Supplementary Material
Temporal Phenotyping using Deep Predictive Clustering of Disease Progression

Changhee Lee [1] **Mihaela van der Schaar** [2 3 1]

A. AC-TPC for Regression and Binary Classification Tasks

As the task changes, estimating the label distribution and calculating the KL divergence in (1) of the manuscript must be redefined accordingly: For regression task, i.e., $\mathcal{Y} = \mathbb{R}$, we modify the predictor as $g_\phi : \mathcal{Z} \to \mathbb{R}$ and replace ℓ_1 by $\ell_1(y_t, \bar{y}_t) = \|y_t - \bar{y}_t\|_2^2$. Minimizing $\ell_1(y_t, \bar{y}_t)$ is equivalent to minimizing the KL divergence between $p(y_t|\mathbf{x}_{1:t})$ and $p(y_t|s_t)$ when we assume these probability densities follow Gaussian distribution with the same variance. For the M-dimensional binary classification task, i.e., $\mathcal{Y} = \{0,1\}^M$, we modify the predictor as $g_\phi : \mathcal{Z} \to [0,1]^M$ and replace ℓ_1 by $\ell_1(y_t, \bar{y}_t) = -\sum_{m=1}^M y_t^m \log \bar{y}_t^m + (1 - y_t^m)\log(1 - \bar{y}_t^m)$ which is required to minimize the KL divergence. Here, y_t^m and $\bar{y}_t^m$ indicate the m-th element of y_t and $\bar{y}_t$, respectively. The basic assumption here is that the distribution of each binary label is independent given the input sequence.

B. Detailed Derivation of (5)

To derive the gradient of the predictive clustering loss in (5) of the manuscript with respect $\omega_A = [\theta, \psi]$, we utilized the ideas from actor-critic models (Konda & Tsitsiklis, 2000) on $\mathcal{L}_A(\theta, \psi, \phi) = \mathcal{L}_1(\theta, \psi, \phi)$:

$$\begin{aligned}\nabla_{\omega_A}\mathcal{L}_A(\theta,\psi,\phi) &= \mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\nabla_{\omega_A}\left(\sum_{t=1}^{T}\mathbb{E}_{s_t\sim Cat(\pi_t)}\big[\ell_1(y_t,\bar{y}_t)\big]\right)\right] + \alpha\nabla_{\omega_A}\mathcal{L}_2(\theta,\psi) \\ &= \mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\sum_{t=1}^{T}\mathbb{E}_{s_t\sim Cat(\pi_t)}\big[\ell_1(y_t,\bar{y}_t)\nabla_{\omega_A}\log\pi_t(s_t)\big]\right] + \alpha\nabla_{\omega_A}\mathcal{L}_2(\theta,\psi),\end{aligned} \tag{S.1}$$

where the second equality comes from the following derivation of the former term:

$$\begin{aligned}\mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\nabla_{\omega_A}\left(\sum_{t=1}^{T}\mathbb{E}_{s_t\sim Cat(\pi_t)}\left[\ell_1(y_t,\bar{y}_t)\right]\right)\right] &= \mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\nabla_{\omega_A}\left(\sum_{t=1}^{T}\sum_{s_t\in\mathcal{K}}\pi_t(s_t)\ell_1(y_t,\bar{y}_t)\right)\right] \\ &= \mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\sum_{t=1}^{T}\sum_{s_t\in\mathcal{K}}\nabla_{\omega_A}\pi_t(s_t)\ell_1(y_t,\bar{y}_t)\right] \\ &= \mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\sum_{t=1}^{T}\sum_{s_t\in\mathcal{K}}\frac{\nabla_{\omega_A}\pi_t(s_t)}{\pi_t(s_t)}\pi_t(s_t)\ell_1(y_t,\bar{y}_t)\right] \\ &= \mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\sum_{t=1}^{T}\sum_{s_t\in\mathcal{K}}\pi_t(s_t)\ell_1(y_t,\bar{y}_t)\nabla_{\omega_A}\log\pi_t(s_t)\right] \\ &= \mathbb{E}_{\mathbf{x},y\sim p_{XY}}\left[\sum_{t=1}^{T}\mathbb{E}_{s_t\sim Cat(\pi_t)}\left[\ell_1(y_t,\bar{y}_t)\nabla_{\omega_A}\log\pi_t(s_t)\right]\right].\end{aligned}$$

C. Pseudo-Code of AC-TPC

As illustrated in Section 3.2, AC-TPC is trained in an iterative fashion. We provide the pseudo-code for optimizing our model in Algorithm 1 and that for initializing the parameters in Algorithm 2.

Algorithm 1 Pseudo-code for Optimizing AC-TPC

```
Input: Dataset $\mathcal{D} = \{(\mathbf{x}_t^n, y_t^n)_{t=1}^{T^n}\}_{n=1}^{N}$, number of clusters $K$, coefficients $(\alpha, \beta)$,
       learning rate $(\eta_A, \eta_C, \eta_E)$, mini-batch size $n_{mb}$, and update step $M$
Output: AC-TPC parameters $(\theta, \psi, \phi)$ and the embedding dictionary $\mathcal{E}$
Initialize parameters $(\theta, \psi, \phi)$ and the embedding dictionary $\mathcal{E}$ via Algorithm 2

repeat
  Optimize the Encoder, Selector, and Predictor
  for $m = 1, \cdots, M$ do
    Sample a mini-batch of $n_{mb}$ data samples: $\{(\mathbf{x}_t^n, y_t^n)_{t=1}^{T^n}\}_{n=1}^{n_{mb}} \sim \mathcal{D}$
    for $n = 1, \cdots, n_{mb}$ do
      Calculate the assignment probability: $\pi_t^n = [\pi_t^n(1) \cdots \pi_t^n(K)] \leftarrow h_\psi(f_\theta(\mathbf{x}_{1:t}^n))$
      Draw the cluster assignment: $s_t^n \sim Cat(\pi_t^n)$
      Calculate the label distributions: $\bar{y}_t^n \leftarrow g_\phi(\mathbf{e}(s_t^n))$ and $\hat{y}_t^n \leftarrow g_\phi(f_\theta(\mathbf{x}_{1:t}^n))$
    end for
    Update the encoder $f_\theta$ and selector $h_\psi$:
        $\theta \leftarrow \theta - \eta_A \left( \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{t=1}^{T^n} \ell_1(y_t^n, \bar{y}_t^n) \nabla_\theta \log \pi_t^n(s_t^n) - \alpha \nabla_\theta \sum_{k=1}^{K} \pi_t^n(k) \log \pi_t^n(k) \right)$
        $\psi \leftarrow \psi - \eta_A \left( \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{t=1}^{T^n} \ell_1(y_t^n, \bar{y}_t^n) \nabla_\psi \log \pi_t^n(s_t^n) - \alpha \nabla_\psi \sum_{k=1}^{K} \pi_t^n(k) \log \pi_t^n(k) \right)$
    Update the predictor $g_\phi$:
        $\phi \leftarrow \phi - \eta_C \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{t=1}^{T^n} \nabla_\phi \ell_1(y_t^n, \bar{y}_t^n)$
  end for

  Optimize the Cluster Centroids
  for $m = 1, \cdots, M$ do
    Sample a mini-batch of $n_{mb}$ data samples: $\{(\mathbf{x}_t^n, y_t^n)_{t=1}^{T^n}\}_{n=1}^{n_{mb}} \sim \mathcal{D}$
    for $n = 1, \cdots, n_{mb}$ do
      Calculate the assignment probability: $\pi_t^n = [\pi_t^n(1) \cdots \pi_t^n(K)] \leftarrow h_\psi(f_\theta(\mathbf{x}_{1:t}^n))$
      Draw the cluster assignment: $s_t^n \sim Cat(\pi_t^n)$
      Calculate the label distributions: $\bar{y}_t^n \leftarrow g_\phi(\mathbf{e}(s_t^n))$
    end for
    for $k = 1, \cdots, K$ do
      Update the embeddings $\mathbf{e}(k)$:
          $\mathbf{e}(k) \leftarrow \mathbf{e}(k) - \eta_E \left( \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{t=1}^{T^n} \nabla_{\mathbf{e}(k)} \ell_1(y_t^n, \bar{y}_t^n) - \gamma \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \nabla_{\mathbf{e}(k)} \ell_1\big(g_\phi(\mathbf{e}(k)), g_\phi(\mathbf{e}(k'))\big) \right)$
    end for
    Update the embedding dictionary: $\mathcal{E} \leftarrow \{\mathbf{e}(1), \dots \mathbf{e}(K)\}$
  end for
until convergence
```

Algorithm 2 Pseudo-code for pre-training AC-TPC

```
Input: Dataset $\mathcal{D} = \{(\mathbf{x}_t^n, y_t^n)_{t=1}^{T^n}\}_{n=1}^{N}$, number of clusters $K$, learning rate $\eta$, mini-batch size $n_{mb}$
Output: AC-TPC parameters $(\theta, \psi, \phi)$ and the embedding dictionary $\mathcal{E}$
Initialize parameters $(\theta, \psi, \phi)$ via Xavier Initializer

Pre-train the Encoder and Predictor
repeat
  Sample a mini-batch of $n_{mb}$ data samples: $\{(\mathbf{x}_t^n, y_t^n)_{t=1}^{T^n}\}_{n=1}^{n_{mb}} \sim \mathcal{D}$
  for $n = 1, \cdots, n_{mb}$ do
    Calculate the label distributions:   $\hat{y}_t^n \leftarrow g_\phi(f_\theta(\mathbf{x}_{1:t}^n))$
  end for
      $\theta \leftarrow \theta - \eta \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{t=1}^{T^n} \nabla_\theta \ell_1(y_t^n, \hat{y}_t^n)$      $\phi \leftarrow \phi - \eta \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{t=1}^{T^n} \nabla_\phi \ell_1(y_t^n, \hat{y}_t^n)$
until convergence

Initialize the Cluster Centroids
Calculate the embedding dictionary $\mathcal{E}$ and initial cluster assignments $c_t^n$
      $\mathcal{E}, \{\{c_t^n\}_{t=1}^{T^n}\}_{n=1}^{N} \leftarrow \texttt{K-means}(\{\{\mathbf{z}_t^n\}_{t=1}^{T^n}\}_{n=1}^{N}, K)$

Pre-train the Selector
repeat
  Sample a mini-batch of $n_{mb}$ data samples: $\{(\mathbf{x}_t^n, y_t^n)_{t=1}^{T^n}\}_{n=1}^{n_{mb}} \sim \mathcal{D}$
  for $n = 1, \cdots, n_{mb}$ do
    Calculate the assignment probability:   $\pi_t^n = [\pi_t^n(1) \cdots \pi_t^n(K)] \leftarrow h_\psi(f_\theta(\mathbf{x}_{1:t}^n))$
  end for
  Update the selector $h_\psi$:
      $\psi \leftarrow \psi + \eta \frac{1}{n_{mb}} \sum_{n=1}^{n_{mb}} \sum_{t=1}^{T^n} \sum_{k=1}^{K} c_t^n(k) \log \pi_t^n(k)$
until convergence
```

D. Details of the Datasets

D.1. UKCF Dataset

UK Cystic Fibrosis registry (UKCF)[1] records annual follow-ups for 5,171 adult patients (aged 18 years or older) over the period from 2008 and 2015, with a total of 25,012 hospital visits. Each patient is associated with 89 variables (i.e., 11 static and 78 time-varying features), including information on demographics and genetic mutations, bacterial infections, lung function scores, therapeutic managements, and diagnosis on comorbidities. The detailed statistics are given in Table S.1.

D.2. ADNI Dataset

Alzheimer's Disease Neuroimaging Initiative (ADNI)[2] study consists of 1,346 patients with a total of 11,651 hospital visits, which tracks the disease progression via follow-up observations at 6 months interval. Each patient is associated with 21 variables (i.e., 5 static and 16 time-varying features), including information on demographics, biomarkers on brain functions, and cognitive test results. The three diagnostic groups were normal brain functioning (0.55), mild cognitive impairment (0.43), and Alzheimer's disease (0.02). The detailed statistics are given in Table S.2.

E. Details of the Benchmarks

We compared AC-TPC in the experiments with clustering methods ranging from conventional approaches based on K-means to the state-of-the-art approaches based on deep neural networks. The details of how we implemented the benchmarks are described as the following:

[1] https://www.cysticfibrosis.org.uk/the-work-we-do/uk-cf-registry

[2] https://adni.loni.usc.edu

Table S.1. Summary and description of the UKCF dataset.

STATIC COVARIATES		Type	Mean		Type	Mean
Demographic	Gender	Bin.	0.55			
Genetic	Class I Mutation	Bin.	0.05	Class VI Mutation	Bin.	0.86
	Class II Mutation	Bin.	0.87	DF508 Mutation	Bin.	0.87
	Class III Mutation	Bin.	0.89	G551D Mutation	Bin.	0.06
	Class IV Mutation	Bin.	0.05	Homozygous	Bin.	0.58
	Class V Mutation	Bin.	0.04	Heterozygous	Bin	0.42

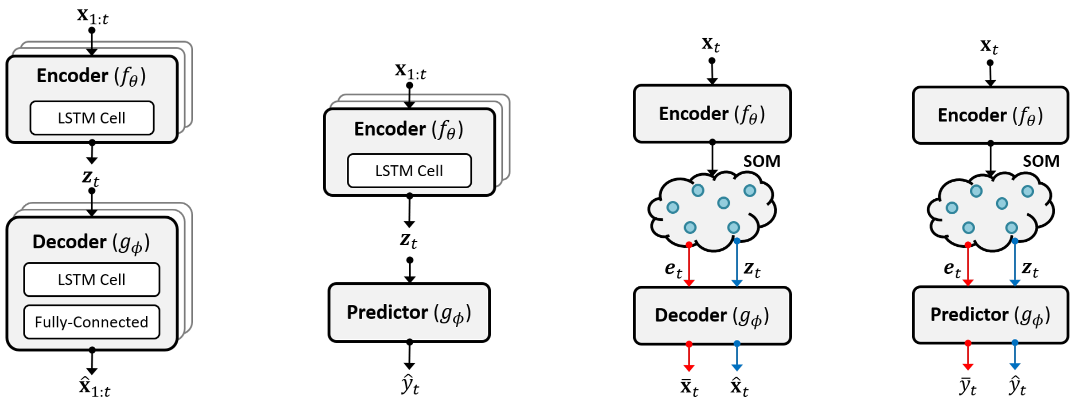
TIME-VARYING COVARIATES		Type	Mean	Min / Max		Type	Mean	Min / Max
Demographic	Age	Cont.	30.4	18.0 / 86.0	Height	Cont.	168.0	129.0 / 198.6
	Weight	Cont.	64.1	24.0 / 173.3	BMI	Cont.	22.6	10.9 / 30.0
	Smoking Status	Bin.	0.1					
Lung Func. Scores	FEV$_1$	Cont.	2.3	0.2 / 6.3	Best FEV$_1$	Cont.	2.5	0.3 / 8.0
	FEV$_1$% Pred.	Cont.	65.1	9.0 / 197.6	Best FEV$_1$% Pred.	Cont.	71.2	7.5 / 164.3
Hospitalization	IV ABX Days Hosp.	Cont.	12.3	0 / 431	Non-IV Hosp. Adm.	Cont.	1.2	0 / 203
	IV ABX Days Home	Cont.	11.9	0 / 441				
Lung Infections	B. Cepacia	Bin.	0.05		P. Aeruginosa	Bin.	0.59	
	H. Influenza	Bin.	0.05		K. Pneumoniae	Bin.	0.00	
	E. Coli	Bin.	0.01		ALCA	Bin.	0.03	
	Aspergillus	Bin.	0.14		NTM	Bin.	0.03	
	Gram-Negative	Bin.	0.01		Xanthomonas	Bin.	0.05	
	S. Aureus	Bin.	0.30					
Comorbidities	Liver Disease	Bin.	0.16		Depression	Bin.	0.07	
	Asthma	Bin.	0.15		Hemoptysis	Bin.	0.01	
	ABPA	Bin.	0.12		Pancreatitus	Bin.	0.01	
	Hypertension	Bin.	0.04		Hearing Loss	Bin.	0.03	
	Diabetes	Bin.	0.28		Gall bladder	Bin.	0.01	
	Arthropathy	Bin.	0.09		Colonic structure	Bin.	0.00	
	Bone fracture	Bin.	0.01		Intest. Obstruction	Bin.	0.08	
	Osteoporosis	Bin.	0.09		GI bleed – no var.	Bin.	0.00	
	Osteopenia	Bin.	0.21		GI bleed – var.	Bin.	0.00	
	Cancer	Bin.	0.00		Liver Enzymes	Bin.	0.16	
	Cirrhosis	Bin.	0.03		Kidney Stones	Bin.	0.02	
Treatments	Dornase Alpha	Bin.	0.56		Inhaled B. BAAC	Bin.	0.03	
	Anti-fungals	Bin.	0.07		Inhaled B. LAAC	Bin.	0.08	
	HyperSaline	Bin.	0.23		Inhaled B. SAAC	Bin.	0.05	
	HypertonicSaline	Bin.	0.01		Inhaled B. LABA	Bin.	0.11	
	Tobi Solution	Bin.	0.20		Inhaled B. Dilators	Bin.	0.57	
	Cortico Combo	Bin.	0.41		Cortico Inhaled	Bin.	0.15	
	Non-IV Ventilation	Bin.	0.05		Oral B. Theoph.	Bin.	0.04	
	Acetylcysteine	Bin.	0.02		Oral B. BA	Bin.	0.03	
	Aminoglycoside	Bin.	0.03		Oral Hypo. Agents	Bin.	0.01	
	iBuprofen	Bin.	0.00		Chronic Oral ABX	Bin.	0.53	
	Drug Dornase	Bin.	0.02		Cortico Oral	Bin.	0.14	
	HDI Buprofen	Bin.	0.00		Oxygen Therapy	Bin.	0.11	
	Tobramycin	Bin.	0.03		O$_2$ Exacerbation	Bin.	0.03	
	Leukotriene	Bin.	0.07		O$_2$ Nocturnal	Bin.	0.03	
	Colistin	Bin.	0.03		O$_2$ Continuous	Bin.	0.03	
	Diabetes Insulin	Bin.	0.01		O$_2$ Pro re nata	Bin.	0.01	
	Macrolida ABX	Bin.	0.02					

ABX: antibiotics

Table S.2. Summary and description of the ADNI dataset.

STATIC COVARIATES		Type	Mean	Min/Max (Mode)		Type	Mean	Min/Max (Mode)
Demographic	Race	Cat.	0.93	White	Ethnicity	Cat.	0.97	No Hisp/Latino
	Education	Cat.	0.23	C16	Marital Status	Cat.	0.75	Married
Genetic	APOE$_4$	Cont.	0.44	0/2				

TIME-VARYING COVARIATES		Type	Mean	Min / Max		Type	Mean	Min / Max
Demographic	Age	Cont.	73.6	55/92				
Biomarker	Entorhinal	Cont.	3.6E+3	1.0E+3 / 6.7E+3	Mid Temp	Cont.	2.0E+4	8.9E+3 / 3.2E+4
	Fusiform	Cont.	1.8E+5	9.0E+4 / 2.9E+5	Ventricles	Cont.	4.1E+4	5.7E+3 / 1.6E+5
	Hippocampus	Cont.	6.9E+3	2.8E+3 / 1.1E+4	Whole Brain	Cont.	1.0E+6	6.5E+5 / 1.5E+6
	Intracranial	Cont.	1.5E+6	2.9E+2 / 2.1E+6				
Cognitive	ADAS-11	Cont.	8.58	0/70	ADAS-13	Cont.	13.61	0/85
	CRD Sum of Boxes	Cont.	1.21	0/17	Mini Mental State	Cont.	27.84	2/30
	RAVLT Forgetting	Cont.	4.19	-12/15	RAVLT Immediate	Cont.	38.25	0/75
	RAVLT Learning	Cont.	4.65	-5/14	RAVLT Percent	Cont.	51.70	-500/100



(a) DCN-S2S (b) DCN-E2P, KM-E2P (c) SOM-VAE (d) SOM-VAE-P

Figure S.1. The block diagrams of the tested benchmarks.

Table S.3. Comparison table of benchmarks.

Methods	**Handling Time-Series**	**Clustering Method**	**Similarity Measure**	**Label Provided**	**Label Associated**
KM-DTW	DTW	K-means	DTW	N	N
KM-E2P ($\mathcal{Z}$)	RNN	K-means	Euclidean in $\mathcal{Z}$	Y	Y (indirect)
KM-E2P ($\mathcal{Y}$)	RNN	K-means	Euclidean in $\mathcal{Y}$	Y	Y (direct)
DCN-S2S	RNN	K-means	Euclidean in $\mathcal{Z}$	N	N
DCN-E2P	RNN	K-means	Euclidean in $\mathcal{Z}$	Y	Y (indirect)
SOM-VAE	Markov model	embedding mapping	reconstruction loss	N	N
SOM-VAE-P	Markov model	embedding mapping	prediction loss	Y	Y (direct)
Proposed	RNN	embedding mapping	KL divergence	Y	Y (direct)

- **Dynamic time warping followed by K-means**[3]: Dynamic time warping (DTW) is utilized to quantify pairwise distance between two variable-length sequences and, then, K-means is applied (denoted as **KM-DTW**).

- **K-means with deep neural networks**: To handle variable-length time-series data, we utilized an encoder-predictor network as depicted in Figure S.1b and trained the network based on (6) for dimensionality reduction; this is to provide fixed-length and low-dimensional representations for time-series. Then, we applied K-means on the latent encodings $\mathbf{z}$ (denoted as **KM-E2P ($\mathcal{Z}$)**) and on the predicted label distributions $\hat{y}$ (denoted as **KM-E2P ($\mathcal{Y}$)**), respectively. We implemented the encoder and predictor of KM-E2P with the same network architectures with those of our model: the encoder is a single-layer LSTM with 50 nodes and the decoder is a two-layered fully-connected network with 50 nodes in each layer.

- **Extensions of DCN**[4] (Yang et al., 2017): Since the DCN is designed for static data, we utilized a sequence-to-sequence model in Figure S.1a for the encoder-decoder network as an extension to incorporate time-series data (denoted as **DCN-S2S**) and trained the network based on the reconstruction loss (using the reconstructed input sequence $\hat{\mathbf{x}}_{1:t}$). For implementing DCN-S2S, we used a single-layer LSTM with 50 nodes for both the encoder and the decoder. And, we augmented a fully-connected layer with 50 nodes is used to reconstruct the original input sequence from the latent representation of the decoder.

  In addition, since predictive clustering is associated with the label distribution, we compared a DCN whose encoder-decoder structure is replaced with our encoder-predictor network in Figure S.1b (denoted as **DCN-E2P**) to focus the dimensionality reduction – and, thus, finding latent encodings where clustering is performed – on the information for predicting the label distribution. We implemented the encoder and predictor of DCN-E2P with the same network architectures with those of our model as described in Section 5.

[3] https://github.com/rtavenar/tslearn
[4] https://github.com/boyangumn/DCN

- **SOM-VAE**[5] (Fortuin et al., 2019): We compare with SOM-VAE – though, this method is oriented towards visualization of input data via SOM – since it naturally clusters time-series data assuming Markov property (denoted as **SOM-VAE**). We replace the original CNN architecture of the encoder and the decoder with three-layered fully-connected network with 50 nodes in each layer, respectively. The network architecture is depicted in Figure S.1c where $\hat{\mathbf{x}}_t$ and $\bar{\mathbf{x}}_t$ indicate the reconstructed inputs based on the encoding $\mathbf{z}_t$ and the embedding $\mathbf{e}_t$ at time t, respectively.

  In addition, we compare with a variation of SOM-VAE by replacing the decoder with the predictor to encourage the latent encoding to capture information for predicting the label distribution (denoted as **SOM-VAE-P**). For the implementation, we replaced the decoder of SOM-VAE with our predictor which is a two-layered fully-connected layer with 50 nodes in each layer to predict the label distribution as illustrated in Figure S.1d. Here, $\hat{y}_t$ and $\bar{y}_t$ indicate the predicted labels based on the encoding $\mathbf{z}_t$ and the embedding $\mathbf{e}_t$ at time t, respectively.

  For both cases, we used the default values for balancing coefficients of SOM-VAE and the dimension of SOM to be equal to K.

We compared and summarized major components of the benchmarks in Table S.3.

F. Additional Experiments

F.1. Contributions of the Auxiliary Loss Functions

As described in Section 3.1, we introduced two auxiliary loss functions – the sample-wise entropy of cluster assignment (3) and the embedding separation loss (4) – to avoid trivial solution that may arise in identifying the predictive clusters. To analyze the contribution of each auxiliary loss function, we report the average number of activated clusters, clustering performance, and prediction performance on the UKCF dataset with 3 comorbidities as described in Section 5.4. Throughout the experiment, we set $K = 16$ – which is larger than C – to find the contribution of these loss functions to the number of activated clusters.

Table S.4. Performance comparison with varying the balancing coefficients α, β for the UKCF dataset.

Coefficients			Clustering Performance			Prognostic Value	
α	β	Activated No.	Purity	NMI	ARI	AUROC	AUPRC
0.0	0.0	16	0.573±0.01	0.006±0.00	0.000±0.00	0.500±0.00	0.169±0.00
0.0	1.0	16	0.573±0.01	0.006±0.00	0.000±0.00	0.500±0.00	0.169±0.00
3.0	0.0	8.4	0.795±0.01	0.431±0.01	0.569±0.01	0.840±0.01	0.583±0.02
3.0	1.0	8	**0.808±0.01**	**0.468±0.01**	**0.606±0.01**	**0.852±0.00**	**0.608±0.01**

As we can see in Table S.4, both auxiliary loss functions make important contributions in improving the quality of predictive clustering. More specifically, the sample-wise entropy encourages the selector to choose one dominant cluster. Thus, as we can see results with $\alpha = 0$, without the sample-wise entropy, our selector assigns an equal probability to all 16 clusters which results in a trivial solution. We observed that, by augmenting the embedding separation loss (4), AC-TPC identifies a smaller number of clusters owing to the well-separated embeddings.

F.2. Additional Results on Targeting Multiple Future Outcomes

Throughout the experiment in Section 5.5, we identified 12 subgroups of patients that are associated with the next-year development of 22 different comorbidities in the UKCF dataset. In Table S.5, we reported 12 identified clusters and the full list of comorbidities developed in the next year since the latest observation and the corresponding frequency which is calculated in a cluster-specific fashion based on the true label.

As we can see in the table, the identified clusters displayed very different label distributions; that is, the combination of comorbidities as well as their frequency were very different across the clusters. For example, patients in Cluster 1 experienced low-risk of developing any comorbidities in the next year while 85% of patients in Cluster 0 experienced diabetes in the next year.

[5] https://github.com/ratschlab/SOM-VAE

Table S.5. The comorbidities developed in the next year for the 12 identified clusters. The values in parentheses indicate the corresponding frequency.

Clusters	Comorbidities and Frequencies			
Cluster 0	Diabetes (0.85)	Liver Enzymes (0.21)	Arthropathy (0.14)	Depression (0.10)
	Hypertens (0.08)	Osteopenia (0.07)	Intest. Obstruction (0.07)	Cirrhosis (0.04)
	ABPA (0.04)	Liver Disease (0.04)	Osteoporosis (0.03)	Hearing Loss (0.03)
	Asthma (0.02)	Kidney Stones (0.01)	Bone fracture (0.01)	Hemoptysis (0.01)
	Pancreatitis (0.01)	Cancer (0.00)	Gall bladder (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 1	Liver Enzymes (0.09)	Arthropathy (0.08)	Depression (0.07)	Intest. Obstruction (0.06)
	Diabetes (0.06)	Osteopenia (0.05)	ABPA (0.04)	Asthma (0.03)
	Liver Disease (0.03)	Hearing Loss (0.03)	Osteoporosis (0.02)	Pancreatitis (0.02)
	Kidney Stones (0.02)	Hypertension (0.01)	Cirrhosis (0.01)	Gall bladder (0.01)
	Cancer (0.01)	Hemoptysis (0.00)	Bone fracture (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 2	ABPA (0.77)	Osteopenia (0.21)	Intest. Obstruction (0.11)	Hearing Loss (0.10)
	Liver Enzymes (0.07)	Diabetes (0.06)	Depression (0.05)	Pancreatitis (0.05)
	Liver Disease (0.04)	Arthropathy (0.04)	Asthma (0.03)	Bone fracture (0.02)
	Osteoporosis (0.02)	Hypertension (0.01)	Cancer (0.01)	Cirrhosis (0.01)
	Kidney Stones (0.01)	Gall bladder (0.01)	Hemoptysis (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 3	Asthma (0.89)	Liver Disease (0.87)	Diabetes (0.29)	Osteopenia (0.28)
	Liver Enzymes (0.24)	ABPA (0.15)	Osteoporosis (0.11)	Hearing Loss (0.06)
	Arthropathy (0.05)	Intest. Obstruction (0.05)	Depression (0.04)	Hypertension (0.03)
	Cirrhosis (0.02)	Kidney Stones (0.02)	Pancreatitis (0.02)	Gall bladder (0.02)
	Cancer (0.01)	Bone fracture (0.00)	Hemoptysis (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 4	Osteoporosis (0.76)	Diabetes (0.43)	Arthropathy (0.20)	Liver Enzymes (0.18)
	Osteopenia (0.15)	Depression (0.13)	Intest. Obstruction (0.11)	ABPA (0.11)
	Hearing Loss (0.09)	Liver Disease (0.08)	Hypertension (0.07)	Cirrhosis (0.07)
	Kidney Stones (0.03)	Asthma (0.02)	Hemoptysis (0.02)	Bone fracture (0.02)
	Gall bladder (0.01)	Pancreatitis (0.01)	Cancer (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 5	Asthma (0.88)	Diabetes (0.81)	Osteopenia (0.28)	ABPA (0.24)
	Liver Enzymes (0.22)	Depression (0.15)	Osteoporosis (0.14)	Intest. Obstruction (0.12)
	Hypertension (0.10)	Cirrhosis (0.10)	Liver Disease (0.09)	Arthropathy (0.08)
	Bone fracture (0.01)	Hemoptysis (0.01)	Pancreatitis (0.01)	Hearing Loss (0.01)
	Cancer (0.01)	Kidney Stones (0.01)	GI bleed – var. (0.01)	Gall bladder (0.00)
	Colonic stricture (0.00)	GI bleed – no var. (0.00)		
Cluster 6	Liver Disease (0.85)	Liver Enzymes (0.37)	Osteopenia (0.27)	ABPA (0.09)
	Arthropathy (0.07)	Diabetes (0.06)	Intest. Obstruction (0.06)	Osteoporosis (0.05)
	Depression (0.03)	Asthma (0.03)	Hearing Loss (0.03)	Cirrhosis (0.02)
	Hemoptysis (0.02)	Hypertension (0.01)	Kidney Stones (0.01)	Pancreatitis (0.00)
	Gall bladder (0.00)	Bone fracture (0.00)	Cancer (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 7	ABPA (0.83)	Diabetes (0.78)	Osteopenia (0.25)	Osteoporosis (0.24)
	Liver Enzymes (0.15)	Intest. Obstruction (0.12)	Liver Disease (0.09)	Hypertension (0.07)
	Hearing Loss (0.07)	Arthropathy (0.06)	Depression (0.04)	Cirrhosis (0.02)
	Asthma (0.01)	Bone fracture (0.01)	Kidney Stones (0.01)	Hemoptysis (0.01)
	Cancer (0.00)	Pancreatitis (0.00)	Gall bladder (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 8	Diabetes (0.94)	Liver Disease (0.83)	Liver Enzymes (0.43)	Osteopenia (0.30)
	Hearing Loss (0.11)	Osteoporosis (0.10)	Intest. Obstruction (0.09)	Cirrhosis (0.08)
	Depression (0.08)	ABPA (0.07)	Arthropathy (0.06)	Hypertension (0.05)
	Kidney Stones (0.05)	Asthma (0.02)	Hemoptysis (0.01)	Bone fracture (0.01)
	Cancer (0.00)	Pancreatitis (0.00)	Gall bladder (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 9	Asthma (0.89)	Osteopenia (0.26)	ABPA (0.19)	Arthropathy (0.14)
	Intest. Obstruction (0.11)	Depression (0.08)	Osteoporosis (0.08)	Diabetes (0.06)
	Liver Enzymes (0.06)	Hemoptysis (0.03)	Hypertension (0.02)	Liver Disease (0.02)
	Pancreatitis (0.02)	Bone fracture (0.01)	Cancer (0.01)	Cirrhosis (0.01)
	Gall bladder (0.01)	Hearing Loss (0.01)	Kidney Stones (0.00)	Colonic stricture (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		
Cluster 10	Osteopenia (0.82)	Diabetes (0.81)	Arthropathy (0.23)	Depression (0.19)
	Liver Enzymes (0.18)	Hypertension (0.16)	Hearing Loss (0.10)	Liver Disease (0.10)
	Osteoporosis (0.10)	Intest. Obstruction (0.09)	ABPA (0.09)	Kidney Stones (0.07)
	Cirrhosis (0.05)	Asthma (0.01)	Cancer (0.00)	GI bleed – var. (0.00)
	Bone fracture (0.00)	Hemoptysis (0.00)	Pancreatitis (0.00)	Gall bladder (0.00)
	Colonic stricture (0.00)	GI bleed – no var. (0.00)		
Cluster 11	Osteopenia (0.77)	Liver Enzymes (0.18)	Arthropathy (0.12)	Depression (0.09)
	Hypertension (0.06)	Diabetes (0.06)	Hearing Loss (0.06)	ABPA (0.05)
	Liver Disease (0.05)	Osteoporosis (0.04)	Intest. Obstruction (0.04)	Cirrhosis (0.02)
	Asthma (0.02)	Pancreatitis (0.02)	Bone fracture (0.01)	Cancer (0.01)
	Kidney Stones (0.00)	Gall bladder (0.00)	Colonic stricture (0.00)	Hemoptysis (0.00)
	GI bleed – no var. (0.00)	GI bleed – var. (0.00)		

References

Fortuin, V., Hüser, M., Locatello, F., Strathmann, H., and Rätsch, G. SOM-VAE: Interpretable discrete representation learning on time series. *In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*, 2019.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. *In Proceedings of the 13th Conference on Neural Information Processing Systems (NIPS 2000)*, 2000.

Yang, B., Fu, X., Sidiropoulos, N. D., and Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. *In Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, 2017.