Journal of Healthcare Informatics Research manuscript No.
(will be inserted by the editor)

LongSil - An Evaluation Metric to Assess Quality of Clustering Longitudinal Clinical Data

Duc Thanh Anh Luong · Prerna Singh · Mahin Ramezani · Varun Chandola

Received: date / Accepted: date

Abstract Longitudinal disease subtyping is an important problem within the broader scope of computational phenotyping. In this article, we discuss several data-driven unsupervised disease subtyping methods to obtain disease subtypes from longitudinal clinical data. The methods are analyzed in the context of Chronic Kidney Disease, one of the leading health problems, both in the US and worldwide. To provide a quantitative comparison of the different methods, we propose a novel evaluation metric that measures the cluster tightness and degree of separation between the various clusters produced by each method. Comparative results for two significantly large clinical datasets are provided, along with key insights that are possible due to the proposed evaluation metric.

Keywords Silhouette coefficient · clustering · disease subtype · evaluation metric · computational phenotyping

1 Introduction

With the increasing availability of Electronic Health Records (EHR) data for research and analysis, computational phenotyping has become an emergent and significant topic in the area of Health Informatics [4]. One approach that has been explored in the context of computational phenotyping is to identify groups of patients that exhibit similar disease progression as captured by the clinical observations present in EHR data. Many recent papers in this area have formulated the task of subgroup identification as an *unsupervised clustering problem* [5]. The objective is to cluster patients into groups based on their longitudinal EHR data, such that each cluster represents a distinct disease progression which can then be studied to identify a common disease mechanism (a phenotype).

However, applying standard clustering algorithms, such as k-means, to EHR data is not straightforward. The primary reason is that most of these algorithms require a *similarity metric* to compare the data for a pair of patients. If one analyzes disease progression in terms of a single disease marker, the problem will become a time series clustering task [9]. Given that such time series are typically sparse, irregularly sampled, and misaligned, (see Figure 1 for an example) applying standard time series similarity measures such as cross-correlation, Dynamic Time Warping (DTW), etc., are ill-suited in this context.

In other words, we want to cluster the patients into groups with similar disease progression and further identify the underlying mechanism in each cluster. Few methods have been proposed to solve this clustering

D. Luong
University at Buffalo
E-mail: ducthanh@buffalo.edu

P. Singh
Johns Hopkins University
E-mail: psingh26@jhu.edu

M. Ramezani
Alzahra University
E-mail: mahin@tamu.edu

V. Chandola
University at Buffalo
E-mail: chandola@buffalo.edu

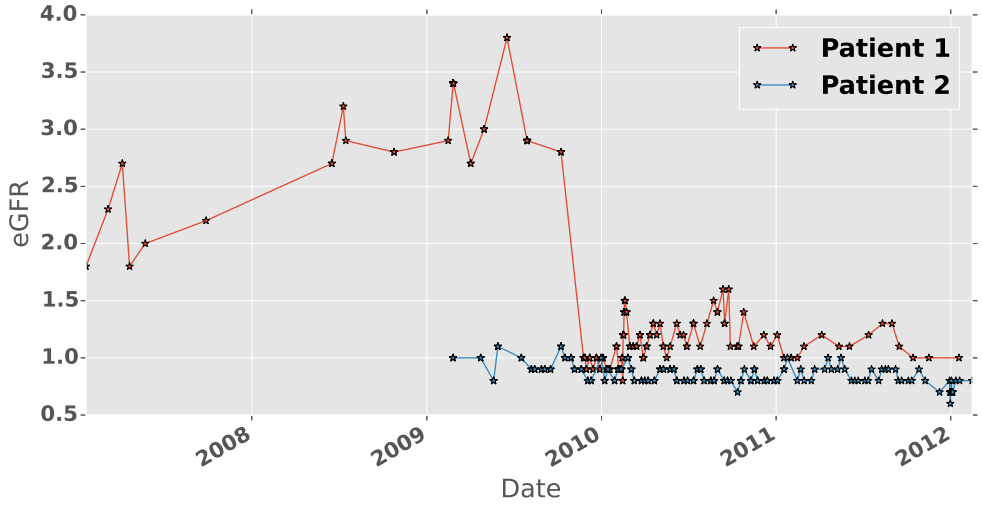


Fig. 1: Disease progression profiles (*Estimated Glomerular Filtration Rate*) of two patients suffering from Chronic Kidney Disease (CKD). Data obtained from DARTNet database [15].

problem. In particular, Schulam et al. proposed Probabilistic Subtyping Model (PSM) [20], a disease progression model that can probabilistically assign patient into clusters (disease subtypes, phenotypes) by analyzing the patient lab measurements and explaining away the effect of different covariates. Luong and Chandola [11] also introduced a k-means clustering approach to cluster patients based on their disease progressions. Singh et al. [21] followed another approach that imputes the missing values in the irregularly sampled time series and used traditional time series clustering method such as Partition Around Medoid (PAM) to group the patient longitudinal profiles into clusters. In another study, Baytas et al. [1] used a sequence model called T-LSTM Auto-encoder to project the set patient longitudinal profiles into an embedded space and apply k-means clustering to cluster the embedded representations. Although these methods can give reasonable clustering results, they all require the number of clusters to be known in advance. In many cases, there is no clear evidence to choose the number of clusters, thus, hindering the use of these methods in practice. In addition, we also need to decide a set of basis or requirements that we can use to assess the quality of a clustering result obtained from a particular disease subtyping model. As a result, there is a rising need for a quantitative and objective approach to judging the quality of clustering result. In particular, given a clustering assignment, we need to measure the "goodness-of-fit" of individual patient profile. In this article, we address these challenges by introducing a quantitative evaluation metric that can help us evaluate the quality of the clustering result, both globally as a set of patient profiles and locally as an individual patient profile.

In the literature of clustering evaluation [22, Chapter 17], there are two main approaches to evaluate the clustering result. One is external evaluation in which we use external information such as ground-truth of cluster assignments or other external characteristics associating with the subjects being clustered to evaluate the purity of each cluster with respect to those characteristics. Another alternative approach is internal evaluation in which we measure the clustering result by the tightness of each cluster as well as the degree of separation between neighboring clusters. In the context of longitudinal disease subtyping, external evaluation of clustering result can be done by examining the distribution of diagnosis codes or medications or demographics of patients assigned to a specific cluster. However, there is a lack of internal evaluation of disease subtypes, in the sense of cluster's tightness and degree of separation between clusters. Our approach aims to fill this missing piece by introducing an internal evaluation metric that can judge the quality of clustering result in terms of the cluster's tightness and degree of separation between clusters.

In traditional clustering settings, Silhouette coefficient is a common quantitative measure of the quality of clustering result [17]. However, this measurement cannot be used directly to evaluate the clusters of disease progression. Therefore, in this article, we develop a longitudinal Silhouette coefficient (abbreviate as *longSil*) based on the concept of Silhouette coefficient, which can help us evaluate the clustering result. It is the key contribution of this article. This evaluation metric also allows us to perform a large-scale comparative study for multiple subtyping methods in the context of Chronic Kidney Disease.

The remainder of the article is structured as follows. In Section 2, we discuss the existing evaluation approaches that have been used to evaluate the quality of longitudinal disease subtypes and explain why we need a new evaluation metric. In Section 3, we propose an evaluation metric, *longSil*, to assess the quality of longitudinal disease subtypes. In Section 4, we provide details of various longitudinal disease subtyping methods that we will use for comparison. In Section 5, we present experimental results obtained by using our evaluation metric to assess the performance of different subtyping methods. In Section 6, we further discuss the advantages and limitations of our proposed metric. Finally, in Section 7, we give a conclusion to our study.

2 Background

In this section, we further explain the need of an evaluation metric for longitudinal disease subtypes by examining evaluation approaches that have been done in prior studies and identifying the missing piece that we plan to fill in this article.

A most common way that people have used to evaluate the resulting disease subtypes is to characterize disease progressions in each subtype and align those progressions with an existing medical understanding of the specific disease [11, 13, 18, 20, 21]. This characterization of disease subtypes often includes narrative descriptions of general trends in the subtypes as well as numerical quantities such as rate of increase/decrease and baseline values. However, besides providing readers better understanding of the discovered subtypes, this characterization approach does not provide enough evidence to evaluate whether a disease subtype is distinctively different from the others and worth further analysis on its clinical relevance.

Another evaluating approach often found in prior studies is by examining the prediction power of the proposed model [1, 11, 19, 20]. It is important to note that although most disease subtyping methods are unsupervised learning models, as they need to take into account the temporal dependency between observations as well as patient covariates, they can be effectively used as prediction models. Evaluating the prediction power of a model can help us assess whether the model correctly captures the characteristics of data. However, from a perspective of evaluating the quality of resulting disease subtypes, the prediction power may not reflect the quality of subtypes in terms of differences between subtypes as well as how distinctive a subtype is in comparison with others.

To capture the difference between one subtype and the others, prior studies have used statistical tests to check whether one subtype is significantly different from the others [1, 13, 18]. These tests use external data such as demographic information or other clinical markers that have not been used in learning the disease subtypes. This approach is useful to understand how a subtype disease is different from others and allows us to have more evidence to further investigate interesting disease subtypes. The only missing piece in this approach is that it only works with external data. It is important to note that sometimes we are more interested in evaluating the disease subtypes using the original data itself rather than the external data. Therefore, the main theme of this article is to propose an evaluation metric of longitudinal disease subtypes by using the original data itself.

3 Evaluation metric

In this section, we discuss an evaluation metric to validate the clusters of longitudinal patient profiles output from any disease subtyping methods. First, in Section 3.1, we briefly review the concept of Silhouette coefficient and its role in assessing clustering result. In Section 3.2, building on top of Silhouette coefficient, we define *longSil* coefficient and provide a formulation for it.

3.1 Review of Silhouette coefficient

Silhouette coefficient was first introduced by Rousseeuw in 1987 as a graphical aid to validate the clustering result [17]. Since then, it has been used substantially as a validation technique to evaluate the quality of the clustering result. In the formulation of Silhouette coefficient, every data point being clustered is evaluated by two measures: (1) tightness - how close a data point is with respect to all other data points in the same cluster and (2) degree of separation - how far away a data point is from the closest neighboring cluster.

In particular, given a set of n subjects with corresponding cluster assignments, Silhouette coefficient measuring the quality of cluster placement for each individual i is computed as $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$ where the

tightness $a(i)$ is measured as the average distance from i to all other subjects in the same cluster and the degree of separation $b(i)$ is measured as the shortest average distance from i to any clusters different from its own cluster. In the above formulation, the range of a Silhouette coefficient is between -1 and 1. $s(i) > 0$ means that the subject i is closer to its assigned cluster than its closest neighboring cluster. On the other hand, $s(i) < 0$ indicates that subject i is closer to the neighboring cluster than its own - an indication of incorrect cluster placement for subject i. When $s(i) = 0$, the subject i lies on the border between its own cluster and its closest neighboring one. Therefore, Silhouette coefficient can be used as a validation measure for clustering placement of each subject i in the dataset.

Besides the use of inspecting clustering quality for each individual subject in the dataset, the Silhouette coefficient can also be used to assess the overall clustering result by computing the average Silhouette coefficient across all subjects. As a result, for any clustering result, we can compute its average Silhouette coefficient and use it to quantify the quality of clustering result.

3.2 *longSil* coefficient

Although Silhouette coefficient works reasonably well for traditional clustering results, it cannot be easily translated to an evaluation of longitudinal disease subtypes because of an assumption of a distance metric between any pairs of subjects. In clinical datasets, the laboratory measurements are sparse and irregularly sampled (see Figure 1). As a result, given two longitudinal patient profiles, there is no trivial way to compute the distance between them. Therefore, we propose an alternative metric, denoted as *longSil* - **long**itudinal **Sil**houette coefficient to assess the quality of longitudinal disease subtypes.

In order to avoid computing pairwise distance between longitudinal patient profiles, one may note that in the measure of tightness $a(i)$ and degree of separation $b(i)$ for individual i, we actually compute the relative distance between subject i and sets of many other subjects. This allows us to take an alternative definition of tightness $a(i)$ and degree of separation $b(i)$ based on the distance between a longitudinal patient profile and a set of multiple profiles.

In a prior study of Luong and Chandola [11], the authors computed the distance between a longitudinal profile and a cluster by representing a cluster of multiple profiles as a common regression line and compute the total of squared vertical distances between clinical observations and corresponding values in the regression line. In this article, we follow a similar strategy of representing a set of multiple longitudinal profiles as a regression line. However, instead of using the total of squared vertical distances as a measure of distance, we use the average of squared vertical distances between observations and corresponding values on the regression line. This allows the contribution of each observation in the patient profile to be treated equally across different patients.

Given a dataset of clinical observations, we denote n_i as the number of observations of lab measurement that patient i takes. The vector of observations of patient i is denoted as $\boldsymbol{x}_i = [x_{i,1}, \cdots, x_{i,n_i}]^T$ while $\boldsymbol{t}_i = [t_{i,1}, \cdots, t_{i,n_i}]^T$ is the vector of corresponding timestamps.

Given a clustering result of N patients into K clusters, we denote $c(i) \in \{1, \cdots, K\}$ as the cluster that the patient i is assigned to. We also define $\bar{C}(i) = \{1, \cdots, K\} \setminus c(i)$ as the set of clusters that patient i is not assigned to.

Each cluster k is represented by a regression line using all observations of all patients assigned to it. This regression line of cluster k is formulated as $f_k(t) = \sum_{l=1}^{L} \beta_l^{(k)} \Phi_l(t)$ where $\boldsymbol{\Phi}(\cdot) = \{\Phi_1(\cdot), \cdots, \Phi_L(\cdot)\}$ is the set of L basis functions and $\boldsymbol{\beta}^{(k)} = \{\beta_1^{(k)}, \cdots, \beta_L^{(k)}\}$ is the set of corresponding coefficients. For a vector of input timestamps $\boldsymbol{t}_i \in \mathbb{R}^{n_i}$ of patient i, we denote $\boldsymbol{\Phi}(\boldsymbol{t}_i) = [\Phi_1(\boldsymbol{t}_i), \cdots, \Phi_L(\boldsymbol{t}_i)] \in \mathbb{R}^{n_i \times L}$ as a matrix in which l^{th} column is obtained by applying basis function $\Phi_l(\cdot)$ to each element of vector $\boldsymbol{t}_i$.

With the above notations, we define the tightness for patient profile i with respect to its own cluster $a(i)$ as the average of squared vertical distances between observations and corresponding values in the regression line:

$$a(i) = \frac{1}{n_i} \left\| \boldsymbol{x}_i - \boldsymbol{\Phi}(\boldsymbol{t}_i) \hat{\boldsymbol{\beta}}^{(c(i))} \right\|_2^2 \tag{1}$$

In the above formula, $\hat{\boldsymbol{\beta}}^{(c(i))}$ is the vector of coefficients of basis functions for the set of patients belonging to the same cluster of patient i (excluding patient i).

We also define the degree of separation for patient profile i as the closest distance from it to all other clusters that it is not assigned to:

$$b(i) = \min_{k \in \bar{C}(i)} \frac{1}{n_i} \left\| \boldsymbol{x}_i - \boldsymbol{\Phi}(\boldsymbol{t}_i)\boldsymbol{\beta}^{(k)} \right\|_2^2 \tag{2}$$

Using these two new formulations of $a(i)$ and $b(i)$, the *longSil* coefficient for patient profile i is defined similarly to the Silhouette coefficient:

$$longSil(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3}$$

Similar to Silhouette coefficient, *longSil* coefficient also ranges between -1 and 1 with higher value indicates better clustering quality.

4 Longitudinal Subtyping Methods

In this section we discuss four longitudinal subtyping methods that we will evaluate in the experiments including (1) Probabilistic Subtyping Model (PSM) [20], (2) Temporal K-means [11], (3) Spline Induced Clustering [21] and (4) Time-Aware LSTM Auto-encoder [1]. These methods all have a set of patient longitudinal profiles as inputs. In particular, a patient longitudinal profile is defined as a set of clinical measurements and corresponding timestamps of these measurements. In the context of Chronic Kidney Disease, we consider eGFR as a main clinical measurement. For PSM, an additional input it has is gender information which is used as a covariate for modeling patient disease progression. Furthermore, we also need to provide a number of subtypes for these three subtyping methods so that they can partition a set of patients into clusters. The output of these four methods are partitions of patients in which each cluster contains patients with similar disease progression. In the case of PSM, it returns probabilistic assignments of patients into clusters. One can effectively convert these probabilistic assignments into partitions by assigning patients into a cluster in which has the highest probability.

4.1 Probabilistic Subtyping Model (PSM)

Schulam et al. [20] first introduced this method to identify disease subtypes by "explaining away" other effects. In particular, this method models patient disease progression as a combination of a few separate effects. At a population level, there is a covariate effect that captures the effect of various types of patient covariates such as gender, age group or smoking behavior. Within the scope of this article, we only consider gender as a relevant covariate to include in the model. At an individual level, there are individual long-term effect and individual short-term effect. The long-term effect is used to model individual long-term health condition. On the other hand, a temporary health condition that may affect clinical measurements is modeled as a short-term effect. We also have a disease subtype effect, which is modeled to capture the effect of disease subtype in which many patients share similar disease progression. The overall disease progression is obtained by adding all the aforementioned effects. As a result, a probabilistic inference problem is solved by using an Expectation-Maximization approach.

4.2 Temporal K-means

Luong and Chandola [11] proposed this algorithm as a variant of k-means clustering method to cluster patients with similar disease progressions. The algorithm starts with an initialization step by randomly assigning patients into k clusters. After that, it iteratively performs two steps: (1) update step - using current cluster assignments to compute the cluster and (2) assignment step - assigning patients into closest clusters. The algorithm stops when it converges or maximum iteration is reached. Although the temporal k-means and original k-means have substantial overlap, temporal k-means has two main differences from the original k-means algorithm: (1) the "cluster centroid" is computed as a regression line fitted by using all observations of all patients currently assigned to the cluster and (2) the distance between a longitudinal patient profile and cluster is measured by the sum of squared vertical distances between observations and the corresponding values

in the regression line which represents the cluster. This total vertical distance is also the quantity which the algorithm optimizes in their objective function. The temporal k-means can also be viewed as a hard-clustering version of PSM when only the subtype effect is in consideration.

One can notice that the quantity in the objective function of temporal k-means is similar to the quantity $a(i)$ in equation (1), except that quantity $a(i)$ in equation (1) uses the average instead of the sum as in temporal k-means. As a result, almost all *longSil* coefficients computed from the result of this method are greater than zero. However, the k-means algorithm, in their formulation, doesn't attempt to optimize the separation between clusters, which sometimes return clusters with little difference.

Similar to the experiment setting in Luong and Chandola's study [11], in our experiment, ten cubic b-spline basis functions are used with an addition of the intercept term. The knots are chosen based on quantiles of the set of all timestamps in the dataset. The temporal k-means algorithm is run with three different random initializations and the best result among the three is returned. For each random initialization, the algorithm is run until convergence achieves or when the number of iteration exceeds 100.

4.3 Spline Induced Clustering

Singh et al. [21] introduced this method to cluster patient disease progression while compensating for missing values which are common in EHR data. By applying data augmentation, missing values in time series of clinical observations are imputed. The algorithm starts with an estimation of values of the clinical marker over the full range of observations by implementing a statistical spline regression. By transforming the longitudinal disease profile into a continuous curve using spline regression, there is an estimation of the clinical marker at every point in time.

Because time series clustering methods require observations to be present at equal distances, the imputed value in a time series was computed from the spline regression in a consistent interval. Next, a dissimilarity matrix between different patient profiles is computed. Each entry in this matrix is obtained by calculating dissimilarity between two time series of imputed values with Euclidean distance as the distance metric.

Using this dissimilarity matrix, Partitioning Around Medoids (PAM) clustering algorithm was used to obtain clusters of time series. PAM is a clustering method similar to K-means. However, PAM works with medoids which represent the dataset in groups while K-means works with centroids which are artificially created entities that represent clusters. The PAM algorithm partitions the dataset of n objects into k clusters by minimizing the distance between points assigned to a cluster and a point evaluated as the center of the cluster (medoid).

4.4 Time-Aware LSTM Auto-encoder

Baytas et al. [1] attempted to summarize the longitudinal patient profiles by projecting them into a latent space and perform k-means algorithm with the new embedded representations of patients. The underlying model of this approach is Time-Aware Long Short-Term Memory (T-LSTM) - a variant of recurrent neural networks which focuses on modeling sequences. The embedded representation is obtained by using two T-LSTM structures, one for encoding patient clinical observations into a latent space and one for decoding this representation to reconstruct the original clinical observations. The parameters of the model are trained so that the reconstruction error is minimized. This is an auto-encoder approach in which longitudinal patient profile is encoded into a latent space and later being reconstructed from the latent representation. In the experiments, we use the hidden unit at the last time step of the encoder to represent the patient profile as given in the original study of Baytas et al. [1]. In addition, we set the dimension of the hidden unit in T-LSTM Auto-encoder to be 64 as suggested in Luong and Chandola's study [12] in which they used the same CKD datasets.

5 Experiments

In this section, we present our experimental results with different subtyping methods. In particular, Section 5.1 explains the datasets that we will use in our experiments. Next, in Section 5.2, we qualitatively evaluate *longSil* in four typical clustering results, each with a subtyping method to see if this measure reflects well the quality of clustering results. Finally, in Section 5.3, we use our proposed quantitative evaluation metric to compare the performance of different subtyping methods.

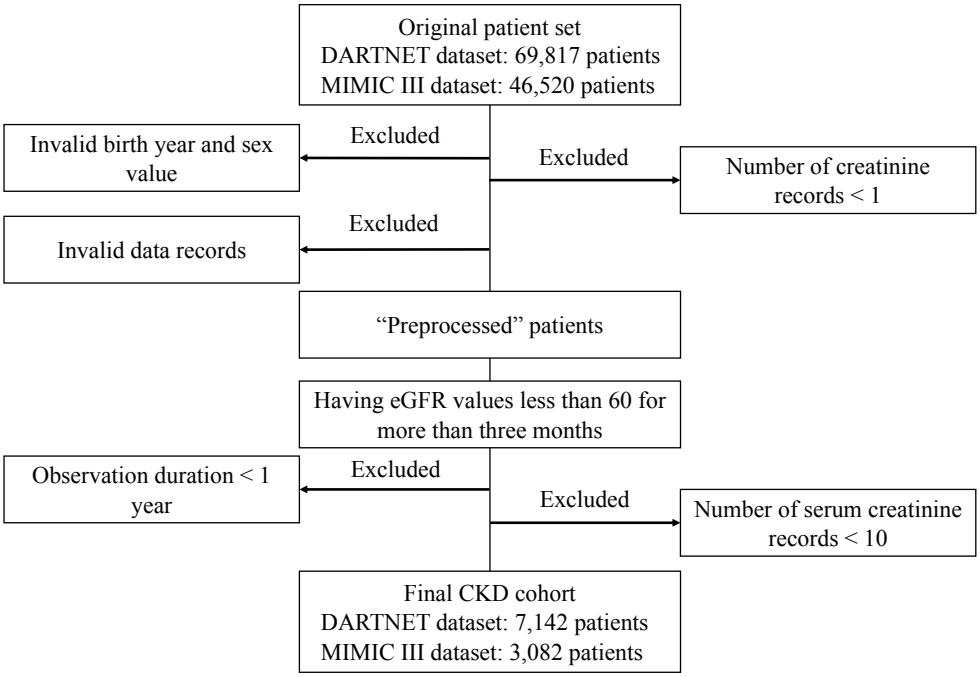


Fig. 2: Preprocessing procedure to obtain CKD cohort

5.1 Data

The datasets we use in this article are subsets of two larger datasets, called DARTNet [15] and MIMIC-III [6]. The DARTNet dataset was collected by a federated network of electronic clinical data from multiple organizations across the United States which contains health information of nearly 70,000 patients having various degree of kidney damage. One the other hand, MIMIC (Medical Information Mart for Intensive Care) is a large, openly available database containing de-identified health record of 46,520 patients who admitted to critical care units at a large tertiary care hospital [6]. From these datasets, we extract sets of patients having Chronic Kidney Disease (CKD) - a rising health problem in both the US and worldwide. As CKD is a chronic disease with heterogeneous disease progressions, extracting the longitudinal disease subtypes will enable further research on the underlying mechanism of each subtype and subsequently tailor the treatment for each subtype. In CKD, one main indicator of disease severity is *estimated Glomerular Filtration Rate* (eGFR) that measures the condition of the kidney [14]. The eGFR value can be estimated using the CKD-EPI equation which takes into account serum creatinine measure as well as patients' age, sex and race [8]. In our experiment, a longitudinal profile of a patient is a set of eGFR observations with the corresponding timestamps in which the first few eGFR observations are under 60 for more than three months - a criterion used for determining stage 3 CKD according to clinical guidelines [14]. This criterion of choosing the beginning of patient longitudinal profile can be considered as selecting the patients who are transitioning to stage 3 CKD as well as the set of existing CKD patients in stage 3, 4 and 5. To ensure that patients have enough eGFR observations within a long enough time span to understand their disease progression, we only retain patients with at least one year of eGFR observations since their first eGFR records and having at least ten serum creatinine observations. Figure 2 shows an outline of preprocessing steps to obtain CKD cohort in DARTNet and MIMIC-III datasets which is similar to one in the study of Luong and Chandola [11]. This preprocessing step results in patient cohorts of $7,142$ patients and $3,082$ patients in DARTNet and MIMIC-III datasets respectively. It is also worth to mention that as MIMIC-III dataset was collected from patients who admitted to critical care units, its data distribution is different from DARTNet dataset. In particular, as shown in Table 1, although DARTNet dataset has more patients than MIMIC-III, DARTNet dataset has fewer eGFR observations in comparison with MIMIC-III and the average time span for each patient's eGFR trajectory in DARTNet dataset is also much longer than the one in MIMIC-III dataset.

5.2 Qualitative evaluation of *longSil*

In this section, we inspect some typical subtyping results and see how *longSil* coefficient reflects the quality of the results. In each of the following sections, a clustering result obtained from each subtyping method is inspected. In order to demonstrate the use of *longSil* in capturing the desired characteristics of clustering result, we chose the number of clusters K in each experiment so that both high quality and low quality clustering results are included in the experiment.

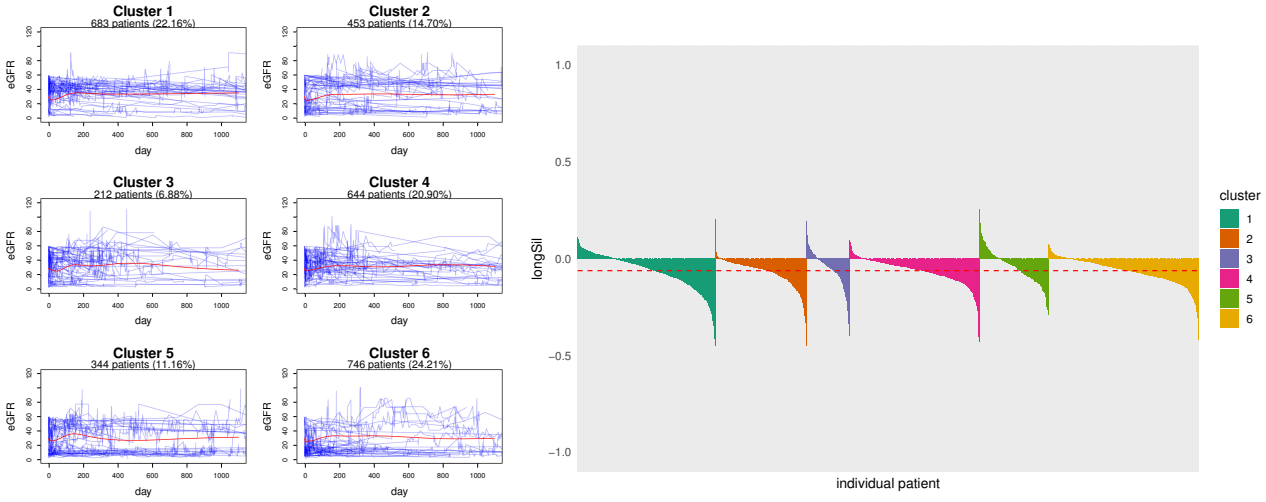
In all of our experiments, we use ten cubic b-splines as the set of basis functions for our *longSil* evaluation metrics. The knots are chosen based on quantiles of the set of all timestamps in the dataset. Although other choices of basis functions can be considered and empirically evaluated, in our experiments, we only consider b-splines as our basis functions. Further evaluation of other basis functions will be a topic for future research.

5.2.1 PSM with 6 clusters on MIMIC-III dataset

In this experiment, we evaluate the clustering result of PSM with 6 clusters on MIMIC-III dataset and assess how the qualitative clustering evaluation is reflected in quantitative measure *longSil*. The resulting subtypes obtained by PSM is shown in Figure 3a. As we can observe from the figure, within each cluster, the longitudinal profiles vary widely and there is no clear trend in each cluster. In addition, based on the clustering result as shown in Figure 3a, there is no clear distinction to distinguish one cluster against another.

As explained earlier, PSM is a probabilistic model in which subtyping effect is only a component of the model. Beside this subtyping effect, there are other effects including demographic effects as well as individual effects coupled within the model. Therefore, the resulting subtypes as shown in Figure 3a may not represent clear trends of progressions in CKD.

In Figure 3b, the distribution of *longSil* for this clustering result is shown. In the figure, there are many negative values of *longSil* coefficients which signals that many longitudinal profiles are closer to other clusters rather than its own. This leads to an overall low average *longSil* coefficient in the result. From this experiment, we see that average *longSil* coefficient can be used as a gauge for the quality of clustering result.



(a) Resulting disease subtypes (only 50 patient profiles are shown for each cluster)

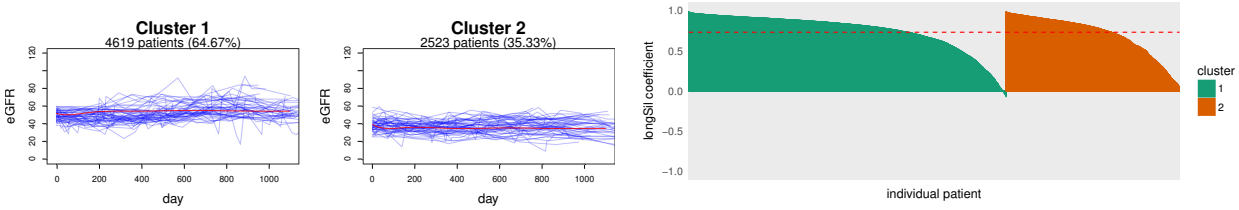
(b) Distribution of *longSil* coefficients (red dotted line indicate the average *longSil* coefficient)

Fig. 3: Subtyping result output from PSM with $k = 6$ in MIMIC-III dataset (Best viewed in color)

5.2.2 Temporal K-means with 2 clusters on DARTNet dataset

The resulting subtypes obtained by performing Temporal K-means on DARTNet dataset with $k = 2$ is shown in Figure 4a. In the figure, cluster 1 has higher eGFR values in comparison with cluster 2. In fact, the mean of eGFR values across all observations in cluster 1 is 53.6 while the corresponding mean for cluster 2 is 34.5. The two resulting clusters as shown in Figure 4a seems to have both qualities that we consider in the evaluation of

clustering result: tightness and separation. The two clusters are both tight in capturing similar progressions. In addition, they are distinctive from each other. This observation is then reflected in the distribution of *longSil* as shown in Figure 4b. In this figure, almost all patients have positive *longSil* value which indicates that they are assigned into correct clusters. This is because the objective function in Temporal K-means has a step that directly optimizes for the tightness of cluster. In particular, the assignment step of Temporal K-means assigns the patient into a cluster such that the total sum of the squared vertical distance between observations and the corresponding values in the regression line of the cluster is minimized.



(a) Resulting disease subtypes (only 50 patient profiles are shown for each cluster)

(b) Distribution of *longSil* coefficients (red dotted line indicate the average *longSil* coefficient)

Fig. 4: Subtyping result output from Temporal K-means with $k = 2$ in DARTNet dataset (Best viewed in color)

5.2.3 Splince Induced Clustering with 6 clusters on MIMIC-III dataset

In this experiment, we use Spline Induced Clustering to cluster CKD patient profiles into four clusters. Figure 5a shows the resulting clusters obtained by this method. Similar to the case of PSM with 6 clusters shown in Section 5.2.1, the resulting clusters do not exhibit any clear trends of progressions in CKD. In addition, the clusters are also not well separated. This may suggest that the missing value imputation step in Spline Induced Clustering may fail to estimate the missing values in patient trajectories and consequently leads to poor clustering result.

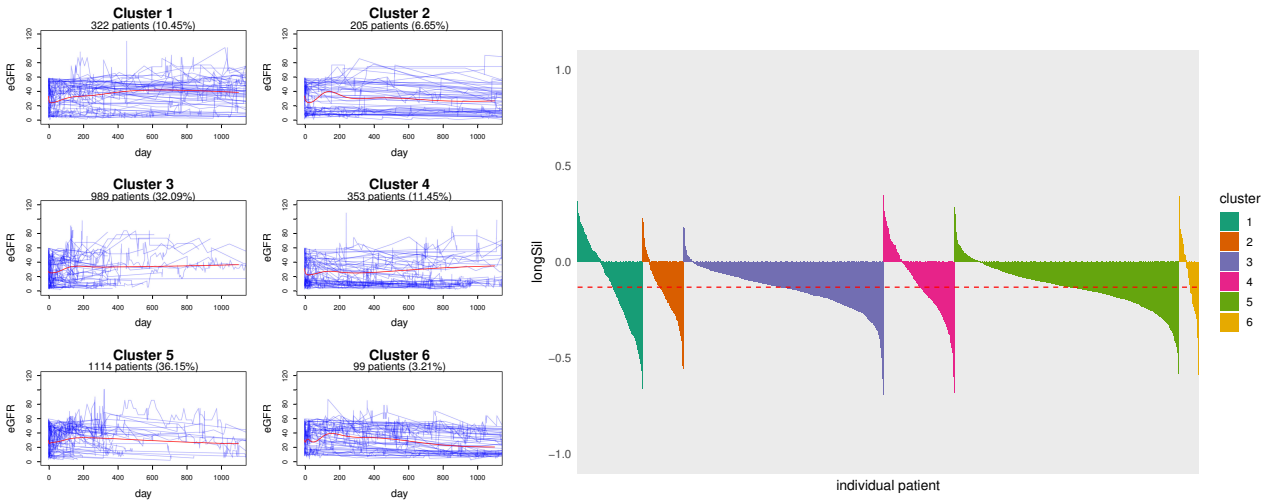
The corresponding *longSil* coefficients of this clustering result are shown in Figure 5b. This reflects well the quality of our clustering result with many negative *longSil* values and the overall negative value of average *longSil* coefficient.

5.2.4 T-LSTM Auto-encoder with 4 clusters on DARTNet dataset

We now examine another clustering result obtained by using T-LSTM Auto-encoder to project longitudinal patient profiles into the latent space and use these embedded representations to perform K-means clustering with 4 clusters to obtain clusters of patient profiles. Figure 6a shows the CKD profiles of four clusters. One can observe from this figure that although the four clusters are highly overlapped with similar overall trends, the trajectories of CKD profiles in each cluster are tight and follow a similar trend. In other words, one may judge that this clustering result has high cluster tightness while suffering from a low degree of separation between clusters. This observation can also be seen in Figure 6b that shows the distribution of *longSil* coefficients for the set of CKD patients. Overall, the average *longSil* coefficient is positive with many individual *longSil* coefficients are positive indicating the majority of CKD are assigned to their closest clusters while there is also a set of patients with negative *longSil* coefficients that are probably incorrectly assigned to the clusters because of overlapping clusters.

5.3 Quantitative comparison of different subtyping methods using *longSil* coefficient

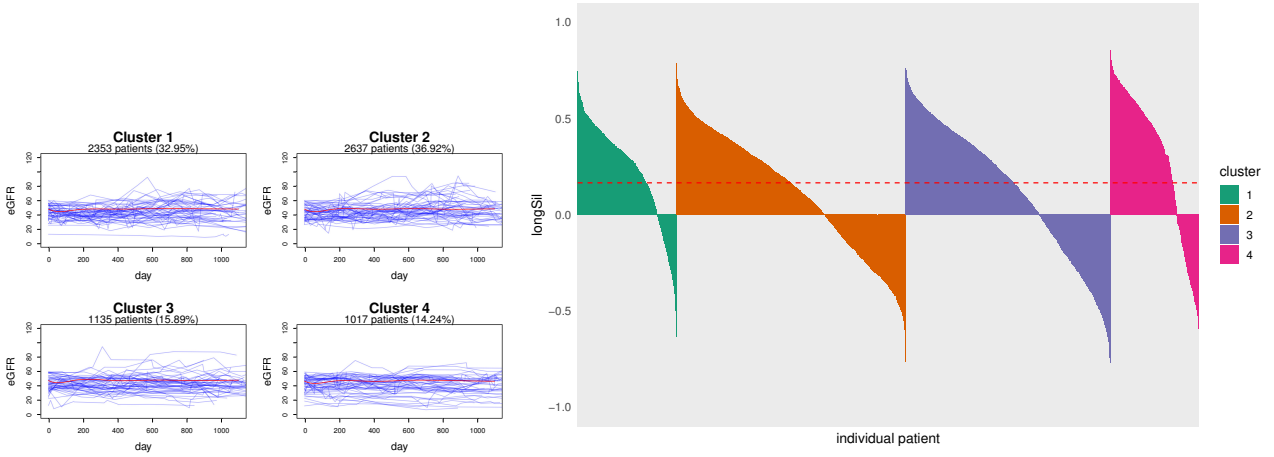
In the previous section, we have observed how well *longSil* coefficient can capture the quality of clustering result in terms of cluster tightness and degree of separation between clusters. Furthermore, the overall distribution of *longSil* can be summarized by average *longSil* coefficient, which captures the overall quality of a clustering result. In this section, we further use this average *longSil* coefficient to compare the clustering results of four



(a) Resulting disease subtypes (only 50 patient profiles are shown for each cluster)

(b) Distribution of *longSil* coefficients (red dotted line indicate the average *longSil* coefficient)

Fig. 5: Subtyping result output from Spline Induced Clustering with $k = 6$ in MIMIC-III dataset (Best viewed in color)



(a) Resulting disease subtypes (only 50 patient profiles are shown for each cluster)

(b) Distribution of *longSil* coefficients (red dotted line indicate the average *longSil* coefficient)

Fig. 6: Subtyping result output from T-LSTM Auto-encoder with $k = 4$ in MIMIC-III dataset (Best viewed in color)

methods as described in Section 4. In our experiment, we let the number of subtypes span from 2 to 15 and compute the average *longSil* coefficient for each case.

Figure 7 shows the comparison between four longitudinal disease subtyping methods in terms of their average *longSil* values in two datasets (DARTNet and MIMIC-III). As a reminder, a higher value of average *longSil* coefficient indicates better result, in terms of overall cluster's tightness and degree of separation. Moreover, average *longSil* coefficient below zero may indicate overall incorrect placement of patient profiles.

As we can observe from Figure 7, Temporal K-means consistently has better results in comparison with the others, in both datasets, across different numbers of clusters. This is not surprising as Temporal K-means shares some common computational objectives with *longSil* - our evaluation metric. Temporal K-means directly optimizes for cluster tightness by assigning patient profile into the closest cluster. In addition, both Temporal K-means and *longSil* represent the cluster by a regression line using all observations of patient profiles assigned to it. Therefore, Temporal K-means always achieves positive values for average *longSil* coefficients as shown in Figure 7.

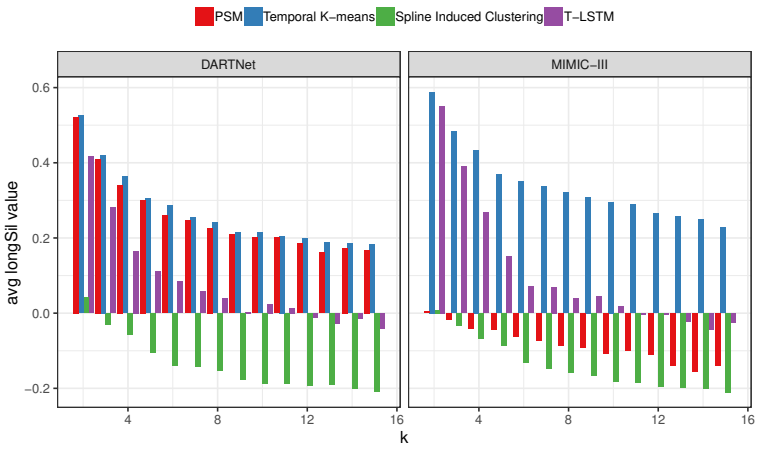


Fig. 7: Comparison of three longitudinal disease subtyping methods with different number of clusters in terms of average *longSil* cofficient (Best viewed in color)

	DARTNet	MIMIC-III
Number of patients	7,142	3,082
Number of eGFR records	113,097	153,471
Average eGFR records per patient	15.8	48.8
Average time span of clinical records of a patient	1744.4 days	826.4 days

Table 1: Some differences between cohort in DARTNet and MIMIC-III dataset

Regarding the performance of PSM in this comparison, one may observe that it performs well in DARTNet dataset with positive average *longSil* coefficients but it fails to obtain good results in MIMIC-III dataset. This contrasting performance between two different datasets may originate from the differences between the two datasets. Table 1 highlights some differences between patient cohort in DARTNet and MIMIC-III datasets. In particular, the cohort in MIMIC-III dataset has a significantly fewer number of patients in comparison with one in DARTNet dataset. In addition, the average of eGFR records per patient in MIMIC-III dataset is 48.8 while its counterpart in DARTNet dataset is only 15.8. Moreover, the average time span of clinical records in MIMIC-III dataset is much shorter (826.4 days) in comparison with DARTNet dataset (1,744.4 days). Therefore, having fewer patients, having more eGFR observations per patient, and having a shorter time span in clinical records are the three potential causes to the undesirable results of PSM in MIMIC-III dataset.

In our comparison, the only method that has consistent negative values of average *longSil* coefficients across two datasets is Spline Induced Clustering method. Probably the approach of imputing missing values by using spline regression is not suitable for CKD datasets as observations are sparse and the number of observations needs to be imputed is too many which results into a poor estimation of the patient eGFR progressions and subsequently poor clustering result.

Among the four methods we used in the comparison, T-LSTM Auto-encoder is the only method that does not attempt to group patients based on their disease progressions directly. Instead, it models the temporality in the sequences of eGFR observations and focuses on finding embedded representations such that they can reconstruct well the original sequences. The representations are later used as inputs of K-means algorithm to find groups of patients that are close to each other in the latent space. It is interesting to examine whether the proximity in the latent space also reflects the proximity in CKD progressions between different patients. As shown in Figure 7, in both DARTNet and MIMIC-III dataset, T-LSTM Auto-encoder represents well the clusters of CKD progressions with positive *longSil* coefficients for the number of clusters from 2 to 10. As the number of clusters increases, the performance deteriorates. When the number of clusters exceeds 8, the average *longSil* coefficient becomes negative as the quality of clustering result no longer reflects well the progression of eGFR values.

In Figure 7, one can observe a general trend of decreasing average *longSil* value as we increase the number of clusters across all four subtyping methods. In the original formulation of Silhouette coefficient in traditional clustering, the average Silhouette coefficient for a clustering result can be peaked at the optimal number of

clusters [10, 17]. However, with the new formulation of *longSil* coefficient for evaluating clustering longitudinal profiles, it seems that the average *longSil* value will follow the decreasing trend as we increase the number of clusters. This decreasing trend makes the choice for an optimal number of clusters less viable with *longSil*. However, given the same number of clusters, *longSil* is still a reliable measure for evaluating the quality of different clustering results as it effectively captures the overall cluster's tightness and degree of separation between clusters.

6 Discussion

The derived longitudinal disease subtypes obtained from subtyping methods are usually difficult to analyze, especially when we want to assess its clinical relevance for further analysis. As shown in the experimental results in Section 5, our proposed evaluation metric gives a good indication of the quality of clustering result. It reflects well the cluster tightness and degree of separation between clusters. It allows us to quickly identify interesting sub-groups of patients that are tight and highly distinctive from other patients. This can be used as an augmentation for traditional cluster evaluation to detect groups of patients with interesting disease progressions.

When using *longSil* to compare different models, it is important to note that we inherently assume the set of basis functions that represent the disease progressions of patients within a cluster. In our experiment, we use splines as the set of basis functions. However, depending on different datasets and different diseases, another set of basis functions may be better used. Also, in our evaluation metric, we only use the progressions of one clinical marker to evaluate the quality of the clustering result. In many diseases, the disease progressions are complex with many underlying factors including demographic, phenotypic physiological characteristics of patients. This requires a more complex evaluation process to identify whether disease subtypes are clinically relevance. In particular, for the case of evaluating the disease subtypes with multiple clinical markers, one can extend the clustering quality measure to cope with multiple clinical markers by computing the clustering quality measure with respect to each single clinical marker and subsequently combine them with some weights indicating the importance of each respective clinical marker in the overall quality measure.

In section 5.2, we have evaluated the quality of *longSil* with some specific clustering results obtained from different algorithms. Another way to evaluate the quality of *longSil* as a clustering evaluation metric is to validate it against external clustering validation which brings external data of patients such as demographics, historical medications, and diagnoses into consideration. However, for the case of CKD disease, it is not trivial to pick relevant external information from patients' medical records to compare against the clustering result. Such validation will require further research to identify relevant causes or underlying mechanisms that enable CKD progression among the patients in the same cluster to be similar. Therefore, within the scope of this article, we cannot adopt this approach of validating *longSil* using external clustering validation. In the future, when we have a further understanding of CKD, we can use that information to further reinforce our understanding of the *longSil* as a computational method to evaluate the groups of patients with similar disease progression.

Besides the use of evaluating the quality of subtyping result obtained from a subtyping method, *longSil* can also be used as a tool to diagnose different problems of the results. In particular, by examining the distribution of degree of separation or cluster tightness, an algorithm designer can have a better understanding of their subtyping model and have a better idea of improving the model.

In the literature of internal clustering validation measure, Silhouette coefficient is only an approach among many quantitative approaches that measure the quality of clustering result based on the concept of cluster tightness and degree of separation between different clusters [10]. As we built our quantitative measure based on a regression line that allows us to represent a cluster of longitudinal profiles and subsequently use it to compute the distance between a longitudinal profile and a cluster, the same methodology can also be used to adapt many other internal clustering validation measures for the problem of clustering longitudinal profiles. In our future works, we will further study the differences between different internal clustering validation measures in the context of clustering longitudinal profiles.

It is also worth to mention that in addition to internal clustering validation, a holistic evaluation of different subtyping methods will require them to evaluate with external data such as medical outcomes, demographic distribution or diagnosis codes. In separate studies [11, 13, 21], three subtyping methods including PSM, Temporal K-means and Spline Induced Clustering have been evaluated in the context of CKD with various external data. However, a comprehensive evaluation across different subtyping methods using external data will

need a different methodology to systematically compare between different methods with different experiment settings. As the focus of this article is internal clustering evaluation, a comparison between different subtyping results with external data will be the topic of future research.

Among four disease subtyping methods we examine in this article, T-LSTM Auto-encoder is the only method that uses deep learning to embed patient longitudinal profiles into a latent space. Beyond T-LSTM Auto-encoder, there are also many other deep learning approaches to embed longitudinal EHR profiles into a latent space. However, many of the approaches use diagnosis codes and medication codes as their primary inputs for the model [2, 3]. Although diagnosis codes and medication codes can also be leveraged in analyzing EHR data, in the context of CKD, as our primary clinical marker is eGFR - a laboratory value that measures of kidney function, the methods of obtaining patient embeddings via diagnosis codes and medication codes of longitudinal EHR data cannot be used directly. In another study of Lasko et al. [7], the author used Gaussian Process Regression to transform a noisy, irregularly sampled and sparse observations of a longitudinal profile into a continuous longitudinal probability distribution. For each longitudinal profile, a set of small patches of mean values of the inferred probability distribution with the same length is extracted to train an auto-encoder to learn the hidden representation of each input patch. Although this method can infer the hidden representation of an input patch, it is not clear how to combine the set of hidden representation of multiple patches to construct a hidden representation for a longitudinal profile. In a different study [16], raw EHR data are represented based on Fast Healthcare Interoperability Resources (FHIR) format and subsequently used for training deep learning models to predict various clinical outcomes. Although its approach can obtain strong performance in multiple supervised tasks, the paper provides little insights on how to use such an approach to represent patient longitudinal profiles.

7 Conclusion

In this article, we have introduced a new quantitative and objective evaluation metric that can assess the quality of longitudinal disease subtypes. This evaluation metric can capture both the notion of the cluster's tightness as well as the degree of separation between different clusters. Moreover, the distribution of individual *longSil* coefficients can also be visualized as a graphical aid for validating longitudinal disease subtypes. Using this proposed evaluation metric, we have performed a comparison between four available longitudinal disease subtyping methods including Probabilistic Subtyping Model, Temporal K-means, Spline Induced Clustering, and Time-Aware LSTM Auto-encoder. The experiments show that our proposed evaluation metric reflects well the quality of result obtained from a subtyping model and allows us to gauge the performance of different subtyping models when applying for Chronic Kidney Disease datasets.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou. Patient subtyping via Time-Aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74. ACM, 2017.
2. E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM, 2016.
3. E. Choi, C. Xiao, W. Stewart, and J. Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *Advances in Neural Information Processing Systems*, pages 4547–4557, 2018.
4. G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2012.
5. A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.
6. A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016.

7. T. A. Lasko, J. C. Denny, and M. A. Levy. Correction: Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLoS one*, 8(8), 2013.
8. A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene, et al. A new equation to estimate glomerular filtration rate. *Annals of internal medicine*, 150(9):604–612, 2009.
9. T. W. Liao. Clustering of time series data - a survey. *Pattern recognition*, 38(11):1857–1874, 2005.
10. Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916. IEEE, 2010.
11. D. T. A. Luong and V. Chandola. A k-means approach to clustering disease progressions. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 268–274, Aug 2017.
12. D. T. A. Luong and V. Chandola. Learning deep representations from clinical data for chronic kidney disease. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 2019.
13. D. T. A. Luong, D. Tran, W. D. Pace, M. Dickinson, J. Vassalotti, J. Carroll, M. Withiam-Leitch, M. Yang, N. Satchidanand, E. Staton, et al. Extracting deep phenotypes for chronic kidney disease using electronic health records. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 5, 2017.
14. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American journal of kidney diseases: the official journal of the National Kidney Foundation*, 39(2 Suppl 1):S1, 2002.
15. W. D. Pace, C. Fox, T. White, D. Graham, L. M. Schilling, and R. David. The DARTNet institute: seeking a sustainable support mechanism for electronic data enabled research networks. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 2(2):6, 2014.
16. A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.
17. P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.
18. P. Schulam and R. Arora. Disease trajectory maps. In *Advances in Neural Information Processing Systems*, pages 4709–4717, 2016.
19. P. Schulam and S. Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756, 2015.
20. P. Schulam, F. Wigley, and S. Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *AAAI*, pages 2956–2964, 2015.
21. P. Singh, V. Chandola, and C. Fox. Automatic extraction of deep phenotypes for precision medicine in chronic kidney disease. In *Proceedings of the 2017 International Conference on Digital Health*, pages 195–199. ACM, 2017.
22. M. J. Zaki and W. Meira Jr. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.