranked by Mean LIME Weight

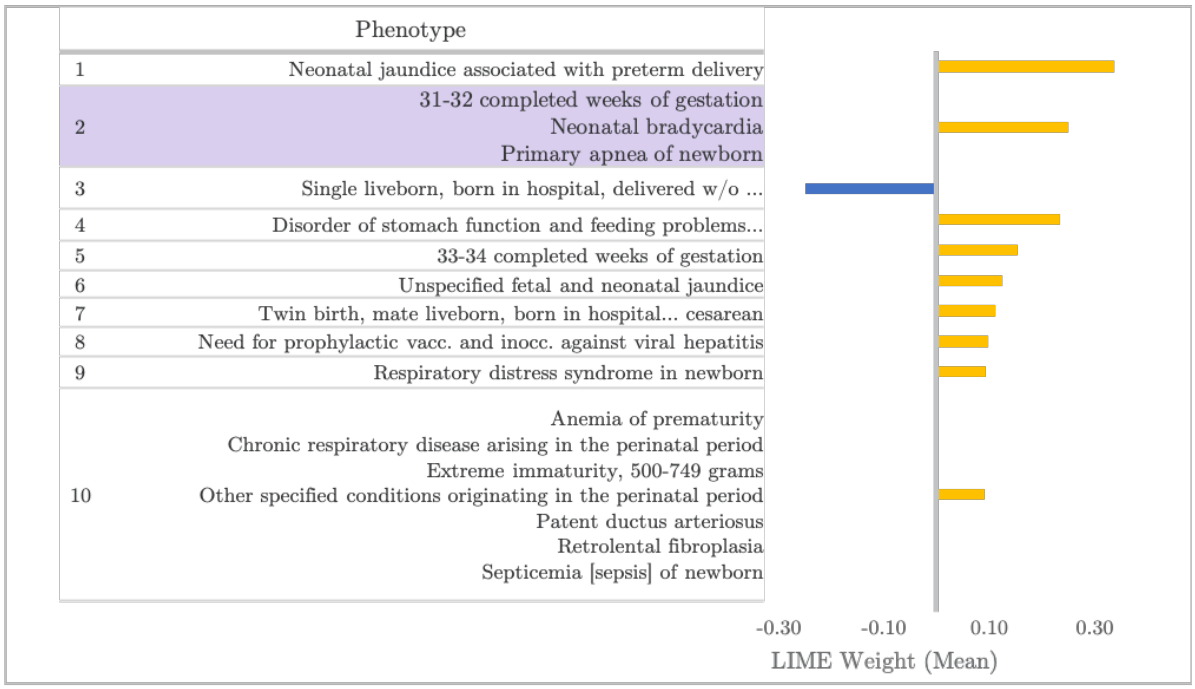


Figure 4.6: Top Ten K-Means(f) Phenotype Prediction Features Among to True Positive Results, Ranked by Mean LIME Weight

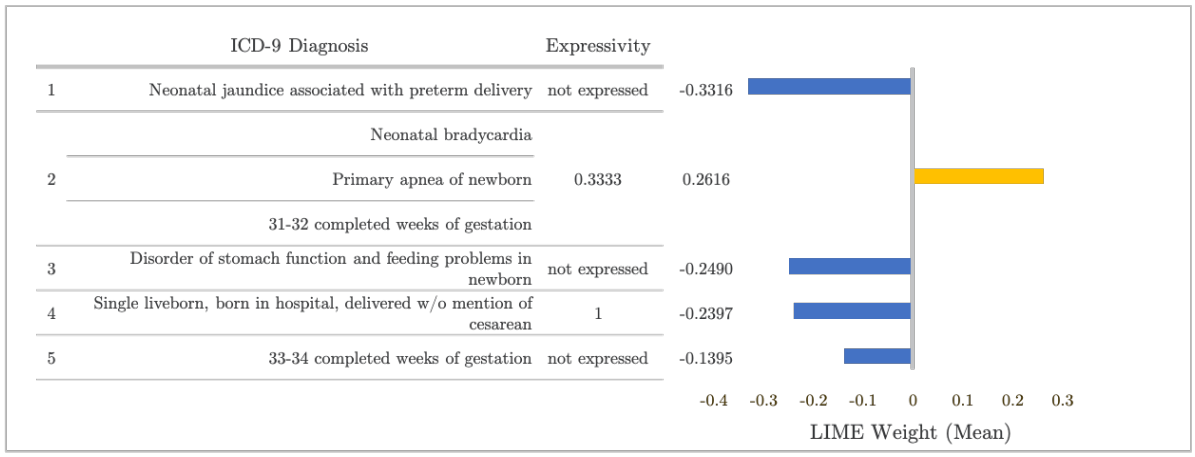


Figure 4.7: LIME Feature Importance for Example False Negative Prediction, K-Means(f) Phenotype-based Prediction

4.6 Computation Time

The Marble approach required significantly greater computation time than did the other approaches, as demonstrated in Figure 4.8. Faster versions of this tensor factorization technique exist, however known examples are not written in Python.[30][31] Because of this issue, Marble(r) results were ignored in analysis of feature importance.

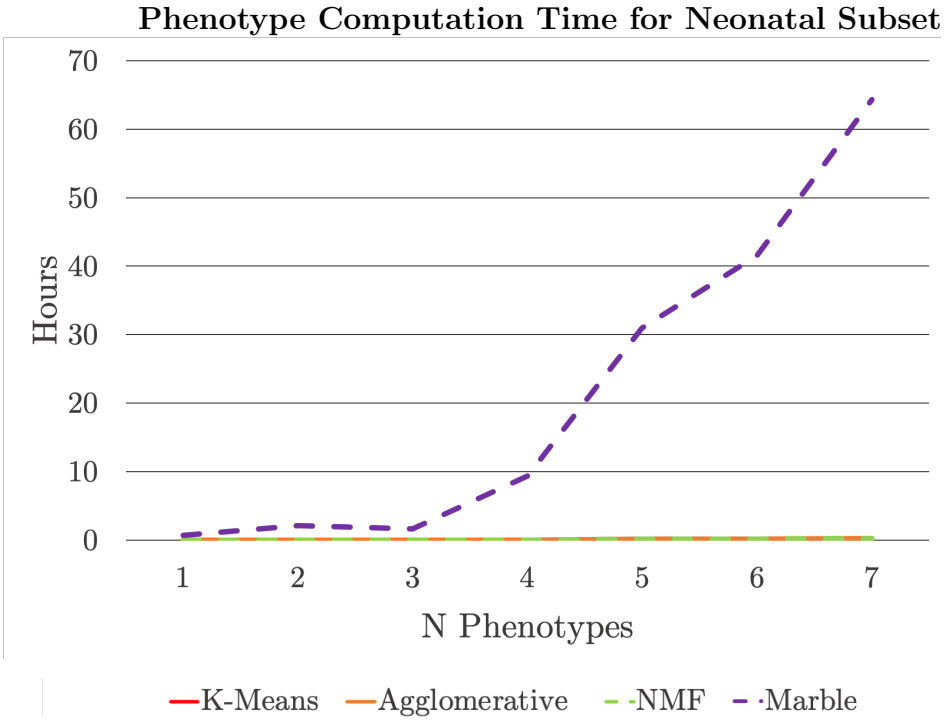


Figure 4.8: Computation Time per Number of Phenotypes

4.7 Discussion

4.7.1 Observations

The comparable prediction performance demonstrated in this work indicates that phenotypes generated by these approaches may be used with no detriment to prediction performance. Table 4.3 shows that with the one exception of recall in the 65+ LOS prediction, by all four

metrics — accuracy, precision, recall, and ROC AUC — all four proposed methods achieve performance that is either comparable to or better than all three baselines. Moreover, because the data subsets used are relatively small, sparse and one-hot encoded, the discriminative burden on the XGBOOST classifier may be relatively low. Hence it seems likely that greater gains in predictive performance are likely to be achieved from these approaches for larger, more complex datasets.

In addition, the proposed K-Means(f) and NMF(r) models were able to capture more information verses the ICD-9 baseline. An example of this is "Phenotype 2", shown in the shaded box of Figure 4.6, which includes both "Neonatal bradycardia" and "Primary apnea of the newborn". Bradycardia means that the heart rate is extremely slow, and primary apnea means a cessation of breathing due to a lack of oxygen.[32] Because bradycardia commonly follows apnea it is no surprise to see these two diagnoses grouped together in the same phenotype. Such a simple connection is easily made with a quick search online, and thus does not require the experience of a clinician. Notably, these two diagnoses appear among the ten most important prediction features both in the ICD-9 baseline (Figure 4.4) and in the K-Means(f) and NMF(r) predictions (Figure 4.6).

Furthermore, K-Means(f) and NMF(r) presented grouped information with greater specificity than did the CCS(b) method. For instance, both diagnoses presented in the previous example, "Neonatal bradycardia" and "Primary apnea of the newborn", are included in the same Level 2 CCS Category, "Other perinatal conditions" (shown in Figure 4.4). Yet, in the Neonatal CCS(b) prediction, "Other perinatal conditions" does not even appear among the top ten most influential categories for true positive prediction (Figure 4.5). Moreover, "Other perinatal conditions" is associated with 127 distinct diagnoses in this dataset, which is far too many for a machine learning specialist to evaluate easily. In contrast, in addition to the prediction importance of "Phenotype 2" as explained in the previous paragraph, the K-Means(f) and NMF(r) phenotypes containing these two diagnoses had only 2 or 3 diagnoses total, leading to far simpler evaluation of results.

However, as demonstrated by the example instance of a false negative prediction (Figure

4.7), the phenotyping results are still not capturing some relevant patterns for this particular LOS prediction[1]. Because the patient in this example expressed only one among the most influential, positively weighted phenotypes (and only partly), the prediction model was unable to distinguish the encounter as a positive event. Predictions for this encounter were incorrect for both the K-Means(f) model and the ICD-9 baseline model; so, additional feature engineering is likely required to improve upon either model.

4.7.2 Limitations

A number of simplifying assumptions have been employed in this investigation. For one, although a focus on only diagnostic information facilitated interpretation of the results, the nuance of many conditions will be lost by grouping diagnosis codes alone. Similarly, the strict limitation to phenotypes with mutually exclusive feature sets may hinder optimal assignment, since common diagnoses are likely to be associated with multiple underlying patterns.

Performance on these particular data subsets would likely have been improved with the use of parameter tuning on either or both of the XGBOOST classifier and the phenotyping approaches. In addition, it seems likely that the size and sparsity of the data contributed to the observed skew in feature membership (Tables 4.4a and 4.4b). It's possible that both the performance and the feature distributions in this experiment could be further improved through additional data cleaning, particularly outlier identification and management.

[1]This assumes that the LIME predicted feature importance for this model and instance is correct, which it may not be.

Chapter 5

CONCLUSION

5.1 Summary of the Thesis

This work discussed the background of predictive health analytics in terms of the complexity and volume of data, as well as the inherent need for explainable results. It then reviewed some of the unsupervised learning methods currently used for data phenotyping, namely clustering and factorization. It then considered alternate forms of these methods, and proposed an Expressivity Score to associate patients with resultant phenotypes. Finally, it showed that the K-Means(f) and NMF(r) phenotyping methods produced at least some clinically relevant groups that were easy to interpret, and that performed at least comparably to predictions using manually-generated methods.

5.2 Future Directions

As this research continues, it will be useful to test additional unsupervised approaches, such as co-clustering and fuzzy clustering. Rather than comparing either patient vectors or feature vectors, co-clustering compares both sets of vectors simultaneously, which may produce better groupings. Alternatively, fuzzy constraints can be applied to clustering algorithms like k-means to allow non-exclusive feature membership.[13] This would enable common diagnoses such as diabetes to be associated with multiple phenotypes, for example.

Moreover, the addition of informed parameter tuning, or even a grid search, to the phenotype creation pipeline may be useful for improving the distribution of feature membership. Finally, it has been shown above that variables may be phenotyped separately potentially without loss of predictive performance. Given that, the Expressivity Score could be expanded for use with alternative phenotyping strategies that are more appropriate for contin-

uous values, such that the same scoring mechanism might be applied to combine disparate approaches.

BIBLIOGRAPHY

[1] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124, 2014.

[2] Healthcare Cost, Agency for Healthcare Research Utilization Project (HCUP), and Quality. Hcupnet. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp. Online; accessed 6-June-2020.

[3] Juan M Banda, Martin Seneviratne, Tina Hernandez-Boussard, and Nigam H Shah. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annual review of biomedical data science*, 1:53–68, 2018.

[4] Pierre-Régis Burgel, Jean-Louis Paillasseur, and Nicolas Roche. Identification of clinical phenotypes using cluster analyses in copd patients with multiple comorbidities. *BioMed research international*, 2014, 2014.

[5] T.J. Pollard and A.E.W. Johnson. The MIMIC-III clinical database. `http://dx.doi.org/10.13026/C2XW26`, 2016.

[6] Prakash J Mathew, Faisal Jehan, Narong Kulvatunyou, Muhammad Khan, Terence O'Keeffe, Andrew Tang, Lynn Gries, Mohammad Hamidi, El-Rasheid Zakaria, and Bellal Joseph. The burden of excess length of stay in trauma patients. *The American Journal of Surgery*, 216(5):881–885, 2018.

[7] BH Tess, HM Glenister, LC Rodrigues, and MB Wagner. Incidence of hospital-acquired

infection and length of hospital stay. *European Journal of Clinical Microbiology and Infectious Diseases*, 12(2):81–86, 1993.

[8] Yang Xie, Günter Schreier, David CW Chang, Sandra Neubauer, Ying Liu, Stephen J Redmond, and Nigel H Lovell. Predicting days in hospital using health insurance claims. *IEEE journal of biomedical and health informatics*, 19(4):1224–1233, 2015.

[9] Jacqueline C Kirby, Peter Speltz, Luke V Rasmussen, Melissa Basford, Omri Gottesman, Peggy L Peissig, Jennifer A Pacheco, Gerard Tromp, Jyotishman Pathak, David S Carrell, et al. Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association*, 23(6):1046–1052, 2016.

[10] Sarah A Pendergrass and Dana C Crawford. Using electronic health records to generate phenotypes for research. *Current protocols in human genetics*, 100(1):e80, 2019.

[11] World Health Organization et al. Classifications: International classification of diseases (icd).

[12] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[13] Rui Xu and Don Wunsch. *Clustering*, volume 10. John Wiley & Sons, 2008.

[14] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[15] John Torous, Patrick Staples, Ian Barnett, Luis R Sandoval, Matcheri Keshavan, and Jukka-Pekka Onnela. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *NPJ digital medicine*, 1(1): 1–9, 2018.

[16] Marina Guisado-Clavero, Albert Roso-Llorach, Tomàs López-Jimenez, Mariona Pons-Vigués, Quintí Foguet-Boreu, Miguel Angel Muñoz, and Concepción Violán. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. *BMC geriatrics*, 18(1):16, 2018.

[17] John E Cornell, Jacqueline A Pugh, John W Williams Jr, Lewis Kazis, Austin FS Lee, Michael L Parchman, John Zeber, Thomas Pederson, Kelly A Montgomery, and Polly Hitchcock Noël. Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database. *Applied multivariate research*, 12(3):163–182, 2008.

[18] Minlei Liao, Yunfeng Li, Farid Kianifard, Engels Obi, and Stephen Arcona. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC nephrology*, 17(1):25, 2016.

[19] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[20] Shalmali Joshi, Suriya Gunasekar, David Sontag, and Joydeep Ghosh. Identifiable phenotyping using constrained non-negative matrix factorization. *arXiv preprint arXiv:1608.00704*, 2016.

[21] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[22] Thomas Papastergiou, Evangelia I Zacharaki, and Vasileios Megalooikonomou. Tensor decomposition for multiple-instance classification of high-order medical data. *Complexity*, 2018, 2018.

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent

Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[24] R Chen. Epic phenotyper. https://bitbucket.org/rchenmit/epic_phenotyper, 2016.

[25] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[26] World Health Organization et al. International classification of diseases: 9th revision, basic tabulation list with alphabetic index. 1978.

[27] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[29] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

[30] Ioakeim Perros, Evangelos E Papalexakis, Haesun Park, Richard Vuduc, Xiaowei Yan, Christopher Defilippi, Walter F Stewart, and Jimeng 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, pages 2080–2089, 2018.

[31] Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. Spartan: Scalable parafac2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 375–384, 2017.

[32] Praveen Chandrasekharan, Munmun Rawat, Anne Marie Reynolds, Kathryn Phillips, and Satyan Lakshminrusimha. Apnea, bradycardia and desaturation spells in premature infants: impact of a protocol for the duration of spell-freeobservation on interprovider variability and readmission rates. *Journal of Perinatology*, 38(1):86–91, 2018.

Appendices

Appendix 1

DISTRIBUTIONS OF ICD-9 DIAGNOSES

ICD-9 Diagnosis	Subset Prevalence
Need for prophylactic vaccination and inoculation against viral hepatitis	71.30%
Observation for suspected infectious condition	68.09%
Single liveborn, born in hospital, delivered without mention of cesarean section	44.01%
Single liveborn, born in hospital, delivered by cesarean section	34.00%
Neonatal jaundice associated with preterm delivery	27.95%
Routine or ritual circumcision	24.89%
Respiratory distress syndrome in newborn	16.20%
Disorder of stomach function and feeding problems in newborn	13.18%
Primary apnea of newborn	12.87%
Twin birth, mate liveborn, born in hospital, delivered by cesarean section	12.31%

Table A.1: Most Common Diagnoses, Neonatal subset

ICD-9 Diagnosis	Subset Prevalence
Unspecified essential hypertension	49.03%
Atrial fibrillation	36.79%
Congestive heart failure, unspecified	33.72%
Coronary atherosclerosis of native coronary artery	32.30%
Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled	23.03%
Other and unspecified hyperlipidemia	22.13%
Acute kidney failure, unspecified	20.33%
Urinary tract infection, site not specified	15.90%
Acute respiratory failure	15.28%
Pure hypercholesterolemia	14.87%

Table A.2: Most Common Diagnoses, 65+ Subset

CCS Level 2 Category	Subset Prevalence
Liveborn	95.78%
Immunizations and screening for infectious disease	71.58%
Factors influencing health care	71.24%
Other perinatal conditions	64.88%
Short gestation; low birth weight; and fetal growth retardation	42.97%
Hemolytic jaundice and perinatal jaundice	36.90%
Respiratory distress syndrome	16.20%
Cardiac and circulatory congenital anomalies	10.99%

Table A.3: Most Common CCS Categories, Neonatal Subset

CCS Level 2 Category	Subset Prevalence
Diseases of the heart	78.10%
Hypertension	66.31%
Diseases of the urinary system	48.38%
Disorders of lipid metabolism [53.]	36.70%
Complications	33.45%
Anemia	32.26%
Fluid and electrolyte disorders [55.]	30.32%
Respiratory failure; insufficiency; arrest (adult)	24.43%
Diabetes mellitus without complication [49.]	24.29%
Diseases of arteries; arterioles; and capillaries	23.17%

Table A.4: Most Common CCS Categories, 65+ Subset

Appendix 2

ADDITIONAL PHENOTYPE AND LOS PREDICTION RESULTS

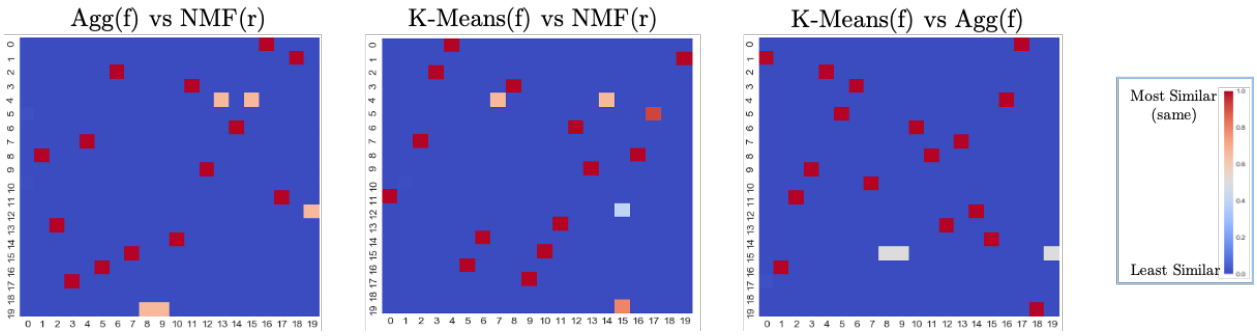


Figure B.1: Feature Membership of Phenotyping Approaches

Example Phenotypes for Neonatal Subset

Phenotype	ICD-9 Diagnosis	CCS Level 2 Category
Example 1	Neonatal bradycardia Primary apnea of newborn	Other perinatal conditions
	31-32 completed weeks of gestation	Short gestation; low birth weight; and fetal growth retardation
Example 2	Septicemia [sepsis] of newborn	Bacterial infection
	Patent ductus arteriosus	Cardiac and circulatory congenital anomalies
	Retrolental fibroplasia	Eye disorders
	Anemia of prematurity Chronic respiratory disease arising in the perinatal period Other specified conditions originating in the perinatal period	Other perinatal conditions
	Extreme immaturity, 500-749 grams	Short gestation; low birth weight; and fetal growth retardation

Table B.1: Comparison of Feature Membership, CCS vs. K-Means(f) Phenotypes

Relative Rank of K-Means(f) Phenotypes for the Neonatal Prediction Problem

(20 Phenotypes)

Phenotype Label	LIME Rank (True Positives)	SHAP Rank (True Positives)	XGBOOST Rank (All Predictions)	K-Means(f) Feature Count
p_5	1	2	4	1
p_3	2	1	9	1
p_15	3	3	2	**3**
p_0	4	4	7	1
p_13	5	7	12	1
p_10	6	12	8	1
p_16	7	9	10	1
p_9	8	11	3	1
p_4	9	8	6	1
p_17	10	10	14	**7**
p_2	11	5	5	1
p_12	12	13	17	1
p_1	13	6	1	**882**
p_18	14	20	20	1
p_11	15	16	13	1
p_19	16	15	18	1
p_8	17	14	16	1
p_6	18	17	11	1
p_14	19	19	19	1
p_7	20	18	15	1

Table B.2: Comparison of Feature Importance Among LIME, SHAP, and XGBOOST Metrics