DOI:
10.1109/JBHI.2021.3129461

Citation for published version (APA):
Zhong, H., Loukides, G., & Pissis, S. P. (Accepted/In press). Clustering demographics and sequences of diagnosis codes. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2021.3129461

Clustering demographics and sequences of diagnosis codes

Haodi Zhong, Grigorios Loukides, *Member, IEEE*, Solon P. Pissis

***Abstract*—A Relational-Sequential dataset (or RS-dataset for short) contains records comprised of a patient's values in demographic attributes and their sequence of diagnosis codes. The task of clustering an RS-dataset is helpful for analyses ranging from pattern mining to classification. However, existing methods are not appropriate to perform this task. Thus, we initiate a study of how an RS-dataset can be clustered effectively and efficiently. We formalize the task of clustering an RS-dataset as an optimization problem. At the heart of the problem is a distance measure we design to quantify the pairwise similarity between records of an RS-dataset. Our measure uses a tree structure that encodes hierarchical relationships between records, based on their demographics, as well as an edit-distance-like measure that captures both the sequentiality and the semantic similarity of diagnosis codes. We also develop an algorithm which first identifies *k* representative records (centers), for a given *k*, and then constructs *k* clusters, each containing one center and the records that are closer to the center compared to other centers. Experiments using two Electronic Health Record datasets demonstrate that our algorithm constructs compact and well-separated clusters, which preserve meaningful relationships between demographics and sequences of diagnosis codes, while being efficient and scalable.**

***Index Terms*—Clustering, Demographics, Diagnosis codes**

TABLE I: A (toy) example of an RS-dataset. *Age*, *Gender*, and *Ethnicity* are demographic attributes; the *Diagnosis codes sequence* attribute is comprised of ICD-9 codes.

Age	Gender	Ethnicity	Diagnosis codes sequence
69	M	Black	(414.01, 250.00, 272.4, 401.9, 412, 696.1)
1	F	White	(765.18, 774.2, 765.27, 769)
67	M	Black	(414.01, 4111, 272.1, 250.00, 401.9)
48	F	White	(441.2, 401.9, 345.90, 414.01)
50	F	White	(414.01, 250.01, 401.9, 412, 2720)
0	F	White	(765.19, 769, 774.2, 779.3, 765.28, 771.7)
61	F	White	(414.01, 424.0, 440.21, 427.89, 250.00, 401.9)
1	F	White	(765.16, 775.6, 765.27, 769)
68	M	Black	(414.01, 411.1, 250.01, 401.9, 272.0)

TABLE II: A clustered RS-dataset produced from the dataset in Table I by our algorithm. *Cluster ID* is for reference.

Cluster ID	Age	Gender	Ethnicity	Diagnosis codes sequence
1	69	M	Black	(414.01, 250.00, 272.4, 401.9, 412, 696.1)
1	67	M	Black	(414.01, 411.1, 272.1, 250.00, 401.9)
1	68	M	Black	(414.01, 411.1, 250.00, 401.9, 272.0)
2	1	F	White	(765.18, 774.2, 765.27, 769)
2	0	F	White	(765.19, 774.6, 765.28, 779.3, 769, 771.7)
2	1	F	White	(765.16, 774.2, 765.27, 769)
3	48	F	White	(441.2, 414.01, 345.90, 440.32, 250.00, 401.9)
3	50	F	White	(414.01, 440.31, 250.01, 401.9, 412, 272.0)
3	61	F	White	(414.01, 424.0, 440.21, 427.89, 250.00, 401.9)

I. Introduction

Electronic Health Records (EHRs) contain a wealth of information (e.g., demographics, diagnoses, medications, and laboratory results) about many patients over long time periods. Data mining techniques are increasingly being employed on EHR data to derive actionable knowledge from such information [1]–[4], which can guide clinical decision support [5], public health monitoring [6], and patient treatment [7].

This work considers clustering, a fundamental data mining task aiming to "partition a set of data points into natural groups, called clusters, such that points within a group are very similar" (i.e., clusters are compact) and "points between different groups are as dissimilar as possible" (i.e., clusters are well-separated) [8]. We focus on clustering an EHR dataset in which every record contains a patient's demographics and a sequence of the patient's diagnosis codes. Such a dataset is termed Relational Sequential dataset (or RS-dataset for short) since demographics are modeled as relational attributes [9]. An example of an RS-dataset is in Table I. The first record corresponds to a 69-year old black male patient who is associated with six diagnoses (ICD-9 codes) [10]: first with 414.01 (coronary atherosclerosis of native coronary artery), then with 250.00 (diabetes mellitus type II without complications), and next with 272.4, 401.9, 412 and 696.1. Clustering an RS-dataset aims to construct clusters with similar values in demographics and also similar diagnosis codes that occur in similar order. An example of a clustering of the dataset in Table I is in Table II. The first cluster (records with *Cluster ID* 1) represents black males over 60 sharing the sequence (414.01, 250.00, 401.9).

After clustering an RS-dataset, one can: (I) discover trends from each cluster, such as disease progression for patients with similar demographics, (II) compare trends across clusters, which could improve diagnosis and treatment decision making, as well as epidemiological analysis and research [11], [12],

H. Zhong and G. Loukides are with King's College London, London, UK. (e-mail: h.zhong@kcl.ac.uk; grigorios.loukides@kcl.ac.uk).

S. P. Pissis is with CWI and the Vrije Universiteit, The Netherlands. (e-mail: solon.pissis@cwi.nl).

or (III) visualize the clusters [13], [14]. Furthermore, one can discover frequently occurring temporal condition patterns [15] from each cluster, which may inform research into causation or other associations [15]. Moreover, clustering an RS-dataset can be applied before: (I) classification to improve the accuracy of a classification model [16], (II) anonymization to enhance data utility [17], or (III) clinical pathway mining [18] to extract pathways for distinct types of patients.

An RS-dataset contains two fundamentally different types of data, namely demographics that are modeled as relational attributes, as well as a sequence of diagnosis codes. Thus, its clustering is particularly challenging. In fact, as it will be explained in Section II, existing clustering algorithms (e.g., that of [5] which clusters a relational dataset comprised of demographics and clinical information) are inappropriate to address this task. Also, it is inappropriate to first convert an RS-dataset into a dataset that can be clustered with an existing clustering algorithm and then clustering it using that algorithm.

Motivated by the usefulness of the task of clustering RS-datasets and the ineffectiveness of existing methods to address it, we propose the first approach for this task.

Our work makes the following specific contributions:

1. The task of clustering an RS-dataset is formalized as a k-center [19], [20] optimization problem.

2. A new distance measure is proposed. The distance measure captures the pairwise similarity between records of an RS-dataset, based on their demographics and sequences of diagnosis codes, in a unified manner. The similarity with respect to demographics is captured using a tree structure that encodes hierarchical relationships between records based on their demographics. The similarity with respect to sequences of diagnosis codes is captured by an edit-distance-like measure that we develop to account for the semantic similarity of diagnosis codes, as specified by well-defined taxonomies.

3. A new clustering algorithm is designed. The algorithm first identifies k records (centers) representing clusters, for a given k, and then constructs k clusters, each containing a center and the records that are closer to this center, with respect to our distance measure. Our algorithm finds centers that are no more than two times worse than the best possible k centers.

4. Experiments, using two EHR datasets, are conducted to show that our algorithm constructs clusters which are: (1) compact (two times more compact on average compared to clusters constructed by state-of-the-art algorithms [9], [21] that we adapt to cluster RS-datasets); (2) well-separated (different clusters contain different values in Age, Gender, and Ethnicity, as well as different frequent sequences of diagnosis codes); and (3) able to preserve meaningful patterns that are documented in the medical literature. Our algorithm is also shown to be efficient and scalable with respect to the number of records in the RS-dataset and the number of clusters.

5. The clusters constructed by our algorithm can be used in analytic tasks, including visualization, classification, clinical pathway mining, and trend discovery. These tasks could help practice review and support decisions and potentially improve patient treatment and care. Also, our work implies the need for developing new methods for clustering complex EHR data.

II. Related Work

Data mining techniques that are being applied to EHR data include sequential pattern mining [22], classification [23], and clustering [24]. For example, sequential pattern mining and clustering have been employed in the task of clinical pathway mining (see e.g., [18]). In the following, we focus on clustering and refer readers to [2] for a survey on EHR data mining.

There is a very large number of clustering algorithms [24] such as k-means [25], hierarchical clustering [26], k-medoids [27], and DBSCAN [28] (see [24] for a survey). There is also much work on EHR data clustering [4], [9], [29], [30], including works for clustering clinical and/or demographic attributes [5]. However, existing clustering algorithms are inappropriate to cluster RS-datasets, as: (I) they assume an input dataset that contains a single attribute type (e.g., atomic or set-valued [9]); and (II) their similarity measures cannot be directly applied to records with attributes of multiple types [8]. Similarly, multiobjective clustering algorithms are inappropriate to cluster RS-datasets because it is difficult to combine similarity measures for demographics with similarity measures for diagnosis codes [17].

One may naturally wonder whether an RS-dataset can be first transformed so that it contains a single attribute type and then clustered using an existing algorithm. For example, a sequence of diagnosis codes in an RS-dataset can be transformed into a set of atomic attributes, each containing the frequency of a q-gram (i.e., a substring of q letters) of the sequence. Considering all distinct q-grams of all sequences and assuming a fixed order for them allows us to find a set of atomic attributes that is common to all sequences of diagnosis codes in an RS-dataset. These attributes can then replace the *sequence of diagnosis codes* attribute in the RS-dataset to transform it into a relational dataset that can be clustered with existing algorithms such as k-medoids [27]. However, this transformation inevitably incurs information loss which harms the quality of clustering (see Section VII). The same holds for transformations which represent demographics as a sequence. For example, applying one-hot encoding to the demographic values in a record of an RS-dataset, as in [9], creates a (binary) sequence to which we can append the sequence of diagnosis codes. This leads to a sequential dataset which can be clustered with existing algorithms (e.g., hierarchical clustering [26]). Yet, this transformation incurs fake ordering information which affects our ability to meaningfully measure similarity between records of the transformed dataset. Specifically, the values of demographics become ordered and precede diagnosis codes. However, such an artificial ordering has no semantic meaning as there is no natural ordering among attributes; only diagnosis codes can be ordered.

One may also wonder whether algorithms designed for RT-datasets [9], in which each record contains demographics and a set of diagnosis codes, are appropriate for clustering RS-datasets. For instance, the state-of-the-art algorithm [9] for clustering an RT-dataset, called MASPC (for Maximal-frequent All-confident pattern Selection with Pattern-based Clustering), works by: (I) projecting an RT-dataset on a number of carefully selected patterns (sets of demographic values and diagnosis codes), and (II) clustering the projected dataset based on a hierarchical clustering algorithm [26]. The patterns are selected so that they are frequent and comprised of correlated diagnosis codes. It is easy to cluster an RS-dataset using MASPC by: (I) keeping only one occurrence of every diagnosis code in every record of the RS-dataset (this converts an RS-dataset into an RT-dataset), (II) applying MASPC on the RT-dataset, and (III) replacing the set of diagnosis codes in every record of the clustered RT-dataset with the sequence of diagnosis codes of its corresponding record in the RS-dataset. However, the resultant clusters contain dissimilar records with respect to their diagnosis codes, as shown in Section VII. This is because MASPC does not consider multiple occurrences of diagnosis codes in a record nor the order in which these diagnosis codes occur in the record.

III. Background

In the following, we summarize the notation used throughout (see Table III) and introduce some preliminary concepts.

TABLE III: Table of notation.

Notation	Definition
$\mathbf{D_{RS}}$	RS-dataset
$A^i, i \in [1, l]$	i-th demographic
s	Sequence of diagnosis codes
r^{dem}	Projection of record r on demographics
r^{seq}	Projection of record r on sequence
d_{JC}, d_{WE}, d_{JCE}, d_J, d_{LCS}	Distance functions
$\mathbf{T_{RS}}$	RS-Tree
w_{dem}, w_{diag}	Weights

RS-dataset. An RS-dataset is denoted by $\mathbf{D_{RS}}$ and contains $l \geq 1$ demographic attributes, $A^1, \ldots, A^l$. Each demographic attribute can be numerical or categorical. Each record in $\mathbf{D_{RS}}$ is a vector containing l values, one in each demographic, and a sequence $s = (s[1]s[2] \ldots s[n])$ comprised of $|s| = n$ diagnosis codes. The projection of a record r on the set of demographic attributes (respectively, sequence s) is denoted by r^{dem} (respectively, r^{seq}). The diagnosis codes in a sequence are drawn from an alphabet $\sum$ and can be represented in different formats, e.g., as ICD-9 codes. In the latter case, $\sum$ is the set of all ICD-9 codes.

Semantic Distance. Jiang-Conrath (JC) distance [31] is a well-established measure to calculate the semantic distance between two concepts, represented as nodes in a tree. JC distance is based on information content (IC) [32], a measure of specificity which we define first. The information content of a node u in a tree with L leaves is defined as $IC(u) = -\log_2((\frac{\text{leafdesc}(u)}{\text{asc}(u)+1} + 1)/(L+1))$ [32], where $\text{asc}(u)$ (respectively, $\text{leafdesc}(u)$) is the number of ascendants (respectively, leaf-level descendants) of u. Thus, nodes close to leaves, which correspond to more specific concepts, get lower information content values.

The JC distance for two nodes u_i, u_j is denoted by $d_{JC}(u_i, u_j)$ and computed using (1):

$$d_{JC}(u_i, u_j) = (IC(u_i) + IC(u_j)) - 2\times\ IC(LCA(u_i, u_j)), \quad (1)$$

where $LCA(u_i, u_j)$ denotes the least common ancestor of u_i, u_j. For example, using the tree of Fig. 1, we have: $L = 9$,

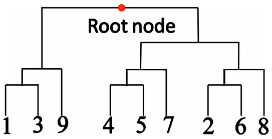


Fig. 1: An example tree.

$\text{leafdesc}(1) = \text{leafdesc}(4) = 0$, $\text{asc}(1) = 3$, $\text{asc}(4) = 4$, $IC(1) = IC(4) = -\log_2(\frac{1}{10})$, and $IC(LCA(1, 4)) = IC(root) = 0$. Thus, $d_{JC}(1, 4) = -2\ \log_2(\frac{1}{10}) \approx 6.64$.

Weighted Edit Distance. Edit distance [33] is commonly used to capture the distance between two sequences s_i, s_j and is expressed as the minimum number of element insertions, deletions, and substitutions needed to transform s_i to s_j. We employ a weighted version of edit distance from [34], which considers semantic similarity. It is denoted by $d_{WE}(s_j, s_i)$ and can be computed using (2):

$$d_{WE}(s_i, s_j) = \begin{cases} |s_i| & \text{if } |s_j| = 0 \\ |s_j| & \text{if } |s_i| = 0 \\ d_{WE}(\text{tail}(s_i), \text{tail}(s_j)) & \text{if } s_i[1] = s_j[1] \\ \min \begin{cases} d_{WE}(\text{tail}(s_i), s_j) + 1 \\ d_{WE}(s_i, \text{tail}(s_j)) + 1 \\ d_{WE}(\text{tail}(s_i), \text{tail}(s_j)) + \text{sub}(s_i[1], s_j[1]) \end{cases} & \text{otherwise.} \end{cases} \quad (2)$$

where $\text{tail}(s_i) = (s_i[2] \ldots s_i[|s_i|])$ and $\text{sub}(x, y) \in [0, 1]$ is the cost of substituting element x with y. d_{WE} differs from edit distance in that the substitution cost is given by sub() instead of being 1. Let $s_i = (a, b)$, $s_j = (a, c)$, $\text{sub}(a, a) = 0$ and $\text{sub}(b, c) = 0.5$. Then, $d_{WE}(s_i, s_j) = 0.5$ as $s_i[1] = s_j[1] = a$, and $d_{WE}(\text{tail}(s_i), \text{tail}(s_j)) = d_{WE}(b, c) = 0.5$.

IV. The RS-Tree and d_{JCE} Measure

A. RS-Tree $\mathbf{T_{RS}}$

An RS-Tree, denoted by $\mathbf{T_{RS}}$, encodes hierarchical relationships between records of an RS-dataset $\mathbf{D_{RS}}$ regarding demographics. $\mathbf{T_{RS}}$ is constructed in two steps: (I) The records in $\mathbf{D_{RS}}$ are partitioned into groups, each containing all records with the same values in all demographics. This is to quickly identify pairs of records that have distance zero based on demographics. (II) Agglomerative average-linkage hierarchical clustering [26] is applied to the groups obtained from step I. The distance between two groups is measured using the Hamming distance [8] computed over demographics. To apply Hamming distance, we discretize numerical demographic attributes (if any) [8].

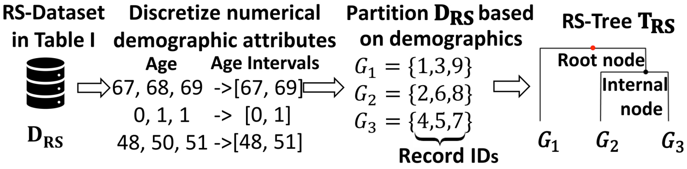


Fig. 2: RS-Tree Construction. Record ID x corresponds to the x-th record in Table I.

The leaves of an RS-Tree represent groups of records with the same demographics and the root a group comprised of all records. Fig. 2 illustrates an RS-Tree constructed from the RS-dataset in Table I, after discretizing *Age*. The three groups of records (leaves in the RS-Tree) are created in step I.

An RS-Tree can be used to measure the distance between two records based on their demographics without requiring user input. This is an important benefit over existing distance measures [17], which need users to set several data-dependent parameters. Also, RS-Tree can consider the hierarchical information of all demographic attributes, which can provide better clustering results as shown in Section VII.

B. The d_{JCE} Measure

Our measure, d_{JCE} (for Jiang-Conrath Edit distance), captures the distance between two records in an RS-Dataset based on both their demographics and sequences of diagnosis codes, in a unified manner. d_{JCE} is computed for a pair of records, r_i, r_j, based on the JC-distance for demographics and the weighted edit distance for sequences of diagnosis codes, as shown in (3):

$$d_{JCE}(r_i, r_j) = \sqrt{w_{dem} \cdot d_{JC}(r_i^{dem}, r_j^{dem}) + w_{diag} \cdot d_{WE}(r_i^{seq}, r_j^{seq})} \quad (3)$$

where w_{dem}, w_{diag} are weights trading off the importance of d_{JC} and d_{WE} in the computation of d_{JCE}. The distance d_{JC} is computed using the RS-Tree $\mathbf{T_{RS}}$ that is constructed from $\mathbf{D_{RS}}$, as explained in Section IV-A.

To effectively use d_{JCE} in our context, we modify it in two ways. First, we define the substitution cost function sub() in d_{WE} to reflect the semantic distance between diagnosis codes. Specifically, for any two ICD-9 codes i, j, we set:

$$\text{sub}(i, j) = d_{JC}(u_i, u_j) / \max_{u_q, u_p \in \mathbf{H}} d_{JC}(u_q, u_p) \quad (4)$$

where u_i and u_j are the nodes in the standard ICD-9 code hierarchy $\mathbf{H}$ [35] that correspond to i and j, respectively. This gives a zero substitution cost when $i = j$ and a smaller cost for semantically similar diagnosis codes. For example, $\text{sub}(401.9, 414.01) < \text{sub}(401.9, 250.00)$ as the first two codes are closer with respect to the ICD-9 hierarchy, since they both represent diseases of the circulatory system. The substitution cost function in d_{WE} captures that two sequences with semantically similar diagnosis codes are more similar. This is not captured by edit distance, which penalizes every pair of different ICD-9 codes equally. Second, we modify d_{JCE} to avoid bias in favor of d_{JC} or d_{WE} by: (I) normalizing d_{JC} (respectively, d_{WE}) by dividing with its maximum possible value $\max_{r_i, r_j \in \mathbf{D_{RS}}} d_{JC}(r_i^{dem}, r_j^{dem})$ (respectively, $\max\left(|r_i^{seq}|, |r_j^{seq}|\right)$), and (II) selecting weights in (3) so that $w_{dem}, w_{diag} \in [0, 1]$ and $w_{dem} + w_{diag} = 1$. Normalization ensures that the values of d_{JCE} are in $[0, 1]$ and that $d_{JCE}(r_i, r_j) = 0$, if and only if $r_i^{dem} = r_j^{dem}$ and $r_i^{seq} = r_j^{seq}$. Of note, d_{JCE} is a metric, since d_{JC} and d_{WE} are metrics [36], which makes it generic enough for other uses (e.g., it can be incorporated into information retrieval algorithms used for clinical reasoning [37]).

V. Problem Definition

We define the following clustering problem:

Problem 1 (RS-Dataset k-Center (RSDC)): Given an RS-dataset $\mathbf{D_{RS}}$ and an integer $k > 0$, find a set of C of k records in $\mathbf{D_{RS}}$, referred to as centers, such that the distance $\max_{r \in \mathbf{D_{RS}}} \min_{c \in C} d_{JCE}(r, c)$ is minimized.

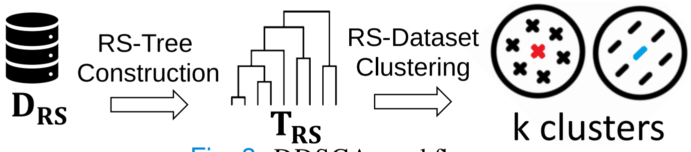


Fig. 3: DDSCA workflow.

```
Algorithm 1 DDSCA ($\mathbf{D_{RS}}$, $w_{dem}$, $w_{diag}$, $k$)
Input: Dataset $\mathbf{D_{RS}}$, $w_{dem}$, $w_{diag}$, the number of clusters $k$
Output: a set of clusters $U$
     // RS-Tree Construction Phase
 1:  $\mathbf{T_{RS}} \leftarrow$ Agglomerative-Average-Linkage($\mathbf{D_{RS}}$)
     // RS-Dataset Clustering Phase
 2:  $c \leftarrow$ arbitrary record from $\mathbf{D_{RS}}$
 3:  $U \leftarrow \{c\}$   // Set of clusters
 4:  $C \leftarrow c$   // Set of centers
 5:  while $|C| < k$ do
 6:      Select a record $c$ from $\mathbf{D_{RS}}$ such that $c \notin C$ and
         $\min_{c' \in C} d_{JCE}(c, c')$ is maximized
 7:      $U \leftarrow U \cup \{c\}$
 8:      $C \leftarrow C \cup c$
 9:  for $r \in \mathbf{D_{RS}}$ do
10:      if $r \notin C$ then
11:          Assign $r$ to the cluster in $U$ whose center $c$ has
             minimum $d_{JCE}(c, r)$
12:  return $U$
```

RSDC aims to find a set C of k centers such that the largest distance of any record that is not in C to its closest center is minimum. RSDC can be formulated as a k-Center problem with d_{JCE} as its objective function (see Supplementary Material Sec. I). After finding C, the final clustering of $\mathbf{D_{RS}}$ is obtained by adding each center to a different cluster and then adding each record that is not a center to the cluster where its closest center is (breaking ties arbitrarily).

We show that it is hard to solve RSDC optimally, or even to obtain a solution that is less than two times worse than the optimal (see Supplementary Material Sec. II).

VI. DDSCA Algorithm

Our Demographics and Diagnosis Sequences Clustering Algorithm (DDSCA) clusters an RS-dataset based on the RSDC problem. DDSCA works in two phases (see Fig. 3): (I) RS-Tree Construction; and (II) RS-Dataset Clustering, which is based on the algorithm of [20]. We now explain each phase (see Algorithm 1 for the pseudocode):

RS-Tree Construction Phase: In this phase (line 1), the RS-Tree is constructed from the input dataset $\mathbf{D_{RS}}$.

RS-Dataset Clustering Phase: The set C of k centers is constructed iteratively (lines 2-8), as follows. An arbitrarily selected record of $\mathbf{D_{RS}}$ becomes the first cluster in the set of clusters U and added into C (lines 2-4). Then, in each iteration, another record whose distance from its closest center is as large as possible is found and added into C and into U (lines 5-8). The process continues until k records are added into C. After that, k clusters are built (lines 9-11). Then, each record that is not in C is added into the cluster whose center is closest to it (line 11). Last, U is returned (line 12).

Despite its simplicity, DDSCA computes the best theoretically possible centers. Specifically, we prove that it computes a set C that is at most 2 times worse with respect to d_{JCE} than the optimal solution to RSDC (see Supplementary Material Sec. III). We also provide the time complexity of DDSCA in Supplementary Material Sec. IV.

VII. Experimental evaluation

A. Data and experimental setup

1) Data: Two RS-datasets, MIMIC [38] and INFORMS [39] were used. Each record in these datasets contains the demographics *Age*, *Gender*, and *Ethnicity* of a patient and a sequence with the patients' ICD-9 codes. A similar dataset to INFORMS but with sets instead of sequences of diagnosis codes was used in [9], [17]. Age in MIMIC and INFORMS was discretized using a standard hierarchy [40], in different ways (see Supplementary Material Sec. V). Table IV summarizes the characteristics of MIMIC and INFORMS.

TABLE IV: Datasets characteristics.

Dataset	$\lvert\mathbf{D_{RS}}\rvert$	# of demographics	$\lvert\Sigma\rvert$	Max # of diag. codes/record	Avg # of diag. codes/record
MIMIC	$37,730$	3	$5,558$	39	9.21
INFORMS	$26,630$	3	545	46	3.63

2) Competitors: We compared DDSCA, in terms of effectiveness and efficiency, against two state-of-the-art clustering methods that we adapted to cluster RS-datasets: AGC (Adaptive Graph Convolution) [21] and MASPC (Maximal-frequent All-confident pattern Selection with Pattern-based Clustering) [9]. AGC and MAPSC were chosen, as they outperformed deep-learning-based methods (e.g., [41]) and a hybrid algorithm along the lines of [17], respectively. Details on AGC and MASPC are in Supplementary Material Sec. VI.

3) Evaluation measures: For evaluating the quality of clusters based on demographics, we used the ASPJ (Average Sum of Pairwise Jaccard distance) measure, defined as follows: ASPJ $= \frac{1}{k}\sum_{c\in C}\sum_{r_i,r_j\in c} d_J(r_i^{dem}, r_j^{dem})$, where C is a set of k clusters and $d_J()$ is Jaccard distance [8]. For evaluating the quality of clusters based on diagnosis codes, we used the ASPWE (Average Sum of Pairwise Weighted Edit distance) measure, defined as follows: ASPWE $= \frac{1}{k}\sum_{c\in C}\sum_{r_i,r_j\in c} d_{WE}(r_i^{seq}, r_j^{seq})$. ASPJ and ASPWE are well-known internal clustering indices [24] measuring how similar are records in clusters with respect to demographics and sequences of diagnosis codes, respectively. We also used ASPLCS, which is similar to ASPWE but instead of d_{WE} is based on the Longest Common Subsequence (LCS) distance measure [18]. The latter is defined as $d_{LCS}(r_i^{seq}, r_j^{seq}) = |r_i^{seq}| + |r_j^{seq}| - 2 \cdot \ell(r_i^{seq}, r_j^{seq})$, where $\ell(r_i^{seq}, r_j^{seq})$ is the longest common subsequence of the diagnosis code sequences r_i^{seq} and r_j^{seq}. Small values in ASPJ, ASPWE and ASPLCS are preferred. Similar measures to ASPLCS but with other string distance functions were also used and analogous results were obtained (see Supplementary Material, Sec. VII). External clustering indices were not used, since our datasets do not have ground-truth clusters.

For evaluating the compactness of clusters, we defined a simple measure, called Average Cluster Compactness (ACC). Let $\mathbf{D_{RS}}$ be an RS-dataset that is clustered into a set C of k clusters and $\{u_1^i, \ldots, u_m^i\}$ be the set of values of an attribute A^i, $i \in [1, l]$, in these clusters. ACC is defined as follows:

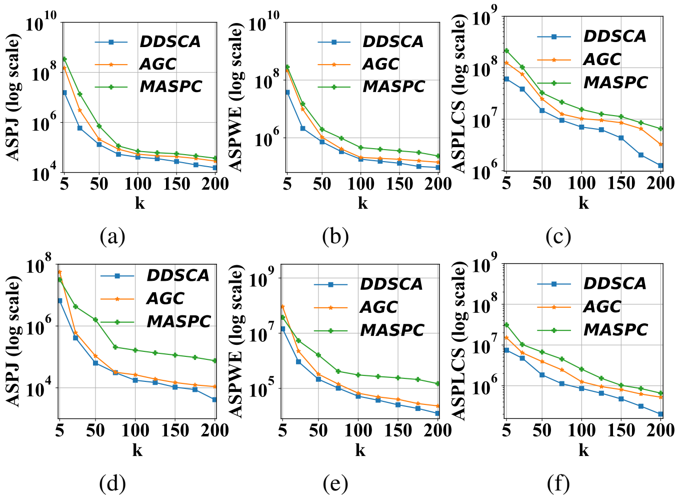
$$\text{ACC}(\{A^1, \ldots, A^l\}) = \frac{1}{k \cdot l}\sum_{i\in[1,l]}\sum_{c\in C}\frac{\max_{j\in[1,m]} RF(u_j^i, c)}{\sum_{j\in[1,m]} RF(u_j^i, c)}, \quad (5)$$

where c is a cluster and $RF(u_j^i, c)$ is the relative frequency of a value in A^i that is contained in c. Large values in ACC are preferred. For example, for a single attribute A^i, ACC$(\{A^i\})$ takes values in $(0, 1]$ and large values indicate that the most frequent value in clusters appears in many records of the clusters. Similarly, ACC$(\{A^i\}) = 1$ means that every cluster contains only one value in A^i and thus the clusters are as compact as possible with respect to this attribute.

4) Environment and code: We implemented DDSCA in Python 3 and used the Python implementations of AGC[1] and MASPC[2]. All experiments were performed on an Intel i9 at 3.70GHz with 64GB RAM. Our source code is available at `https://bitbucket.org/EHR_Clustering/ddsca/src/master/`. Unless stated otherwise, we used $k = 50$ (selected by the Elbow Method [42]) for all methods and $w_{dem} = w_{diag} = 0.5$ for our method. This weight configuration was chosen to treat demographics and diagnosis codes equally, as the competitors do. Also, it resulted in very compact clusters compared to alternative configurations (see Supplementary Material Sec. VIII). All parameters for AGC and MASPC were set to their default values from [9], [21].

B. Clustering effectiveness

Fig. 4 shows that DDSCA outperformed both competitors in all tested cases, according to all measures. For example, it created clusters with 184% (respectively, 167%) lower ASPWE (respectively, ASPLCS) and 178% lower ASPJ on average compared to AGC in MIMIC. This demonstrates that DDSCA created more compact clusters, which allow meaningful analyses based on demographics and diagnosis codes. AGC did not perform well because it has to transform sequences of diagnosis codes into vectors of q-gram frequencies (see Supplementary Material Sec. VI), which inevitably affects sequence similarity measurement and does not consider the semantic distance between diagnosis codes. For example, the sequences $(414.01, 250.00, 414.01)$ and $(250.00, 414.01, 250.00)$ are considered maximally similar by AGC, since they have the same 2-gram frequencies and hence the same vector representation. Yet, they are dissimilar when the ordering of diagnosis codes and their semantic similarity is considered. MASPC did not perform well because it operates on sets of diagnosis codes.



(a) (b) (c)

(d) (e) (f)

Fig. 4: (a) ASPJ, (b) ASPWE, and (c) ASPLCS vs. *k* for MIMIC. (d) ASPJ, (e) ASPWE, and (f) ASPLCS vs. *k* for INFORMS.

TABLE V: ACC for: (a) MIMIC and (b) INFORMS. ICD Chapter sequences are produced by replacing every ICD-9 code with its corresponding ICD Chapter.

Methods	ACC (*Age*)	ACC (*Gender*)	ACC (*Ethnicity*)	ACC (*ICD Chapter seq.*)	ACC (all attributes)
DDSCA	0.725	0.924	0.749	0.893	0.823
AGC	0.412	0.432	0.514	0.341	0.425
MASPC	0.315	0.354	0.415	0.385	0.367

(a)

Methods	ACC (*Age*)	ACC (*Gender*)	ACC (*Ethnicity*)	ACC (*ICD Chapter seq.*)	ACC (all attributes)
DDSCA	0.595	0.847	0.805	0.831	0.769
AGC	0.334	0.563	0.601	0.423	0.480
MASPC	0.336	0.412	0.445	0.407	0.400

(b)

C. Compactness of clusters

Table V shows that, when applied to either MIMIC or INFORMS, DDSCA creates much more compact clusters than the competitors with respect to all attributes and also with respect to each of the attributes separately. For example, the ACC of DDSCA over all attributes was 1.94 (respectively, 1.6) times better than that of AGC, the best competitor, in MIMIC and (respectively, INFORMS). The reasons that AGC and MASPC do not perform well are the same as those discussed in the experiment above.

Fig. 5 shows the distribution of values in each demographic and the distribution of ICD Chapters [3] [43] in ten clusters constructed by applying DDSCA with $k = 50$ on MIMIC. The reported clusters were selected randomly among the best 25 clusters with respect to ACC, computed over all attributes. For the results of all clusters, see Supplementary Material Sec. IX. The corresponding results of AGC and MASPC are not reported as they were much worse, as expected by the ACC measures for these algorithms. Most records in each cluster have the same value in each demographic attribute, and their top-2 frequent diagnosis codes belong to two ICD Chapters. This implies that clusters are compact. For example, in Cluster 1, 93% of patients are over 80 years old, 100% are male, and 99% are white. Meanwhile, 98% of patients in Cluster 1 have at least one diagnosis code in Chapter 7 (Diseases of the Circulatory System), and 71% have at least one diagnosis code in Chapter 3 (Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders). Similar results were obtained for INFORMS (see Supplementary Material Sec. IX).

D. Separability of clusters

Table VI shows that, when applied to MIMIC, DDSCA created well-separated clusters (i.e., clusters which differ with respect to their most frequent values in demographics and ICD Chapters, and with respect to frequent sequential patterns mined from them). For example, Clusters 4, 5, and 6 have the same most frequent values in Age, Gender, and Ethnicity but are different with respect to their top-3 frequent sequential patterns. Well-separated clusters are more interpretable (e.g., they can be concisely described based on frequent patterns [9]). Note that some patterns appear in more than one clusters. This is expected because they are comprised of diagnosis codes associated with patients that have different demographics. For example, the pattern (401.9, 427.31) which appears in four clusters implies that many middle-aged, or older patients, in each cluster were associated first with hypertension and then with atrial fibrillation. This is expected because hypertension is a risk factor of atrial fibrillation, and these diagnoses are common among middle-aged and aged patients [44]. Similar results to those of Table VI for INFORMS are in Supplementary Material Sec. IX. The competitors created significantly less separable clusters compared to our method (e.g., their top-3 frequent sequential patterns have very low frequency), so we do not report detailed results for them.

E. The medical relevance of top-3 frequent sequential patterns in clusters

Having shown that our algorithm outperforms the competitors, we now show that it creates clusters in which frequent sequential patterns capture relationships between diagnosis codes that are documented in the medical literature. Specifically, we discuss the top-3 frequent sequential patterns in the clusters of Table VI that were created by our method (see Supplementary Material Sec. IX for results on INFORMS). These patterns were discovered using the algorithm in [45]. Since the clusters are also compact and well-separated, they allow discovering potentially useful patterns associated with patients with similar demographics.

Cluster 1: Atrial fibrillation (427.31) that appears in all patterns frequently co-exists with congestive heart failure (428.0) [46], or with hypertension (401.9) [44], or with coronary artery disease (such as "coronary atherosclerosis of native coronary artery" denoted by 414.01) [47].

Cluster 2: Hypertension (401.9) frequently co-exists with coronary artery disease (414.01) [48], or with atrial fibrillation (427.31) [44], while coronary artery disease (414.01)

[1] https://github.com/karenlatong/AGC-master

[2] https://bitbucket.org/EHR_Clustering/maspc/src/master/

[3] We used first 17 Chapters, as the data does not have E and V codes.

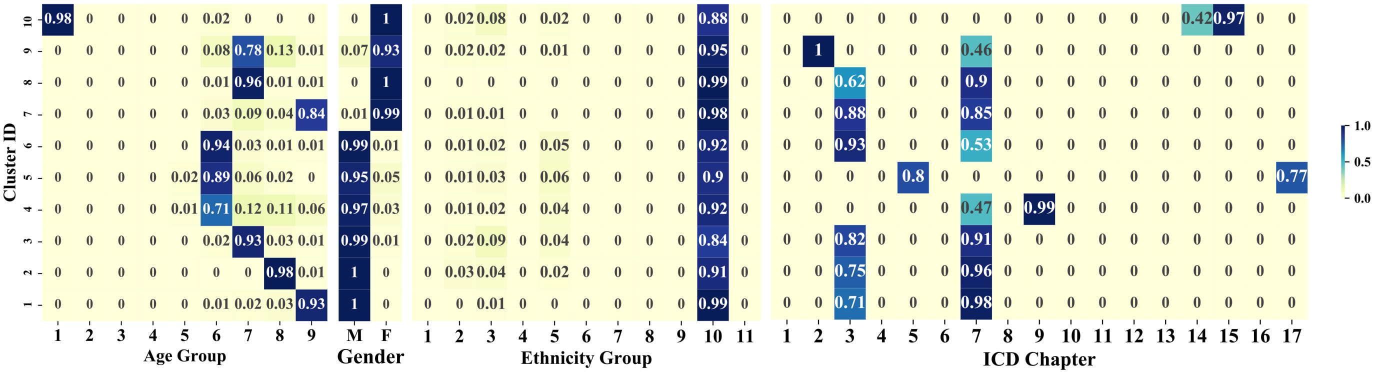


Fig. 5: Heat map of MIMIC for *Age*, *Gender*, *Ethnicity*, and ICD Chapter. The description of values in Age and Ethnicity groups are in Supplementary Material Sec. V. The values in the cells are ratios of records in a cluster (e.g., 0.98 of records in cluster with ID 10 have their Age values in Age Group 1 corresponding to Newborns). The sum of ratios for a cluster over the ICD Chapters is not 1, since a record can contain diagnosis codes belonging into multiple ICD Chapters.

TABLE VI: The top-1 (i.e., most) frequent value in each demographic, top-2 frequent ICD Chapters, and top-3 frequent sequential patterns of each cluster. A top-3 frequent sequential pattern in a cluster c is a sequence of ICD-9 codes appearing in the first, second, or third largest number of records in c.

Cluster ID	Gender	Age	Ethnicity	ICD Chapters	Top-3 Frequent Sequential Patterns		
1	M	Over 80	White	{3, 7}	(427.31, 428.0)	(401.9, 427.31)	(414.01, 427.31)
2	M	Aged	White	{3, 7}	(401.9, 414.01)	(414.01, 427.31)	(401.9, 427.31)
3	M	Middle aged	White	{3, 7}	(401.9, 414.01)	(272.4, 401.9)	(272.0, 401.9)
4	M	Adult	White	{7, 9}	(401.9, 530.81)	(414.01, 530.81)	(272.4, 530.81)
5	M	Adult	White	{5, 17}	(305.1, 401.9)	(305.00, 305.1)	(305.1, 518.81)
6	M	Adult	White	{3, 7}	(276.2, 518.81)	(276.2, 584.9)	(518.81, 584.9)
7	F	Over 80	White	{3, 7}	(427.31, 428.0)	(428.0, 584.9)	(401.9, 427.31)
8	F	Middle aged	White	{3, 7}	(401.9, 414.01)	(272.4, 401.9)	(401.9, 427.31)
9	F	Middle aged	White	{2, 7}	(198.3, 198.5)	(197.7, 198.5)	(197.0, 198.5)
10	F	Newborn	White	{14, 15}	(769, 774.2)	(774.2, 770.81)	(774.2, 779.3)

frequently co-exists with atrial fibrillation (427.31) [47]. Interestingly, the number of patients with atrial fibrillation in Cluster 1 (containing patients aged 80 or over) is larger than that of Cluster 2 (containing patients aged 65 to less than 80). This is explained by the fact that the prevalence of atrial fibrillation increases with age [49].

Cluster 3: Hypertension (401.9) that appears in all patterns frequently co-exists with coronary artery disease (414.01) [48], or with hyperlipidemia (272.4) [50], or with pure hypercholesterolemia (272.0) [51].

Cluster 4: Esophageal reflux or Gastroesophageal reflux disease (530.81) that appears in all patterns frequently co-exists with hypertension (401.9) [52], or with coronary artery disease (414.01) [53], or with hyperlipidemia (272.4) [54].

Cluster 5: Tobacco use disorder (305.1) that appears in all patterns frequently co-occurs with hypertension (401.9) [55], or with non-dependent alcohol abuse (305.00) [56], or with diseases of the lung (such as acute respiratory failure denoted by 518.81) [57].

Cluster 6: Acidosis (276.2) frequently co-occurs with acute respiratory failure (518.81) [58], or with acute kidney failure (584.9) [59].

Cluster 7: The patients in this cluster are very similar to those in Cluster 1 regarding Age and Ethnicity but differ in Gender. Thus, two of the three patterns in Cluster 7 also appear in Cluster 1. Furthermore, all patients are very old (the most frequent Age category is “Aged, 80 and over”), thus it is expected to have frequent patterns with coronary artery disease (414.01) (see Cluster 1), congestive heart failure (428.0) (see Cluster 7), and hypertension (401.9) (see Clusters 1 and 7). The reason is that ageing predisposes to a high incidence and prevalence of coronary artery disease, congestive heart failure, and hypertension (401.9) in both males and females [60], [61].

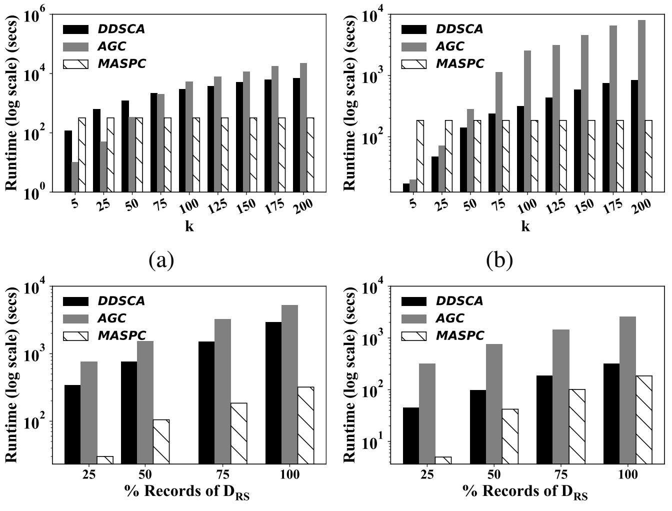
Cluster 8: The patients in this cluster are very similar to those in Cluster 3 regarding Age, Ethnicity, ICD Chapters but differ in Gender. Thus, two of the three patterns in Cluster 8 also appear in Cluster 3. In both of these clusters, many patients have hypertension (401.9), but the number of patients with hypertension in Cluster 3 is three times larger than that in Cluster 8. This is expected because hypertension is more prevalent among males (99% of patients in Cluster 3 are male) than females (all patients in Cluster 8 are female) [62].

Cluster 9: The patterns in this cluster all contain secondary malignant neoplasm of bone and bone marrow (198.5). This diagnosis code frequently co-exists with secondary malignant neoplasm of brain and spinal cord (198.3), or with secondary malignant neoplasm of liver (197.7), or with secondary malignant neoplasm of lung (197.0). Since secondary bone cancer occurs when cancers that develop elsewhere spread, or metastasize, to the bones [63], it is expected that 198.5 frequently co-occurs with secondary cancers in other organs (198.3, or 197.7, or 197.0).

Cluster 10: Neonatal jaundice (774.2) that appears in all patterns frequently co-occurs with respiratory distress syndrome (769) [64], or with primary apnea of newborn (770.81) [65], or with feeding problems (779.3) [66].

F. Runtime

Fig. 6 shows the runtime of all methods for varying number of clusters k and varying number of records. DDSCA scaled better than AGC with respect to both criteria and was faster in most cases. MASPC was the fastest (except for $k < 50$), mainly because it directly added a large percentage (32% to 45%) of records into a single cluster without considering their similarity to other records. This helped runtime but harmed effectiveness. As discussed above, MASPC was the least effective method. The runtime of MASPC is not affected by k, as it uses hierarchical clustering, which is insensitive to k.



(a)

(b)

(c)

(d)

Fig. 6: Runtime vs. k: (a) MIMIC and (b) INFORMS. Runtime vs. % of records in the input dataset for: (c) MIMIC and (d) INFORMS.

VIII. Limitations

Our study is limited in three aspects which provide opportunities for future work.

First, our algorithm does not deal with datasets in which diagnosis codes are associated with dates of visits, which are helpful in longitudinal studies [67]. One way to deal with such datasets is to treat the date of visit corresponding to the diagnosis codes in a record in the same way as a demographic. However, this is not appropriate when there are multiple dates, each corresponding to some of the diagnosis codes in a record, due to the curse of dimensionality [8]. In this case, a fundamentally new approach is required.

Second, our algorithm was evaluated using data containing ICD-9 codes. Although our distance measure can easily be modified to deal with other types of diagnosis codes such as ICD-10 codes [68] (or, in general, any sequence of events), further work is needed to evaluate its effectiveness in such settings. Similarly, it would be interesting to extend our algorithm to consider sequences comprised of other patient information, such as medications and lab results. This requires further work, as capturing the similarity of multiple interrelated sequences is challenging.

Third, it is useful to evaluate the effectiveness of our algorithm in tasks beyond frequent sequential pattern mining. An example of such a task is causal relationship discovery [69], for which we have obtained some preliminary results (see Supplementary Material Sec. X). Other examples of such tasks are classification [16] and anonymization [17], which we also leave for future work.

IX. Conclusion

The task of clustering an RS-dataset is important for several analyses, but existing algorithms are inappropriate to deal with it. Thus, in this work, we formalized this task as an optimization problem. We proved that the problem is computationally hard, and we proposed an effective and efficient algorithm to address it. Our experiments demonstrate that our algorithm can construct compact and well-separated clusters. which preserve meaningful relationships between demographics and sequences of diagnosis codes. In addition, they show that our algorithm is efficient and scalable.

References

[1] D. Gotz, J. Sun, N. Cao, and S. Ebadollahi, "Visual cluster analysis in support of clinical decision intelligence," in *AMIA Annu. Symp. proc.*, vol. 2011. American Medical Informatics Association, 2011, p. 481.

[2] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining electronic health records (ehrs) a survey," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–40, 2018.

[3] A. Wright, E. S. Chen, and F. L. Maloney, "An automated technique for identifying associations between medications, laboratory results and problems," *J. Biomed. Inform.*, vol. 43, no. 6, pp. 891–901, 2010.

[4] L. Ohno-Machado, "Mining electronic health record data: finding the gold nuggets," *J. Am. Med. Inform. Assoc.*, vol. 22, no. 5, pp. 937–937, 2015.

[5] D. Gartner and R. Padman, "Mathematical modelling and cluster analysis in healthcare analytics-the case of length of stay management," in *Proc. Int. Con. on Information Systems*, 2016.

[6] J. M. Sanderson, D. C. Proops, L. Trieu, E. Santos, B. Polsky, and S. D. Ahuja, "Increasing the efficiency and yield of a tuberculosis contact investigation through electronic data systems matching," *J. Am. Med. Inform. Assoc.*, vol. 22, no. 5, pp. 1089–1093, 2015.

[7] R. J. Carroll, A. E. Eyler, and J. C. Denny, "Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis," *Expert Rev. Clin. Immunol.*, vol. 11, no. 3, pp. 329–337, 2015.

[8] M. J. Zaki, W. Meira Jr, and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[9] H. Zhong, G. Loukides, and R. Gwadera, "Clustering datasets with demographics and diagnosis codes," *J. Biomed. Inform.*, vol. 102, p. 103360, 2020.

[10] International Classification of Diseases - Ninth Revision, https://www.cdc.gov/nchs/icd/icd9cm.htm, Last accessed 2021-06-01.

[11] Y. Wang *et al.*, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *J. Biomed. Inform.*, vol. 102, p. 103364, 2020.

[12] P. M. Schnell, Q. Tang, W. W. Offen, and B. P. Carlin, "A bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects," *Biometrics*, vol. 72, no. 4, pp. 1026–1036, 2016.

[13] D. Chushig-Muzo, C. Soguero-Ruiz, A. P. Engelbrecht, P. D. M. Bohoyo, and I. Mora-Jiménez, "Data-driven visual characterization of patient health-status using electronic health records and self-organizing maps," *IEEE Access*, vol. 8, pp. 137 019–137 031, 2020.

[14] R. Guo *et al.*, "Comparative visual analytics for assessing medical records with sequence embedding," *Vis. Inform.*, vol. 4, no. 2, pp. 72–85, 2020.

[15] E. A. Campbell, E. J. Bass, and A. J. Masino, "Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 4, pp. 558–566, 2020.

[16] N. Sokolovska, O. Cappé, and F. Yvon, "The asymptotics of semi-supervised learning in discriminative probabilistic models," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 984–991.

[17] G. Poulis, G. Loukides, S. Skiadopoulos, and A. Gkoulalas-Divanis, "Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints," *J. Biomed. Inform.*, vol. 65, pp. 76–96, 2017.

[18] Y. Zhang, R. Padman, and N. Patel, "Paving the COWpath: Learning and visualizing clinical pathways from electronic health record data," *J. Biomed. Inform.*, vol. 58, pp. 186–197, 2015.

[19] H. Calik, M. Labbé, and H. Yaman, *p-Center Problems*. Springer International Publishing, 2015, pp. 79–92.

[20] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theor. Comput. Sci.*, vol. 38, pp. 293–306, 1985.

[21] X. Zhang, H. Liu, Q. Li, and X.-M. Wu, "Attributed graph clustering via adaptive graph convolution," in *Proc. IJCAI Int. Jt. Conf. Artif. Intell.*, 2019, p. 4327–4333.

[22] P. Fournier-Viger, J. C.-W. Lin, R.-U. Kiran, Y.-S. Koh, and R. Thomas, "A survey of sequential pattern mining," *Data Science and Pattern Recognition*, vol. 1, no. 1, p. 54–77, 2017.

[23] C. C. Aggarwal, Ed., *Data Classification: Algorithms and Applications*. CRC Press, 2014. [Online]. Available: http://www.crcnetbase.com/doi/book/10.1201/b17320

[24] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. SIAM, 2020.

[25] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *J. R. Stat. Soc.*, vol. 28, no. 1, pp. 100–108, 1979.

[26] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics*, vol. 17, no. suppl_1, pp. S22–S29, 2001.

[27] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.

[28] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, vol. 96, no. 34, 1996, pp. 226–231.

[29] F. Doshi-Velez, Y. Ge, and I. Kohane, "Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis," *Pediatrics*, vol. 133, no. 1, pp. e54–e63, 2014.

[30] R. A. Hubbard, J. Xu, R. Siegel, Y. Chen, and I. Eneli, "Studying pediatric health outcomes with electronic health records using bayesian clustering and trajectory analysis," *J. Biomed. Inform.*, vol. 113, p. 103654, 2021.

[31] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. Int. Conf. Research in Computational Linguistics*, 1997.

[32] D. Sánchez, M. Batet, and D. Isern, "Ontology-based information content computation," *Knowl-Based. Syst.*, vol. 24, no. 2, pp. 297–303, 2011.

[33] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, 2001.

[34] N. Miura and T. Takagi, "Wsl: sentence similarity using semantic distance between words," in *Proc. 9th Int. Workshop on Semantic Evaluation*, 2015, pp. 128–131.

[35] ICD 9 Code Hierarchy, https://en.wikipedia.org/wiki/List_of_ICD-9_codes, Last accessed 2021-06-01.

[36] S. Avgustinovich and D. Fon-Der-Flaass, "Cartesian products of graphs and metric spaces," *Eur. J. Comb.*, vol. 21, no. 7, pp. 847–851, 2000.

[37] L. Wirbka, W. E. Haefeli, and A. D. Meid, "A framework to build similarity-based cohorts for personalized treatment advice–a standardized, but flexible workflow with the r package simbaco," *PloS one*, vol. 15, no. 5, p. e0233686, 2020.

[38] A. E. Johnson *et al.*, "Mimic-iii, a freely accessible critical care database," *Sci. data*, vol. 3, no. 1, pp. 1–9, 2016.

[39] INFORMS, "Informs data mining contest," https://sites.google.com/site/informsdataminingcontest/data/, Last accessed 2021-06-01.

[40] M. Kastner, N. L. Wilczynski, C. Walker-Dilks, K. A. McKibbon, and B. Haynes, "Age-specific search strategies for medline," *J. Med. Internet Res.*, vol. 8, no. 4, p. e25, 2006.

[41] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, "Adversarially regularized graph autoencoder for graph embedding," in *Proc. IJCAI Int. Jt. Conf. Artif. Intell.*, 2018, p. 2609–2615.

[42] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: an analysis and critique," *Strateg. Manag. J.*, vol. 17, no. 6, pp. 441–458, 1996.

[43] ICD-9-CM Chapters, https://icd.codes/icd9cm, Last accessed 2021-06-01.

[44] M. S. Dzeshka, A. Shantsila, E. Shantsila, and G. Y. Lip, "Atrial fibrillation and hypertension," *Hypertension*, vol. 70, no. 5, pp. 854–861, 2017.

[45] J. Pei *et al.*, "Prefixspan,: mining sequential patterns efficiently by prefix-projected pattern growth," in *Proc. 17th Int. Conf. on Data Engineering*, 2001, pp. 215–224.

[46] S. A. Lubitz, E. J. Benjamin, and P. T. Ellinor, "Atrial fibrillation in congestive heart failure," *Heart Fail. Clin.*, vol. 6, no. 2, pp. 187–200, 2010.

[47] E. Michniewicz, E. Mlodawska, P. Lopatowska, A. Tomaszuk-Kazberuk, and J. Malyszko, "Patients with atrial fibrillation and coronary artery disease–double trouble," *Adv. Med. Sci.*, vol. 63, no. 1, pp. 30–35, 2018.

[48] T. Weber *et al.*, "Hypertension and coronary artery disease: epidemiology, physiology, effects of treatment, and recommendations," *Wien. Klin. Wochenschr.*, vol. 128, no. 13, pp. 467–479, 2016.

[49] C. Wilkinson, O. Todd, A. Clegg, C. P. Gale, and M. Hall, "Management of atrial fibrillation for older people with frailty: a systematic review and meta-analysis," *Age Ageing*, vol. 48, no. 2, pp. 196–203, 2019.

[50] R. P. Ames, "Hyperlipidemia in hypertension: causes and prevention," *Am. Heart J.*, vol. 122, no. 4, pp. 1219–1224, 1991.

[51] B. Ivanovic and M. Tadic, "Hypercholesterolemia and hypertension: two sides of the same coin," *Am. J. Cardiovasc. Drugs*, vol. 15, no. 6, pp. 403–414, 2015.

[52] Z.-t. Li, F. Ji, X.-w. Han, L. Wang, Y.-q. Yue, and Z.-g. Wang, "The role of gastroesophageal reflux in provoking high blood pressure episodes in patients with hypertension," *J. Clin. Gastroenterol.*, vol. 52, no. 8, p. 685, 2018.

[53] M. H. Mellow, A. G. Simpson, L. Watt, L. Schoolmeester, and O. L. Haye, "Esophageal acid perfusion in coronary artery disease: induction of myocardial ischemia," *Gastroenterology*, vol. 85, no. 2, pp. 306–312, 1983.

[54] J. P. P. Moraes-Filho, T. Navarro-Rodriguez, J. N. Eisig, R. C. Barbuti, D. Chinzon, and E. M. Quigley, "Comorbidities are frequent in patients with gastroesophageal reflux disease in a tertiary health care hospital," *Clinics*, vol. 64, no. 8, pp. 785–790, 2009.

[55] K. Gao, X. Shi, and W. Wang, "The life-course impact of smoking on hypertension, myocardial infarction and respiratory diseases," *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, 2017.

[56] D. J. Drobes, "Concurrent alcohol and tobacco dependence: mechanisms and treatment," *Alcohol Res Health.*, vol. 26, no. 2, p. 136, 2002.

[57] B. Balbi *et al.*, "Smoking-related lung diseases: a clinical perspective," *Eur. Respir. J.*, vol. 17, p. 122–132, 2010.

[58] L. R. Engelking, *Textbook of Veterinary Physiological Chemistry, Updated 2/e*. Academic Press, 2010.

[59] P. K. Moore, R. K. Hsu, and K. D. Liu, "Management of acute kidney injury: core curriculum 2018," *Am. J. Kidney Dis.*, vol. 72, no. 1, pp. 136–148, 2018.

[60] P. Kazemian, G. Oudit, and B. I. Jugdutt, "Atrial fibrillation and heart failure in the elderly," *Heart Fail. Rev.*, vol. 17, no. 4-5, pp. 597–613, 2012.

[61] M. V. Madhavan, B. J. Gersh, K. P. Alexander, C. B. Granger, and G. W. Stone, "Coronary artery disease in patients $\geq$ 80 years of age," *J. Am. Coll. Cardiol.*, vol. 71, no. 18, pp. 2015–2040, 2018.

[62] B. Everett and A. Zajacova, "Gender differences in hypertension and hypertension awareness among young adults," *Biodemogr. Soc. Biol.*, vol. 61, no. 1, pp. 1–17, 2015.

[63] C. Canal, R. Fontelo, I. Hamouda, J. Guillem-Marti, U. Cvelbar, and M.-P. Ginebra, "Plasma-induced selectivity in bone cancer cells death," *Free Radic. Biol. Med.*, vol. 110, pp. 72–80, 2017.

[64] C. Weir and W. S. Millar, "The effects of neonatal jaundice and respiratory complications on learning and habituation in 5-to 11-month-old infants," *J. Child Psychol. Psychiatry Allied Discip.*, vol. 38, no. 2, pp. 199–206, 1997.

[65] S. B. Amin and H. Wang, "Unbound unconjugated hyperbilirubinemia is associated with central apnea in premature infants," *J. Pediatr.*, vol. 166, no. 3, pp. 571–575, 2015.

[66] P. A. Dennery, D. S. Seidman, and D. K. Stevenson, "Neonatal hyperbilirubinemia," *N. Engl. J. Med.*, vol. 344, no. 8, pp. 581–590, 2001.

[67] A. Tamersoy, G. Loukides, M. E. Nergiz, Y. Saygin, and B. Malin, "Anonymization of longitudinal electronic medical records," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 413–423, 2012.

[68] International Classification of Diseases - Tenth Revision, https://www.cdc.gov/nchs/icd/icd10cm.htm, Last accessed 2021-06-01.

[69] P. Bühlmann, M. Kalisch, and M. H. Maathuis, "Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm," *Biometrika*, vol. 97, no. 2, pp. 261–278, 2010.

Supplementary Material for “Clustering demographics and sequences of diagnosis codes”

I. Mathematical Formulation

Define a binary variable y_j s.t. $y_j = 1$ if record $r_j \in \mathbf{D_{RS}}$ is selected as a center and 0 otherwise. Also, define binary variables x_{ij} s.t. $x_{ij} = 1$ if $r_i \in \mathbf{D_{RS}}$ is closest to center $r_j \in \mathbf{D_{RS}}$ and 0 otherwise. Then, RSDC can be formulated as follows (the formulation is similar to the k-center formulation in [1]):

$$\begin{aligned}
&\min \quad z && && \text{(1a)}\\
&\text{s.t.} \quad \sum_{j\in[1,|\mathbf{D_{RS}}|]} d_{JCE}(r_i,r_j)\cdot x_{ij} \le z && \forall i \in [1,|\mathbf{D_{RS}}|] && \text{(1b)}\\
&\sum_{j\in[1,|\mathbf{D_{RS}}|]} x_{ij} = 1 && \forall i \in [1,|\mathbf{D_{RS}}|] && \text{(1c)}\\
&x_{ij} \le y_{ij} && \forall i,j \in [1,|\mathbf{D_{RS}}|] && \text{(1d)}\\
&\sum_{j\in[1,|\mathbf{D_{RS}}|]} y_j = k && && \text{(1e)}\\
&y_j \in \{0,1\} && \forall j \in [1,|\mathbf{D_{RS}}|] && \text{(1f)}\\
&x_{ij} \in \{0,1\} && \forall i,j \in [1,|\mathbf{D_{RS}}|] && \text{(1g)}
\end{aligned}$$

Eqs. 1a and 1b ensure that the objective value is no less than the maximum record-to-center distance. Eq. 1c assigns each record to exactly one center. Eq. 1d ensures that no record assigns to j unless there is a center at j. Eq. 1e ensures there are k centers selected and Eqs. 1f and 1g are the binary restrictions.

II. Hardness

Theorem 1: The RSDC problem is NP-hard.

Proof. The decision version of RSDC asks whether there exists a set C of k centers such that $\max_{r\in\mathbf{D_{RS}}}\min_{c\in C} d_{JCE}(r,c) \le B$ for a given real number B. The decision version of RSDC is clearly in NP. Thus, it suffices to show that the decision version of RSDC can be reduced from the decision version of the NP-complete k-Center problem [2], defined as follows: Given a metric space (P,d), where P is a set of n points and $d: P\times P \rightarrow R^+$ is a distance function, an integer $k \in [1,n]$, and a real number B, decide whether there exists a subset $S \subseteq P$ of k points s. t. $\max_{p\in P}\min_{s\in S} d(p,s) \le B$. For the reduction, we first construct an instance I_{RSDC} of the decision version of RSDC in polynomial time from any instance I_{kC} of the decision version of k-Center by: (I) adding a record r into an RS-dataset $\mathbf{D_{RS}}$ for each point $p \in P$, (II) setting $d_{JCE}(r,r') = d(p,p')$ for every pair $p,p' \in P$ that corresponds to a pair of records $r,r' \in \mathbf{D_{RS}}$, and (III) setting the parameters k (respectively, B) in I_{RSDC} to the value of the parameter k (respectively, B) in I_{kC}. We then observe that if I_{RSDC} has a positive answer, then there is a set C of k records such that $\max_{r\in\mathbf{D_{RS}}}\min_{c\in C} d_{JCE}(r,c) \le B$, which correspond to a subset S of k points in P for which $\max_{p\in P}\min_{s\in S} d(p,s) \le B$. Thus, I_{kC} has a positive answer. It is easy to see that the converse also holds.

Theorem 2: The RSDSC problem can be approximated within a factor of 2.

Proof. We reduce RSDC to the k-Center problem, which can be approximated within a factor of 2 [3]. We first construct an instance I_{kC} of k-Center from any given instance I_{RSDC} of RSDC in polynomial time by: (I) adding a point p into the set of points P, for each record r in the RS-dataset $\mathbf{D_{RS}}$, (II) setting $d(p,p') = d_{JCE}(r,r')$ for every pair of records $r,r' \in \mathbf{D_{RS}}$ that corresponds to a pair of points $p,p' \in P$, and (III) setting the parameters k (respectively, B) in I_{kC} to the value of the parameter k (respectively, B) in I_{RSDC}. We then prove the correspondence between a solution S_{kC} to I_{kC} and a solution S_{RSDC} to I_{RSDC}. If S_{kC} is a solution to I_{kC}, then there is a subset S of k points in P for which $\max_{p\in P}\min_{s\in S} d(p,s)$ is minimum. These points correspond to a set C of k records s.t. $\max_{r\in\mathbf{D_{RS}}}\min_{c\in C} d_{JCE}(r,c)$ is minimum. Thus, S_{RSDC} is a solution to RSDC. Clearly, the converse also holds.

Theorem 3: The RSDSC problem cannot be approximated within a factor of $2-\epsilon$, for any $\epsilon > 0$, in polynomial time unless P=NP.

Proof. We reduce the k-Center problem, which cannot be approximated within a factor of $2-\varepsilon$, for any $\varepsilon > 0$ [3], to RSDC. Given any instance I_{kC} of k-Center, we construct an instance I_{RSDC} of RSDC in polynomial time in the size of I_{kC} by: (I) adding a record r into an RS-dataset $\mathbf{D_{RS}}$ for each point $p \in P$, (II) setting $d_{JCE}(r,r') = d(p,p')$ for every pair $p,p' \in P$ that corresponds to a pair of records $r,r' \in \mathbf{D_{RS}}$, and (III) setting the parameters k (respectively, B) in I_{RSDC} to the value of the parameter k (respectively, B) in I_{kC}. Since we have proved the correspondence between a solution S_{kC} to I_{kC} and a solution S_{RSDC} to I_{RSDC} above, we have a reduction from k-Center problem to RSDC.

III. Approximation guarantee of DDSCA

The fact that DDSCA finds a set C of k centers that is at most 2 times worse with respect to d_{JCE} than the best possible set of k centers follows from the following facts: (1) DDSCA selects centers with a strategy following that of the algorithm of Gonzalez et al. [3]. (2) The algorithm of [3] has an approximation ratio of 2. (3) The RSDC problem can be approximated within a factor of 2, which we proved above.

IV. The time complexity of DDSA

DDSA takes $O(((\max_{r^{seq}\in\mathbf{D_{RS}}}|r^{seq}|)^2 + |\mathbf{T_{RS}}|)k|\mathbf{D_{RS}}|)$ time, where $|\mathbf{T_{RS}}|$ is the number of nodes in $\mathbf{T_{RS}}$, k is the number of clusters, and $|\mathbf{D_{RS}}|$ is the number of records in $\mathbf{D_{RS}}$. This means that it scales linearly with the number of records in the input dataset and the number of clusters.

V. Categories for demographics

TABLE I: (a) Age and (b) Ethnicity group categories.

MIMIC		INFORMS	
Age Group	Definition	Age Group	Definition
1	Birth to 1 month (Newborn)	1	Birth to 1 month (Newborn)
2	1 month to <24 months (Infant)	2	2 to <6 years (Preschool)
3	2 to <6 years (Preschool)	3	6 to <13 years (Child)
4	6 to <13 years (Child)	4	13 to <19 years (Adolescent)
5	13 to <19 years (Adolescent)	5	19 to <45 years (Adult)
6	19 to <45 years (Adult)	6	45 to <65 years (Middle aged)
7	45 to <65 years (Middle aged)	7	65 to <80 years (Aged)
8	65 to <80 years (Aged)	8	80 years (Aged, 80 and over)
9	80 years (Aged, 80 and over)		

(a)

MIMIC		INFORMS	
Ethn. Group	Definition	Ethn. Group	Definition
1	AMERICAN	1	WHITE
2	ASIAN	2	BLACK
3	BLACK	3	AMERICAN INDIAN/ALASKA NATIVE
4	CARIBBEAN	4	ASIAN
5	HISPANIC	5	NATIVE HAWAIIAN / PACIFIC ISLANDER
6	MIDDLE EASTERN	6	MULTI RACE
7	MULTI RACE		
8	PACIFIC		
9	PORTUGUESE		
10	WHITE		
11	SOUTH AMERICAN		

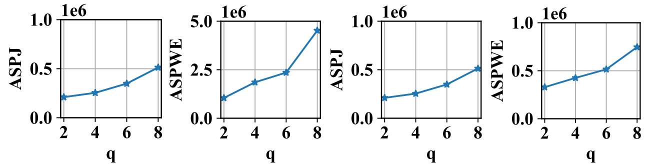
(b)

Age was discretized using a well-defined taxonomy [4] (see Table Ia). We have also experimented with two different discretizations of *Age* based on the same taxonomy to demonstrate that our method still outperforms the competitors (see [5]).

VI. Detailed description of competitors

AGC [6]: AGC gets as input a graph whose nodes have vectors as attributes, and it outputs a partition of the nodes of the graph into k groups. In our context, the graph is a tree, $\mathbf{T_{AGC}}$, that is similar to the RS-tree $\mathbf{T_{RS}}$ of DDSCA. The difference between $\mathbf{T_{AGC}}$ and $\mathbf{T_{RS}}$ and is that: (1) Each record r in any leaf-level node of $\mathbf{T_{AGC}}$ does not contain the sequence of diagnosis r^{seq} but a vector with the frequencies of all q-grams (i.e., substrings of length q) of r^{seq}, and (2) each record r in a non-leaf node of $\mathbf{T_{AGC}}$ contains a dummy vector of zeros. The differences are because each node in an input graph of AGC must contain a vector of values. AGC uses graph convolution and spectral clustering [7] to construct the groups of $\mathbf{T_{AGC}}$. However, some groups may contain non-leaf nodes, which contain no useful information about sequences. Thus, we remove such nodes from the groups. This does not affect the quality of clustering. Next, we obtain the clusters by going over each group and adding into a cluster all the records in $\mathbf{D_{RS}}$ corresponding to the nodes in the group.

We also examined the impact of q for AGC. Figure 1 below shows that cluster quality in MIMIC and INFORMS became worse as q increased This is because the number of distinct q-grams increases, and thus ASPWE and ASPJ increased as well. Therefore, the default value q=2 is a suitable choice.



(a) (b) (c) (d)

Fig. 1: (a) ASPJ and (b) ASPWE vs. q for MIMIC. (c) ASPJ and (d) ASPWE vs. q for INFORMS.

MASPC [8]: MASPC gets as input a dataset in which each record contains values in demographics, as well as a set of diagnosis codes, and it outputs k clusters of the dataset. This dataset is produced by an RS-dataset and final clusters are produced, as described in Section II of the main paper. If there are records that are not part of the projection, MASPC adds them into a single cluster, without considering how similar they are to other records.

VII. Effectiveness based on string distances

TABLE II: Effectiveness based on various popular string distances for $k = 50$ on: (a) MIMIC and (b) INFORMS.

Method	ASPLCS	ASPL	ASPJW	ASPNW
DDSCA	14,743,919	694,237	656,737	683,299
AGC	24,752,163	1,266,537	1,138,542	1,244,528
MASPC	32,584,215	2,216,213	2,344,153	2,041,415

(a)

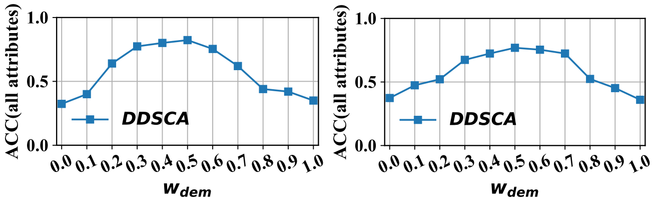
Method	ASPLCS	ASPL	ASPJW	ASPNW
DDSCA	1,842,278	247,428	233,382	184,711
AGC	3,928,452	397,642	381,475	310,142
MASPC	6,715,423	1,814,728	1,764,532	1,423,324

(b)

Table II shows that, for $k = 50$, DDSCA outperformed both competitors with respect to the ASPLCS, ASPL, ASPJW, and ASPNW measures. These measures are similar to ASPWE but use one of the LCS, Levenshtein, Jaro-Winkler[1], or Needleman-Wunsch distances [2] (see [9] for definitions and analysis). Analogous results for varying k are in [5].

VIII. Impact of weights

Fig. 2 shows that the most compact clusters with respect to ACC are produced when equal weights are used. Analogous results for ASPJ and ASPWE are in [5].



(a) (b)

Fig. 2: ACC vs. w_{dem} for (a) MIMIC and (b) INFORMS.

IX. Compactness, separability, and medical relevance of top-3 frequent sequential patterns

Compactness results for MIMIC are in Fig. 3. Separability results for INFORMS are in Table III, while a discussion of cluster compactness and the medical relevance of top-3 frequent sequential patterns for INFORMS are in [5].

TABLE III: The top-1 (i.e., most) frequent value in each demographic, top-2 frequent ICD Chapters, and top-3 frequent sequential patterns, for ten clusters in INFORMS for DDSCA with k=50.

ID	Gender	Age	Ethnicity	Chapter	Top 3 Sequential Patterns		
1	M	Aged	White	{7,13}	(401,716)	(272,401)	(272,716)
2	M	Aged	White	{3,7}	(272,401)	(401,250)	(401,780)
3	M	Aged	White	{3,7}	(250,401)	(250,272)	(250,272,401)
4	M	Middle aged	White	{8,17}	(401,460)	(272,460)	(460,724)
5	M	Middle aged	Black	{3,7}	(250,401)	(272,401)	(250,272,401)
6	M	Middle aged	White	{7,16}	(272,401)	(401,780)	(530,401)
7	M	Adult	Black	{5,13}	(300,311)	(311,401)	(311,477)
8	F	Middle aged	White	{3,8}	(272,401)	(272,477)	(250,272)
9	F	Middle aged	White	{7,13}	(272,401)	(401,716)	(250,401)
10	F	Adult	White	{8,13}	(401,724)	(311,724)	(716,724)

X. Causal relationship discovery

Causal relationships discovered in clusters obtained by our method are in [5].

References

[1] *Center Problems.* John Wiley & Sons, Ltd, 1995, ch. 5, pp. 154–197.
[2] D. S. Hochbaum, "When are np-hard location problems easy?" *Ann. Oper. Res.*, vol. 1, no. 3, pp. 201–214, 1984.
[3] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theor. Comput. Sci.*, vol. 38, pp. 293–306, 1985.
[4] M. Kastner, N. L. Wilczynski, C. Walker-Dilks, K. A. McKibbon, and B. Haynes, "Age-specific search strategies for medline," *J. Med. Internet Res.*, vol. 8, no. 4, p. e25, 2006.
[5] https://bitbucket.org/EHR_Clustering/ddsca/src/master/Additional_results/.
[6] X. Zhang, H. Liu, Q. Li, and X.-M. Wu, "Attributed graph clustering via adaptive graph convolution," in *Proc. IJCAI Int. Jt. Conf. Artif. Intell.*, 2019, p. 4327–4333.
[7] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[8] H. Zhong, G. Loukides, and R. Gwadera, "Clustering datasets with demographics and diagnosis codes," *J. Biomed. Inform.*, vol. 102, p. 103360, 2020.
[9] N. Gali, R. Mariescu-Istodor, D. Hostettler, and P. Fränti, "Framework for syntactic string similarity measures," *Expert Syst. Appl.*, vol. 129, pp. 169–185, 2019.

[1] The parameter prefix scale p is set to 0.1.

[2] The gap cost is set to -1, match score is set to 1, and mismatch score is set to 0.

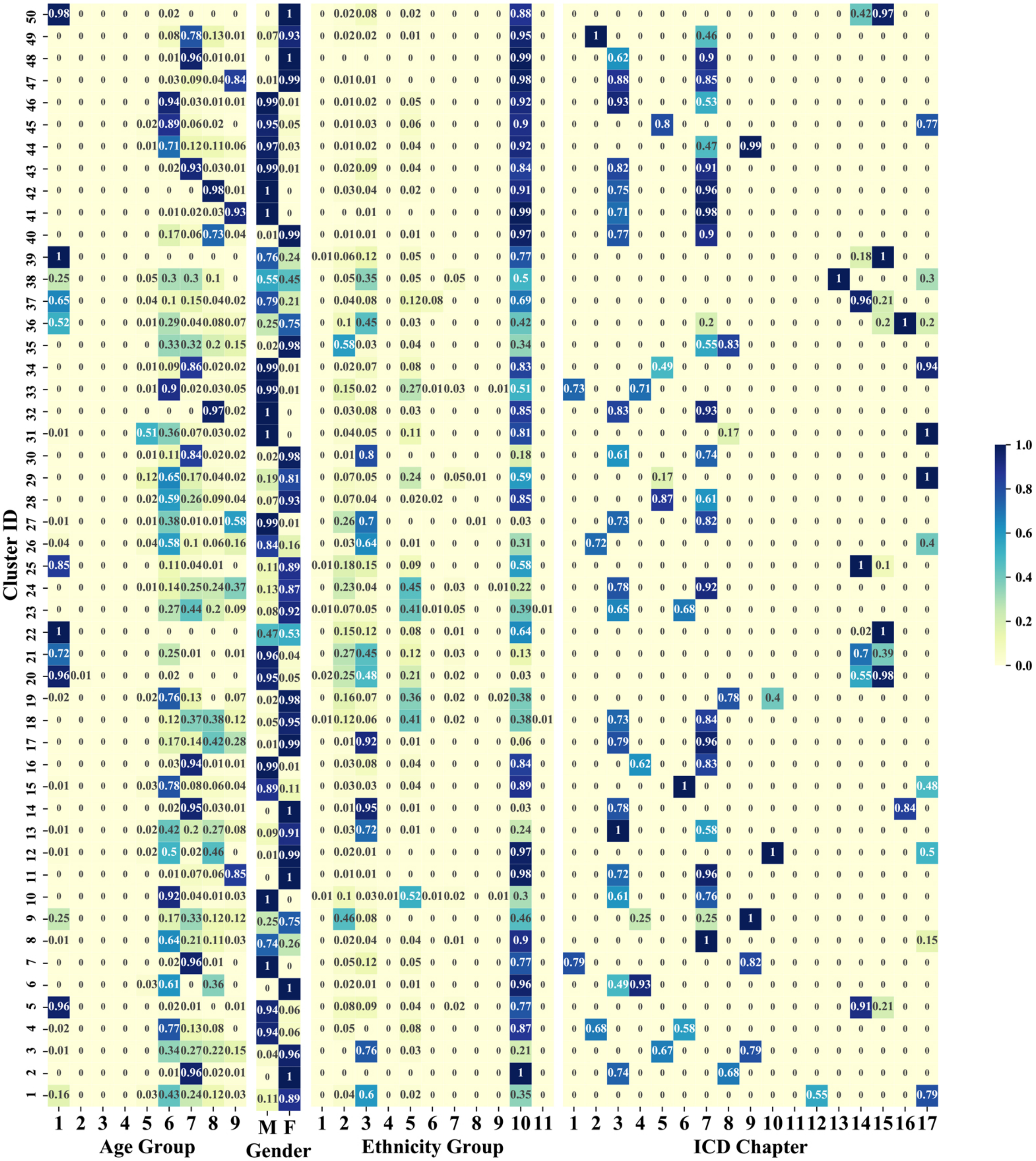


Fig. 3: Heat map of MIMIC for all clusters when k is 50.