Interpretable Phenotyping for Electronic Health Records

Christine Allen*, Juhua Hu†, Vikas Kumar*, Muhammad Aurangzeb Ahmad*, Ankur Teredesai†
* *KenSci Inc., Seattle, WA, USA*
† *Center for Data Science, School of Engineering and Technology, University of Washington, Tacoma, WA, USA*
christine.allen@kensci.com, juhuah@uw.edu, vikas@kensci.com, maahmad@uw.edu, ankurt@uw.edu

***Abstract*—Datasets from Electronic Health Records (EHRs) are increasingly large and complex, creating challenges in their use for predictive modeling. The two major challenges are large-scale and high-dimensionality. One of the common way to address the large-scale challenge is through use of *data phenotypes*: clinically relevant characteristic groupings that can be expressed as logical queries (e.g., "senior patients with diabetes"). With the increasing use of machine learning across the continuum of care, phenotypes play an important role in modeling for population management, clinical trials, observational and interventional research, and quality measures. Yet, phenotype interpretation can often be difficult and require post-hoc clarifications from experienced clinicians. For example, detailed analysis may be needed to find that all patients in a a phenotype are diabetic seniors with complications from previous surgery. Moreover, the high-dimensionality problem is often addressed either separately or simultaneously with phenotyping by dimension reduction methods that may further hinder interpretability. In this paper, we introduce the notion of interpretable data phenotypes generated by an unsupervised learning technique. Methods are designed to disambiguate relative feature memberships, thus facilitating general clinical validation, and alleviating the problem of high-dimensionality. The empirical study applies the proposed unsupervised interpretable phenotyping method to a real world healthcare dataset (MIMIC), then uses hospital length of stay as a reference prediction task. The results demonstrate that the proposed method produces phenotypes with improved interpretability and without diminishing the quality of prediction results.**

***Index Terms*—EHRs, High-Dimensionality, Data Phenotyping, Unsupervised Learning, Interpretable Phenotyping.**

I. Introduction

Machine Learning (ML) problems in the healthcare space are often impeded by the curse of dimensionality: there can be tens of thousands of potentially influential features for a given prediction, many or most of which are largely missing. In collaboration with medical experts, standard feature engineering methods are typically able to identify the salient features for a given problem. However, through that approach it is easy to eliminate sparse features for which certain values happen to be critical for evaluating special cases, and particularly so when those cases involve complex interactions which are not yet well understood. Moreover, reliance on manual feature curation and medical experience can slow model development, particularly for outsourced analyses or other high-throughput model types. For these reasons, unsupervised ML has become an increasingly popular solution to automatically extract clinically relevant data patterns to enhance predictions.

However, predictive healthcare problems often require model transparency and interpretability. Let length of stay in hospital (LOS) prediction serve as one example of such an application. If a prolonged LOS is predicted, it may indicate that a patient is likely to suffer from complications due to existing comorbidities, which would require medical intervention. On the other hand, it may otherwise indicate operational delays due to either insufficient staffing or weekend delays [1]–[3], which might conversely require operational changes, or possibly no intervention at all. Hence, transparency (i.e., the linkage of individual predictions to data inputs) and interpretability (i.e., the ability to explain why a model made a given prediction) are critical components of predictive healthcare models. This need for semantic interpretability is not limited to the LOS use-case. Rather, interpretability is becoming increasingly important across the continuum of care for hundreds of predictive modeling use-cases.

Considering that typical deployments of EHR systems across the world capture tens of thousands of raw features (e.g., diagnosis codes and procedure codes) for millions of encounters every year, the curse of dimensionality becomes a major obstacle in developing highly effective and interpretable ML models [4]. *Data phenotyping* (a.k.a., *electronic phenotyping*) has emerged to describe data groupings that are associated with clinically relevant concepts which can capture semantics for interpretation. For example, when evaluating a prediction, comparing a large set of independent features such as {"excess growth", "low blood sugar", etc.} requires time and significant medical understanding to interpret. In contrast, a single summarizing phenotype like "complications resulting from diabetes" is quickly understandable even to a layperson [5].

Unsupervised ML approaches like clustering [6], [7] and matrix/tensor factorization [5], [8] have been applied to generate data phenotypes with demonstrable success. However, existing investigations of these approaches typically focus either on uncovering novel health states [9] or on highly accurate detection of specific, well-known diseases [10], for which the ease of interpretability is not addressed. Because of this, phenotypes generated through these approaches often have nebulous feature memberships: each feature is associated with each phenotype via continuous values that are difficult to translate in real world terms (as shown in Fig. I). Take non-negative matrix factorization (NMF) [8] as an example. This

	Feature A	Feature B	Feature C	Feature D	Feature E	Feature F
Phenotype 1	0.54	1.2	0.00374	1.23	0.00076	0.24
Phenotype 2	1.97	0.002	0	0.1	0.008	1.9
Phenotype 3	0.003	2.6	0.00274	0.1	0.004	0.37

Fig. 1. An example of NMF feature assignment.

unsupervised method is very powerful in addressing both the large-scale and high-dimensionality problems. NMF can be used to represent each patient in a reduced feature space (e.g., Phenotype 1 to 3 in Fig. I) instead of the original feature space (e.g., Feature A to F in Fig. I). However, for interpretable modeling, the relationship between each phenotype and the set of original features with semantic meanings can be unintuitive. This is especially true when the raw features are binary-valued or categorical. For example in Fig. I, if Feature A represents a diagnosis code such as, "has unspecified anemia", it is difficult to interpret what the value of "1.97" means for Phenotype 2.

Interpretable ML has become a prominent subject in recent years as ML systems have started impacting the lives of billions of people. In the context of ML, "interpretablity" is defined (perhaps most generally) as the ability to explain a model in human terms (Doshi-Velez and Kim [11]). Kim et al. [12] define it more specifically as "the degree to which a human can consistently predict the model's result". For this work, interpretability may include the ability for a clinician to understand why certain features are grouped together physiologically. However, we focus on the ability for a data scientist to explain why any individual prediction was made, particularly when the machine logic relies on rare combinations of features.

Interpretability of ML models is especially relevant, for example, where fairness is critical, where the consequences for ML-based decisions are far reaching, or where the cost of mistakes is high. This is especially true at point of care, where decision-making can literally be a matter of life and death [13]. Thus interpretable ML is crucial in the construction of decision support systems in healthcare. There is a large body of work on interpretable ML in supervised learning [14]. However, the literature in unsupervised learning mainly focuses on classical clustering techniques [15]. This work compares classical clustering with other techniques from unsupervised ML.

To provide clearly interpretable phenotypes that can be used in generic ML tasks, we propose a new unsupervised phenotyping framework that also can alleviate the curse of dimensionality before the data are used in any potential ML tasks. Specifically, unsupervised ML methods can be used to form phenotypes with clear feature memberships by grouping features rather than patients. Then, using a proposed *expressivity score*, the set of phenotypes can be used to replace the set of features describing each patient. Depending on the number of phenotypes determined, the dimensionality of each observation can thus be significantly reduced. And, because each phenotype thus has a clear, boolean inclusion of original features, this simplifies semantics for interpretation. After applying the proposed framework to data from the Medical Information Mart for Intensive Care III (MIMIC-III) database [16], we evaluate how the proposal will affect the hospital length of stay prediction problem on two different sub-tasks (i.e., Neonatal and Geriatric). The empirical study demonstrates that this framework can be used, with no detriment to the tasks' performance metrics, while providing improved interpretability. In summary, the main contributions of this work are as follows:

- We propose a new unsupervised phenotyping framework to generate easily interpretable phenotypes as new features to describe each observation that can then be used for generic ML tasks of EHRs.
- This proposed framework can not only alleviate problems from high-dimensionality often faced in EHR applications, but also provide improved interpretability which is crucial for healthcare.
- We evaluate phenotypes generated through the proposed framework on two LOS prediction problems, for which improved interpretability can be observed without significantly diminishing the LOS prediction performance.

II. Related Work

A. Data Phenotyping

Expanding upon the concept of biological phenotypes, data phenotypes are groupings of EHR data that expose clinically relevant underlying patterns. Having emerged from concepts of genetic expression, these groupings have historically been used to describe disease patterns or to create patient cohorts for medical and epidemiological study. These applications historically relied on rule-based and statistical methods ("manual methods") [9]. However, with the burgeoning development of ML methods in recent years, unsupervised learning approaches are now used to uncover novel patterns in EHR data.

Non-negative Matrix Factorization (NMF) and Non-negative Tensor Factorization (NTF) have shown particular utility in EHR pattern discovery. For example, Joshi et al. [8] demonstrated that the use of NMF in combination with the *bag-of-words* NLP technique is helpful in deriving clinically valid comorbidity phenotypes from clinical notes. Marble [5] is another example using a series of tensor-factorization-based approaches designed to find phenotypes in categorical health data. Both approaches demonstrated competitively predictive phenotypes deemed to be interpretable by clinical practitioners.

However, in healthcare AI, models and features are not designed by clinical practitioners alone. Rather, they are often

generated by data scientists with limited clinical background, and optimized for speedy feature validation to enable rapid development. Ideally, both clinical and data science subject matter experts should be able to vet the resultant phenotypes without requiring continual interpretation from each other on the semantic logic. Thus, while NMF/NTF methods are able to produce "high-throughput" phenotypes (i.e., faster than manual methods) [5], the validation process remains a bottleneck for deployment in automated ML pipelines. Moreover, feature membership in NMF/NTF-generated phenotypes is ill-defined, since it may be the case that any given phenotype discovered by these techniques is associated with all possible features by some fractional value. This leads to arbitrary cutoffs, as well as feature-phenotype associations that are not semantically translatable and thus introducing subjectivity.

Clustering is another approach to EHR data grouping, commonly used to generate patient cohorts for epidemiological study. The k-means algorithm, for example, is one of the most popular point-assignment based clustering algorithms in use. Among recent work, [6] used k-means to identify 3 groups of patients suffering from chronic obstructive pulmonary disease (COPD) and [7] used k-means to group patients by an MCA-reduced set of boolean diagnoses. Another popular clustering approach is hierarchical clustering [17]. [18], for example, used agglomerative hierarchical clustering to produce groups of like-illnesses relevant to chronic disease management in primary care. However, clustering approaches almost invariably generate phenotypes by clustering patients. And while interpretation of clustering results may be easier relative to the aforementioned factorization methods, once again any given phenotype may be associated with all possible features and medical experience is often required.

In this work, we aim to generate data phenotypes that are interpretable in their formation by disambiguating feature membership. These interpretable phenotypes can then be used to add interpretability to other generic ML tasks (e.g., additional clustering or supervised learning).

B. Feature Grouping

There are alternative strategies for grouping clinical features without unsupervised learning, aside from standard feature engineering. One prominent example is the Phenotype Knowledgebase (PheKB) [19]. PheKB is a tool and public repository for community-generated, pre-validated rule-based phenotyping algorithms. While crowdsourcing rule definitions in this way reduce the burden of validation in some contexts, the database is not comprehensive. Also, importantly, the PheKB algorithms may not perform well across data from different healthcare systems [20] and they can be cumbersome to produce.

Another alternative strategy is the use of medical surveillance and billing codes as grouping units, such as the Healthcare Cost and Utilization Project's (HCUP) Clinical Classification Software (CCS) [21] codes. These codes are based on the International Classification of Diseases (ICD) [22] codes, which are often used as the raw features that are manually designed for medical reporting. Since ICD codes are highly granular, they can present an unmanageable number of codes for the problems under discussion. HCUP's CCS codes group individual ICD codes into more manageable categories, organized by higher-level concepts such as "body systems". Level 2 CCS categories are often used in health informatics for data interpretation. However, these manually generated groups do not capture variability in the specific datasets, nor do they encapsulate the intricate interactions among the EHR features such as labs and medications with diagnosis codes. Moreover, because both the ICD and CCS systems are pre-specified, they cannot be expanded to include custom features (e.g., specific biometric ranges), and are hence not suitable surrogates for interpretation.

C. Interpretable ML

We refer the reader to surveys of interpretable supervised learning models such as [14] for a comprehensive overview. [13] describes specific challenges in healthcare ML interpretability. For unsupervised learning, prominent examples of interpretable models include constraint-based clustering models that use decision trees [23], models that first separate the feature space using a decision tree by greedily maximizing a heterogeneity measure [24], and probablistic models that produce clustering rules for each cluster [15]. An active learning model approach is introduced by Gilbert et al. [25] for discovering rule sets for describing clusters. Interpretable ML models such as these often focus on one specific ML task (e.g., supervised learning or clustering), whereas in this work we aim to prepare the EHR data with interpretable phenotypes to be used in any generic ML tasks (in addition to addressing the high-dimensionality challenge).

III. The Proposed Method

To facilitate interpretation, we aim to find phenotypes as clinically relevant feature groupings that are distinct and each contain a concise feature subset to semantically show what the phenotype is expressing.

A. Phenotype Generation

Consider an EHR data set $X \in \mathcal{R}^{n \times d}$ with n patients, each described by d features. We aim to find k ($k \ll d$) phenotypes $P = \{P_1, P_2, \cdots, P_k\}$, where each phenotype P_i contains m_i features such that $\sum_i^k m_i = d$ and $P_i \cap P_j = \emptyset$ for any $1 \leq i < j \leq k$. The main goal is to ensure that all features within each group are as similar as possible considering all patients. Thus, the problem is to partition d features, each represented by n patients as $F = X^\top \in \mathcal{R}^{d \times n}$ (i.e., the transpose of X), into k clusters in which each feature $\mathbf{f} \in \mathcal{R}^{1 \times n}$ belongs to the cluster with the nearest mean (i.e., cluster center or cluster centroid). Therefore, we can formulate the problem as

$$\arg\min_{P} \sum_{i=1}^{k} \sum_{\mathbf{f} \in P_i} dis(\mathbf{f}, \mu_i) \tag{1}$$

where μ_i is the center of the i-th cluster, and $dis(\cdot, \cdot)$ calculates the distance between two features using one of many

popular distance functions (for example, Euclidean distance or Manhattan distance). This becomes a standard clustering problem and many traditional clustering methods (e.g., k-means, agglomerative clustering, or spectral clustering [17]) can be applied to group features into k clusters.

B. Representation in Phenotype Space

Given k cluster-generated feature groups P, where $P = \{P_1, P_2, \cdots, P_k\}$, we propose to use phenotyping as a dimension reduction process and to represent patients in the phenotype space for any generic tasks on the data. In other words, each phenotype becomes a new single feature to describe a patient. Thus, we can reduce the size of the patient matrix from $X^{n\times d}$ to $Y^{n\times k}$, where the curse of dimensionality can be alleviated for any future tasks.

The challenge then is to calculate the new feature values $Y^{n\times k}$ of each patient with respect to each phenotype generated, since each phenotype may consist of multiple original features. We propose an *expressivity score* to calculate each element of Y, that is, each Y_{ij} for the i-th patient with respect to the j-th phenotype. Considering that each phenotype P_j contains m_j original features as

$$\{\mathbf{f}_{j1}, \mathbf{f}_{j2}, \cdots, \mathbf{f}_{jm_j}\} \tag{2}$$

many strategies can be adopted to represent the phenotype with respect to different data types. For example, the average value of each feature over all patients can be used to represent the phenotype's corresponding feature value as

$$P_j = [avg(\mathbf{f}_{j1}), avg(\mathbf{f}_{j2}), \cdots, avg(\mathbf{f}_{jm_j})] \tag{3}$$

Then, the expressivity score Y_{ij} can be calculated based on the distance between the i-th patient described by the features in P_j and the j-th phenotype as

$$Y_{ij} = sim(X_i^{P_j}, P_j) \tag{4}$$

where $0 \leq sim(.,.) \leq 1$, and $X_i^{P_j}$ means the original feature values of the i-th patient described by the features in j-th phenotype and different similarity functions can be adopted for sim considering the specific application scenario. Thereafter, $Y^{n\times k}$ can be obtained for generic tasks, which not only overcomes the high-dimensionality problem in the original feature space, but also provides clear semantics of each phenotype (i.e., each new feature) for interpretation using a subset of original features. The overall procedure of the proposed interpretable phenotyping framework is summarized in Alg. 1.

Algorithm 1 Interpretable Phenotyping

Input: EHR data $X \in \mathcal{R}^{n\times d}$, the number of phenotypes k.
Obtain the transpose of X as $F = X^\top$
Clustering features in F into k disjoint clusters
for each cluster **do**
  Obtain representation using original features as Eqn. 3
end for
for $i = 1$ **to** n **do**
  Compute the expressivity score using Eqn. 4
end for
return F for interpretation and $Y^{n\times k}$ for generic ML

IV. Experiments

To evaluate the proposed phenotyping framework for EHR data, we apply it to two cohorts from the MIMIC-III [16] data and use the LOS problem as an example of generic ML task for evaluation.

A. Data

MIMIC-III is a large, freely-available, and de-identified database of critical care encounters at the Beth Israel Deaconess Medical Center in Boston from 2001 to 2012 [16]. We selected two subset cohorts. The first, "Neonatal" subset contains all encounters of patients who were transferred to a neonatal Intensive Care Unit (ICU). The second, "Geriatric" subset is that of all encounters for patients aged 65 and above. Diagnoses and procedures are encoded from the 9$^{\text{th}}$ edition of the ICD, hence referenced as "ICD-9" [22]. ICD-9 code distributions for both subsets are listed in Table I, where an encounter means an interaction between a patient and healthcare provider(s) for the purpose of providing healthcare service(s) or assessing the health status of a patient. Therefore, each patient may have multiple encounters due to different interactions. Using only the diagnosis and procedure codes as features, the value of each feature is binary, meaning present or not.

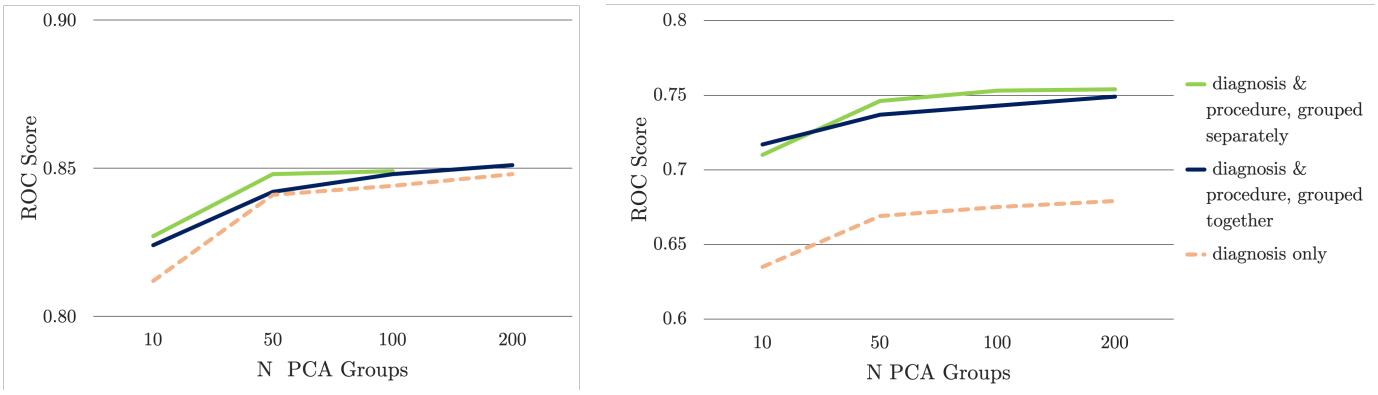
B. Phenotyping Setup

We apply the proposed phenotyping method to obtain new data representations for each cohort without including additional information. To comprehensively evaluate the proposed framework, we apply several standard clustering methods to group features as described in the following list, then compare the relative performance. Given the binary values of the features, Manhattan distance is used to measure the distance between two features.

- **K-Means(f)**: k-means feature clustering [17]. K-means clustering is applied directly to group features (rather than patients), where each feature is assigned to only one phenotype.
- **Agg(f)**: hierarchical (agglomerative) feature clustering [17]. Agglomerative clustering is applied directly to group features, similar to K-Means(f).
- **NMF(r)**: a restricted variant of NMF [8]. As shown in Fig. I, with the standard NMF, each feature can be associated with multiple phenotypes of different strengths, which makes interpretation hard. In this variant, each feature is assigned to a single phenotype of the strongest association value. For example, feature F is restricted to be in Phenotype 2 only as shown in Fig. I.

TABLE I
MIMIC-III DATA SUBSETS

Subset	Encounters	ICD-9 Diagnoses	ICD-9 Procedures	Positive	Distribution
Neonatal	8,101	927	164	LOS $>$ 4 days	P(+) = 0.53
Geriatric	26,074	4,889	1,581	LOS $\geq$ Mean (10.0467 days)	P(+) = 0.33



(a) Neonatal

(b) Geriatric

Fig. 2. Comparison of binary LOS classification performance (ROC AUC) with vs. without procedure codes besides diagnosis codes when the number of phenotypes are increasing using PCA.

- **Marble(r)**: a restricted variant of Marble [5]. Similar to NMF(r), each feature is assigned to a single phenotype of the strongest association value.

After we obtain the phenotypes using the above mentioned methods where each feature is assigned to only one phenotype, we treat each phenotype as one new feature. We calculate the new data representation, that is the expressivity score Y_{ij} of the i-th patient for the j-th phenotype, using the Manhattan distance since all feature values are binary to calculate the similarity in Eqn. 4. This can be written as

$$Y_{ij} = 1 - \frac{Manh(X_i^{P_j}, P_j)}{m_j} \tag{5}$$

where P_j is a simple m_j-length vector of 1's (i.e., the case when every feature is present).

C. Task Setup

For each cohort selected, we evaluate the task performance through a binary LOS prediction problem. To evaluate the proposed framework for different application scenarios, we set $LOS > 4$ as the positive class for the Neonatal cohort to balance the positive and negative classes, while setting the average (mean) LOS as the threshold for the Geriatric cohort for a general setting where the positive and negative classes are imbalanced as shown in Table I.

Besides the proposed phenotyping framework, we also apply some state-of-the-art dimension reduction methods for comparison.

- **ICD-9**: the original data presentation without dimension reduction.
- **CCS** [21]: the manually-generated level-2 categorizations.
- **PCA** [26]: a standard unsupervised dimension reduction method.
- **NMF** [8]: the standard NMF method.
- **Marble** [5]: the standard Marble method.

Based on the data representations obtained using the above methods, XGBoost [27] using default parameters is applied for the binary LOS classification problem to generate predictions at the encounter level (for each patient-clinic interaction), since different interactions from the same patient can also result in different LOS. Then, four evaluation metrics are used to compare relative performance over the binary LOS classification problem, including accuracy, precision, recall, and ROC AUC score, averaged over 10 trials of 80-20 holdout validation. Each method is tested on the same set of 10 holdout datasets.

D. Ablation Study

In this ablation study, we evaluate whether the inclusion of procedure codes (i.e., ICD-9 codes) besides the diagnosis codes can significantly affect the LOS prediction performance. We also evaluate how the number of phenotypes can affect said performance using PCA as a standard dimension reduction method. Considering that we can apply PCA separately or together to diagnosis codes and procedure codes, we compare three different settings, i.e., diagnosis only, diagnosis and procedure grouped separately, and diagnosis and procedure grouped together.

As shown in Fig. 2, for the Neonatal LOS problem no significant difference can be observed with the inclusion of procedure codes, regardless of how variables are combined. In contrast, there is a significant difference for the Geriatric LOS problem, where the inclusion of procedure codes leads to significantly better performance. A potential reason for this

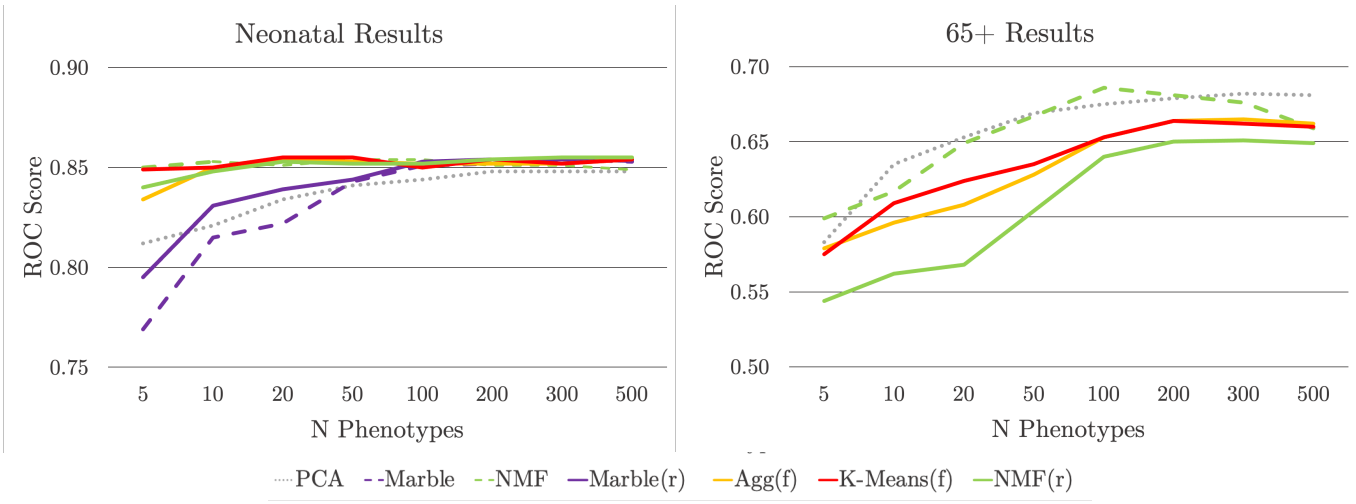


Fig. 3. LOS prediction performance comparison: ROC AUC relative to the number of phenotypes. (Left: Neonatal; Right: Geriatric (i.e., 65+))

is that we have a relatively larger variety in procedures for the cohort of Geriatric, which can result in different LOS. Notably, the method of combination makes no significant difference. Moreover, the trend in performance relative to the number of phenotypes is similar regardless of which code types are included. Fewer than 100 phenotypes are sufficient for prediction regardless of the inclusion of procedure codes. Thus, we use only diagnosis codes in the following experiments for the sake of simplicity.

E. Task Performance

In this subsection, we compare the LOS prediction performance based on the data representations obtained through different phenotyping methods applied to the diagnosis only data.

As illustrated in Fig. 3, the performance converges for both cohorts when the number of phenotypes reaches around 200. Compared to $13,000+$ individual diagnosis codes, 200 phenotypes is significantly fewer (i.e., dimension is well reduced) and thus reduce the cost for any further generic ML task and facilitate the model interpretation. It should be noted that Marble results are not included for the Geriatric task. This is because that the computational time of Marble increases dramatically relative to an increasing number of desired phenotypes, rendering it intractable for the Geriatric problem, as shown in Fig. 4.

From Fig. 3, it can be observed that when the number of phenotypes is small, the standard NMF consistently provides the best performance, while the performance of PCA is not consistent for these two cohorts (i.e., good for Geriatric but not for Neonatal). However, both standard NMF and PCA lack clear feature assignments for interpretation. Yet, the more interpretable, restricted NMF (i.e., NMF(r)) has degraded performance. K-Means(f), on the other hand, demonstrates similar performance to NMF but with clear feature assignments for interpretation. When the number of phenotypes is large enough, all phenotyping methods provide similar performance,

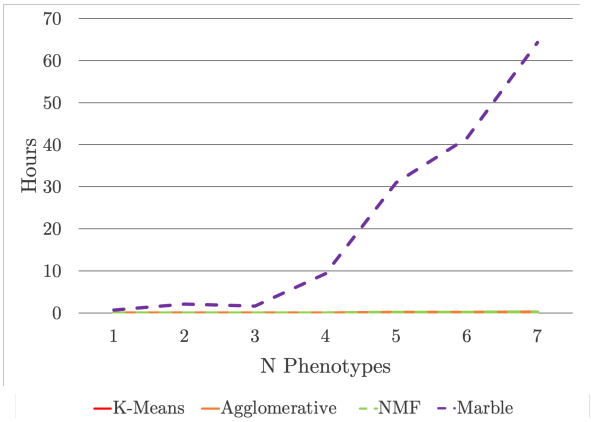


Fig. 4. Computation time comparison per number of phenotypes on Neonatal

and thus methods with better interpretability (e.g., K-Means(f)) are preferred.

To succinctly compare all methods including the two baselines, ICD-9 and CCS, we take the case of 200 phenotypes as an example for each cohort and compare their performance in Table II, where the best performance of each evaluation metric is emphasized in bold. It should be noted that the average performance of 10 trails is reported with its standard deviation in the brackets as "mean (std)". For the Neonatal LOS problem, the CCS baseline yields the best precision, yet is the worst performer by a significant margin for all other performance metrics. In contrast, the other approaches yield similar precision, with comparable performance across all other metrics, although K-Means(f) yields both the best Accuracy and best ROC AUC score. For the Geriatric LOS problem, all approaches, including the CCS baseline, yield similar performance across metrics, with the exception of recall that has larger variance. Therefore, it appears that the proposed phenotyping framework will not significantly sacrifice task performance (and in some cases may improve

TABLE II
COMPARISON OF LOS PREDICTION PERFORMANCE WITH 200 PHENOTYPES
(MEAN (STD) OF 10 TRIALS AND THE BEST METHOD FOR EACH METRIC IS IN BOLD)

Metric	ICD-9	CCS	PCA	K-Means(f)	Agg(f)	NMF	NMF(r)	Marble	Marble(r)
Neonatal									
Accuracy	0.8565 (0.0055)	0.7944 (0.0088)	0.8529 (0.0057)	**0.8579 (0.0055)**	0.8564 (0.0054)	0.8554 (0.0046)	0.8572 (0.0061)	0.8559 (0.0057)	0.8574 (0.0052)
Precision	0.828 (0.0063)	**0.8555 (0.0099)**	0.8189 (0.0089)	0.8302 (0.0057)	0.8275 (0.0054)	0.8222 (0.0065)	0.8321 (0.0068)	0.829 (0.0059)	0.8309 (0.0057)
Recall	0.9198 (0.0121)	0.7357 (0.0152)	0.9272 (0.0047)	0.9195 (0.0099)	0.9203 (0.0081)	**0.9274 (0.0101)**	0.9147 (0.01)	0.9168 (0.0088)	0.9172 (0.0088)
ROC AUC	0.8526 (0.0053)	0.7980 (0.0086)	0.8483 (0.0061)	**0.8541 (0.0055)**	0.8524 (0.0054)	0.8510 (0.0047)	0.8536 (0.0061)	0.8522 (0.0057)	0.8537 (0.0052)
Geriatric									
Accuracy	0.7499 (0.0033)	0.75 (0.0032)	0.7541 (0.0043)	0.7557 (0.0031)	0.7553 (0.0046)	**0.7635 (0.0045)**	0.7479 (0.0045)	0.748 (0.0028)	0.7482 (0.0026)
Precision	**0.7543 (0.0115)**	0.7064 (0.0116)	0.6947 (0.0097)	0.7387 (0.0098)	0.7366 (0.0118)	0.7372 (0.0102)	0.7339 (0.0136)	0.7332 (0.0071)	0.7349 (0.0072)
Recall	0.3535 (0.0092)	0.4091 (0.0067)	**0.4479 (0.0101)**	0.3961 (0.0113)	0.3963 (0.0112)	0.4343 (0.0126)	0.3643 (0.0103)	0.3654 (0.0083)	0.3645 (0.0081)
ROC AUC	0.6486 (0.0044)	0.6629 (0.0033)	0.6758 (0.0055)	0.6638 (0.0048)	0.6635 (0.0061)	**0.6793 (0.0063)**	0.6498 (0.0057)	0.6502 (0.0041)	0.6501 (0.0038)

TABLE III
COMPARISON OF NEONATAL LOS PREDICTION PERFORMANCE WITH 20 PHENOTYPES
(MEAN (STD) OF 10 TRIALS AND THE BEST METHOD FOR EACH METRIC IS IN BOLD)

Metric	ICD-9	CCS	PCA	K-Means(f)	Agg(f)	NMF	NMF(r)	Marble	Marble(r)
Accuracy	0.8565 (0.0055)	0.7944 (0.0088)	0.8454 (0.0063)	**0.8583 (0.0053)**	0.8569 (0.0054)	0.8553 (0.0049)	0.8568 (0.0057)	0.8417 (0.0066)	0.8503 (0.0062)
Precision	0.8280 (0.0063)	**0.8555 (0.0099)**	0.8140 (0.0080)	0.8300 (0.0060)	0.8286 (0.0059)	0.8234 (0.0061)	0.8329 (0.0059)	0.8102 (0.0071)	0.8214 (0.0074)
Recall	0.9198 (0.0121)	0.7357 (0.0152)	0.9176 (0.0081)	0.9207 (0.0104)	0.9198 (0.0082)	**0.9249 (0.0097)**	0.9125 (0.0103)	0.9162 (0.0075)	0.9166 (0.0099)
ROC AUC	0.8526 (0.0053)	0.7980 (0.0086)	0.8410 (0.0065)	**0.8544 (0.0052)**	0.8530 (0.0054)	0.8510 (0.0050)	0.8534 (0.0057)	0.8371 (0.0067)	0.8463 (0.0062)

performance), while reducing the dimensionality of the problem. More importantly, the proposed phenotyping framework can facilitate interpretation as described in the following subsection.

F. Model Explanation

In this subsection, we evaluate how the proposed framework can help in model explanation. To explain the prediction model's behavior, one way is to check which features are used heavily to make the prediction decision, for example via Local Interpretable Model-agnostic Explanations (LIME) [28] and SHapley Additive exPlanations (SHAP) [29]. We take the Neonatal task with 20 phenotypes as a case study for which a performance comparison between different methods can be found in Table III, where the average performance of 10 trails and its standard deviation in the brackets as "mean (std)" are reported. Similar to the case with 200 phenotypes, values for performance metrics are generally similar. While the CCS approach provides the worst performance in terms of ROC AUC, other approaches yield similar performance, and K-Means(f) yields the best.

One potential reason that the interpretable phenotyping methods (K-Means(f), Agg(f) and NMF(r)) yield similar prediction performance on this Neonatal cohort is that the methods produce similar phenotypes, as observed in Fig. 5. Concretely, when comparing feature membership between pairs of these 20 phenotypes (pairs of phenotypes from two different phenotyping methods), at least 14 pairs are essentially identical between any two of the methods. For example, phenotype 14 of Agg(f), phenotype 10 of NMF(r), and phenotype 15 of K-Means(f) are essentially the same (see Fig. 5). This demonstrates that different traditional clustering methods can be adopted in our framework to generate interpretable phenotypes for the sake of different application purposes, while more efficient and easy-to-use ones like K-Means(f) are preferred.

To avoid potential bias based on our choice of interpretation method, we applied both LIME [28] and SHAP [29] to evaluate the interpretable methods (ICD-9, CCS, K-Means(f), and NMF(r)). It should be noted that, SHAP can provide both global and instance-level interpretation, while LIME is an instance-level prediction explainer. To compare between these two interpreters, we also obtain the global feature importance for LIME by aggregating over instance-level explanations. Concretely, global feature importance is assessed as the aggregate mean importance of all prediction instances across all 10 holdout trials. Since SHAP can be applied to each holdout trial, the average importance values were computed for both LIME and SHAP, and then validated across the two methods. Because this assessment yields consistent top phenotypes between the

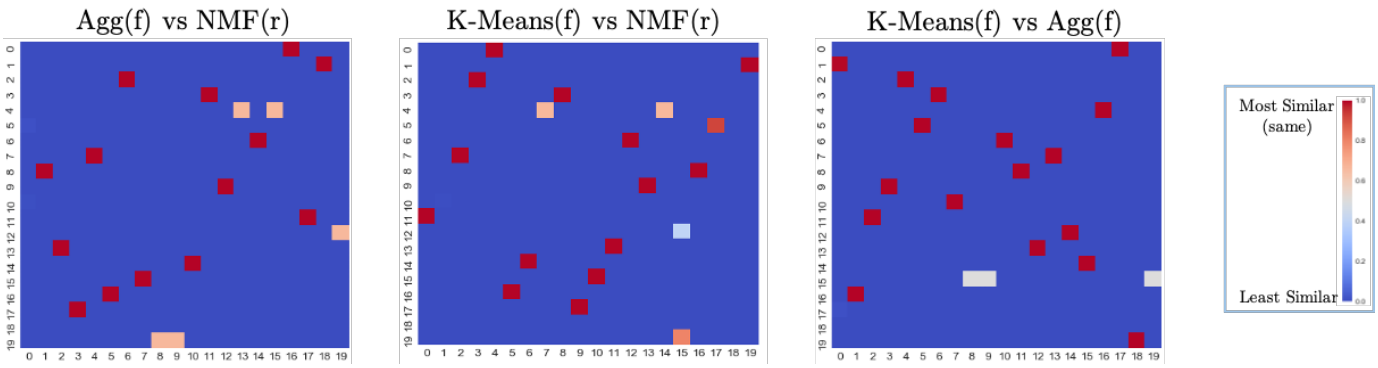


Fig. 5. Feature membership comparison between different phenotyping methods for 20 phenotypes on Neonatal data

two interpretation methods as shown in Table IV (e.g., top 4 phenotypes are the same although the ranking is slightly different for the true positive predictions), we present in the following only the results from LIME for discussion.

TABLE IV
COMPARISON OF FEATURE IMPORTANCE BETWEEN LIME AND SHAP FOR TRUE POSITIVE PREDICTIONS WITH 20 PHENOTYPES ON NEONATAL

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

Here we take the top five phenotypes used for true positive prediction as an example (Fig. 6). As shown in Fig. 6, we can observe that CCS uses very different sets of ICD-9 codes relative to the other methods. This explains why CCS yields different prediction performance compared to the rest, and it shows the weakness of the manually created categories when considering that the performance was reduced.

Additionally, the proposed K-Means(f) and NMF(r) are able to capture more information relative to the ICD-9 baseline and this additional information can be easily validated by clinicians. An example of this is the set of phenotypes at rank-2 importance, which include 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 [30]. 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 even require the experience of a clinician.

Furthermore, the proposed framework presents the phenotype information with greater specificity than the CCS method. For instance, although "Neonatal bradycardia" and "Primary apnea of the newborn" are included in the same Level 2 CCS category ("Other perinatal conditions"), that category is associated with 127 distinct diagnoses in this dataset. This is far from easy interpretation and explains why they are not among the top most important phenotypes for true positive prediction.

A benefit of using LIME is that we can use its instance-level interpretation to understand why one instance is misclassified by a trained prediction model, so as to identify the model's potential weaknesses. For the example shown in Fig. 7, we can observe that this model placed significant negative weight on several phenotypes that the patient did not express, with significant positive weight placed on a phenotype that was only somewhat expressed. Whereas, the one fully expressed phenotype, containing "Single liveborn...", has significant negative weight, which results in the false negative decision. This indicates some potential bias or weakness of the trained model that could be improved.

In summary, the proposed unsupervised phenotyping framework can be used to prepare data for generic ML tasks with no detriment to performance, but clearly reduces the dimensionality and provides improved model interpretability.

V. CONCLUSIONS AND FUTURE WORK

Considering the complexity and volume of data, as well as the inherent need for explainable results in predictive healthcare analytics, this work proposes an unsupervised interpretable phenotyping framework to prepare EHR data for generic healthcare tasks. Specifically, we propose to group high-dimensional features into a smaller number of disjoint groups as phenotypes. Then, each phenotype will be treated as a new feature to describe each encounter, in which we propose an expressivity score to describe an encounter using the new

Top Features in True Positive Predictions

Importance	ICD-9	CCS	K-Means(f)	NMF(r)
1	- Neonatal jaundice associated with preterm delivery	"Other perinatal conditions" (includes 100+ ICD-9 Codes)	- Neonatal jaundice associated with preterm delivery	- Neonatal jaundice associated with preterm delivery
2	- Single liveborn, born in hospital, delivered w/o mention of cesarean	"Short gestation; low birth weight; and fetal growth retardation" (includes 40+ ICD-9 Codes)	- 31-32 completed weeks of gestation - Neonatal bradycardia - Primary apnea of newborn	- Neonatal bradycardia - Primary apnea of newborn
3	- Disorder of stomach function and feeding problems in newborn	"Hemolytic jaundice and perinatal jaundice" (includes 20+ ICD-9 Codes)	- Single liveborn, born in hospital, delivered w/o mention of cesarean	- Disorder of stomach function and feeding problems in newborn
4	- Primary apnea of newborn	"Immunizations and screening for infectious diseas" (includes 40+ ICD-9 Codes)	- Disorder of stomach function and feeding problems in newborn	- Single liveborn, born in hospital, delivered w/o mention of cesarean
5	- Congenital pneumonia	- Respiratory distress syndrome of the newborn	- 33-34 completed weeks of gestation	- Anemia of prematurity - Other specified conditions originating in the perinatal period - Chronic respiratory disease arising in the perinatal period - Patent ductus arteriosus - Retrolental fibroplasia - Septicemia [sepsis] of newborn

Labels are ICD-9 descriptions except where indicated by quotations.

Fig. 6. Top features for each model as identified by LIME

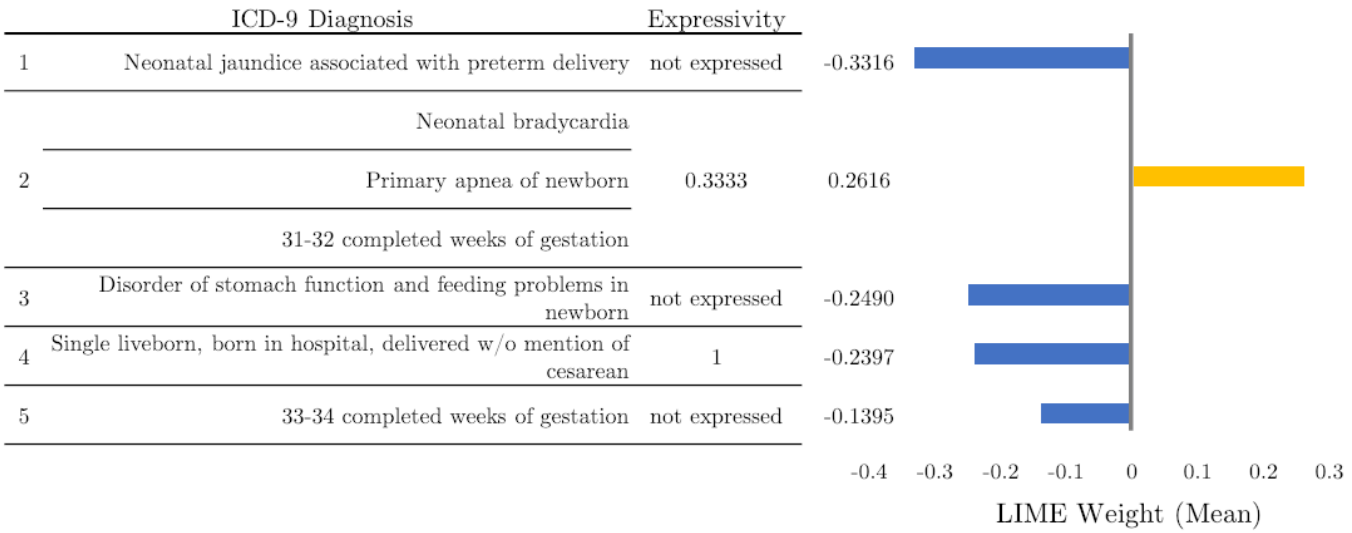


Fig. 7. LIME feature importance for one false negative prediction using K-Means(f) phenotypes

feature. Thereafter, the curse of high dimensionality from EHR can be alleviated. More importantly, each phenotype has a clear assigment of original features for easy interpretation. With application to two different cohorts of the LOS prediction problem, the proposed framework can provide clinically relevant groups of features as phenotypes that are easy to interpret without sacrificing the prediction performance on the task.

One limitation of this work is that we have restricted phenotypes to be mutually exclusive, such that each feature is a member of only a single phenotype. However, this restriction is not inherent to the framework, and future experimentation with fuzzy constraints ("fuzzy clustering") to lift the restriction would allow phenotypes to capture more nuanced feature interactions. It may also lead to improved performance or potential knowledge discovery. Another clustering technique of note for future experimentation is co-clustering, which would compare both the patient vectors and the feature vectors simultaneously. Apparently, co-clustering can help find phenotypes shared by a certain group of patients, where groups of patients are also identified. Therefore, this is another interesting future direction to see how co-clustering would compare in terms of the quality of feature groupings and the predictive performance.

REFERENCES

[1] P. J. Mathew, F. Jehan, N. Kulvatunyou, M. Khan, T. O'Keeffe, A. Tang, L. Gries, M. Hamidi, E.-R. Zakaria, and B. Joseph, "The burden of excess length of stay in trauma patients," *The American Journal of Surgery*, vol. 216, no. 5, pp. 881–885, 2018.

[2] B. Tess, H. Glenister, L. Rodrigues, and M. Wagner, "Incidence of hospital-acquired infection and length of hospital stay," *European Journal of Clinical Microbiology and Infectious Diseases*, vol. 12, no. 2, pp. 81–86, 1993.

[3] Y. Xie, G. Schreier, D. C. Chang, S. Neubauer, Y. Liu, S. J. Redmond, and N. H. Lovell, "Predicting days in hospital using health insurance claims," *IEEE journal of biomedical and health informatics*, vol. 19, no. 4, pp. 1224–1233, 2015.

[4] J. Hu and J. Pei, "Subspace multi-clustering: a review," *Knowl. Inf. Syst.*, vol. 56, no. 2, pp. 257–284, 2018.

[5] J. C. Ho, J. Ghosh, and J. 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*, 2014, pp. 115–124.

[6] J. Garcia-Aymerich, F. P. Gómez, M. Benet, E. Farrero, X. Basagaña, À. Gayete, C. Paré, X. Freixa, J. Ferrer, A. Ferrer *et al.*, "Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (copd) subtypes," *Thorax*, vol. 66, no. 5, pp. 430–437, 2011.

[7] M. Guisado-Clavero, A. Roso-Llorach, T. López-Jimenez, M. Pons-Vigués, Q. Foguet-Boreu, M. A. Muñoz, and C. Violán, "Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis," *BMC geriatrics*, vol. 18, no. 1, p. 16, 2018.

[8] S. Joshi, S. Gunasekar, D. Sontag, and J. Ghosh, "Identifiable phenotyping using constrained non-negative matrix factorization," *arXiv:1608.00704*, 2016.

[9] J. M. Banda, M. Seneviratne, T. Hernandez-Boussard, and N. H. Shah, "Advances in electronic phenotyping: from rule-based definitions to machine learning models," *Annual review of biomedical data science*, vol. 1, pp. 53–68, 2018.

[10] M. Deliu, M. Sperrin, D. Belgrave, and A. Custovic, "Identification of asthma subtypes using clustering methodologies," *Pulmonary therapy*, vol. 2, no. 1, pp. 19–41, 2016.

[11] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2017.

[12] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in neural information processing systems*, 2016, pp. 2280–2288.

[13] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[14] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *Proceedings of the 41st International convention on information and communication technology, electronics and microelectronics*, 2018, pp. 0210–0215.

[15] J. Chen, "Interpretable clustering methods," Ph.D. dissertation, Northeastern University, 2018.

[16] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[17] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[18] J. E. Cornell, J. A. Pugh, J. W. Williams Jr, L. Kazis, A. F. Lee, M. L. Parchman, J. Zeber, T. Pederson, K. A. Montgomery, and P. H. Noël, "Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database," *Applied multivariate research*, vol. 12, no. 3, pp. 163–182, 2008.

[19] J. C. Kirby, P. Speltz, L. V. Rasmussen, M. Basford, O. Gottesman, P. L. Peissig, J. A. Pacheco, G. Tromp, J. Pathak, D. S. Carrell *et al.*, "Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability," *Journal of the American Medical Informatics Association*, vol. 23, no. 6, pp. 1046–1052, 2016.

[20] 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, p. e80, 2019.

[21] HCUP, "HCUPnet," https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp, online; accessed 6-June-2020.

[22] WHO *et al.*, "International classification of diseases: 9th revision, basic tabulation list with alphabetic index," 1978.

[23] B. Liu, Y. Xia, and P. S. Yu, "Clustering through decision tree construction," in *Proceedings of the ninth international conference on Information and knowledge management*, 2000, pp. 20–29.

[24] R. Fraiman, B. Ghattas, and M. Svarc, "Interpretable clustering using unsupervised binary trees," *Advances in Data Analysis and Classification*, vol. 7, no. 2, pp. 125–145, 2013.

[25] K. Gibert and U. Cortés, "Clustering based on rules and knowledge discovery in ill-structured domains," 1998.

[26] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[28] M. T. Ribeiro, S. Singh, and C. 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*, 2016, pp. 1135–1144.

[29] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[30] P. Chandrasekharan, M. Rawat, A. M. Reynolds, K. Phillips, and S. Lakshminrusimha, "Apnea, bradycardia and desaturation spells in premature infants: impact of a protocol for the duration of 'spell-free'observation on interprovider variability and readmission rates," *Journal of Perinatology*, vol. 38, no. 1, pp. 86–91, 2018.