Clinical Analysis from Pattern Disentanglement Insight

Anonymous ACL submission

Abstract

Diagnosis of a clinical condition can help medical professionals save time in the decision-making and prevent overlooking risks. Several machine learning models have been developed to predict clinical conditions, however, many existing models may have ineffective interpretability which is often desirable. In this paper, we explore the problem of text interpretability using free-text medical notes recorded in electronic health records (EHR). We propose an algorithm combining text mining and pattern discovery solution to discover strong association patterns between patient discharge summaries and the code of international classification of diseases (ICD9 code). The proposed approach offers a straightforward interpretation of the underlying relation of patient characteristics in an unsupervised machine learning setting and also outperforms the baseline clustering algorithm and is comparable to baseline supervised methods.

1 Introduction

If Artificial Intelligence is to play a significant role in support of the automatic decision process, it is essential for the users to gain trust (Kim, 2021). Hence, besides the outcomes of the decisions, interpretability with specific statistical support is of ample importance to enable humans to understand the reasons behind the machine learning decision. Hence, in this study, we focus on interpreting the diagnostic characteristics/patterns from the electronic health records (EHR).

Topic modeling (Blei et al., 2003) has been applied to the unstructured notes of EHRs to predict clinical outcomes without focusing upon interpretability (Bright et al., 2021; Huang et al., 2015; Wang et al., 2020). Recently, methods in interpretability such as attention and saliency have had questions raised about their effectiveness (Bastings and Filippova, 2020) and security (Zhang et al., 2021). Meanwhile, other NLP methods such as minimal contrastive editing are computationally expensive (Ross et al., 2020) or require intrinsic implementations via prompts (Sun and Marasović, 2021).

Hence, to address the issue of interpretability of EHR, we created a novel two-stage algorithm, leveraging interpretable feature engineering of text such as topic models (Chen et al., 2019) and pattern discovery techniques (Wong et al., 2021), to discover strong association patterns from patient profiles and discharge summaries to reveal their relationships with the diagnosed disease [1], and clustering patients into specific groups. The output is clustering groups and an interpretable Knowledge Base.

The contributions of the paper are three folds: 1) Interpretability: a novel algorithm focusing on white-box model interpretation for free-text clinical notes; 2) Unsupervised Learning: the grouping of records based on the discovered associations revealing characteristics of records via unsupervised learning; 3) All-In-One Knowledge-Base: generating an all-in-one knowledge base to link the knowledge (hierarchical clusters), patterns (characteristics of records), and data (patients' records) together to show "what" (disease), "who/where" (tracking patient records back) and "why" (discovered patterns) to interpret clinical notes for better clinical decision making.

2 Material: MIMIC-III Data Description

MIMIC-III is a de-identified relational clinical database containing observations from over 40,000 patients in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). Our present study utilizes clinical notes, found in the NOTEEVENTS table, and diagnoses, found in the DIAGNOSES_ICD table.

[1] ICD9 code, which is the code of international classification of diseases

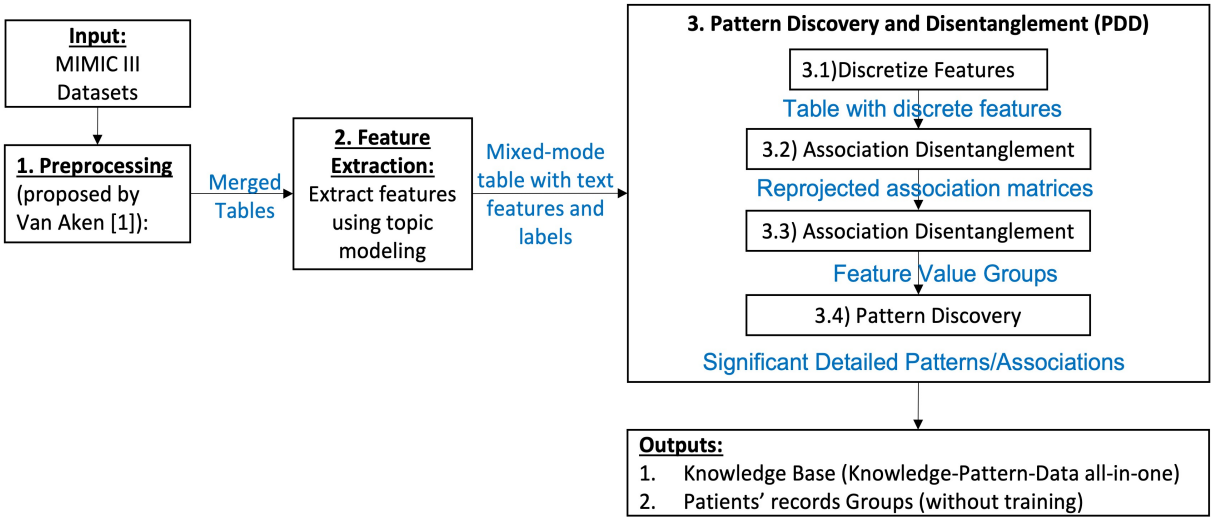

Figure 1: The overview of the proposed algorithm

Our final data contains 11,537 patient records and corresponds with the top four classes/diseases represented by the ICD9 code, which are: 414 - chronic ischemic heart disease, 038 - septicemia, 410 - acute myocardial infarction, and 424 - diseases of the endocardium. The four classes were slightly imbalanced, with 3502(30.35%), 3184(27.6%), 3175(27.52%), and 1676(14.53%) observations, respectively. We chose to include only the top 4 most common codes to highlight the pattern-discerning capability of the proposed algorithm, as including many codes (especially those with fewer observations) would decrease the interpretability and performance even for supervised learning models.

3 Methodology

In this section, we present the proposed methodology applied to the MIMIC-III dataset. The algorithm proposes tasks in three main steps: preprocessing, feature extraction, and pattern discovery. The overview of the proposed algorithm is shown in Figure 1. We first apply a preprocessing pipeline proposed by Van Aken et al. (2021) to clean and merge the dataset.

3.1 Feature Extraction

we further extract features from the clean dataset using topic modeling (Jelodar et al., 2019). The values of the features are represented by the probabilities of topics (group of words) occurring in the records. Labels (i.e. ICD9 code) are then merged with the features for unsupervised exploration. The optimal number of topics computed using coherence of the topic cluster instance (Röder et al., 2015) is 5, 20, and 30 - and therefore we create topic models with those respective parameters.

3.2 Pattern Discovery and Disentanglement

The dataset can be represented as a $M \times N$ matrix, where M represents the number of patients' records and N represents the number of extracted features [2].

Step 1. Pattern Disentanglement. First, we convert the values of numerical features into categorical features by using the Equal Frequency discretization. We denote categorical values of feature as Attribute Value (AV) (Wong et al., 2021). Second, In order to measure the strength of the association between each pair of AVs (i.e. the specific values of one attribute co-occurring with the value of another attribute), we construct an association matrix using the value of adjusted standardized residual (Wong et al., 2021). Then, we use Principal Component Analysis (PCA) to decompose the association matrix into principal components that are ranked according to the weights of the associations (eigenvalues). We then reproject the principal components onto the association matrix again. We refer to the reprojected association matrix as disentangled space. The above process is called *Pattern Disentanglement* which allows us to take the reprojected components/vectors from PCA and use the reprojected values as new measurements/criteria to represent the strength of associations between AVs in different orthogonal disentangled spaces.

[2] In pattern discovery, we use the term attribute instead of feature

Lastly, in order to obtain only the significant pairs of AV associations, we filter out statistical residual values greater than 1.96 in our newly reprojected association matrix (i.e. association matrix with disentangled associations)

Step 2. Pattern Clustering. In an unsupervised manner, we cluster the associations. Typically the number line of one projected principal component has two opposite sets of AV. However, when such opposing sets do not exist, we only use AV sets from one side of the PC. Furthermore, in order to reveal further characteristics of the records of the disentangled patterns, we separate the above sets into several subsets by clustering them. The similarity measure we used for clustering is the percentage of the overlapping records covered by each AV subcluster and we denote each AV subgroup by a three-digit code [#PC, #Group, #SubGroup]. The AV sets or subsets can reveal the characteristics of the records corresponding to disentangled patterns in order to provide statistical evidence for downstream clustering or prediction. Furthermore, patient records are obtained according to their particular characteristics (disentangled patterns) from the AV groups or subgroups.

The output of PDD is organized into an all-in-one representational framework (PDD Knowledge Base) with three parts: a Knowledge Section showing the hierarchical clusters such that each cluster unveils distinct characteristics of a related group of records; a Pattern Section listing patterns showing detailed associations between AVs; and the Data Section listing the record ID which link the patient to the knowledge and pattern sections.

4 Experimental Result

We present our results in Table 1 and knowledge base in Figure 2.

4.1 Comparison of Unsupervised and Supervised Learning

Given the imbalanced nature of our dataset (Zhou and Wong, 2021), we followed the same evaluation method in (Van Aken et al., 2021), *balanced accuracy* (Balanced Acc. in Table 1) and *weighted F1-scores* (Weighted F1 in Table 1), to evaluate performance of both supervised and unsupervised results. We compared the clustering results of PDD with K-mean, as the baseline, and also two supervised learning algorithms: Random Forest (Breiman, 2001) and CNN (Kalchbrenner et al., 2014) [3].

As the baseline comparison for features, we also applied all supervised and unsupervised learning algorithms on the dataset with words extracted using TFIDF (Jones, 1972). To make the interpretation meaningful, we selected the top 40 words in TFIDF with a feature selection algorithm by Random Forest.

The comparison results are shown in Table 1. It is interesting to observe that PDD outperformed other K-means. However, both supervised learning algorithms, Random Forest and CNN perform better on the TFIDF dataset. The reason should be that the top 40 words (feature) are selected based on classification results.

When topic modeling results are used as a dataset, PDD outperforms K-means and even the two other supervised learning algorithms when only 5 topics are used. As for Random Forest, it performs better when applied to the topic modeling results with 20 topics than the two experiments running on 5 topics and 30 topics. While as for CNN, the results of experiments on 30 topics are slightly better than the results on 20 topics.

One important notion we would like to bring forth is that, even if the accuracy score reflects the algorithm performance to some extent, class labels may not always be reliable in supervised classification algorithms. On the contrary, clustering merely recognizes patterns in the data and holds no such risk.

4.2 Discussion on Topic Modeling

From a clinical perspective, the generated topic models correspond reasonably well with each ICD9 diagnosis. In the 20-topic model, septicemia - a widespread infection of the body, was predicted by topics containing relevant words such as "infection", “bacteria", and "culture". Conversely, topics that contained cardiovascular-related terms such as "ventricular" or "aorta" predicted the heart-related diagnoses. Additionally, the algorithm was able to discern the heart-related diagnoses from one another: dividing acute myocardial infarction (410) from the more chronic and congenital diseases (414, 424). The algorithm may have discerned that words representing severe prognoses or procedures, such as "angioplasty", "emergency", and "death" were more correlated with acute myocardial infarction.

[3] further experimental details in appendix

Unsupervised Learning								
Features	$TFIDF_{40}$		TM_5		TM_{20}		TM_{30}	
Algorithms	K-mean	PDD	K-mean	PDD	K-mean	PDD	K-mean	PDD
Acc.	0.49	0.50	0.59	**0.78**	0.56	0.72	0.58	0.70
Balanced Acc.	0.48	0.45	0.62	**0.78**	0.50	0.74	0.51	0.73
Precision	0.48	0.75	0.58	**0.84**	0.47	0.73	0.50	0.73
Recall	0.49	0.45	0.62	**0.78**	0.50	0.74	0.51	0.73
Weighted F1	0.42	0.41	0.57	**0.78**	0.54	0.72	0.56	0.71
Avg. F1	0.44	0.38	0.57	**0.78**	0.48	0.71	0.50	0.70
Supervised Learning								
Features	$TFIDF_{40}$		TM_5		TM_{20}		TM_{30}	
Algorithms	RF	CNN	RF	CNN	RF	CNN	RF	CNN
Acc.	0.82	**0.84**	0.66	0.67	0.74	0.72	0.74	0.73
Balanced Acc.	0.81	**0.85**	0.62	0.62	0.72	0.70	0.71	0.70
Precision	0.82	**0.84**	0.64	0.67	0.74	0.72	0.74	0.73
Recall	0.81	**0.84**	0.62	0.67	0.71	0.72	0.71	0.73
Weighted F1	0.82	**0.84**	0.65	0.66	0.74	0.72	0.73	0.72
Avg. F1	0.82	**0.84**	0.63	0.67	0.72	0.72	0.72	0.73
AUC.	0.95	**0.96**	0.87	0.88	0.91	0.90	0.91	0.91

Table 1: Experimental Result Comparison.

PDD Knowledge Base

Knowledge Space				Pattern Space									Data Space
				Attributes (i.e. Topics in this study)									
PC	Group	SubGroup	Residual	ICD9	Topic 0	Topic 1	Topic 2	...	Topic 16	Topic 17	Topic 18	Topic 19	Records ID
1	1	1	19.76	424	[0.01 0.42]	[0.03 0.54]	[0.03 0.44]	...					#1, #9, #13,...
1	1	2	9.39	410	[0.01 0.42]		[0.03 0.44]	...		[0.07 0.45]			#2, #4, #5, #7,...
1	1	3	26.59	414	[0.01 0.42]		[0.03 0.44]	...					#3, #6, #16,...
1	2	1	50.27	38	[0.00 0.01)	[0.00 0.01)	[0.00 0.03)	...	[0.00 0.02)		[0.00 0.01)		#9, #12, #16,...
2	1	1	24.46	424	[0.01 0.42]		[0.00 0.03)	...	[0.02 0.05)			[0.02 0.04)	#1, #9, #13,...
2	1	2	33.81	414	[0.01 0.42]	[0.03 0.54]	[0.00 0.03)	...	[0.02 0.05)		[0.01 0.03)	[0.02 0.04)	#3, #6, #16,...
2	2	1	15.28	410		[0.00 0.01)	[0.03 0.44]	...					#2, #4, #5, #7,...

Note: PC=Principal Component; Group=Attribute Value Group; SubGroup = Attribute Value Sub-Group;

Figure 2: The PDD Knowledge Base when Top 20 topics are used as input.

4.3 Discussion on Interpretability

Figure 2 shows the partial knowledge base on 20 topics dataset. As same with the above results, in the first principal component, two opposite groups are discovered: one where ICD9=4XX (heart diseases), and the other where ICD9 = 038 (septicemia). But the difference is that three subgroups (i.e. 424, 414, 410) are further detected related to three different ICD9 codes. The discovered significant patterns are summarized for 20 topics as below.

ICD9=424 (diseases of the endocardium) and 414 (chronic ischemic heart disease) shows similar patterns, for example: i) **high** probabilities appear in the topics 1,2(Cardiovascular/Surgery),5,16; ii) and topics with **low** probabilities are topics 6, 7 (Status/Consciousness), 8 (Lung disease), 9. ICD9=038 (septicemia) shows opposite patterns, for example: i) topics with **high** probabilities are topics 3, 4 (Intensive care/Infection), 7 (Status/Consciousness), 8 (Lung disease); ii)and **low** probabilities appear in the topics 0(Heart anatomy) 1, 2 (Cardiovascular/Surgery), 5, 12 (Cardiovascular), 16, 18.

5 Conclusion

In this work, we propose a novel two-step algorithm, using interpretable NLP features with unsupervised pattern discovery to solve clinical text analysis. PDD performs better than K-means, especially when applied to the dataset extracted by topic modeling. Clustering results of PDD based on the discovered patterns may reflect the functional sources of the original dataset instead of class labels. In addition, our method is a global interpretable white-box model (from the input, throughput to the output) to provide an explainable All-in-One Knowledge Base (KB) that synchronizes self-correcting classification and clustering results in summarized/comprehensive forms to provide interpretability and traceability.

References

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Roselie A Bright, Summer K Rankin, Katherine Dowdy, Sergey V Blok, Susan J Bright, and Lee Anne M Palmer. 2021. Finding potential adverse events in the unstructured text of electronic health care records: Development of the shakespeare method. *JMIRx Med*, 2(3):e27017.

Jinying Chen, John Lalor, Weisong Liu, Emily Druhl, Edgard Granillo, Varsha G Vimalananda, and Hong Yu. 2019. Detecting hypoglycemia incidents reported in patients' secure messages: using cost-sensitive learning and oversampling to reduce data imbalance. *Journal of medical Internet research*, 21(3):e11990.

Zhengxing Huang, Wei Dong, and Huilong Duan. 2015. topic model for clinical risk stratification from electronic health records. *Journal of Biomedical Informatics*, 58:28–36.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

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. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:16.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Been Kim. 2021. Interpretability.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. 2019. Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2690.

Kimberly J O'malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2):1620–1639.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Alexis Ross, Ana Marasović, and Matthew E Peters. 2020. Explaining nlp models via minimal contrastive editing (mice). *arXiv preprint arXiv:2012.13985*.

Kaiser Sun and Ana Marasović. 2021. Effective attention sheds light on interpretability. *arXiv preprint arXiv:2105.08855*.

Betty Van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A Gers, and Alexander Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. *arXiv preprint arXiv:2102.04110*.

Yanshan Wang, Yiqing Zhao, Terry M Therneau, Elizabeth J Atkinson, Ahmad P Tafti, Nan Zhang, Shreyasee Amin, Andrew H Limper, Sundeep Khosla, and Hongfang Liu. 2020. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. *Journal of biomedical informatics*, 102:103364.

Andrew KC Wong, Ho Yin Sze-To, and Gary L Johanning. 2018. Pattern to knowledge: Deep knowledge-directed machine learning for residue-residue interaction prediction. *Scientific reports*, 8(1):1–14.

Andrew KC Wong, Pei-Yuan Zhou, and Zahid A Butt. 2021. Pattern discovery and disentanglement on relational datasets. *Scientific reports*, 11(1):1–11.

Die Zhang, Huilin Zhou, Hao Zhang, Xiaoyi Bao, Da Huo, Ruizhao Chen, Xu Cheng, Mengyue Wu, and Quanshi Zhang. 2021. Building interpretable interaction trees for deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14328–14337.

Pei-Yuan Zhou, Gary CL Li, and Andrew KC Wong. 2016. An effective pattern pruning and summarization method retaining high quality patterns with high area coverage in relational datasets. *IEEE access*, 4:7847–7858.

Pei-Yuan Zhou and Andrew KC Wong. 2021. Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement. *BMC medical informatics and decision making*, 21(1):1–15.

A Materials and Methods

An EHR is a digital collection of medical information about a person, which includes information about a patient's health history, such as diagnoses, medicines, tests, allergies, immunizations, and treatment plans. The MIMIC-III (Medical Information Mart of Intensive Care) is an openly available extensive database comprising de-identified information relating to patients admitted to critical care units at a large tertiary care hospital (Johnson et al., 2016). Data primarily stores both structured (e.g. MIMIC-III medications, laboratory results are stored in the table with columns as features and rows as records) and unstructured data (e.g. MIMIC-III clinical notes, discharge summaries are stored in the format of free text). The discharge summary of patients is free text, thus making interpreting it a challenge. Hence, the first step is transforming free text into a structured dataset formatting as a table with columns as features and rows as records. The second step is discovering patterns and grouping patients' records based on patterns in an unsupervised manner.

We presented the detailed steps of the proposed algorithm as below (Figure 1).

A.1 Feature Extraction

Topic modelling (Jelodar et al., 2019) is described as a method for finding a group of words (i.e topic) from a collection of documents that best represents the information in the collection. Hence, we extract features from the clean dataset using topic modelling. The value of the features is represented by the probabilities of topics occurring in the records. Labels are then merged with the features for unsupervised exploration; in this case, the label is the ICD9 code - the diagnostic code indicating categories of disease. We use LDA (Latent Dirichlet Allocation) for the topic model because it identifies topics best describing distinct subsets of documents within a corpus (Jelodar et al., 2019). To determine the ideal number of topics, we choose the optimal number of topics by computing the coherence of the topic cluster instance (Röder et al., 2015). We find that the coherence score peaks when the number of topics is 5, 20, and 30 - and therefore we create topic models with those respective parameters. The output of our coherence scores is shown as Figure 3.

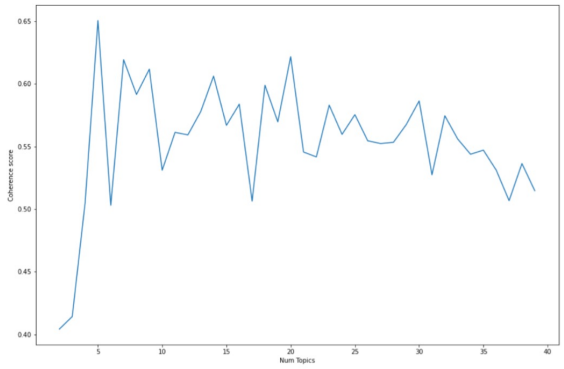

Figure 3: Optimal number of topics by coherence of the topic cluster

A.2 Pattern Discovery and Disentanglement

After preprocessing and extracting features from the text, the dataset has been transformed into a structured table of patients' records in rows and features in columns, which is represented as a $M \times N$ matrix, where M represents the number of patients' records and N represents the number of extracted features [4].

A.2.1 Discretize Numerical Feature Values

The output matrix in the last step contains probabilities of topics or extracted words, which are all numerical values. Due to infinite degrees of freedom of numerical features, it is hard to correlate features with the target variable and interpret the associations. Hence, we discretize features into event-based/discrete features. To detect event-based patterns, we convert the values of numerical features into categorical features by using the Equal Frequency discretization which distributes the values into equal size bins, so that numerical feature values are converted into discrete values referred to as "feature value" (meaning the discrete value for that feature). To be consistent with the study of PDD (Wong et al., 2021), we use the term Attribute Value (AV) instead.

A.2.2 Association Disentanglement

In order to measure the association between a pair of AVs (i.e. certain values of one attribute co-occurs with the value of another attribute), we use the statistical measure of adjusted standardized residual, abbreviated by SR, to represent the statistical weights of the AV pair, which is denoted as SR($AV_1 \leftrightarrow AV_2$) (shorten as SR($AV_{12}$)) and

[4] In pattern discovery, we use the term attribute instead of feature.

calculated by Eqn. (1) below.

$$SR(AV_{12}) = \frac{Occ(AV_{12}) - Exp(AV_{12})}{\sqrt{Exp(AV_{12})}} \times (1 - \frac{Occ(AV_1)}{T}\frac{Occ(AV_2)}{T}) \quad (1)$$

where $Occ(AV_1)$ and $Occ(AV_2)$ are the number of occurrences of AV; $Occ(AV_{12})$ is the total number of co-occurrence for two AVs in a AV pair; and $Exp(AV_{12})$ is the expected frequency and T is the total number of records.

An association matrix, treated as a vector space, is then generated to represent the strength of associations between each pair of AVs. Each row of the matrix, corresponding to a distinct AV, represents an AV-vector with SRs between that AV associated with all other AVs corresponding to the column vectors as its coordinates. We call the matrix the SR Vector Space (SRV). SRV is an *N* dimensional vector space consisting of *N* distinct AV-vectors.

We then use PCA to decompose SRV (Wong et al., 2021) (Wong et al., 2018) into principal components to reveal AV associations orthogonal to others AV associations, i.e. PC=PC_1, PC_2,... PC_k which are ranked according to the weights of the associations (eigenvalues). We then reproject the projections of AV-vectors on the principal components onto the SRV again, to obtain a set of reprojected-SRVs (abbreviated by RSRV). We refer to the PC together with its RSRV as a disentangled space.

The above process is called *Pattern Disentanglement* which allows us to take the reprojected components/vectors from PCA and use the reprojected values as new measurements/criteria to represent the strength of associations between AVs in different orthogonal disentangled spaces.

A.2.3 Pattern Clustering

In an RSRV, after screening in the statistical residual values (referred to as RSR) greater than 1.96, only the significant pairs of AV associations remain. Statistically, under the null hypothesis that the two AVs are independent, the adjusted residuals will have a standard normal distribution. So, an adjusted residual that is more than 1.96 (2.0 is used by convention) indicates the association is significantly greater than what would be expected (with a significance level of 0.05 or 95% confidence level) if the hypothesis were true. We can also set a threshold as 1.44 with 85% confidence, or 1.28 with 80% confidence level.

As an unsupervised learning approach, on each RSRV, we generate AV groups such that each group contains a set of AVs. We build the set of AVs up iteratively by adding AVs that are associated with AVs in the set. That is to say, an AV (e.g., AV_i) that is significantly associated with another AV (e.g. AV_j) in the group will join the group, otherwise, a new AV group is generated for AV_i. Theoretically, in one projected principal component, usually two AV groups on the opposite sides are generated as two opposite groups. When such opposite groups do not exist, we may obtain AV groups only on one side of the PC. The output of this step is one or two AV groups, and each group contains a set of AVs.

Furthermore, to obtain detailed separated groups, several AV subgroups can be generated for each AV group using a similarity measure such that the similarity between two AV subclusters is specified as the percentage of the overlapping records covered by each AV subcluster. We denote each AV subgroup by a three-digit code [#PC, #Group, #SubGroup]. The AV groups or subgroups can reveal the characteristics of the records at specific groups with disentangled patterns to provide statistical evidence for further clustering or prediction. Furthermore, patient record groups are obtained according to their specific characteristics (disentangled patterns) discovered in the AV groups or subgroups.

Traditional pattern clustering algorithm (Zhou et al., 2016), without PCA, can group patterns based on their "similarity", which is limited and time-consuming. In this case, after disentanglement and generating AV groups/subgroups, only a few AVs remain to be candidate patterns, which can reduce time consumption when high-order patterns are growing. The high-order pattern describes a statistically significant association among more than two AVs.

A.2.4 Pattern Discovery

So far, each AV subgroup contains a set of AVs considered as candidate patterns. We then test the candidates from order > 2 (i.e. consisting of more than 2 AVs) to high order sets to determine their pattern status. Hence, we obtain a compact set of patterns which are statistically significant and interpretable. Hence PDD reduces the computational

complexity drastically and produces very small and succinct pattern sets for interpretation and tracking. The disease related record groups of patients can then be explicitly revealed.

A.3 Output

The output of PDD is organized into an all-in-one representational framework known as PDD Knowledge Base. It consists of three parts: a Knowledge Section showing the hierarchical clusters such that each cluster unveil distinct characteristics of a related group of records; a Pattern Section listing the discovered patterns showing detailed associations between AVs; and the Data Section listing the record ID's, the knowledge source and pattern(s) associated with each patient by linking the patient to the Knowledge and Pattern Sections.

B Parameter Setting

To classify the dataset, the data were split into 70% training and 30% for testing. We used default parameter settings for K-means and random forest available in sklearn package for Python 3.0.

For CNN (LeCun et al., 1995), we trained a CNN model with the input layer as a reshaped cleaned dataset with probabilities of topics or extracted words and ICD9 labels. The architecture is as follows: a 1D CNN layer, followed by batch normalization, then a dropout layer for regularization (Li et al., 2019), and finally a 1D max-pooling layer. After the CNN and pooling, the learned features are flattened to one long vector and passed through a fully connected layer before the output layer for prediction. We used the Adam optimizer with a learning rate of 0.001 trained on 25 epochs with a batch size of 32.

C Additional Experimental Results

In the knowledge base shown as Figure **??**, the first three columns show the knowledge space, which describes clustering results of PDD and statistical measurement of each pattern. The clusters are identified by a three-digital code [#PC, #Group, #Subgroup] (PC: Principal Component, Group: pattern groups in the same principal component, Subgroup: pattern Sub-group in the same pattern group). We observe that, in the first principal component, two opposite groups are discovered: one where ICD9=4XX, and the other where ICD9 = 038. All ICD9=4XX are diseases related to heart disease, while ICD9=038 is related to Septicemia, so these are two opposite groups. Then in the second principal component, ICD9=424 (diseases of the endocardium) was separated, still showing opposite patterns with ICD9=38. Finally, in the third principal component, ICD9=424 was separated from ICD9=410 (acute myocardial infarction). To be more specific, the unveiled knowledge can be summarized below. ICD9=424 (diseases of the endocardium), 414 (chronic ischemic heart disease), and 410 (acute myocardial infarction) show similar patterns. For example, **low** probabilities appear in the topic0 (Medication). ICD9=424 and 414 show more closed patterns compared to 410 (acute myocardial infarction). For example, **low** probabilities appear in the topic4 (Intensive Care/Infection). And ICD9=38(septicemia) shows opposite characteristics compared to ICD9=4XX. For example, **high** probabilities appear in topic 0 (Medication); **low** probability appears in topic2 (Cardiovascular 2); and **high** probabilities appear in topic4 (Intensive Care/Infection). The data space shows the IDs of the records that are covered by the patterns. For example, the first association pattern listed in the first row of the knowledge base can be covered by the records with ID = 2,11,44,53,63, and so on. And all the above records belong to the group labeled as ICD9=410, which is the same as the discovered pattern

D Limitations

This study has the following limitations. First, to prove the concept of the PDD algorithm, only records with the four most common ICD9 codes are selected. Second, PDD, used as an interpretable clustering algorithm in this study, accepts limited selected features. When too many features are included, acquired data leads to high time complexity, and overwhelming pattern number and redundancy, making interpretability very difficult. For future work, we will enlarge the dataset and the number of features to investigate their impact on the performance of the algorithm. Finally, as the predicted label is ICD9 code, we presume it to be the ground truth for diagnosis. However, ICD9 is used for billing purposes and therefore may not accurately reflect a patient's true condition (O'malley et al., 2005).