Journal of the American Medical Informatics Association, 28(12), 2021, 2641–2653
doi: 10.1093/jamia/ocab203
Advance Access Publication Date: 27 September 2021
Research and Applications

Research and Applications

Deep significance clustering: a novel approach for identifying risk-stratified and predictive patient subgroups

Yufang Huang,[1] Yifan Liu,[1] Peter A.D. Steel,[2] Kelly M. Axsom,[3] John R. Lee,[4] Sri Lekha Tummalapalli,[1,4] Fei Wang,[1] Jyotishman Pathak,[1] Lakshminarayanan Subramanian,[5,6] and Yiye Zhang[1,2]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA, [2]Department of Emergency Medicine, Weill Cornell Medicine, New York, New York, USA, [3]Department of Medicine, Columbia University Vagelos College of Physicians and Surgeons, New York, New York, USA, [4]Department of Medicine, Weill Cornell Medicine, New York, New York, USA, [5]Courant Institute of Mathematical Sciences, New York University, New York, New York, USA, and [6]Department of Population Health, New York University School of Medicine, New York, New York, USA

Corresponding Author: Yiye Zhang, PhD, MS, Department of Population Health Sciences, Weill Cornell Medicine, 425 East 61st Street, New York, NY 10065, USA; yiz2014@med.cornell.edu

Received 27 May 2021; Revised 4 August 2021; Editorial Decision 31 August 2021; Accepted 2 September 2021

ABSTRACT

Objective: Deep significance clustering (DICE) is a self-supervised learning framework. DICE identifies clinically similar and risk-stratified subgroups that neither unsupervised clustering algorithms nor supervised risk prediction algorithms alone are guaranteed to generate.
Materials and Methods: Enabled by an optimization process that enforces statistical significance between the outcome and subgroup membership, DICE jointly trains 3 components, representation learning, clustering, and outcome prediction while providing interpretability to the deep representations. DICE also allows unseen patients to be predicted into trained subgroups for population-level risk stratification. We evaluated DICE using electronic health record datasets derived from 2 urban hospitals. Outcomes and patient cohorts used include discharge disposition to home among heart failure (HF) patients and acute kidney injury among COVID-19 (Cov-AKI) patients, respectively.
Results: Compared to baseline approaches including principal component analysis, DICE demonstrated superior performance in the cluster purity metrics: Silhouette score (0.48 for HF, 0.51 for Cov-AKI), Calinski-Harabasz index (212 for HF, 254 for Cov-AKI), and Davies-Bouldin index (0.86 for HF, 0.66 for Cov-AKI), and prediction metric: area under the Receiver operating characteristic (ROC) curve (0.83 for HF, 0.78 for Cov-AKI). Clinical evaluation of DICE-generated subgroups revealed more meaningful distributions of member characteristics across subgroups, and higher risk ratios between subgroups. Furthermore, DICE-generated subgroup membership alone was moderately predictive of outcomes.
Discussion: DICE addresses a gap in current machine learning approaches where predicted risk may not lead directly to actionable clinical steps.
Conclusion: DICE demonstrated the potential to apply in heterogeneous populations, where having the same quantitative risk does not equate with having a similar clinical profile.

Key words: machine learning, predictive clustering, risk stratification

INTRODUCTION

Background and significance

Risk stratification involving clinical and sociodemographic factors is crucial to the management of disease in medicine. Risk stratification is often implemented in clinical pathways in directing care to distinct subgroups of patients according to risk status.[1–4] While risk stratification has been particularly successful within specific disease or outcome contexts, clinical pathways that address risk in a broad cohort of patients with heterogenous sociodemographic and clinical profiles are more complex to implement due to the need to identify interventions specific to risk levels and patient subgroups.[5–9] For example, heart failure (HF) impacts nearly 6 million Americans where more than 80% of individuals suffer from 3 or more comorbidities.[10] The complexity due to frequent comorbidity and the lack of guidelines that incorporate heterogeneity present challenges in the discovery of patient strata to assist with clinical decision-making.[11] Another motivating example is acute kidney injury among COVID-19 patients (Cov-AKI),[12–14] where the initial kidney recovery during admission ranges from 30% to 75%.[12,14–16] The high degree of heterogeneity potentially originate from different pathophysiologic mechanisms such as volume depletion, acute tubular necrosis leading to fibrosis, and cardiometabolic disease leading to the incident cardiorenal syndrome.[12,14–16] Effective treatment strategies against Cov-AKI may benefit from risk stratification that targets each stratum.

Machine learning has been widely explored for risk stratification in medicine,[17] with supervised algorithms showing great potential in predicting individual risks. However, in a heterogenous population, patients may have the same risk levels while exhibiting different disease manifestations and thus requiring different interventions. Thus, to support the use in real patient care, there remains a gap between predicted risks and the next reasonable clinical actions. From an opposite angle, unsupervised machine learning algorithms have been used in previous literature to identify patient subgroups who do exhibit similar disease manifestations and thus requiring similar interventions.[18–22] However, the lack of supervision may lead to patient subgroups derived as clusters without actually stratifying patients based on the outcome of interests.[23–25] Existing clustering algorithms are also not designed to be predictive, limiting the utility of applying to unseen patients. Thus, distinctively partitioned patient subgroups, or precisely predicted individualized risks, without a bridge to the next clinical steps, may still bear limited translational values.[26–29] Yet, few existing clustering and risk prediction algorithms jointly achieve outcome-driven clustering in an end-to-end fashion for clinical applications.[23–25,30]

This gap between practical needs in medicine and existing machine learning solutions inspired deep significance clustering (DICE), an end-to-end, risk-stratifying, and predictive clustering algorithm. By jointly training representation learning, clustering, and classification, DICE identifies deep representations that generate outcome-driven cluster membership as subgroups. Patients within each subgroup are intended to have similar levels of risk of an outcome, as well as similar clinical needs. The novelty and feasibility of DICE originate from the use of a combined objective function including a constraint requiring significantly different outcome distributions across clusters. This framework design enforces backpropagation through the representation, clustering, and outcome prediction components. In addition, this design allows unseen patients to be predicted into risk-stratified subgroups trained in DICE as a multiclass classification task. Lastly, DICE performs neural architecture search (NAS) designed with an alternative grid search strategy over the number of clusters and representation dimension size to heuristically optimize outcome prediction. The architecture of DICE is illustrated in Figure 1. Supplementary Figure S1 provides an illustration of DICE using a simple example to provide the motivation for its development.

DICE is customized to medicine by considering statistical significance, a concept familiar to many medical researchers, into a machine learning framework. Previous work on risk stratification and subtyping has commonly conducted *post hoc* analysis on variable significance,[22,31] whereas DICE directly incorporates the statistical significance as a constraint. For evaluation, we applied DICE on 2 real-world electronic health record (EHR) datasets to compare the performance of DICE to baseline methods through extensive experiments, ablation studies, and fairness evaluation. Baseline methods compared include principal component analysis (PCA),[32] as well as autoencoder (AE),[33] *k*-means clustering, and logistic regression performed in separate steps without having the statistical significance constraint. Since the ground truth for stratification is unknown, we used Silhouette score,[34] Calinski-Harabasz index,[35] and Davies-Bouldin index[36] to evaluate the clustering performance. We also computed the relative risk ratios across the subgroups to assess the associations between subgroups and the outcomes. In addition, we evaluated the predictiveness of the DICE-learned representation by area under the ROC curve (AUC).

MATERIALS AND METHODS

Related work

Unsupervised learning is a fundamental topic in machine learning and has been widely applied to medical data.[19,22,37,38] Clustering algorithms such as *k*-means and hierarchical clustering separate a population based on the similarities of input variables. For example, *k*-means algorithm determines the cluster centroids by iterating between selecting centroids according to the assignment of data points to clusters, and assigning data points to clusters according to current centroids, until stopping criteria are met.[39] The cluster assignment is mainly driven by the cluster purity in terms of distances within or between clusters, but not by whether the distribution of one target variable differs across clusters. There are also semisupervised learning algorithms that make use of a small amount of labeled data with a large amount of unlabeled data.[40] Neither purely unsupervised learning nor semisupervised learning directly address the need for risk-stratified clustering of patients.

Most related to our proposed methodology is self-supervised learning,[41] and in particular, previous work on outcome-driven, or predictive, clustering.[23–25,42] Xia et al[25] applied *k*-means clustering on the learned representation from multitask classification model. Liu et al[23] applied agglomerative hierarchical clustering based on a distance metric that best suits the patient population. In their experiment, a linear discriminant analysis was chosen to learn a generalized Mahalanobis distance metric. These 2 methods are 2-stage, with clustering process independent from the representation learning process or metric learning. Locally Supervised Metric Learning proposed by Sun et al minimizes the distance of neighborhoods with the same class label while maximizing the distance of neighborhoods with different class labels. In addition, Lee et al proposed an actor-critic approach for predictive clustering by minimizing the Kullback-Leibler (KL) divergence between a predictor's output given learned representations and that given the assigned centroids. This is to ensure that patients in the same cluster share similar future outcomes.

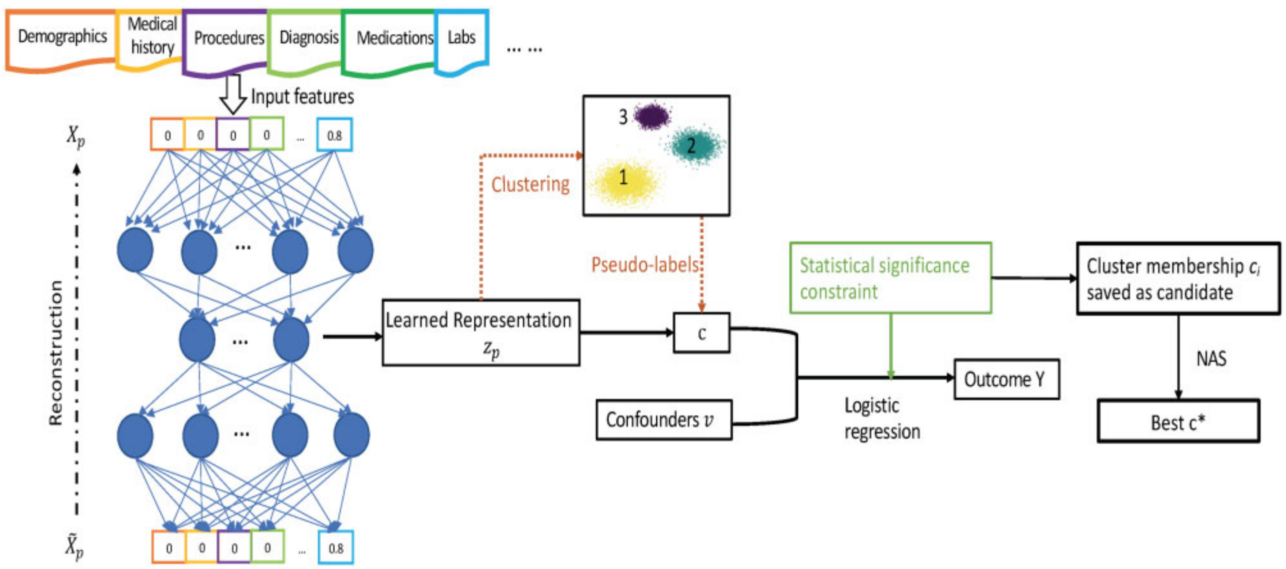

Figure 1. The framework of the proposed DICE. Clustering is applied to the representation $\mathbf{z}_p$. A statistical significance constraint is explicitly added to ensure the association of the clustering membership **c** and outcome *y*, which facilitates the learning of discriminative representations $\mathbf{z}_p$. DICE: deep significance clustering; NAS: neural architecture search.

Lastly, Zhang et al[43] add a constraint on a centroid-based probability distribution. Different from previous works, DICE proposes a "back-propagation" through the cluster membership classification component, to use cluster membership probabilities as input to predict the outcome to ensure that patients in the same cluster have similar outcome distribution. Importantly, DICE proposes a novel constraint to obtain ensure the outcome distribution is statistically significantly different across clusters.

To select variables that most contribute the outcome-driven stratification, DICE has a deep representation learning step to compress input data prior to clustering.[43–47] The representation learning component of DICE is related to other data transformation approaches which map the raw data into a new feature set such as PCA[32] and AE.[33] To reduce the dimensionality of input data, PCA identifies principal components that most well explain the data by computing eigenvectors and eigenvalues of the covariance matrix, regardless of whether a target outcome variable represents the input data. Recent deep clustering approaches are learning-based and conduct inference in one-shot, consisting of 2 stages, such as deep representation learning followed by various clustering models.[47] Caron et al[48] jointly learned the parameters of a deep network and the cluster assignments of the resulting representation. Deep clustering via a Gaussianmixture variational autoencoder with Graph embedding (DGG) uses Gaussian mixture variational AEs and graph embedding to improve the clustering and data representation abilities.[49] Yang et al use alternating stochastic optimization to update clustering centroids and representation learning parameters iteratively. Different from these methods, DICE constructs a clustering prediction network and updates representation learning parameters through self-supervised learning by considering cluster memberships as pseudolabels of the clustering prediction network.

In addition to machine learning approaches, DICE has similar objectives to statistical approaches including finite mixture model,[50,51] Gaussian Mixture Models (GMM),[39] kernel methods,[52] model-based clustering,[53,54] and spectral methods.[55,56] Compared to these models, DICE does not have distribution assumptions on observations[54] and can handle high computational complexity on large-scale datasets.[57] Jagabathula et al[58] proposed a conditional gradient approach for nonparametric estimation of mixing distributions. However, clustering of high-dimensional heterogeneous data remains challenging because of inefficient data representation.

NAS is a technique to find the network architecture with the best performance on the validation set. Early NAS conducted architecture optimization and network learning in a nested manner.[59–61] These works typically used reinforcement learning or evolution algorithms to explore the architecture search space $\mathcal{A}$. A recent work decoupled architecture search and weight optimization in a one-shot NAS framework and uses evolutionary architecture search to find candidate architectures after training.[62] EfficientNet and EfficientDet[63,64] further used grid search to balance network depth, width, and resolution and achieve state-of-the-art results on the ImageNet and COCO datasets, respectively.[65,66] We propose an alternative grid search to optimize the number of clusters and other hyperparameters in the DICE framework.

Representation learning

Given a dataset $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_P\}$ with P subjects, we denote each subject as a sequence of events $\mathbf{X}_p = \left(\mathbf{x}_p^1, \mathbf{x}_p^2, \ldots, \mathbf{x}_p^{n_p}\right)$ of length np. A multivariate feature vector $\mathbf{x}_p^t = \left[x_{p,1}^t, x_{p,2}^t, \ldots, x_{p,F}^t\right] \in \mathbb{R}^F$ is the tth instance of subject p in sequence $\mathbf{X}_p$, where F is the number of features at each timestamp. We have an outcome y_p for each subject p. The first step is to transform discrete sequences into latent continuous representations, followed by clustering and outcome prediction. The latent representation learning for each subject is performed by a long short-term memory (LSTM) AE.[33] The AE consists of 2 parts, the encoder and the decoder, denoted as $\mathcal{E}$ and $\mathcal{F}$, respectively. Given the pth input sequence $\mathbf{X}_p = \left(\mathbf{x}_p^1, \mathbf{x}_p^2, \ldots, \mathbf{x}_p^{n_p}\right)$, the encoder can be formulated as $\mathbf{z}_p = \mathcal{E}(\mathbf{X}_p; \theta_{\mathcal{E}})$, where $\mathbf{z}_p \in \mathbb{R}^d$ is the representation, d is the dimension of representation, and $\mathcal{E}$ is an LSTM network with parameter $\theta_{\mathcal{E}}$.[67] We choose the last hidden state $\mathbf{z}_p$ of LSTM to be the representation of the input $\mathbf{X}_p$. The decoder can be formulated as $\widetilde{\mathbf{X}}_p = \mathcal{F}(\mathbf{z}_p; \theta_{\mathcal{F}})$, and $\mathcal{F}$ is the other LSTM network with parameter $\theta_{\mathcal{F}}$. The representation learning is achieved by minimizing the reconstruction error

$$min_{\theta_{\mathcal{E}},\theta_{\mathcal{F}}}\ \mathcal{L}_{AE} = \frac{1}{P}\sum\nolimits_{p=1}^{P} \parallel \mathcal{F}(\mathcal{E}(\mathbf{X}_p;\theta_{\mathcal{E}});\theta_{\mathcal{F}}) - \mathbf{X}_p \parallel_{L_2}^2, \tag{1}$$

where we use L_2 norm in the loss.

Self-supervised clustering

The obtained representations $\mathbb{Z} = \{\mathbf{z}_p\}_{p=1}^{P}$ can be employed for clustering with K clusters,

$$\begin{gathered} min_{\mathbf{M},\{\mathbf{c}_p\}_{p=1}^{P}}\ \mathcal{L}_{clustering} = \sum\nolimits_{p=1}^{P} \parallel \mathbf{z}_p - \mathbf{M}\mathbf{c}_p \parallel_2^2 \\ \text{s.t. } 1^T\mathbf{c}_p = 1, c_p^k \in \{0,1\}, \\ \forall p \in \{1,2,\ldots,P\},\ k \in \{1,2,\ldots,K\} \end{gathered} \tag{2}$$

where K is a hyperparameter of total number of clusters to tune, $\mathbf{c}_p = [c_p^1,\ldots,c_p^K]$, c_p^k is the cluster membership of cluster k, $\mathbf{M} \in \mathbb{R}^{d\times K}$ and the k-th columns of $\mathbf{M}$ is the centroid of the k-th cluster. To enable fast inference and learn representation with the driven of outcome, we build a cluster classification network for deep clustering based on self-supervision from $\mathbf{c}_p$ in Equation (2). We employ the *a priori* clustering results $\{\mathbf{c}_p\}_{p=1}^{P}$ in Equation (2) as pseudolabels to update the parameters of the encoder $\mathcal{E}$ and $\mathcal{F}$. The cluster membership assignment can be formulated as a classification network,

$$\hat{\mathbf{c}_p} = g(\mathbf{z}_p;\theta_1),\ \ min\ _{\theta_1} \mathcal{L}_1 = -\sum\nolimits_{p=1}^{P}\sum\nolimits_{k=1}^{K} c_p^k \log(\hat{c_p}\, k), \tag{3}$$

where $\hat{\mathbf{c}_p} = [\hat{c_p}\, 1,\ldots,\hat{c_p}\, K]$ is the predicted cluster membership from the cluster classification network $g(\cdot;\theta_1)$, θ_1 is the parameter in the cluster classification network, $\mathcal{L}_1$ is the negative log-likelihood loss for multiclass cluster classification.

Outcome prediction

After obtaining cluster membership $\{\hat{\mathbf{c}_p}\}_{p=1}^{P}$ for K clusters, we use the cluster membership and other confounders such as demographics to predict the outcome, formulated as:

$$\begin{gathered} \hat{y_p} = g([\hat{\mathbf{c}_p}, \mathbf{v}_p];\theta_2), \\ min_{\theta_2}\ \mathcal{L}_2 = -\sum\nolimits_{p=1}^{P}(y_p\log(\hat{y_p}) + (1-y_p)\log(1-\hat{y_p})), \end{gathered} \tag{4}$$

where $\mathbf{v}_p$ represents confounders to adjust in testing the significance, $[\cdot,\cdot]$ denotes the concatenation of cluster membership feature and confounders. $g(\cdot;\theta_2)$ is the logistic regression for the outcome prediction, and $\mathcal{L}_2$ is the negative log-likelihood loss for the classification. This approach partially addresses interpretability in the application of deep learning methods in medicine. Using the cluster membership from the learned representation as the input to predict the outcome allows us to infer a broad theme (ie, risk-stratified stratum) with a set of learned representations. Interpretability is further enhanced by enforcing the following statistical significance constraint to the cluster membership with respect to the outcome.

Statistical significance constraint

The main novelty of DICE is the introduction of a statistical significance constraint to the cluster membership with respect to the outcome distribution to drive the deep clustering process. This step also drives the interpretation of the representation learning. After obtaining cluster memberships $\{\hat{\mathbf{c}_p}\}_{p=1}^{P}$ for K clusters, we require that the association between the cluster membership and outcome be statistically significant while adjusting for relevant confounders. To quantify the significant difference of cluster k_1 and cluster k_2 $(k_1 \neq k_2)$, we use likelihood-ratio test to calculate the P value of variable $\hat{c}k_2$ when considering cluster $\hat{c}k_1$ as the reference, where $\hat{c}k$ refers to the cluster membership belonging to cluster k, formulated as,

$$G_{k_1,k_2} = -2\log\left[\frac{\mathcal{L}_2\left(g\left(\left[\frac{\hat{\mathbf{c}}}{\hat{c}k_1}, \hat{c}k_2, \mathbf{v}\right];\theta_2\right), y\right)}{\mathcal{L}_2\left(g\left(\left[\frac{\hat{\mathbf{c}}}{\hat{c}k_1}, \mathbf{v}\right];\theta_2\right), y\right)}\right]. \tag{5}$$

Then we obtain the P value from Chi-square distribution, denoted as S_{k_1,k_2}. A predefined threshold of significance α (equivalently, $G_{k_1,k_2} > \alpha_G$) is used to measure significance. In this paper, we use $\alpha = .05$. In implementation, we design a mask technique to remove variables of input $\widehat{\mathbf{c}}$, corresponding to cluster k_1 and cluster k_2, in Equation (5), then calculate the likelihood ratio G_{k_1,k_2}, and add significance constraint to the likelihood-ratio G_{k_1,k_2}, that is, $G_{k_1,k_2} > \alpha_G, \forall k_1 \neq k_2$.

Objective function

The neural weights optimization is denoted as:

$$\begin{aligned} \min_{\theta} \mathcal{L}(\mathcal{N}(K,d,\theta)) = {} & \min_{\theta} \lambda_1\mathcal{L}_{AE} + \mathcal{L}_{clustering} + \lambda_2\mathcal{L}_1 + \lambda_3\mathcal{L}_2 \\ & + \lambda_4(\alpha_G - G_{k_1,k_2}) \end{aligned} \tag{6}$$

$$\begin{gathered} \text{s.t.} \qquad 1^T\mathbf{c}_p = 1,\ c_p^j \in \{0,1\}, \\ \forall p \in \{1,\ldots,P\}, k_1 \in \{1,\ldots,K\}, k_2 \in \{1,\ldots,K\}, k_1 \neq k_2, \end{gathered}$$

where λ_1, λ_2, λ_3, and λ_4 are tradeoffs for $\mathcal{L}_{AE}$, $\mathcal{L}_1$, $\mathcal{L}_2$, and the statistical significance constraint. We iteratively optimize deep clustering and the other components with the statistical significance constraint. We firstly employ *a priori*, such as k-means,[68] to obtain pseudolabels for the cluster classification network. Then, we can optimize $\mathcal{L}_{AE}$ for the representation learning network, $\mathcal{L}_1$ for cluster classification network, $\mathcal{L}_2$ for outcome prediction network, and the statistical significance constraint jointly. The algorithm is elaborated in Algorithm 1.

Algorithm 1. DICE: Deep significance clustering

```
Input: $\mathbb{X}, \{\mathbf{v}\}, K, d$
Output: $\{\mathbf{z}_p\}_{p=1}^{P}, \{\mathbf{c}_p\}_{p=1}^{P}$
Initialize the AE of representation learning through $\mathcal{L}_{AE}$;
Extract representations $\{\mathbf{z}\}$;
For $i = 1\ :\ n_{iter}$ do
    Optimize $\mathcal{L}_{clustering}$ by $k$-means;
    Calculate the cluster membership;
     Use the cluster memberships as pseudo-labels for
     cluster classification network in $\mathcal{L}_1$;
    For $j = 1\ :\ n_{epoch}$ do
        Jointly optimize $\mathcal{L}_{AE}$, $\mathcal{L}_1$, $\mathcal{L}_2$, and $G_{k_1,k_2}$;
    end
    Extract representations $\{\mathbf{z}\}$;
end
return $\{\mathbf{z}_p\}_{p=1}^{P}, \{\mathbf{c}_p\}_{p=1}^{P}$
```

Architecture search

We utilize NAS to optimize the network hyperparameters in the DICE: the hyperparameter in the clustering and the network hyperparameters in the representation learning. NAS conducts 2 processes sequentially. The first is the neural weights optimization of a given network architecture with the fixed number of clusters K and hidden state dimension d in DICE. The second is the NAS process. NAS is conducted in the search space to select the combination of hyperparameters and has no direct link to the cost function of neural weights optimization. We choose the network architecture which is trained on the training set and has the best evaluation performance on the validation set, that is

$$(K^*, d^*) = \underset{K,d}{argmax}\, AUC_{val}(\mathcal{N}(K, d, \theta)), \qquad (7)$$

where $AUC_{val}(\cdot)$ is the AUC score on the validation set.

Experimental setting

Data

Study data included HF and COVID-19 patients treated in the inpatient and emergency department (ED) settings in 2 hospitals of an urban academic center, respectively. EHR variables extracted include information on sociodemographics, vital signs, diagnoses, therapeutics orders, medication prescriptions, laboratory test orders and test results, and census-tract level social determinants of health (SDOH). The sociodemographic information included age, gender, race, marital status, preferred language, and insurance payor. Diagnoses were extracted using International Classification of Diseases, Ninth/Tenth Revision, Clinical Modification (ICD-9/10-CM) codes.[69] Outcomes are defined as discharged to home among HF patients and Cov-AKI among COVID-19 patients. Continuous variables for each patient were represented as normalized vectors, and they were normalized with mean of 0 and standard deviation of 1. Categorical variables were converted to binary vectors, whose values were represented as 1 or 0, using one-hot encoding. Missing values in laboratory and SDOH variables were imputed with mean values.

HF data include adult patients from years 2014 to 2018 who were treated on the inpatient Medicine services. Only those patients whose initial (acute, ED, admitting) and principal diagnoses both contained HF codes were included to ensure that HF was the working diagnosis throughout the hospital stay and was being treated from the beginning of the encounter. In the HF data, variables were timestamped into day intervals since ED arrival and used as sequential features. Figure 2 describes the datasets and the inclusion/exclusion criteria for the HF cohort. HF definitions in ICD-9/10-CM are listed in Supplementary Table S4. COVID-19 was defined by a positive polymerase chain reaction test. COVID-19 data included adult patients who were admitted to the hospital in March and April 2020 from the ED.[14] We define baseline creatinine to be the closest creatinine obtained prior to March 2020, and alternatively, if not available, the earliest creatinine at the time of ED presentation. AKI was defined by the Kidney Disease Improving Global Outcomes criteria.[70,71] It is defined as an increase in creatinine of 0.3 or greater from the baseline creatinine during the hospitalization, or in an increase of creatinine greater than 1.5 times the baseline creatinine during the hospitalization, or initiation of renal replacement therapy. Furthermore, this definition of AKI was verified by manual chart reviews led by an MD coauthor.[14] In COVID-19 data, demographic variables (age, gender, race), chronic conditions, and the first values of commonly ordered laboratory tests obtained within 12 h of ED presentation were included as one-time features. Variables used are listed in Supplementary Table S5.

Baselines

We compared our method with baseline methods including (1) PCA for representation learning followed by k-means clustering (PCA [k-means]), (2) AE for representation learning followed by k-means

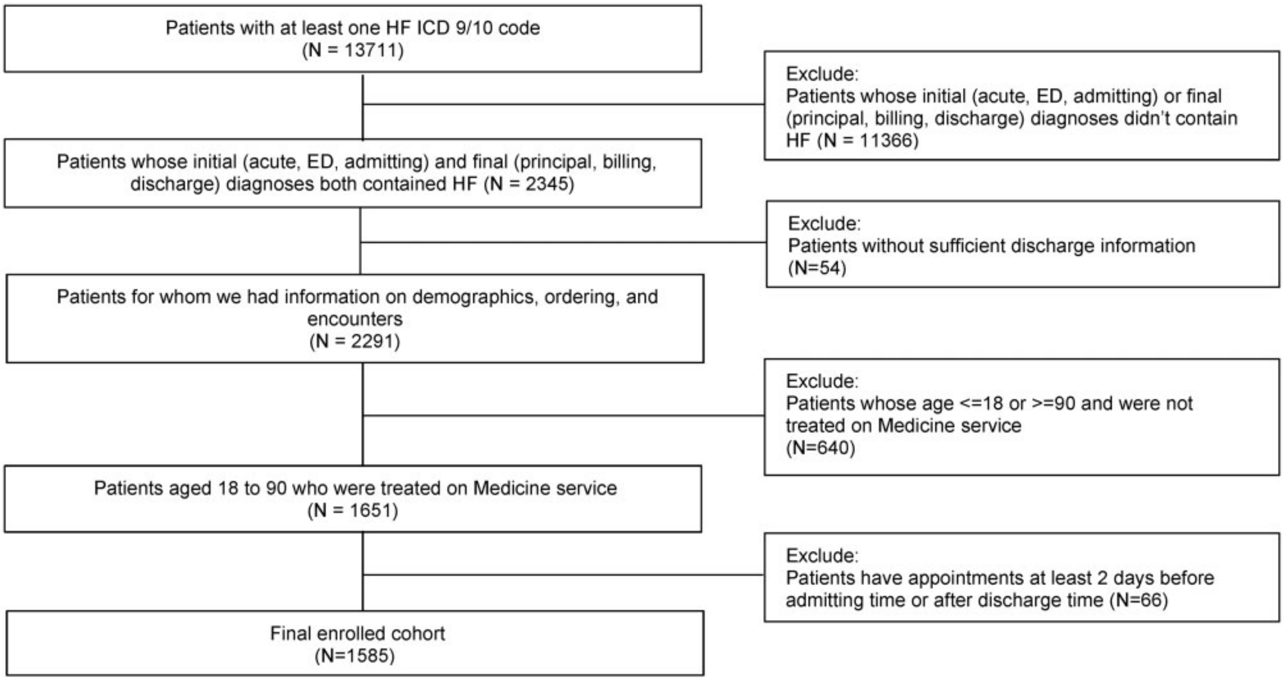
Patients with at least one HF ICD 9/10 code
(N = 13711)
Exclude:
Patients whose initial (acute, ED, admitting) or final (principal, billing, discharge) diagnoses didn't contain HF (N = 11366)
Patients whose initial (acute, ED, admitting) and final (principal, billing, discharge) diagnoses both contained HF (N = 2345)
Exclude:
Patients without sufficient discharge information
(N=54)
Patients for whom we had information on demographics, ordering, and encounters
(N = 2291)
Exclude:
Patients whose age <=18 or >=90 and were not treated on Medicine service
(N=640)
Patients aged 18 to 90 who were treated on Medicine service
(N = 1651)
Exclude:
Patients have appointments at least 2 days before admitting time or after discharge time (N=66)
Final enrolled cohort
(N=1585)

Figure 2. Inclusion and exclusion criteria for HF cohort. HF: heart failure.

clustering (AE [*k*-means]), and (3) AE for representation learning with classification followed by *k*-means clustering (AE w/class [*k*-means]). For baseline (1), we treated sequential data as one-time features in HF dataset to learn PCA representations, followed by *k*-means clustering. In (2), *k*-means clustering was applied directly to representations learned from AE.[33] In (3), we first jointly trained AE and outcome prediction with representation learned from AE as the input for outcome prediction, then applied *k*-means clustering to the final learned representation. We report the results of these baseline methods of the same hyperparameters with DICE. Supplementary Table S6 lists the baseline methods against DICE.

Training

Based on the dataset size and the number of features, the number of clusters experimented was set to 2 through 5. The sizes of the representation dimension were 20 through 100 for HF and 10–20 for COVID-19, respectively. Experiments were conducted in PyTorch[72] on NVIDIA GeForce RTX 2070. We initialized the AE with one epoch training. We set P value $\alpha = 0.05$ which leads to $\alpha_G = 3.841$, $n_{iter} = 150$, $n_{epoch} = 1$. Parameters λ_1, λ_2, λ_3, λ_4 were set as 1.0, 10, 1.0, 1.0 for HF and COVID-19 based on the accuracy on the validation set. HF and COVID-19 datasets were split into training, validation, and test sets in a 4:1:1 ratio.

Evaluation

DICE was compared against 3 baseline methods with respect to AUC on the outcome prediction, Silhouette score,[34] Calinski-Harabasz index,[35] and Davies-Bouldin index.[36] Silhouette score, Calinski-Harabasz index, and Davies-Bouldin index are normalized metrics, and therefore, allow us to evaluate the cluster goodness across methods regardless of the input representation scale. To evaluate the outcome-driven nature of the clusters, we computed risk ratios between each cluster and the cluster with the lowest incidence as CI_{c_i}/CI_{c_r} where C_i is the cumulative incidences of clusters i and C_r the reference cluster, r.

Code availability

Source code is available in https://github.com/YiyeZhangLab/DICE.

RESULTS

HF data contained 1585 patients, of whom 36.8% of the patients were discharged to home (Figure 2). Supplementary Tables S1 and S7 describe the demographic information in the data. Among the 1002 COVID-19 patients, 30.3% of the patients developed AKI subsequently during hospitalization. The network hyperparameters chosen were $K = 4$, $d = 35$ for discharged to home among HF patients and $K = 3$ and $d = 16$ for AKI among COVID-19 patients.

Table 1, displaying performance on the test set, shows that DICE can generate more distinctive clusters as subgroups. Experiments and analyses demonstrate that DICE obtained better performance than baseline methods in deriving subgroups that have higher risk ratios in comparing the reference (lowest risk) with the other subgroups. We further demonstrate the clustering separation across the 2 datasets through the t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations in Figures 3 and 4. Compared with baselines shown in Figure 3B–D for HF, the 4 subgroups in Figure 3A discovered by DICE displayed tighter separation (Silhouette score = 0.48, Calinski-Harabasz index = 212, Davies-Bouldin index = 0.86). In order of outcome rates, subgroups 1–4 had 79.9%, 38.8%, 29.7%, and 8.6%, respectively. The baseline AE with classification also discovered 4 subgroups with the outcome ratio in each subgroup ranging from 72.2% to 5.8%, but the cluster purity metrics were lower (Silhouette score = 0.35, Calinski-Harabasz index = 200, Davies-Bouldin index = 1.30). PCA (*k*-means) and AE (*k*-means) did not discover subgroups as clearly separated and outcome-driven as DICE. Examining the visualizations for the COVID-19 dataset in Figure 4, AE (k-means) achieved pure cluster metrics (Silhouette score = 0.46, Calinski-Harabasz index = 163, Davies-Bouldin index = 0.84). However, the subgroups were similar with respect to the outcome rates, ranging from 46.9% to 23.9% (risk ratios 1.88 and 1.10), showing that cluster purity does not necessarily guarantee risk-stratified subgroups.

Table 2, displaying predictive performance on the test set, shows that DICE-learned representations are predictive. To evaluate the learned representation by DICE, we used the representations for outcome prediction using L1-regularized logistic regression. DICE outperformed the baselines in AUC, true positive rate (TPR), false negative rate, positive predictive value (PPV), and negative predictive value (NPV) in both HF dataset and COVID-19 dataset. Relatedly, we evaluated the AUC for outcome prediction using DICE subgroup membership alone. Notably, the DICE subgroup membership alone achieved moderately high prediction of the outcome (AUC = 0.772 for HF, AUC = 0.627 for COVID-19). Supplementary Table S8 describes the predictiveness of the DICE cluster membership alone, including AUC with confidence bounds, accuracy (ACC), TPR, true negative rate (TNR), PPV, and NPV.

Using HF data, we examined the advantage of the statistical significance constraint as well as algorithm fairness. Figure 5 illustrates

Table 1. Clustering performance evaluation on the test set

	Model	Silhouette score↑	Calinski-Harabasz index↑	Davies-Bouldin index↓	Risk ratio*↑
HF	PCA (*k*-means)	0.097	16.1	2.609	1.54, 1.50, 1.39
	AE (*k*-means)	0.281	68.1	1.744	1.91, 1.48, 1.43
	AE w/class. (*k*-means)	0.346	200.0	1.304	**9.2, 4.56, 2.65**
	DICE	**0.484**	**212.2**	**0.864**	6.77, 3.32, 2.94
COVID-19	PCA (*k*-means)	0.188	30.0	1.840	1.88, 1.10
	AE (*k*-means)	0.462	162.8	0.841	1.11, 1.54
	AE w/class. (*k*-means)	0.266	92.4	1.124	2.82, 1.39
	DICE	**0.514**	**253.6**	**0.664**	**5.06, 1.02**

Notes :↑: higher values are superior; ↓: lower values are superior. *displaying risk ratio between each subgroup and the subgroup with the lowest incidence as reference group. Bold values denote $P < 0.05$.

AE: autoencoder; DICE: deep significance clustering; PCA: principal component analysis.

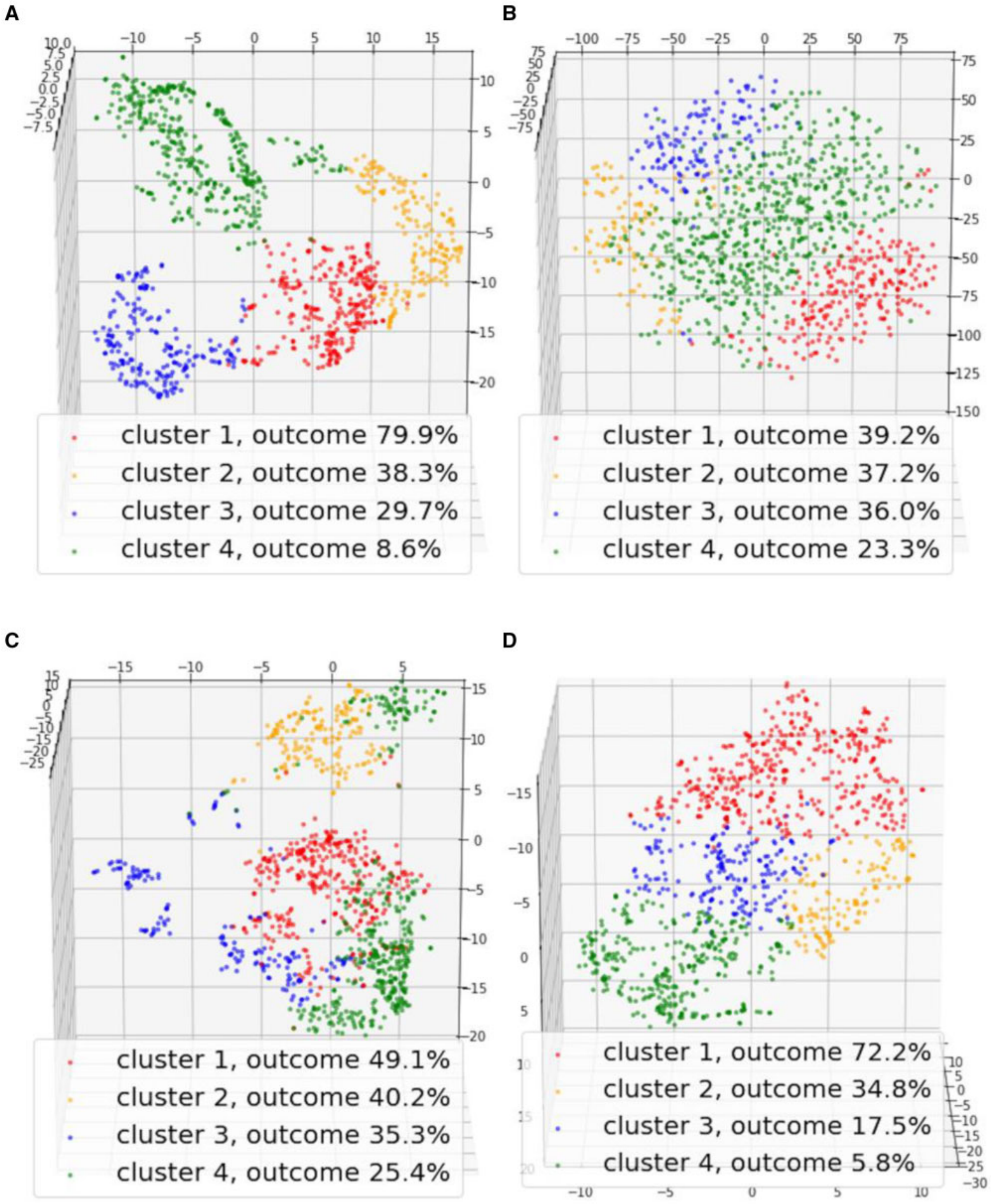

Figure 3. Visualization of patient subtyping results by various methods on HF dataset. (**A**) DICE. (**B**) PCA (k-means). (**C**) AE (k-means). (**D**) AE w/class. (k-means). AE: autoencoder; DICE: deep significance clustering; PCA: principal component analysis.

the AUC on the HF validation set across different neural network architecture on the *Y*-axis and representation dimension *d* on the *X*-axis. At each fixed cluster size and representation dimension, the architecture network that met the statistical constraint achieved higher AUC than those that did not. We further conducted ablation studies to gauge the effect of the statistical significance constraint. When we disabled the statistical significance constraint, 2 clusters were outputted by NAS, compared to the 4-level separation as reported in

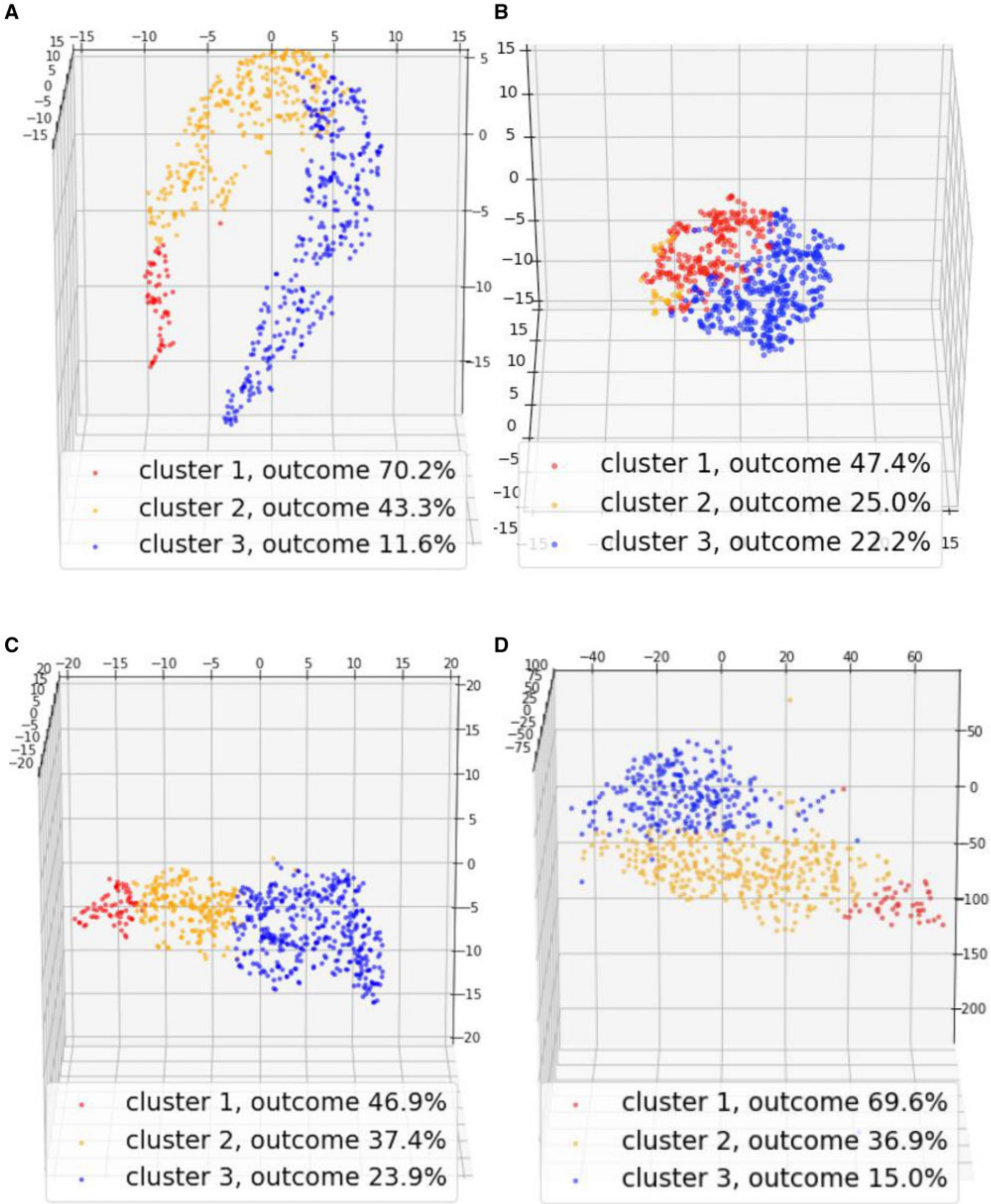

Figure 4. Visualization of patient subtyping results by various methods on COVID-19 dataset. (**A**) DICE. (**B**) PCA (k-means). (**C**) AE (k-means). (**D**) AE w/class. (k-means). AE: autoencoder; DICE: deep significance clustering; PCA: principal component analysis.

Table 1. In addition, the percentage of neural networks that passed the significance constraint in NAS decreased from 82.4% to 64.7% when cluster size was set to 5 in the ablation study. The AUC was lower when the statistical significance constraint was not met. These results suggest that the statistical significance constraint contributes to better stratification especially as we increase the number of clus-

Table 2. Outcome prediction comparison on the test set

		AUC	TPR	TNR	PPV	NPV
HF	PCA (*k*-means)	0.773 ± 0.061	0.598	0.778	0.611	0.769
	AE (*k*-means)	0.712 ± 0.067	0.433	**0.850**	0.627	0.721
	AE w/class. (*k*-means)	0.818 ± 0.058	0.794	0.746	0.647	0.862
	DICE	**0.834** ± 0.054	**0.845**	0.743	**0.656**	**0.892**
COVID-19	PCA (*k*-means)	0.738 ± 0.087	0.647	0.724	0.508	0.824
	AE (*k*-means)	0.686 ± 0.091	0.647	0.716	0.500	0.822
	AE w/class (*k*-means)	0.734 ± 0.087	0.667	0.698	0.493	0.827
	DICE	**0.777** ± 0.083	**0.726**	**0.737**	**0.544**	**0.861**

Bold values denote $P < 0.05$. AE: autoencoder; AUC: area under the ROC curve; DICE: deep significance clustering; NPV: negative predictive value; PCA: principal component analysis; PPV: positive predictive value; TPR: true positive rate.

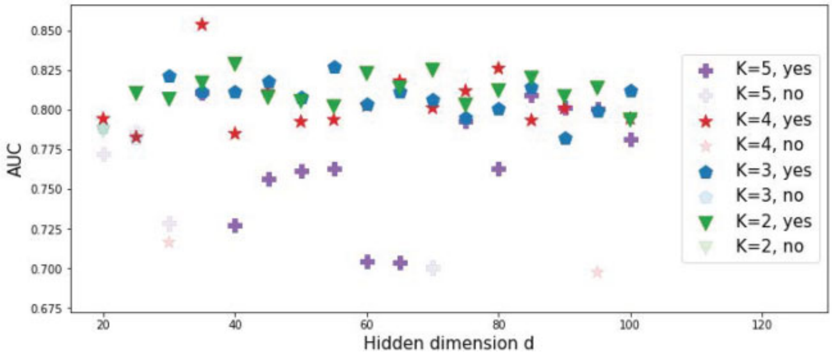

Figure 5. The model selection on HF dataset. "yes" represents that the architecture network met the significance constraint, and "no" otherwise. HF: heart failure.

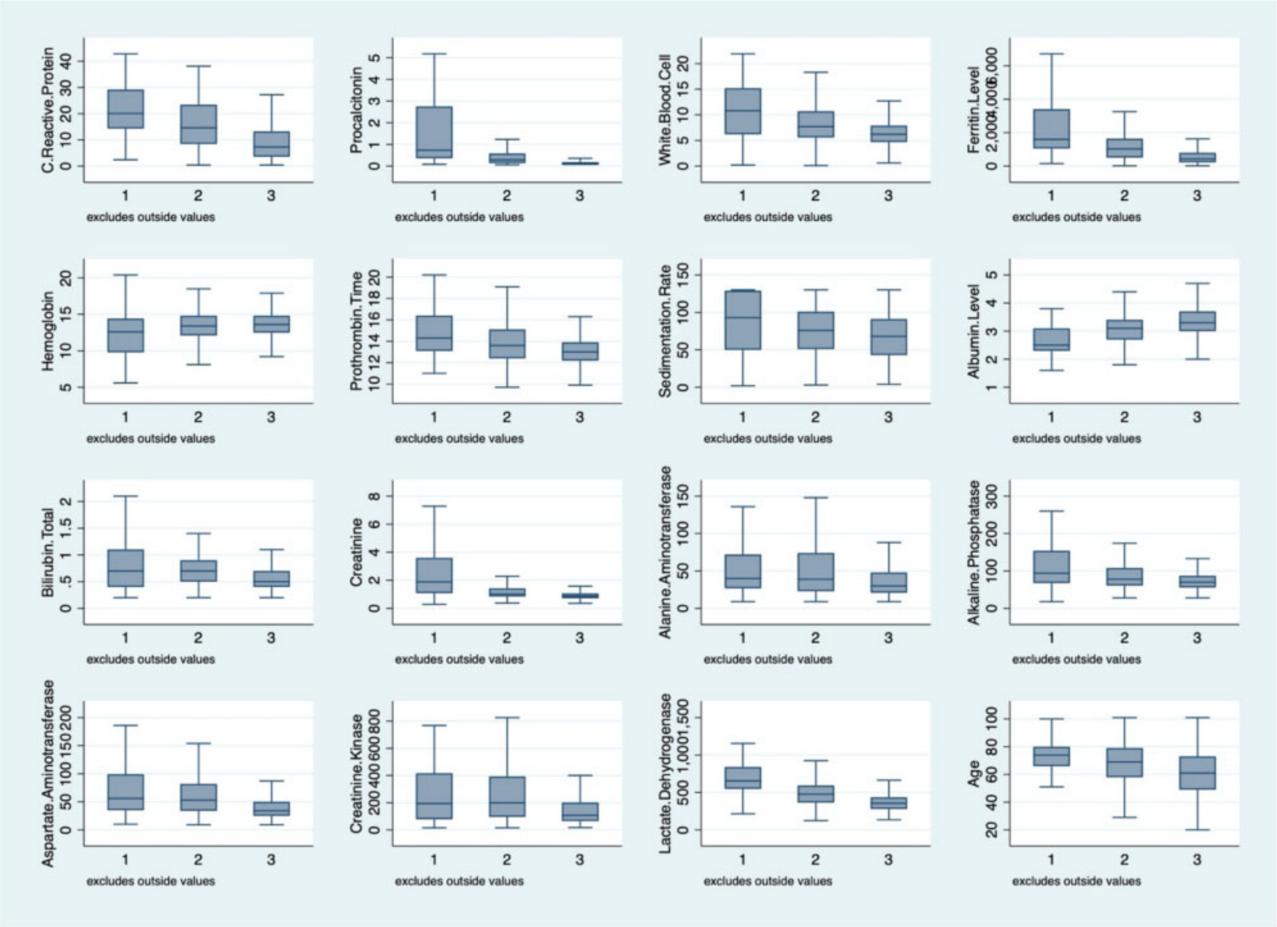

Figure 6. Distribution of the laboratory variables across clusters in COVID-19 cohort. 1: high risk, 2: medium risk, and 3: low risk.

Table 3. Characteristics of the clusters generated by DICE in HF cohort

	Lowest (*N* = 572, 10.8% discharged home)	Low (*N* = 308, 31.8% discharged home)	High (*N* = 248, 35.9% discharged home)	Highest (*N* = 457, 73.1% discharged home)
Female*	334 (58.4)	151 (49.0)	87 (35.1)	132 (28.9)
Diagnosis				
Anemia*	148 (25.9)	66 (21.4)	33 (13.3)	50 (10.9)
Chronic kidney disease*	305 (53.3)	144 (46.8)	87 (35.1)	136 (29.8)
Obesity*	31 (5.4)	11 (3.6)	7 (2.8)	7 (1.5)
Medication				
Diatrizoate meglumine and sodium oral liq*	101 (17.7)	22 (7.1)	15 (6.0)	23 (5.0)
Heparin flush*	105 (18.4)	26 (8.4)	16 (6.5)	27 (5.9)
Metoprolol tartrate inj*	89 (15.6)	42 (13.6)	26 (10.5)	29 (6.3)
Acetaminophen tab*	395 (69.1)	181 (58.8)	136 (54.8)	225 (49.2)
Calcium gluconate inj*	116 (20.3)	36 (11.7)	15 (6.0)	25 (5.5)
Alteplase cath clearance +R+*	59 (10.3)	12 (3.9)	7 (2.8)	5 (1.1)
Potassium chloride oral liq*	229 (40.0)	87 (28.2)	47 (19.0)	80 (17.5)
Bumetanide inj*	192 (33.6)	56 (18.2)	34 (13.7)	60 (13.1)
Lidocaine inj 1%*	122 (21.3)	33 (10.7)	25 (10.1)	26 (5.7)
Haloperidol inj*	39 (6.8)	17 (5.5)	11 (4.4)	2 (0.4)
Ondansetron inj*	111 (19.4)	47 (15.3)	25 (10.1)	34 (7.4)
Guaifenesin + dextromethorphan oral liq*	58 (10.1)	27 (8.8)	19 (7.7)	23 (5.0)
Piperacillin tazobactam inj*	62 (10.8)	33 (10.7)	16 (6.5)	17 (3.7)
Insulin reg inj (humulin R)*	69 (12.1)	30 (9.7)	11 (4.4)	18 (3.9)
Morphine sulfate inj*	121 (21.2)	40 (13.0)	27 (10.9)	21 (4.6)
Vancomycin IVPB (initial-72 h stop)*	100 (17.5)	24 (7.8)	14 (5.6)	11 (2.4)
Lorazepam inj*	104 (18.2)	33 (10.7)	22 (8.9)	24 (5.3)
Potassium chloride inj*	194 (33.9)	42 (13.6)	31 (12.5)	42 (9.2)
Amiodarone inj*	78 (13.6)	13 (4.2)	10 (4.0)	15 (3.3)
Procedure				
Social work nursing referral*	255 (44.6)	119 (38.6)	88 (35.5)	144 (31.5)
O_2 via—nasal cannula*	243 (42.5)	111 (36.0)	82 (33.1)	131 (28.7)
Indwelling urinary catheter (Foley)*	196 (34.3)	61 (19.8)	42 (16.9)	26 (5.7)
Central venous line care*	68 (11.9)	17 (5.5)	6 (2.4)	9 (2.0)

*$P < 0.05$.

DICE: deep significance clustering.

Table 4. Characteristics of the clusters generated by DICE in COVID-19 cohort

	High risk (*N* = 75, 69.3% AKI)	Medium risk (*N* = 444, 41.9% AKI)	Low risk (*N* = 483, 13.7% AKI)
Age*	74.0 (66.50, 80.0)	69.0 (58.0, 79.0)	61.0 (49.50, 73.0)
Gender: male*	54 (72.0)	328 (73.9)	237 (49.1)
Alanine aminotransferase*	40.0 (28.0, 71.50)	39.0 (23.0, 74.0)	30.0 (21.0, 48.0)
Albumin level*	2.50 (2.30, 3.10)	3.10 (2.70, 3.40)	3.30 (3.0, 3.70)
Alkaline phosphatase*	94.0 (68.50, 153.50)	78.0 (61.0, 108.0)	70.0 (56.0, 87.0)
Aspartate aminotransferase*	56.0 (35.0, 98.0)	53.0 (34.0, 82.0)	34.0 (25.0, 50.0)
Bilirubin total*	0.70 (0.40, 1.10)	0.70 (0.50, 0.90)	0.50 (0.40, 0.70)
C-reactive protein*	20.05 (14.43, 28.93)	14.60 (8.40, 23.40)	7.30 (3.60, 13.15)
Creatine kinase*	194.0 (79.0, 417.0)	199.0 (96.0, 388.50)	107.50 (65.25, 200.75)
Creatinine*	1.89 (1.09, 3.58)	1.02 (0.85, 1.43)	0.88 (0.73, 1.07)
D-dimer*	2876.0 (880.0, 11470.0)	564.0 (339.0, 1099.50)	335.0 (216.50, 573.50)
Ferritin level*	1595.50 (1047.65, 3368.18)	1025.65 (513.13, 1635.38)	430.60 (227.95, 795.25)
Hemoglobin*	12.60 (9.90, 14.25)	13.40 (12.10, 14.80)	13.60 (12.50, 14.80)
Lactate dehydrogenase*	653.50 (551.75, 839.75)	477.0 (365.75, 591.25)	356.0 (280.25, 435.0)
Lactic acid level*	3.20 (1.70, 4.40)	1.70 (1.30, 2.50)	1.30 (1.0, 1.70)
Procalcitonin*	0.73 (0.37, 2.57)	0.29 (0.14, 0.58)	0.10 (0.06, 0.18)
Prothrombin time*	14.30 (13.10, 16.40)	13.60 (12.40, 15.08)	13.0 (12.20, 13.90)
Sedimentation rate*	93.0 (50.0, 129.0)	76.0 (51.0, 101.0)	68.0 (43.0, 91.0)
Troponin-I*	0.10 (0.04, 0.44)	0.03 (0.03, 0.06)	0.03 (0.03, 0.03)
White blood cell*	10.80 (6.30, 15.20)	7.70 (5.60, 10.70)	6.20 (4.75, 7.90)
Urine protein: positive*	58 (77.3)	278 (62.6)	168 (34.8)
Urine blood: positive*	58 (77.3)	231 (52.0)	72 (14.9)

*$P < 0.05$.

DICE: deep significance clustering.

ters. To evaluate the fairness of the algorithm, we report the predictive performance of DICE across racial subgroups. Using DICE subgroup membership as predictors, the AUCs for unknown, Asian, others, Black, and White patient population were 0.905, 0.882, 0.856, 0.832, and 0.847, respectively. When learned representation was used as the predictor, the AUCs for unknown, Asian, others, Black, and White are 0.863, 0.829, 0.782, 0.854, and 0.853, respectively.

Subgroups generated by DICE were evaluated for their clinical relevance. Figure 6 illustrates the distribution of relevant laboratory variables across subgroups in the COVID-19 cohort. DICE discovered 3 subgroups that had high (69%), medium (42%), and low (14%) incidence of AKI. Distributions of laboratory measurements across subgroups had linear trends from high- to low-risk subgroups. The distributions were consistent with clinically expected risk factors of AKI among COVID patients ,[73] including older age, higher value of alkaline phosphatase, C-reactive protein, D-dimer, Ferritin; and lower values of hemoglobin and albumin, corresponding to severe illness and higher risk of AKI.[73] Other baseline techniques were unable to detect these AKI-focused subgroups (Figure 6). Table 3 to Supplementary Table S2 and Table 4 to Supplementary Table S3 show the notable characteristics across subgroups in the 2 datasets, respectively, comparing DICE and baseline AE w/class (k-means). Only variables with a linear trend observed across the highest- to lowest-risk clusters are displayed. For example, we observe a linear trend in comorbidity (chronic kidney disease and obesity) across the clusters, where clusters with the lowest and the highest percentages of patients discharged to home displaying the highest and lowest percentages of comorbidity, respectively. Similarly, we observe trends in the use of medications that are indicative of disease severity and complexity, such as Bumetanide and Haloperidol being more prevalent in the cluster with the lowest percentages of patients discharged to home. This cluster of patients also has the highest needs for social work referral as observed in the orders placed. *P* values were calculated using Kruskal-Wallis rank-sum test for continuous variables and using Chi-square/Fisher's exact test for categorical variables.

DISCUSSION

DICE was motivated to join concepts of machine learning and statistics as a customized machine learning algorithm for medicine. It is intended to create risk-stratified and predictive subgroups to facilitate risk-stratified intervention designs. These features of DICE were demonstrated in the evaluation using EHR datasets with different sizes, variable types, incidence, and clinical areas. Compared to DICE, applying baseline methods in COVID-19 data, we observed that subgroups had good cluster purity but not clearly stratified by the risk level of the outcome. In HF data, we observed that DICE achieved cluster purity while the cluster membership also served a predictive purpose. Evaluation results suggest that DICE has certain advantage over baseline methods particularly when the characteristics indicative of the outcome risk, or root causes, are heterogenous, rendering outcome prediction challenging.

Beyond patient populations evaluated in this paper, DICE may have the potential to be used in other clinical areas to facilitate subgroup-specific care and clinical pathways for clinical decision support. In this study, DICE jointly trained AE for representation learning, *k*-means for clustering, and logistic regression for prediction.[74] Depending on the data structure, we can revise DICE to replace *k*-means with other clustering algorithms, and similarly, logistic regression with other prediction algorithms. Moreover, if clinical notes were used as input, Transformers may serve as the encoder and decoder in representation learning.[75] Future studies may also evaluate additional statistical concepts to better ensure the outcome separation using metrics such as Tukey's Honestly Significant Difference and Cochran-Armitage test for trend to increase risk ratio across clusters.[76] In summary, DICE offers a flexible framework and a conceptual innovation that may drive meaningful application machine learning in the EHR.

CONCLUSION

This paper demonstrated DICE, an outcome-driven clustering algorithm for risk-stratifying patients. Compared to baseline methods, DICE is optimized to cluster patients based on both the risk level of an outcome and on the input clinical features. Because of this feature, we propose that DICE may be used to identify subgroups of patients who require risk-stratified interventions in a heterogeneous population, who are similar in ways that allow them to respond to the similar treatments against their risk of an outcome. Beyond the datasets used in this paper, DICE has the potential to be used in other clinical areas.

FUNDING

This study was supported by NLM K01LM013257-01 (PI Zhang) and Center for Transportation, Environment, and Community Health (CTECH) New Research Initiatives Fund (PI Zhang).

AUTHOR CONTRIBUTIONS

YZ and YH designed the overall study in consultation with LS, PADS, KMA, JRL, and SLT. YZ, YH, and YL performed data analysis. PADS, KMA, JRL, and SLT provided clinical inputs, and interpretation to the study. YH and YZ wrote the manuscript with inputs from all the authors. FW, JP, LS, and YL provided suggestions to the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

YZ and JP have equity ownership at Iris OB Health. JRL has filed patent US-2020-0048713-A1 titled "Methods of detecting cell-free DNA in biological samples" and has received a grant from BioFire Diagnostics LLC.

DATA AVAILABILITY

The data generated and/or analyzed during the current study cannot be shared publicly due to its inclusion of patient health information protected by the Health Insurance Portability and Accountability Act, but will be shared on reasonable request to the corresponding author.

REFERENCES

1. Lauck SB, Wood DA, Achtem L, *et al.* Risk stratification and clinical pathways to optimize length of stay after transcatheter aortic valve replacement. *Can J Cardiol* 2014; 30 (12): 1583–7.
2. Shaheen AA, Riazi K, Medellin A, *et al.* Risk stratification of patients with nonalcoholic fatty liver disease using a case identification pathway

in primary care: a cross-sectional study. *CMAJ Open* 2020; 8 (2): E370–76.

3. Pillay SM, Oliver B, Butler L, Kennedy HG. Risk stratification and the care pathway. *Ir J Psychol Med* 2008; 25 (4): 123–7.
4. Olin SS, McCord M, Stein REK, *et al.* Beyond screening: a stepped care pathway for managing postpartum depression in pediatric settings. *J Womens Health (Larchmt)* 2017; 26 (9): 966–75.
5. Suh EH, Bodnar DJ, Melville LD, Sharma M, Farmer BM. Crisis clinical pathway for COVID-19. *Emerg Med J* 2020; 37 (11): 700–4.
6. Geleris P, Boudoulas H. Problems related to the application of guidelines in clinical practice: a critical analysis. *Hellenic J Cardiol* 2011; 52 (2): 97–102.
7. Six A, Backus B, Kelder J. Chest pain in the emergency room: value of the HEART score. *Neth Heart J* 2008; 16 (6): 191–6.
8. Navi BB, Kamel H, Shah MP, *et al.* Application of the ABCD2 score to identify cerebrovascular causes of dizziness in the emergency department. *Stroke* 2012; 43 (6): 1484–9.
9. Littlejohn LA, Gibbs J, Jordan LB, *et al.* Assessing the effectiveness of NICE criteria for stratifying breast cancer risk in a UK cohort. *Eur J Hum Genet* 2018; 26 (4): 599–603.
10. Savarese G, Lund LH. Global public health burden of heart failure. *Card Fail Rev* 2017; 3 (1): 7–11.
11. Cubbon RM, Witte KK, Kearney LC, *et al.* Performance of 2014 NICE defibrillator implantation guidelines in heart failure risk stratification. *Heart* 2016; 102 (10): 735–40.
12. Chan L, Chaudhary K, Saha A, *et al.*; Mount Sinai COVID Informatics Center (MSCIC). AKI in hospitalized patients with COVID-19. *J Am Soc Nephrol* 2021; 32 (1): 151–60.
13. Hirsch JS, Ng JH, Ross DW, *et al.*; Northwell Nephrology COVID-19 Research Consortium. Acute kidney injury in patients hospitalized with COVID-19. *Kidney Int* 2020; 98 (1): 209–18.
14. Lee JR, Silberzweig J, Akchurin O, *et al.* Characteristics of acute kidney injury in hospitalized COVID-19 patients in an Urban Academic Medical Center. *Clin J Am Soc Nephrol* 2021; 16 (2): 284–6.
15. Ng JH, Hirsch JS, Hazzan A, *et al.* Outcomes among patients hospitalized with COVID-19 and acute kidney injury. *Am J Kidney Dis* 2021; 77 (2): 204–15.e1.
16. Fisher M, Neugarten J, Bellin E, *et al.* AKI in hospitalized patients with and without COVID-19: a comparison study. *J Am Soc Nephrol* 2020; 31 (9): 2145–57.
17. Beaulieu-Jones BK, Yuan W, Brat GA, *et al.* Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ Digit Med* 2021; 4 (1): 62.
18. Lakshmanan GT, Rozsnyai S, Wang F. Investigating clinical care pathways correlated with outcomes. *Lect Notes Comput Sci* 2013; 8094: 323–38.
19. Zhang Y, Padman R, Patel N. Paving the COWpath: learning and visualizing clinical pathways from electronic health record data. *J Biomed Inform* 2015; 58: 186–97.
20. Chaudhary K, Vaid A, Duffy A, *et al.* Utilization of deep learning for subphenotype identification in sepsis-associated acute kidney injury. *Clin J Am Soc Nephrol* 2020; 15 (11): 1557–65.
21. Xu Z, Chou J, Zhang XS, *et al.* Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. *J Biomed Inform* 2020; 102: 103361.
22. Zhang X, Chou J, Liang J, *et al.* Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study. *Sci Rep* 2019; 9 (1): 797–12.
23. Liu H, Li X, Xie G, *et al.* Precision cohort finding with outcome-driven similarity analytics: a case study of patients with atrial fibrillation. *Stud Health Technol Inform* 2017; 245: 491–5.
24. Lee C, Van Der Schaar M, eds. Temporal phenotyping using deep predictive clustering of disease progression. In: International Conference on Machine Learning. PMLR; 2020.
25. Xia E, Du X, Mei J, et al., eds. Outcome-Driven Clustering of Acute Coronary Syndrome Patients Using Multi-Task Neural Network with Attention. MedInfo; 2019.
26. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178 (11): 1544–7.
27. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med* 2018; 378 (11): 981–3.
28. Sarwar S, Dent A, Faust K, *et al.* Physician perspectives on integration of artificial intelligence into diagnostic pathology. *NPJ Digit Med* 2019; 2 (1): 28–7.
29. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA* 2018; 319 (1): 19–20.
30. Liang J, Chen K, Lin M, Zhang C, Wang F. Robust finite mixture regression for heterogeneous targets. *Data Min Knowl Disc* 2018; 32 (6): 1509–60.
31. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. *J Affect Disord* 2021; 279: 1–8.
32. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst* 1987; 2 (1–3): 37–52.
33. Sutskever I, Vinyals O, Le QV, eds. Sequence to sequence learning with neural networks. Advances in neural information processing systems; 2014.
34. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987; 20: 53–65.
35. Calinski T, Harabasz J. A dendrite method for cluster analysis. *Comm Stats Theory Methods* 1974; 3 (1): 1–27.
36. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 1979; 1 (2): 224–7.
37. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. *J Am Med Inform Assoc* 2014; 21 (e2): e304–11.
38. Deo RC. Machine learning in medicine. *Circulation* 2015; 132 (20): 1920–30.
39. Svensén M, Bishop CM. *Pattern Recognition and Machine Learning*. New York: Springer; 2007.
40. Zhu X, Goldberg AB. Introduction to semi-supervised learning. *Synth Lect Artif Intell Mach Learn* 2009; 3 (1): 1–130.
41. Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans Pattern Anal Mach Intell* 2020. doi: 10.1109/TPAMI.2020.2992393.
42. Sun J, Wang F, Hu J, Edabollahi S. Supervised patient similarity measure of heterogeneous patient records. *Sigkdd Explor Newsl* 2012; 14 (1): 16–24.
43. Zhang H, Basu S, Davidson I. A framework for deep constrained clustering—algorithms and advances. arXiv preprint arXiv:190110061; 2019.
44. Hershey JR, Chen Z, Le Roux J, Watanabe S, eds. Deep clustering: discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2016.
45. Li F, Qiao H, Zhang B. Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognit* 2018; 83: 161–73.
46. Xie J, Girshick R, Farhadi A, eds. Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning. PMLR; 2016.
47. Yang B, Fu X, Sidiropoulos ND, Hong M, eds. Towards k-means-friendly spaces: simultaneous deep learning and clustering. In: International Conference on Machine Learning. PMLR; 2017.
48. Caron M, Bojanowski P, Joulin A, Douze M, eds. Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018.
49. Yang L, Cheung N-M, Li J, Fang J, eds. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019.
50. McLachlan GJ, Peel D. *Finite Mixture Models*. New York: John Wiley\& Sons; 2004.
51. Wedel M, DeSarbo WS. A review of recent developments in latent class regression models. In: Bagozzi R, ed., *Advanced Methods of Marketing Research*. Blackwell Pub; 1994: 352–88.

52. Hofmann T, Scholkopf B, Smola AJ. Kernel methods in machine learning. *Ann Statist* 2008; 36 (3): 1171–220.
53. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002; 97 (458): 611–31.
54. Zhong S, Ghosh J. A unified framework for model-based clustering. *J Mach Learn Res* 2003; 4 (Nov): 1001–37.
55. Ng AY, Jordan MI, Weiss Y, eds. On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems; 2002.
56. Von Luxburg U. A tutorial on spectral clustering. *Stat Comput* 2007; 17 (4): 395–416.
57. Min E, Guo X, Liu Q, Zhang G, Cui J, Long J. A survey of clustering with deep learning: from the perspective of network architecture. *IEEE Access* 2018; 6: 39501–14.
58. Jagabathula S, Subramanian L, Venkataraman A. A conditional gradient approach for nonparametric estimation of mixing distributions. *Manag Sci* 2020; 66 (8): 3635–56.
59. Baker B, Gupta O, Naik N, Raskar R. Designing neural network architectures using reinforcement learning. arXiv preprint arXiv:161102167; 2016.
60. Zoph B, Le QV. Neural architecture search with reinforcement learning. arXiv preprint arXiv:161101578; 2016.
61. Zoph B, Vasudevan V, Shlens J, Le QV, eds. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018.
62. Guo Z, Zhang X, Mu H, *et al.* Single path one-shot neural architecture search with uniform sampling. arXiv preprint arXiv:190400420; 2019.
63. Tan M, Le Q, eds. EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR; 2019.
64. Tan M, Pang R, Le QV. Efficientdet: scalable and efficient object detection. arXiv preprint arXiv:191109070; 2019.
65. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L, eds. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009.
66. Lin T-Y, Maire M, Belongie S, et al., eds. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer; 2014.
67. Hochreiter S, Schmidhuber J, rgen. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
68. MacQueen J, ed. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967; Oakland, CA.
69. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care* 2009; 47 (6): 626–33.
70. Kdigo A. Work group: section 2: AKI definition. *Kidney Int Suppl* 2012; 2: 19–36.
71. Palevsky PM, Liu KD, Brophy PD, *et al.* KDOQI US commentary on the 2012 KDIGO clinical practice guideline for acute kidney injury. *Am J Kidney Dis* 2013; 61 (5): 649–72.
72. PyTorch. Secondary. https://pytorch.org Accessed September 14, 2021.
73. Kumar MP, Mishra S, Jha DK, *et al.* Coronavirus disease (COVID-19) and the liver: a comprehensive systematic review and meta-analysis. *Hepatol Int* 2020; 14 (5): 711–22.
74. Masci J, Meier U, Cireşan D, Schmidhuber J, eds. Stacked convolutional auto-encoders for hierarchical feature extraction. In: International Conference on Artificial Neural Networks. Springer; 2011.
75. Vaswani A, Shazeer N, Parmar N, *et al.*, eds. Attention is all you need. In: Advances in neural information processing systems; 2017.
76. Lee S, Lee DK. What is the proper way to apply the multiple comparison test? (vol 71, pg 353, 2018). *Korean J Anesthesiol* 2020; 73 (6): 572.