Medical Image Analysis 97 (2024) 103231

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Interpretable deep clustering survival machines for Alzheimer's disease subtype discovery

Bojian Hou [a,1], Zixuan Wen [a,1], Jingxuan Bao [a], Richard Zhang [a], Boning Tong [a], Shu Yang [a], Junhao Wen [c], Yuhan Cui [b], Jason H. Moore [d], Andrew J. Saykin [e], Heng Huang [f], Paul M. Thompson [c], Marylyn D. Ritchie [a], Christos Davatzikos [b], Li Shen [a,*], for the Alzheimer's Disease Neuroimaging Initiative [2]

[a] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
[b] Center for Biomedical Image Computing and Analytics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
[c] Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA 90007, USA
[d] Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA 90069, USA
[e] Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA
[f] Department of Computer Science, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Survival analysis
Alzheimer's disease
Subtype discovery
Interpretability

ABSTRACT

Alzheimer's disease (AD) is a complex neurodegenerative disorder that has impacted millions of people worldwide. The neuroanatomical heterogeneity of AD has made it challenging to fully understand the disease mechanism. Identifying AD subtypes during the prodromal stage and determining their genetic basis would be immensely valuable for drug discovery and subsequent clinical treatment. Previous studies that clustered subgroups typically used unsupervised learning techniques, neglecting the *survival information* and potentially limiting the insights gained. To address this problem, we propose an interpretable survival analysis method called Deep Clustering Survival Machines (DCSM), which combines both *discriminative* and *generative* mechanisms. Similar to mixture models, we assume that the timing information of survival data can be *generatively* described by a mixture of parametric distributions, referred to as *expert distributions*. We learn the weights of these expert distributions for individual instances in a *discriminative* manner by leveraging their features. This allows us to characterize the survival information of each instance through a weighted combination of the learned expert distributions. We demonstrate the superiority of the DCSM method by applying this approach to cluster patients with mild cognitive impairment (MCI) into subgroups with different risks of converting to AD. Conventional clustering measurements for survival analysis along with genetic association studies successfully validate the effectiveness of the proposed method and characterize our clustering findings.

1. Introduction

According to the World Health Organization (WHO)[3], dementia has affected 55 million people worldwide in 2023. This number could increase to 139 million by 2050 as more people age. Alzheimer's disease (AD) is the most common cause of dementia, accounting for over two-thirds of the cases. However, as a complex and heterogeneous brain disorder, AD remains poorly understood. Finding subtypes of AD and their genetic factors could help develop new drugs and guide treatments in the prodromal stage for this critical condition.

Previous works usually use unsupervised learning methods such as KMeans (Hartigan and Wong, 1979), GMM (Reynolds et al., 2009), DBSCAN (Ester et al., 1996) etc. to stratify AD patients into different clusters/subtypes (Alashwal et al., 2019; Feng et al., 2022). They merely utilize the feature information to find different groups with

* Corresponding author.
E-mail address: li.shen@pennmedicine.upenn.edu (L. Shen).

[1] The first two authors contribute to this paper equally.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[3] https://www.who.int/news-room/fact-sheets/detail/dementia

https://doi.org/10.1016/j.media.2024.103231
Received 4 October 2023; Received in revised form 4 May 2024; Accepted 3 June 2024
Available online 14 June 2024

distinct characteristics. However, these approaches ignore the fact that dementia patients often visit the hospital multiple times to track their disease progression. These results in longitudinal trajectories that capture additional survival information about the patients such as the probability of conversion from mild cognitive impairment (MCI) to AD. Thus, it would be valuable if we can leverage the survival information to facilitate AD subtype discovery.

There is now a growing field of literature concerning with disease progression modeling and subtyping. These models estimate distinct data-driven disease timelines for multiple subgroups, essentially estimating subtypes and disease progressions simultaneously. The most popular one is SuSTaIn (Young et al., 2018; Aksman et al., 2021; Fonteijn et al., 2012; Young et al., 2014). Nonetheless, our current emphasis lies in survival analysis and stratifying subjects into groups with different risk levels, rendering these studies potentially less applicable to our specific focus.

The goal of survival analysis (Flynn, 2012) is to learn a survival model to build the bridge between features and survival information. When given the feature of a subject, survival models can predict the risk of death or occurrence of some event such as MCI converting to AD. In this way, we have an opportunity to utilize the predicted risk to stratify the MCI patients into different subgroups with different risks of converting to AD. It would thus enable customized treatment for different patients with different risks. In this study, we aim to conduct subtypes discovery for Alzheimer's disease from the survival analysis perspective, which has the potential to improve clinical decision-making by identifying high-risk MCI patients who may require more care or early treatment. It is worth noting that a big challenge in survival analysis is *censoring*, which means the target event of a subject is unobservable after a period of time or no event happens during the monitoring. Therefore, many subjects do not possess complete survival information and we face the problem of semi-supervised learning or weak supervised learning (Hou et al., 2017; Zhou, 2018) instead of fully supervised learning. Given these circumstances, it becomes impractical to directly employ survival information for the purpose of subgroup stratification. Thus, the development of an effective clustering technique that is capable of leveraging partial survival information becomes imperative.

There are many survival models that have been proposed to predict the risk of an event happening, also known as "time-to-event prediction" (Kvamme et al., 2019). The most classic method is called the Cox PH model (Cox, 1972). It assumes a constant hazard rate over time for every subject, known as the proportional hazard (PH) assumption. However, this PH assumption may not hold in practice thus hindering Cox model's performance. There are other methods that do not make any assumptions about the underlying distribution of survival times, such as Kaplan–Meier (Bland and Altman, 1998), Nelson-Aalen (Klein, 1991), and Life-Table (Tarone, 1975). Nevertheless, these methods struggle with high-dimensional data. Machine learning techniques can help overcome this high-dimensional challenge and can learn the association between features and survival outcomes efficiently. For example, Deep Survival Machines (DSM) (Nagpal et al., 2021) uses deep neural networks to learn the compact representation of the features and uses the negative log-likelihood as the loss to learn all the parameters, showing promising results in prediction accuracy. Deep Cox (Katzman et al., 2018) utilizes the derived Cox PH loss to optimize the parameter learning of deep neural networks.

Nevertheless, all the survival models aforementioned are not specifically designed for clustering. They are mainly used to do risk prediction. To leverage them to do clustering, we need to set a threshold for the predicted risks to artificially cluster them into subgroups, such as two groups where one is with high risk and the other is with low risk. Survival Clustering Analysis (SCA) (Chapfuwa et al., 2020) and Variational Deep Survival Clustering (VaDeSC) (Manduchi et al., 2021) are two recent works that can do both risk prediction and clustering. However, SCA cannot control the number of clusters because it utilizes the truncated Dirichlet process to realize the automatic identification of the cluster numbers, and VaDeSC as a fully generative method is restricted to a specific distribution of features. Neural Survival Clustering (Jeanselme et al., 2022) is another recent model developed to perform clustering and time-to-event prediction simultaneously. It learns the survival probability for each instance by learning fixed neural networks without any assumptions in a thorough discriminative manner. However, the model requires a considerable amount of training data to avoid overfitting, which may not be feasible for small-scale datasets, particularly in the medical field.

In this study, we propose a hybrid method that leverages both the discriminative and generative strategies to perform clustering and risk prediction simultaneously. Specifically, we assume that there are a certain number of expert distributions in a latent space and each expert distribution can be modeled by parameterized distributions in a generative way. The survival function for each instance is a weighted combination of all the expert distributions and the weight for each instance is learned by a multi-layer perceptron (MLP) directly from the features in a discriminative manner. Consequently, we can naturally cluster all the instances according to how the weights are allocated to different expert distributions for each instance. We demonstrate the advantage of our method by evaluating not only the conventional clustering measurements for survival analysis but also the genetic association discrepancies between different groups of patients with different risks of converting from MCI to AD. In summary, our contributions are five-fold:

- We propose a hybrid survival analysis method called Deep Clustering Survival Machines (DCSM) that integrates the advantages of the discriminative and generative ideas and can perform both clustering and time-to-event prediction simultaneously.
- We apply our method to Alzheimer's imaging data to discover AD subtypes. LogRank results and their differential genetic associations validate the effectiveness of the proposed method.
- To further validate the effectiveness of the proposed method, we also conduct experiments on several real-world benchmark datasets. The results show promising clustering results as well as competitive time-to-event prediction performance.
- Our method is interpretable in that the expert distributions are constant for all the instances. Different weightings signify different attention to the expert distributions and thus we can easily tell which subgroup the instance belongs to.
- We perform feature importance for different regions of the brain to interpret what regions the proposed model pays more attention to when clustering the patients into low and high risks. The identified important brain regions show a strong relationship to AD.

2. Related work

Clustering. Clustering is the most relevant topic to our paper. Clustering is a concept from the machine learning community that involves grouping similar data points together based on certain characteristics. In the medical domain, people may use stratification or subtype discovery to describe similar problems. These terms refer to the process of identifying subgroups within a larger population that share similar characteristics or traits. The paper will use these terms interchangeably. Traditional clustering methods such as KMeans (Hartigan and Wong, 1979), GMM (Reynolds et al., 2009), DBSCAN (Ester et al., 1996) usually use the sample features only to calculate the similarity or distance between samples to discover the subtype among populations (Alashwal et al., 2019; Feng et al., 2022). Considering that patients can visit hospitals several times and thus render longitudinal information, we can also use survival analysis techniques to predict the risk of getting AD for each patient and stratify them using the predicted risk in a post-hoc way. This risk prediction in survival analysis is also called "time-to-event-prediction".

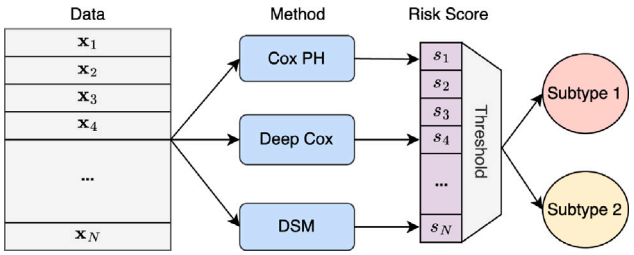


Fig. 1. Post-hoc clustering mechanism of Cox PH, Deep Cox, and DSM. In this work, we adopt median as the threshold. The patients whose predicted risk scores lower than the median are clustered as low-risk group, while the others are clustered as high-risk group

Time-to-event prediction. There are many time-to-event prediction methods that have been proposed. One of the most prevalent methods is the Cox proportional hazards regression model (Cox PH) (Cox, 1972). It assumes that the relative proportional hazard between subjects is constant over time. In other words, if subject *A* has a higher risk of death or another event at some time point than another subject *B*, then *A*'s risk will always be higher than *B*'s. Although Cox PH has achieved many successes in survival analysis, it still has a narrow application since its assumption of proportional hazard is too strong to satisfy in reality. Traditional methods including Kaplan–Meier (Bland and Altman, 1998), Nelson-Aalen (Klein, 1991), and Life-Table (Tarone, 1975) are also useful to do survival analysis. However, they can hardly scale to large dimensionality. Recently, many machine learning methods, especially deep learning methods, have been proposed to do survival analysis. Most of them are dedicated to improving upon Cox PH. A common idea is to use deep neural networks to learn the nonlinear relationship between the explanatory variables and outcome by optimizing the partial likelihood of Cox PH (Katzman et al., 2018). However, this is still restricted to the strong assumption of proportional hazard. Recently, a fully parametric method utilizing deep learning called Deep Survival Machines (DSM) (Nagpal et al., 2021) has attracted much attention. It does not make the PH assumption and can achieve competitive predictive performance compared to state-of-the-art methods. However, DSM learns different base distributions for each instance, making the model hard to interpret (Hou and Zhou, 2018, 2020).

Clustering with time-to-event prediction. There are a few other methods that perform both clustering and time-to-event prediction simultaneously. For example, Survival Clustering Analysis (SCA) (Chapfuwa et al., 2020) assumes that the latent space is a mixture of distributions and uses the truncated Dirichlet process to automatically identify the number of clusters. However, SCA cannot control the number of clusters and thus cannot validate its advantages compared to post-hoc methods. Variational deep survival clustering (VaDeSC) (Manduchi et al., 2021), as a fully generative method, uses a Gaussian mixture distribution to model the features in a latent space and uses the Weibull distribution to model the survival timing information. This builds a good bridge between the features and survival information by jointly optimizing both likelihoods. However, there is a trade-off between the discriminative and generative learning paradigms. A fully generative framework may not be a good fit for all types of data since it is difficult to let both the features and survival information obey the prior assumption of their distributions at the same time. Neural Survival Clustering (Jeanselme et al., 2022) attempts to learn the survival probability for each instance by learning fixed neural networks. This framework brings more flexibility due to the lack of assumptions. However, the model needs large amounts of training data to prevent overfitting. Thus, the model may not be applicable to small-scale data, especially in the medical domain.

3. Preliminaries

3.1. Basic notation

Survival analysis aims to estimate the probability of an event of interest happening after a certain time t based on the features X of individual subjects (Flynn, 2012). This probability can be modeled by a survival function $S(\cdot|X) = P(T > t|X)$. The data we tackled are assumed to be right-censored. This means our dataset $\mathcal{D}$ consists of tuples $\{\mathbf{x}_i, t_i, \delta_i\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^d$ is the d dimensional feature vector for the ith subject, t_i is the last time we followed the ith subject, δ_i is the event indicator, and N is the number of all subjects. If $\delta_i = 1$, the event will occur at time t_i for the ith subject (this means the ith subject is uncensored). If $\delta_i = 0$, we lose the ith subject at time t_i, or the monitoring ended before the occurrence of the event (this means the ith subject is censored). We denote $\mathcal{D}_U$ as the uncensored subset and $\mathcal{D}_C$ as the censored subset.

3.2. Cox PH

Cox PH model (Cox, 1972) is the most popular and conventional survival model. It assumes proportional hazards, meaning the ratio of hazards of two subjects does not change over time. Suppose the parameter of the Cox PH model is $\boldsymbol{\beta} \in \mathbb{R}^d$, then the hazard function at time t for the ith subject with feature vector $\mathbf{x}_i$ is

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t)\exp(\boldsymbol{\beta}^\top \mathbf{x}_i), \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard that describes how the risk of event per time unit changes over time at baseline level of features. To estimate the parameter $\boldsymbol{\beta}$, Cox PH model uses the maximum likelihood estimation (MLE) to maximize the Cox partial likelihood:

$$L(\beta) = \prod_{i\in\mathcal{D}_U} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}{\sum_{j:t_j\geq t_i}\exp(\boldsymbol{\beta}^\top \mathbf{x}_j)}. \tag{2}$$

By calculating the first and second derivative of this partial likelihood function, $\boldsymbol{\beta}$ can be obtained using the Newton–Raphson algorithm (Akram and Ann, 2015).

3.3. Deep Cox

The Deep Cox model (also called DeepSurv) (Katzman et al., 2018) is another solution for the Cox PH idea. As opposed to following the traditional solution of maximizing the partial likelihood mentioned above, Deep Cox attempts to leverage deep neural networks to learn a nonlinear mapping $\phi_\theta(\cdot)$ parameterized by θ for the features $\mathbf{x}_i$ and try to get the optimal predictive model by minimizing the negative logarithm of the partial likelihood as the loss function:

$$\ell(\boldsymbol{\theta}) = -\sum_{i\in\mathcal{D}_U}\left(\phi_\theta(\mathbf{x}_i) - \log\sum_{j:t_j\geq t_i}\exp(\phi_\theta(\mathbf{x}_j))\right). \tag{3}$$

3.4. Clustering mechanism

The clustering mechanism of *post-hoc* survival models is illustrated in Fig. 1. Cox PH, Deep Cox and DSM are used as *post-hoc* clustering methods in this paper. To cluster MCI patients, we first obtain their predicted risk scores, and then we set a threshold (usually the median or the mean of the whole risk scores) to get the subtypes. In our experiment, we choose to use median as the threshold for the three baselines. This is because using the median as the threshold generates better results for them (refer to Table E.10 in the Appendix). This presents a challenge for our method, but we will show in our experiment that our method still outperforms the improved versions of the baselines. In our paper, it is important to emphasize that we concentrate on two distinct clusters: one associated with low risk and the other with high

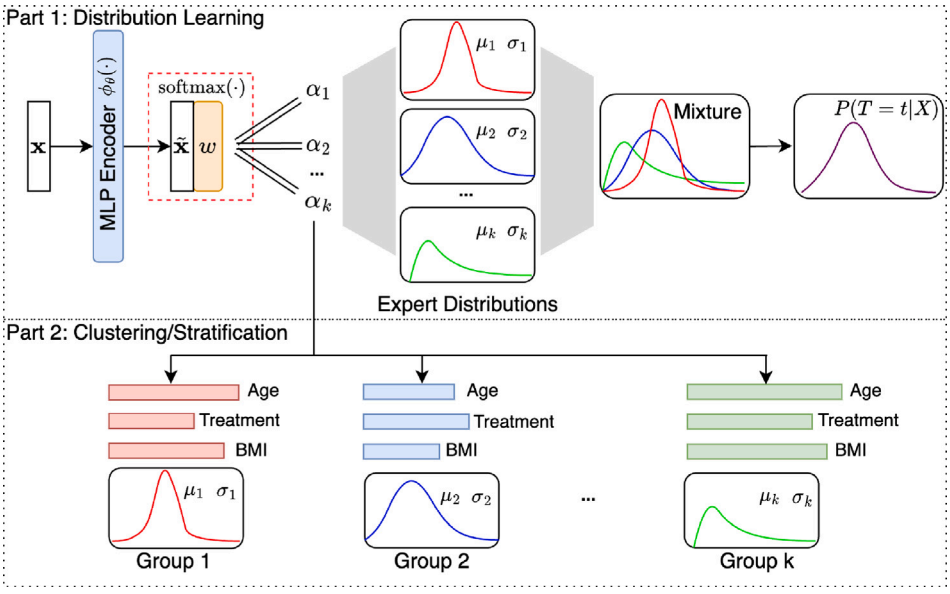


Fig. 2. Mechanism of Deep Clustering Survival Machines (DCSM). In Part 1, DCSM is designed to learn a conditional distribution $P(T = t|X)$, which is a weighted combination over all K constant expert distributions (Weibull distributions). Part 2 illustrates that DCSM cluster each instance/subject according to the weight α_k that is allocated to each expert distribution.

risk. Our primary objective is to provide valuable insights for patients and clinicians regarding the likelihood of developing AD (Alzheimer's disease). By identifying individuals at high risk, we can emphasize the need for early intervention or more customized treatment strategies. While high and low-risk clusters hold significant relevance, including an additional cluster such as medium risk would only introduce confusion for patients and clinicians, thereby detracting from the overall clarity and effectiveness of our findings. Despite these reasons, we still conduct the experiments for $K = 3$ to illustrate the robustness of our method (refer to Table D.7 in the Appendix).

4. Method

In Part 1 of Fig. 2, the deep clustering survival machines (DCSM) is designed to learn a conditional distribution $P(T|X = \mathbf{x})$ by optimizing the maximum likelihood estimation (MLE) of the time T. Similar to the mixture model learning paradigm, the conditional distribution $P(T|X = \mathbf{x})$ is characterized by learning a mixture over K well-defined parametric distributions, referred to as *expert distributions*. In order to use gradient-based methods to optimize MLE, we choose the Weibull distributions as the expert distributions that are flexible to fit various distributions and have closed-form solutions for the PDF and CDF:

$$\text{PDF}(t) = \frac{\mu}{\sigma}\left(\frac{t}{\sigma}\right)^{\mu-1} e^{-\left(\frac{t}{\sigma}\right)^{\mu}}, \ \text{CDF}(t) = e^{-\left(\frac{t}{\sigma}\right)^{\mu}}, \tag{4}$$

where μ and σ are the shape and scale parameters separately.

Part 1 of Fig. 2 indicates that we firstly need to learn an encoder for the input features $\mathbf{x} \in \mathbb{R}^d$ to obtain a compact representation $\tilde{\mathbf{x}} \in \mathbb{R}^{d'}$. Here we use a multi-layer perceptron (MLP) $\phi_\theta(\cdot)$ parameterized by θ as the backbone model. This representation will be multiplied by a parameter $\boldsymbol{w} \in \mathbb{R}^{d' \times K}$ with *softmax* to obtain the mixture weight α_k, $k = 1, \ldots, K$ with respect to each (kth) expert distribution that is parameterized by μ_k and σ_k. The final survival distribution for the time T conditioned on each instance is a weighted combination over all K constant expert distributions. Eventually, we have a set of parameters $\Theta = \{\theta, \boldsymbol{w}, \{\mu_k, \sigma_k\}_{k=1}^K\}$ to learn during the training process. Because μ_k and σ_k are the same for different input instances, we can cluster each instance/subject according to the weight α_k that is allocated to each expert distribution, as illustrated in Part 2 of Fig. 2. Specifically, we assign an subgroup/cluster indicator k to each instance when the instance's corresponding weight α_k is the largest among all K weights.

According to the framework of MLE, our goal is to maximize the likelihood with respect to the timing information T conditioned on $\mathbf{x}$. Given that the likelihood functions are different for uncensored and censored data, we calculate them separately. For the uncensored data, the log-likelihood of T is computed as follows, where **ELBO** is the lower bound of the likelihood derived by Jensen's Inequality:

$$\begin{aligned}
&\ln \mathbb{P}(\mathcal{D}_U|\Theta) = \ln\left(\Pi_{i=1}^{|\mathcal{D}_U|}\mathbb{P}(T = t_i|X = \mathbf{x}_i, \Theta)\right) \\
&= \sum_{i=1}^{|\mathcal{D}_U|} \ln\left(\sum_{k=1}^{K}\mathbb{P}(T = t_i|\alpha_k, \mu_k, \sigma_k)\mathbb{P}(\alpha_k|X = \mathbf{x}_i, \boldsymbol{w})\right) \\
&= \sum_{i=1}^{|\mathcal{D}_U|} \ln\left(\mathbb{E}_{\alpha_k\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})}[\mathbb{P}(T = t_i|\alpha_k, \mu_k, \sigma_k)]\right) \\
&\geq \sum_{i=1}^{|\mathcal{D}_U|} \left(\mathbb{E}_{\alpha_k\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})}[\ln \mathbb{P}(T = t_i|\alpha_k, \mu_k, \sigma_k)]\right) \\
&= \sum_{i=1}^{|\mathcal{D}_U|} \left(\text{softmax}_K(\ln \text{PDF}(t_i|\mu_k, \sigma_k))\right) = \textbf{ELBO}_U(\Theta).
\end{aligned} \tag{5}$$

Similarly, the log-likelihood of T for the censored data is:

$$\begin{aligned}
&\ln \mathbb{P}(\mathcal{D}_C|\Theta) = \ln\left(\Pi_{i=1}^{|\mathcal{D}_C|}\mathbb{P}(T > t_i|X = \mathbf{x}_i, \Theta)\right) \\
&\geq \sum_{i=1}^{|\mathcal{D}_C|} \left(\mathbb{E}_{\alpha_k\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})}[\ln \mathbb{P}(T > t_i|\alpha_k, \mu_k, \sigma_k)]\right) \\
&= \sum_{i=1}^{|\mathcal{D}_C|} \left(\text{softmax}_K(\ln \text{CDF}(t_i|\mu_k, \sigma_k))\right) = \textbf{ELBO}_C(\Theta).
\end{aligned} \tag{6}$$

In addition, to stabilize the performance, we incorporate prior knowledge for μ_k and σ_k. Specifically, we minimize the prior loss L_{prior} to make them as close as possible to the μ and σ from the prior model:

$$L_{prior} = \sum_{k=1}^{K} \|\mu_k - \mu\|_2^2 + \|\sigma_k - \sigma\|_2^2. \tag{7}$$

where the prior model is learned by the same MLE framework with a single expert distribution that is still Weill distribution. In this context, the prior μ and σ are not traditional prior distributions for μ_k and σ_k as required by the maximum a posteriori (MAP) estimation. Instead, they are specific values derived from a pre-trained prior model, which follows a similar framework to ours but with only a single expert

Algorithm 1 DCSM Training, Time-to-event Prediction and Clustering

Input: Dataset $\mathcal{D}$ consists of tuples $\{\mathbf{x}_i, t_i, \delta_i\}_{i=1}^N$
Output: Trained model $f = \{\phi_\theta, \boldsymbol{w}, \{\mu_k, \sigma_k\}_{k=1}^K\}$, the estimated risk r_i and the cluster label k for each subject i
Split $\mathcal{D}$ into a training set $\mathcal{D}_{\text{tr}}$ and a testing set $\mathcal{D}_{\text{te}}$
Training Phase...
Initialize a prior-model $f_{\text{prior}} = \{\phi_{\theta_{\text{prior}}}, \boldsymbol{w}_{\text{prior}}, \mu, \sigma\}$
Pre-train the model f_{prior} with only one expert distribution parameterized by μ and σ by maximizing (5)+(6) on $\mathcal{D}_{\text{tr}}$ where the trained prior-model f_{prior} will be used in (7)
Initialize a formal model $f = \{\phi_\theta, \boldsymbol{w}, \{\mu_k, \sigma_k\}_{k=1}^K\}$
Train the model f with multiple expert distributions parameterized by μ_k and σ_k by minimizing (8) on $\mathcal{D}_{\text{tr}}$ and obtain the trained model f
Time-to-event Prediction Phase...
Calculate the weights $\{\alpha_k\}_{k=1}^K$ of all K expert distributions for the ith subject by (9)
Obtain the risk r_i of the ith subject by (10)
Clustering Phase...
Obtain the cluster label k for the ith subject based on the largest α_k

distribution. During the official training of our model, we introduce regularization terms in the loss function to prevent the parameters of the expert distributions from deviating excessively from the prior values obtained from the pre-trained model. This approach deviates from the conventional MAP estimation, where prior distributions are typically used. The final objective L_{all} is the sum of the negative of the log-likelihoods of both the uncensored and censored data in addition to the prior loss where λ is a trade-off hyperparameter:

$$L_{all} = L_{prior} - \mathbf{ELBO}_U(\Theta) - \lambda \cdot \mathbf{ELBO}_C(\Theta). \tag{8}$$

The implementing details are as follows. First, we split the dataset $\mathcal{D}$ into a training set $\mathcal{D}_{\text{tr}}$ and a testing set $\mathcal{D}_{\text{te}}$. In the training phase, we first initialize a prior-model $f_{\text{prior}} = \{\phi_{\theta_{\text{prior}}}, \boldsymbol{w}_{\text{prior}}, \mu, \sigma\}$ where the prior-model only contains one expert distribution parameterized by μ and σ. In our implementation, we use PyTorch (Paszke et al., 2019) to conduct the model initialization. Then we pre-train the prior-model f_{prior} by maximizing the likelihood (5)+(6) on $\mathcal{D}_{\text{tr}}$. In this way, the learned μ and σ from the prior model can be used in (7). Then we initialize the formal model $f = \{\phi_\theta, \boldsymbol{w}, \{\mu_k, \sigma_k\}_{k=1}^K\}$ and train it by minimizing (8) on $\mathcal{D}_{\text{tr}}$. After we obtain the trained model f, we can conduct time-to-event prediction and clustering simultaneously. To do that, we first need to calculate the weights $\{\alpha_k\}_{k=1}^K$ of all K expert distribution for the ith subject by the *softmax* on $\boldsymbol{w}^\top \phi_\theta(\mathbf{x}_i)$:

$$\alpha_k = \frac{\exp((\boldsymbol{w}^\top \phi_\theta(\mathbf{x}_i))_k)}{\sum_{j=1}^K \exp((\boldsymbol{w}^\top \phi_\theta(\mathbf{x}_i))_j)} \tag{9}$$

For time-to-event prediction, we use the weights to conduct weighted combination for all the CDF value given a specific time t which is the time horizon t_{max} in our case. Then the risk for the ith subject r_i is estimated by

$$\begin{aligned} r_i &= 1 - \sum_{k=1}^K \mathbb{P}(T \le t_{max} | \alpha_k, \mu_k, \sigma_k) \\ &= 1 - \sum_{k=1}^K \alpha_k \text{CDF}(t_{max}) = 1 - \sum_{k=1}^K \alpha_k \exp\left(-\left(\frac{t_{max}}{\sigma_k}\right)^{\mu_k}\right). \end{aligned} \tag{10}$$

For clustering, we just assign the index k to the ith subject if α_k is the largest among all the K weights. The algorithm is summarized in Algorithm 1.

5. Experiments

In this section, we first introduce the datasets we used in the experiments. This includes three popular AD-related datasets, four benchmark datasets, and a group of genotyping data relevant to AD. Then, we introduce the metrics, baselines, and settings in our experiments. Finally, we report and discuss the LogRank, C-index, genetic association analysis, and interpretability results. Code is available at https://github.com/BojianHou/DCSM.

5.1. Datasets

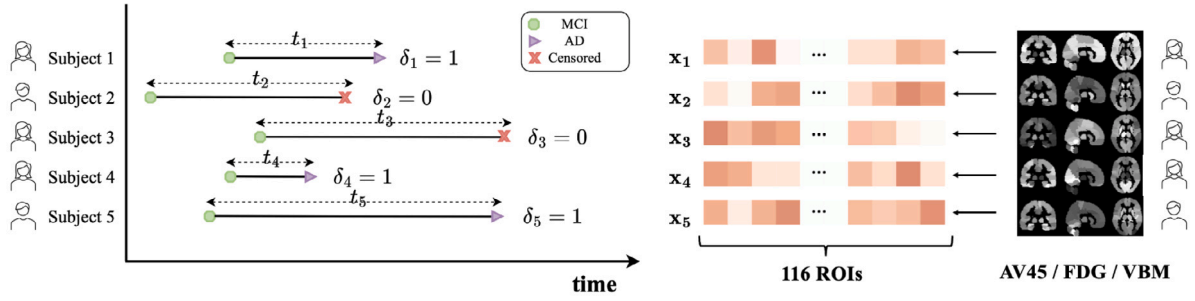
The genotyping data, demographic data and imaging data used in our experiments were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Weiner et al., 2017). ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Up-to-date information about the ADNI database is available at www.adni-info.org.

We focused our analysis on three ADNI imaging modalities:

- AV45: Florbetapir PET (Jagust et al., 2015) that measures amyloid burden.
- FDG: Fluorodeoxyglucose PET (Jagust et al., 2010) which measures glucose metabolism.
- VBM: Structural magnetic resonance imaging (MRI) (Jack et al., 2015) measuring brain morphometry.

Structural MRI scans were processed with voxel-based morphometry (VBM) using the statistical parametric mapping software tool (Ashburner and Friston, 2000). The MarsBaR region of interest (ROI) toolbox (Tzourio-Mazoyer et al., 2002) was used to extract mean gray matter density, amyloid, and FDG-PET glucose utilization values for each ROI. After extraction, there were 116 ROI-level measures for each modality. All the MCI participants with no missing data were analyzed, and we had $N = 466$ for AV45, $N = 467$ for FDG, and $N = 462$ for VBM. We use AV45 (FDG or VBM) as observed covariates as shown in Fig. 3(b). Their characteristics are summarized in Table 1.

As mentioned in Section 3.1, in survival analysis, the label of each subject consists of two items. One is the time t which is the time duration we followed the subject. The other is the event indicator δ that is to indicate whether an event such as death or disease onset happens to the subject. Originally, the subjects from the ADNI dataset did not have such information. What they have are the several visits and which disease they are diagnosed at each certain visit. To obtain the label, we first selected the subjects with MCI status at their first visit and recorded the date of their imaging modalities visit as the *initial time*. Note that in our work, we are interested in the risk of MCI patients converting to AD. Thus, we only focus on the subjects who are MCI patients originally. Then, among these subjects, we selected the ones who ended up converting to AD and recorded this visit time as the *event time*, and their event indicator is 1. We call these patients uncensored patients. It is worth noting that there are reverting issues among these uncensored patients indicating that they can revert back to MCI. We included these patients since we only recorded the label until the first time they converted to AD (event happens), which meets the definition of uncensored data. There are 3 such kind of patients in VBM and 4 for both AV45 and FDG. For the remaining subjects who did not change their status over all the visits, their event indicator is 0. Their event time will be the time of their final visit. The time t is the difference between initial time and event time. The feature vector of each subject is collected from the features at their first visit. All the other feature vectors of that subject generated in the subsequent visits are ignored. Fig. 3(a) illustrates the survival information for five subjects where Subjects 1, 4 and 5 converted to AD, and Subjects 2 and 3 were lost to follow-up.



(a) Survival information (b) Covariates

Fig. 3. Illustration of how our features and labels look like. In this figure, there are 5 subjects in total. (a) illustrates the labels including time duration t and event indicator δ. All five subjects started from MCI. During the monitoring, the event happened to Subjects 1, 4 and 5 thus their event indicator δ equals to 1. Subject 2 quit the monitoring and there was no event happened to Subject 3 until the end of the monitoring. Thus their event indicators are both 0. (b) shows the features (covariates) of the five subjects. All the features are extracted from the 116 brain regions for each modality.

Table 1
Statistics of AD-related datasets used in the experiments. N refers to number of subjects. m refers to the number of brain regions. The time range t_{max} is noted in days. Education is noted in years.

Dataset	N	m	Event (%)	t_{max}	Age	Gender(M/F)	Education	# of overlapping
AV45	466	116	25.10	2258	72.38 ± 7.81	267/199	16.20 ± 2.71	
FDG	467	116	24.62	2254	72.36 ± 7.81	266/201	16.18 ± 2.71	432
VBM	462	116	24.24	2275	72.42 ± 7.91	265/197	16.18 ± 2.74	

Table 2
Statistics of benchmark datasets used in the experiments. The time range t_{max} in PBC is noted in years while others are noted in days. "FRAM" refers to "FRAMINGHAM".

Dataset	SUPPORT	PBC	FRAM	FLCHAIN
Events (%)	68.11	37.28	30.33	30.07
N	9105	1945	11 627	6524
d (categorical)	44 (26)	25 (17)	18 (10)	8 (2)
t_{max}	2029	14.31	8766	5167

For the genetics data, we first downloaded genotyping data from ADNI 1, GO, 2, and 3 studies from the ADNI database (Shen et al., 2014; Saykin et al., 2015; Shen and Thompson, 2020). Then, McCarthy Group Tools (https://www.well.ox.ac.uk/~wrayner/tools/) were used for alignment. We aligned the genotyping data to the Homo Sapiens (human) genome assembly NCBI37 (hg19) genome builder, according to 1000 Genome phase 3 dataset (1000 Genomes Project Consortium et al., 2015). To complement the missing data, we imputed those genotypes using the Michigan Imputation Server (Das et al., 2016) with the 1000 Genome phase 3 reference panel of European ancestry. We annotated our imputed genotyping data using ANNOVAR (Wang et al., 2010). After alignment and imputation, we performed the quality control (QC) using the following criteria: genotyping call rate > 98%, minor allele frequency > 0.1%, Hardy-Weinberg Equilibrium > 1e-6, missingness per individual < 5%. All the QC was performed using PLINK 1.9 (Chang et al., 2015).

To further validate the effectiveness of the proposed method, we also conducted experiments on four public benchmark datasets that are all real-world datasets:

- SUPPORT (Knaus et al., 1995): The SUPPORT dataset is sourced from a study conducted by Vanderbilt University that aims to estimate the survival rate of seriously ill adults who are hospitalized.
- PBC (Fleming and Harrington, 2013): The dataset known as Primary Biliary Cirrhosis is commonly used to assess the performance of survival analysis models that incorporate time-dependent covariates.
- FRAMINGHAM (Dawber et al., 1951): The Framingham dataset consists of 4,434 participants from the well-known and ongoing Framingham Heart study. This dataset is utilized for studying the epidemiology of hypertensive and arteriosclerotic cardiovascular disease.
- FLCHAIN (Kyle et al., 2006): This dataset is a mix of different people, and half of the subjects from a study analyzing the correlation between serum free light chain (FLC) and mortality. The original sample covers around two-thirds of Olmsted County's residents aged 50 or above.

The statistics of the four benchmark datasets are summarized in Table 2.

5.2. Metrics, baselines and settings

Metric We use "LogRank" to evaluate the clustering performance of all the methods. LogRank is a statistical test that compares the survival curves of two or more groups of subjects (Mantel et al., 1966) and is popular and widely used for survival analysis. It tests whether or not there is a significant difference in survival between the groups. It is calculated by comparing the observed and expected number of events in each group under the null hypothesis of no difference. To further validate the efficacy of our framework, we also use genetic association analysis to evaluate the clustering results. We want to identify the genetic basis of different AD subtypes to validate our subtype findings.

Rather than the LogRank, we also incorporate "Concordance Index" (C-Index) (Harrell et al., 1982) as an additional metric to evaluate the time-to-event prediction performance. The C-Index is a widely used measure in survival analysis. It measures how well the order of survival times matches with predictions made by models. Note that the time-to-event prediction is actually not our goal. Our DCSM method is specifically designed for clustering and it is sufficient if we our method can achieve state-of-the-art clustering results and behave reasonably well regarding time-to-event prediction.

Baseline For the clustering task, we compare our method with seven baseline methods, which are either conventional or state-of-the-art:

- KMeans (Hartigan and Wong, 1979): a traditional and popular clustering method that iteratively updates the clustering means and cluster assignments.
- Cox PH (Cox, 1972): a classic survival model for survival risk prediction, which assumes that the hazard rate for each instance, known as the proportional hazard (PH), is constant over time.

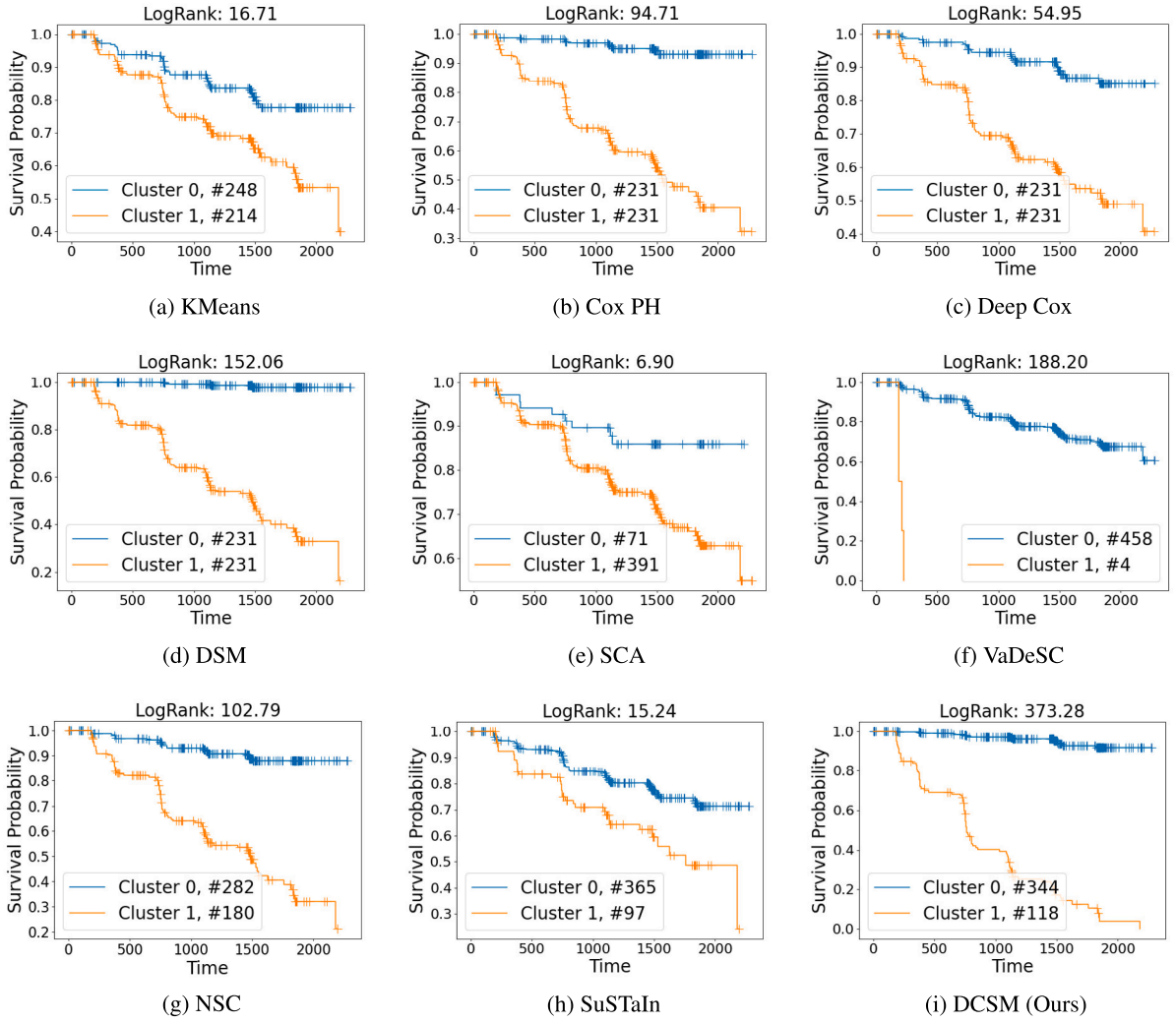


Fig. 4. The Kaplan–Meier plots of KMeans, Cox PH, Deep Cox, DSM, SCA, Vadesc, NSC, SuSTaIn and our DCSM on dataset VBM. The cross mark on the curve means censoring. Cluster 0 means low-risk group while Cluster 1 means high-risk group.

- Deep Cox (Katzman et al., 2018): a deep learning variant of Cox PH that uses Cox PH loss to optimize the parameters of deep neural networks, also called DeepSurv.
- DSM (Nagpal et al., 2021): the Deep Survival Machines model learns different base distributions for different instances using specified prior distributions.
- SCA (Chapfuwa et al., 2020): the Survival Clustering Analysis model assumes that the latent space is a mixture of distributions and uses the truncated Dirichlet process to realize the automatic identification of the number of clusters.
- VaDeSC (Manduchi et al., 2021): the Variational Deep Survival Clustering model uses a Gaussian mixture distribution to model the features in a latent space and uses the Weibull distribution to model the survival timing information.
- NSC (Jeanselme et al., 2022): Neural Survival Clustering is a discriminative variant of DSM that uses neural networks to model the base distributions for each instance.
- SusTaIn (Young et al., 2018): the Subtype and Stage Inference method is designed to identify phenotypes with distinct temporal progression patterns.

For Cox PH, Deep Cox, and DSM, we use the median of all the risk scores as the threshold. In order to demonstrate the difference between the clustering effectiveness of unsupervised learning and semi-supervised learning methods, we also incorporate KMeans (Hartigan and Wong, 1979) as the baseline besides the six survival models to discover the subtypes. For the time-to-event prediction task, we only report the results of all the survival models mentioned above without KMeans.

Setting To obtain the clustering results, we follow the common practice of training and testing on the entire dataset, similar to how other typical clustering methods operate. After training, we down-sampled the testing data (95% are sampled randomly) five times and obtained the average LogRank values along with the standard deviation over the five runs. For the time-to-event prediction task, we split the dataset using the same random seed for all the methods to maintain consistent training (70%) and testing (30%) datasets across experiments. The training set is further divided into validation (1/7) and training (6/7) subsets. Then the hyper-parameters are tuned exclusively on the validation set. Specifically, we perform a grid search for the hyperparameters, including the trade-off parameter (also called discount) ([0.5, 0.75, 1]), the learning rate ([1e-2, 1e-3, 1e-4]), and the number of MLP layers together with the output size ([[], [50], [50, 50]]), based on the C-index performance on the validation set. We then train the model using the optimal parameters on the entire training subset (excluding validation data). We repeat the experiment with five different random seeds to ensure statistical robustness. Finally, we report the mean and standard deviation (std) over the five results from the five different testing sets.

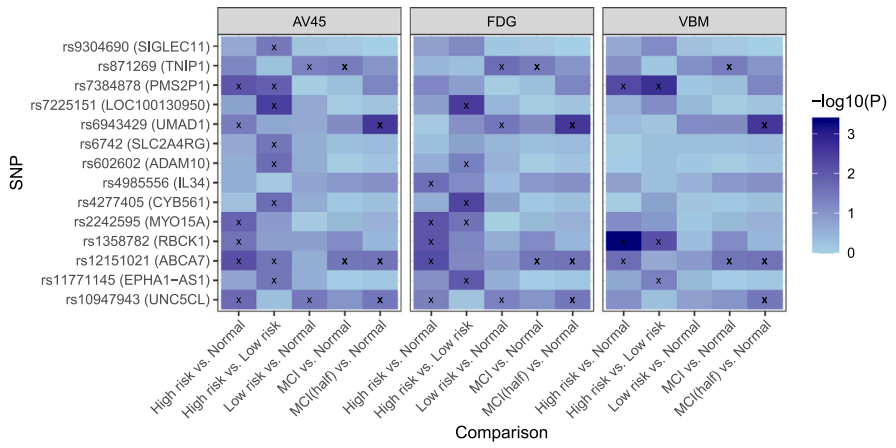


Fig. 5. Targeted genetic association results of DCSM. In the parentheses of the vertical label is the name of the gene closest to the corresponding SNP. The darkness of each blue patch represents the negative logarithm of *p-value* regarding each SNP. The darker the patch is, the more significant the SNP is. "x" marker on the patch means the SNP is statistically significant after FDR correction. This figure indicates the discrepancy between the high-risk group and the normal group is as large as the discrepancy between the high-risk group and the low-risk group, which reassures the performance of our DCSM algorithm.

Table 3
LogRank results comparison between KMeans, Cox PH, Deep Cox, DSM, SCA, VaDeSC, NSC, SuSTaIn and DCSM on datasets AV45, FDG and VBM. The best one in each block (for specific metric and dataset) is bold.

Methods	AV45	FDG	VBM
KMeans	66.96 ± 1.75	28.84 ± 1.36	16.17 ± 1.01
Cox PH	133.60 ± 3.65	117.80 ± 1.88	89.07 ± 3.97
Deep Cox	121.49 ± 10.99	95.39 ± 16.07	63.49 ± 2.76
DSM	160.62 ± 3.79	124.26 ± 1.74	120.41 ± 2.47
SCA	40.10 ± 26.47	18.15 ± 8.94	4.71 ± 3.41
VaDeSC	108.43 ± 131.86	133.69 ± 203.21	282.08 ± 157.12
NSC	160.88 ± 52.13	213.61 ± 82.60	65.52 ± 27.65
SuSTaIn	34.35 ± 13.94	21.19 ± 9.09	16.30 ± 2.83
DCSM	**317.84 ± 31.89**	**384.62 ± 24.03**	**369.29 ± 26.87**

Table 4
C-Index results comparison between Cox PH, Deep Cox, DSM, SCA, VaDeSC, NSC and DCSM on datasets AV45, FDG and VBM. The best one in each block (for specific metric and dataset) is bold.

Methods	AV45	FDG	VBM
Cox PH	0.6365 ± 0.0305	0.6780 ± 0.0811	0.5981 ± 0.0453
Deep Cox	**0.7566 ± 0.0131**	0.7657 ± 0.0545	0.6423 ± 0.0286
DSM	0.7507 ± 0.0262	**0.7928 ± 0.0406**	0.6492 ± 0.0105
SCA	0.5175 ± 0.1136	0.5096 ± 0.1494	0.5122 ± 0.0462
VaDeSC	0.4168 ± 0.0475	0.4194 ± 0.0917	0.5169 ± 0.0632
NSC	0.6884 ± 0.0768	0.7699 ± 0.0498	0.5986 ± 0.0733
DCSM	0.7502 ± 0.0394	0.7770 ± 0.0250	**0.6549 ± 0.0317**

5.3. Results on ADNI

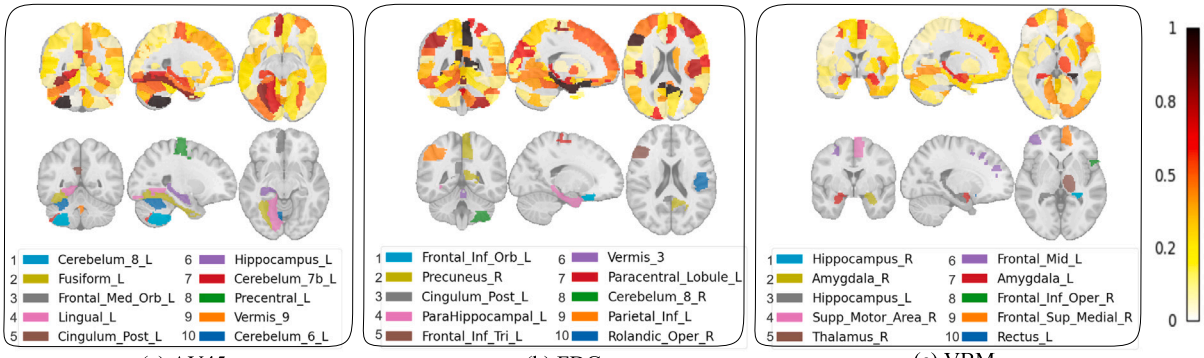
5.3.1. LogRank results

Table 3 gives the LogRank comparison between KMeans, Cox PH, Deep Cox, DSM, SCA, VaDeSC, NSC, SuSTaIn and DCSM on the AV45, FDG, and VBM datasets. We can see that KMeans, which merely uses the feature information, achieves the lowest LogRank compared to other survival models excluding SCA. SCA cannot explicitly control the number of clusters. It usually sets the upper bound of the number of clusters as 25 and uses the Dirichlet process to automatically identify the number of clusters, which makes it difficult to compare to other baselines. To have a fair comparison, we set the upper bound of the number of clusters as 2 for SCA which results in poor clustering performance. Overall, the results in Table 3 demonstrate that incorporating partial survival information allows survival models to better stratify the groups such that the survival differences between the subgroups are larger than their unsupervised counterparts. DCSM, specifically designed for survival clustering, obtains the highest LogRank results out of all the baseline methods.

To make the results more intuitive, we also provide the corresponding Kaplan–Meier (KM) plots that are shown in Fig. 4. We only provide the KM plots on the VBM dataset. Other similar results are deferred to Appendix A. In general, the smaller the LogRank, the closer the two survival curves. From Fig. 4 we can see that the two survival curves of KMeans and SuSTaIn are closer to each other compared to survival models. This shows that survival models can more effectively stratify individuals into different subgroups concerning the risk of MCI converting to AD. Despite the success of SuSTaIn in identifying phenotypes with distinct temporal progression patterns, it is not designed for differentiating risk levels and thus performing inferior to survival models. Note that the two curves of DCSM are farthest apart from each other compared to the other methods with the exception of VaDeSC. The two curves of VaDeSC are farther from each other than DCSM, but the number of patients in Cluster 1 (with high risk) is only 4 which means that the two clusters are very imbalanced and thus not effectively stratified. In addition, the cross marks on the curves indicate censoring, i.e., we do not observe the event of MCI converting to AD. Thus it should be with high probability that the censored samples have low risk to convert to AD. Therefore, we expect that a good clustering results should stratify more censored individuals to the low-risk group. We can see that on the curve of Cluster 1 (with high risk) of DCSM, the number of cross marks is small while most of the censored samples are on the low-risk curve (Cluster 0). This validates the effectiveness of DCSM.

5.3.2. C-index results

Our primary focus lies in clustering and subtype discovery, but we also evaluate the time-to-event prediction performance of DCSM in comparison to other methods with the exception of KMeans which is not able to do time-to-event (risk) prediction. The results are summarized in Table 4. Notably, our method demonstrates superior performance compared to other approaches on the VBM dataset, while maintaining competitive results on the other datasets. It is important to emphasize that our method is not specifically optimized for risk prediction; rather, our primary goal is to enhance clustering performance. In this context, the observed outcome is considered acceptable.



(a) AV45 (b) FDG (c) VBM

Fig. 6. Brain heat map of AV45, FDG, and VBM using the coefficient of our DCSM model. Each dataset includes three slices to display the brain region extensively. The heat maps in the first row showcase all the brain regions of interest, where darker colors represent greater importance for risk prediction. The second row highlights the top ten significant brain regions in each dataset, which are annotated in the legends below them.

5.3.3. Genetic association results

To biologically validate our clustering findings derived from the proposed DCSM, we carried out targeted genetic association analyses. Specifically, we first prioritized 57 single nucleotide polymorphism (SNP) variants from ADNI 1, GO, 2, 3 studies using the list of AD loci with genetic evidence compiled by the Alzheimer's Disease Sequencing Project (ADSP) Gene Verification Committee.[4] These genetic variants have been well-studied and verified to affect AD risk. Then, we conducted genetic association analyses on the subtypes of "High risk vs. Normal", "High risk vs. Low risk", "Low risk vs. Normal", "MCI vs. Normal", and "MCI (half) vs. Normal", where we use half of the MCI subjects (randomly selected) to make the comparison fairer, as the number of subjects in this group is closer to the high-risk, low-risk, and normal groups. The term "Normal" refers to the group of individuals who are cognitively normal. In our approach, we employed univariate logistic regression models, designating the risk of AD onset as the dependent variables and individual autosomal SNP variants as independent variables. To account for the potential confounding factors, we adjust our model with age, sex, and population structure captured by the first 10 principal components obtained from genotyping data. We use the false discovery rate (FDR) to correct the multiple comparison problems. The software used to perform the association analyses was PLINK1.9 (Chang et al., 2015).

As shown in Fig. 5, we find more significant SNPs in the first two columns than in the last three columns of each dataset. This indicates that the discrepancy between the high-risk group and the normal group is as large as the discrepancy between the high-risk group and the low-risk group. In contrast, the discrepancy between the low-risk group and the normal group, as well as between the whole MCI group and the normal group, is much smaller. This is reasonable since the high-risk group should be significantly different from other groups such as the low-risk group or the normal group, while the discrepancy between other groups should be small. These results validate the efficacy of our DCSM clustering framework, where we successfully identified the high-risk group that should be treated more carefully. Our study underscores the robustness of our clustering findings and contributes to a better understanding of the genetic foundations of AD risk subtypes.

We also showcase the targeted genetic association analysis results of all the methods in Fig. B.12 for a comprehensive comparison in Appendix A. We expect to identify more significant SNPs in the first two columns of each block for each dataset. To clearly compare them, we sum up the number of "x" markers in the first two columns for each method. Based on this, KMeans has 24, Cox PH has 20, Deep Cox has 23, DSM has 22, SCA has 16, Vadesc has 0, NSC has 29 and our DCSM has 30, which demonstrates the superiority of the proposed method.

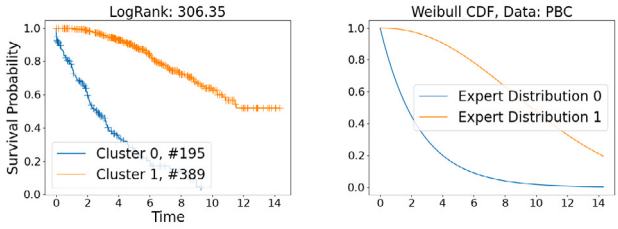
[4] https://adsp.niagads.org/gvc-top-hits-list/

5.4. Interpretability study

Fig. 6 shows the important brain regions that DCSM pays attention to. We compute the importance of each brain region by calculating the dot product of all the parameter coefficients with respect to the final weight of each expert distribution. In particular, we first extract the coefficients from the MLP Encoder $\phi_\theta(\cdot)$ whose input dimension is 116 and the output dimension is 50 in our experiment. Next, we extract the parameter coefficients from $\boldsymbol{w}$ as shown in Fig. 2, whose dimension is 50×2. Finally, we do the dot product of these two coefficients and obtain our desired importance weight of brain regions whose dimensions are 116×2 for high- and low-risk groups respectively. In Fig. 6, we only illustrate the importance weight for the high-risk group in which we are more interested. When plotting the feature importance using weights, it is important to take the absolute value of each weight to indicate the impact on the final prediction, regardless of its sign. After obtaining the absolute values, the weights should be normalized by subtracting the smallest value and dividing by the range of the weights (biggest weight minus smallest weight). The importance weights for the low-risk group are found to be the exact opposite of the weights for the high-risk group. Normalizing them reveals that the final importance weight for the low-risk group is the same as that for the high-risk group, indicating a focus on the same brain regions for predicting risk in both groups. As can be seen, the first row shows the AD risk importance of all the regions, the darker the color is, the more important. The second row highlights the top ten important regions for each modality. For AV45 data, our method demonstrates high importance for AD risk prediction on the left fusiform, medial orbitofrontal, posterior cingulate, and precentral gyri. A significant abnormality in amyloid levels has been observed in the medial orbitofrontal cortex in AD patients (Collij et al., 2020). This region is associated with the episodic memory and simulation network and is very susceptible to aging (Fjell et al., 2014). Additionally, amyloid accumulation and gray matter atrophy occur simultaneously in AD patients in the fusiform gyrus (Chang et al., 2016). Furthermore, many studies have reported a significantly increased amyloid level in posterior cingulate (Huang et al., 2013; Collij et al., 2020), and the amyloid accumulation is related to the executive function and memory decline (Ali et al., 2022). It is evident from the FDG data that the posterior cingulate and the precuneus gyri are two potential AD biomarkers, which is consistent with prior studies showing severe hypometabolism reductions in MCI and AD patients (Bailly et al., 2015). Moreover, the reduction of metabolism caused by AD starts from the posterior cingulate cortex and gradually spreads to the frontal lobe such as the orbitofrontal gyrus (Mosconi, 2005; Bailly et al., 2015). For VBM data, the right hippocampus demonstrates the highest importance, and bilaterally hippocampus and amygdala regions all show a high risk of AD. The hippocampal neurons register places and people in memory, while the amygdala activates related cortical areas

Table 5
C-Index and LogRank results compared to Cox PH, Deep Cox, DSM, SCA, VaDeSC, NSC and DCSM on benchmark datasets. The best ones are bold.

Metric	Method	SUPPORT	PBC	FRAMINGHAM	FLCHAIN
C Index	Cox PH	0.8401 ± 0.0070	**0.8476 ± 0.0126**	0.7580 ± 0.0063	0.7984 ± 0.0046
	Deep Cox	0.8053 ± 0.0058	0.8474 ± 0.0181	**0.7612 ± 0.0057**	0.7893 ± 0.0063
	DSM	0.8300 ± 0.0045	0.8363 ± 0.0133	0.7593 ± 0.0050	**0.8009 ± 0.0036**
	SCA	0.8203 ± 0.0121	0.8251 ± 0.0258	0.5311 ± 0.1235	0.7467 ± 0.0091
	VaDeSC	**0.8419 ± 0.0041**	0.8278 ± 0.0085	0.5802 ± 0.0406	0.7886 ± 0.0100
	NSC	0.8146 ± 0.0072	0.8178 ± 0.0275	0.7396 ± 0.0175	0.7980 ± 0.0052
	DCSM (Ours)	0.8305 ± 0.0028	0.8359 ± 0.0109	0.7530 ± 0.0053	0.7916 ± 0.0074
LogRank	Cox PH	500.3282 ± 60.4977	198.2686 ± 17.3940	576.1450 ± 22.9621	399.0243 ± 25.7657
	Deep Cox	326.1931 ± 54.7026	203.3091 ± 22.8343	593.7317 ± 14.4697	403.4643 ± 35.8034
	DSM	563.4841 ± 0.0045	196.0912 ± 0.0133	587.5718 ± 0.0050	406.4549 ± 0.0036
	SCA	212.5712 ± 26.2629	260.5682 ± 67.4875	278.3525 ± 51.1866	536.1056 ± 109.1680
	VaDeSC	196.8495 ± 19.6887	118.9605 ± 77.4716	348.5500 ± 697.1000	95.5291 ± 108.9488
	NSC	416.4572 ± 31.9528	300.5617 ± 21.3671	313.3190 ± 41.8324	**713.7871 ± 40.9787**
	DCSM (Ours)	**1067.6184 ± 271.6551**	**302.5395 ± 30.1043**	**751.9770 ± 48.9725**	571.0441 ± 99.0101



(a) KM plot of DCSM (b) Expert distribution of DCSM

Fig. 7. (a) The Kaplan–Meier plots of DCSM on dataset PBC. The cross mark indicates censoring. The learned expert distributions are shown in (b). The shape of the two expert distributions resembles our Kaplan–Meier curves, facilitating effective data stratification.

to modulate recognition (Petrovic et al., 2008). The atrophy of the hippocampus is widely considered as the AD biomarker for early detection citedhikav2011potential, and the study found that amygdala atrophy is related to hippocampus atrophy and global AD severity (Poulin et al., 2011). These results also provide directions for discovering potential AD biomarkers.

As for the interpretability from the expert distribution perspective, we can interpret that each expert distribution represents a cluster center. The Kaplan–Meier plot of DCSM on the PBC benchmark dataset is displayed in Fig. 7. The plot is accompanied by CDF (Cumulative Density Function) curves for expert distributions. Notably, the shapes of the two expert distributions closely resemble the Kaplan–Meier curves, offering an intuitive insight into how the expert distributions within the DCSM model effectively steer the patient clustering process and validate the robustness of data stratification. This alignment between the expert distributions and the Kaplan–Meier curves further reinforces the efficacy of our approach.

5.5. Results on benchmark datasets

Table 5 shows the C Index values on real data, including the average results of five independent runs and their standard deviations. These results indicate that our method achieves competitive performance compared to other baselines. Although our model's performance was not the best on some datasets, the difference between the results of DCSM and the best-performing model are not significant at a 95% confidence interval.

Table 5 also summarizes the results of the LogRank tests. The LogRank statistic evaluates how well the clustering results with regard to the survival information. A larger value indicates better performance. The results demonstrate that our method outperforms all the baselines on SUPPORT, PBC and FRAMINGHAM and secure the second position on FLCHAIN. This could be more useful than the time-to-event prediction because such information can facilitate personalized treatment planning. Clinicians may not be able to decide how to provide customized treatment merely based on the risk predictions. Instead, clustering results more intuitively differentiate the patients into groups, enabling more informed clinical decision-making.

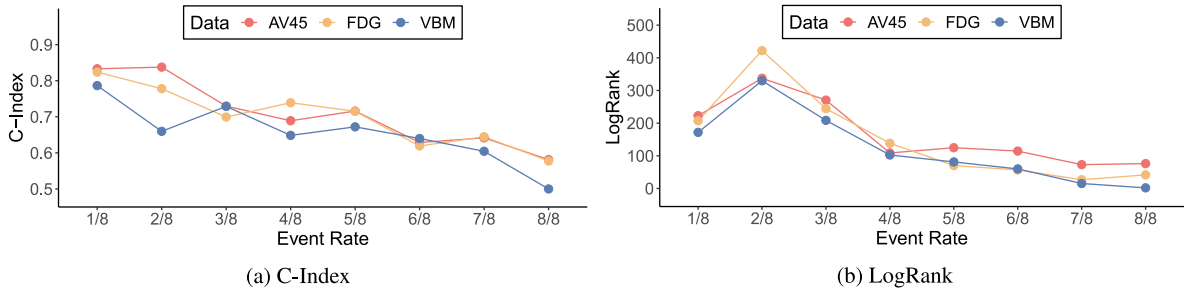
5.6. Sensitive and ablation analysis of DCSM model

In this section, we conduct a sensitive analysis and ablation study for our DCSM model. We study the performance change with different event rates, learning rates, trade-off parameters (discounts), and numbers of layers.

Event Rate To investigate the benefits of incorporating censored data, we conducted extensive experiments with different event rates, ranging from 1/8 to 1, by varying the number of censored samples. A smaller event rate corresponds to a higher number of censored samples. When the event rate is 1, it indicates that no censored data is included in the dataset. For the time-to-event prediction task, the modified data with specific event rates were divided into a 70% training set and a 30% testing set. Fig. 8(a) shows that as the event rate increases (less censored data included), the C-Index performance worsens, highlighting the importance of censored data for accurate time-to-event prediction. For the clustering task, the training data consists of modified data with specific event rates, while the testing data comprises the original entire dataset to ensure a fair comparison across different cases. Fig. 8(b) shows that a higher event rate will cause a lower LogRank performance, demonstrating the benefit of incorporating censored data. The LogRank performance outperforms the others when the event rate is 2/8 (1/4) because the event rate of the original dataset is exactly 1/4. Maintaining the same distribution between the training and testing results in the best performance.

Learning rate and discount In Fig. 9A, D and Fig. 9B, E, we show the sensitive analysis to different learning rates and discounts. We can see that our DCSM model is not sensitive to these two hyperparameters, especially for C-Index, which illustrates the robustness of the proposed method.

Layer We test our model's performance using four different layer settings: [], [50], [50,50], and [50,50,50] for the MLP component. As can be seen from Fig. 9C and Fig. 9F, for LogRank performance, there is a significant performance improvement if we choose an appropriate MLP while the empty bracket achieves the lowest performance demonstrating that the design of the MLP plays an important role in clustering. On the other hand, the C-Index performance does not change too much showing that C-Index is robust to different model structures. It is worthy to note that in all the baselines, only DSM (Nagpal et al., 2021) and NSC (Jeanselme et al., 2022) share the same network structure as ours. However, different numbers of layers or different output sizes can yield different results. We conducted the same hyperparameter search



(a) C-Index (b) LogRank

Fig. 8. The performance of C-Index and LogRank for different event rate based on our DCSM model. The larger the event rate, the worse the performance indicating that the censored data plays an important role in our DCSM's performance.

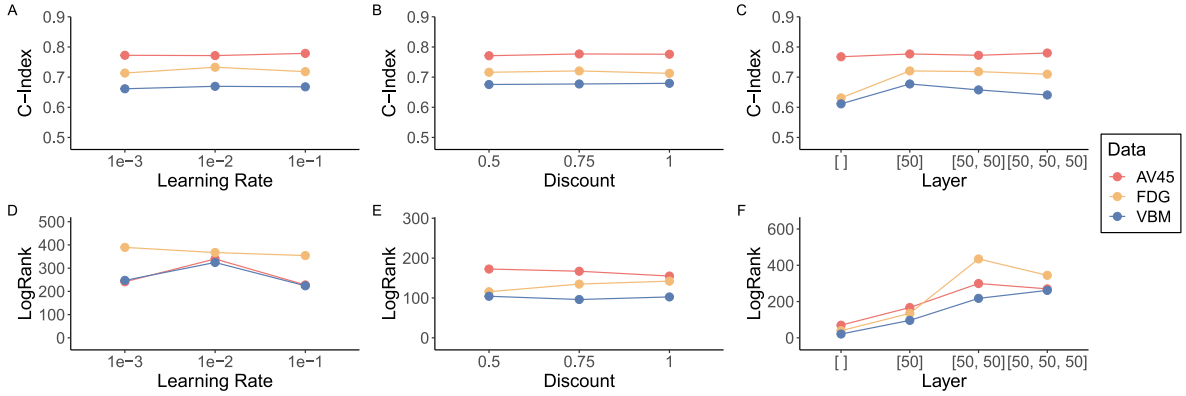


Fig. 9. The performance of C-Index and LogRank for different hyper-parameters of DCSM model. Our model is robust to learning rate and discount. LogRank performance is sensitive to layer indicating that the design of MLP plays an important role in our DCSM's clustering performance.

protocol for all the baselines and selected the best performance for each baseline to ensure a fair comparison. Thus, we did not set them with the same structure.

6. Discussion

Generative and discriminative approaches each have their strengths and weaknesses. Typically, a generative model assumes that the data follows a specific distribution, like Gaussian or Weibull distributions. This means that the model can be trained effectively even with a relatively small amount of data, as long as the data matches the assumed distribution. However, if the actual data distribution does not match this assumption, the resulting model could be biased and lose its ability to work well in different situations. For instance, consider the VaDeSC (Manduchi et al., 2021) model mentioned in our paper. It is a fully generative model that makes assumptions about both the features and the survival information in the data. If either of these assumptions is not satisfied, the model will be inferior. Discriminative models, on the other hand, do not make any assumptions about the data distribution. This means that they can learn the true pattern from the data, even if the data is not well-behaved. However, this approach often requires larger amounts of data to find the right parameters for the model. An example of a discriminative model is NSC (Jeanselme et al., 2022). NSC is a fully discriminative method that does not make any assumptions about the data. However, if the sample size is small, the model may not be able to learn the true pattern from the data.

A combination of discriminative and generative mechanisms can mitigate the drawbacks of both approaches. In our case, we do not make any assumptions about the features, as we believe that the distribution of features can vary from case to case. Making assumptions about the features can limit the model's capability. Instead, we assume that the survival information follows the Weibull distribution, which is a commonly used distribution for modeling survival timing data. The Weibull distribution is flexible in that it can mimic different distributions with different shapes and scales. For example, when the shape parameter is close to 3, the Weibull distribution approximates a normal distribution. When the shape parameter is equal to 1, the Weibull distribution is equivalent to a two-parameter exponential distribution. Therefore, the Weibull distribution is a suitable choice for modeling survival timing information. Our method also has its limitations. For example, if the number of samples is limited, the feature information may not be learned well. This is a common problem in the medical domain, where data are often scarce. This may be the reason why DCSM does not outperform the baseline methods in time-to-event prediction tasks. Additionally, if the survival timing data are too skewed to be modeled by the Weibull distribution, the model's performance may also be degraded. There is no perfect method that can fit all cases. However, by combining discriminative and generative mechanisms, we can develop more robust models that can handle a wider range of data distributions.

In addition, we want to emphasize that we only focus on two clusters for MCI subtype discovery in this study. One subtype is associated with a low risk of developing AD, and the other is associated with a high risk. Our goal is to help patients and clinicians understand their risk of developing AD. By identifying people who are at high risk, we can emphasize the need for early intervention or more customized treatment. While the high- and low-risk clusters are important, adding a medium-risk cluster would only confuse patients and clinicians. This would make our findings less clear and effective.

We intend to utilize the overlapping examples from the three modalities of ADNI data to build multi-modality survival model to improve both clustering and time-to-event prediction in the future work. The demographic information (age, sex, headsize, etc.) may have some effect on both the prediction and the imaging modalities. Thus, our future work are going to explore the potential of multi-modality study by incorporating various modalities such as demographic information,

genetics data, and other AD related biomarkers. One approach involves selecting overlapping samples from different modalities and combining them to create a more comprehensive representation for each sample, which we believe could enhance performance. Additionally, we aim to explore and develop more effective multi-modal strategies such as using attention mechanism from Transformer (Vaswani et al., 2017) or Canonical Correlation Analysis (Zhou et al., 2024) to fuse the modalities to further improve the clustering and time-to-event prediction ability of our method.

7. Conclusion

In this paper, we have proposed a deep hybrid method that integrates the discriminative and generative strategies into one framework. Assuming the survival function for each instance is a weighted combination of constant expert distributions, our method is capable of learning the weight for each expert distribution discriminatively and the distribution of the survival information generatively. We demonstrate our method's superiority by applying it to Alzheimer's disease subtype discovery. Genetic association studies along with feature importance analysis further validate the effectiveness of our proposed method.

CRediT authorship contribution statement

Bojian Hou: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Zixuan Wen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Investigation, Data curation. **Jingxuan Bao:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Data curation. **Richard Zhang:** Writing – original draft, Visualization, Software, Investigation. **Boning Tong:** Writing – original draft, Visualization, Software, Formal analysis. **Shu Yang:** Writing – review & editing, Validation. **Junhao Wen:** Writing – review & editing, Data curation. **Yuhan Cui:** Writing – review & editing, Data curation. **Jason H. Moore:** Writing – review & editing, Resources. **Andrew J. Saykin:** Writing – review & editing, Resources. **Heng Huang:** Writing – review & editing, Resources. **Paul M. Thompson:** Writing – review & editing, Resources. **Marylyn D. Ritchie:** Writing – review & editing, Resources. **Christos Davatzikos:** Writing – review & editing, Resources. **Alzheimer's Disease Neuroimaging Initiative:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is the extension of the conference version (Hou et al., 2023) and was supported in part by the National Institutes of Health, United States grants U01 AG068057, U01 AG066833, RF1 AG063481, R01 LM013463, R01 AG071470, R01 AG058854 and S10 OD023495, and the National Science Foundation, United States grant IIS 1837964. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, United States, the National Institute of Biomedical Imaging and Bioengineering, United States, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

In this appendix, we provide additional experimental details and results including the Kaplan–Meier plots for dataset AV45 and FDG Appendix A, the targeted genetic association analyses for all the methods on all the ADNI datasets Appendix B, the statistics regarding Age, Education and Sex Appendix C, the results on three clusters and different event rates Appendix D, and the impact of different threshold on the cluster results of Cox PH, Deep Cox and DSM Appendix E.

Appendix A. KM plot on AV45 and FDG

In Figs. A.10 and A.11 we provide the Kaplan–Meier (KM) plots that corresponds to Table 3. The smaller the LogRank, the closer the two survival curves. As can be seen, the two survival curves of KMeans and SuSTaIn are closer to each other compared to survival models. This shows that survival models can more effectively stratify people into different subgroups with respect to MCI converting to AD. Especially, the two curves of DCSM are farthest apart from each other. You may notice that the two curves of VaDeSC are farther to each other than ours, but the number of patients in cluster 1 is only 5 and 6 in Fig. A.10 and Fig. A.11 respectively, which means that the two clusters are very imbalanced. In addition, the cross mark on the curve indicates censoring and it indicates that we do not observe the event of MCI converting to AD happens. Thus it is with high probability that the censored samples have low risk to convert to AD. We can see that in the curve of Cluster 1 of DCSM, the number of cross marks is small while most of the censored samples are on the low risk data. This validates the effectiveness of DCSM as well.

Appendix B. Targeted genetic association analyses for other clustering methods

In Fig. B.12, we provide the targeted genetic association analysis results of all the methods. We expect to identify more significant SNPs in the first two columns of each block for each dataset. To clearly compare them, we sum up the number of "x" markers in the first two columns for each method. Based on this, KMeans has 24, Cox PH has 20, Deep Cox has 23, DSM has 22, SCA has 16, Vadesc has 0, NSC has 29 and our DCSM has 30, which demonstrates the superiority of the proposed method.

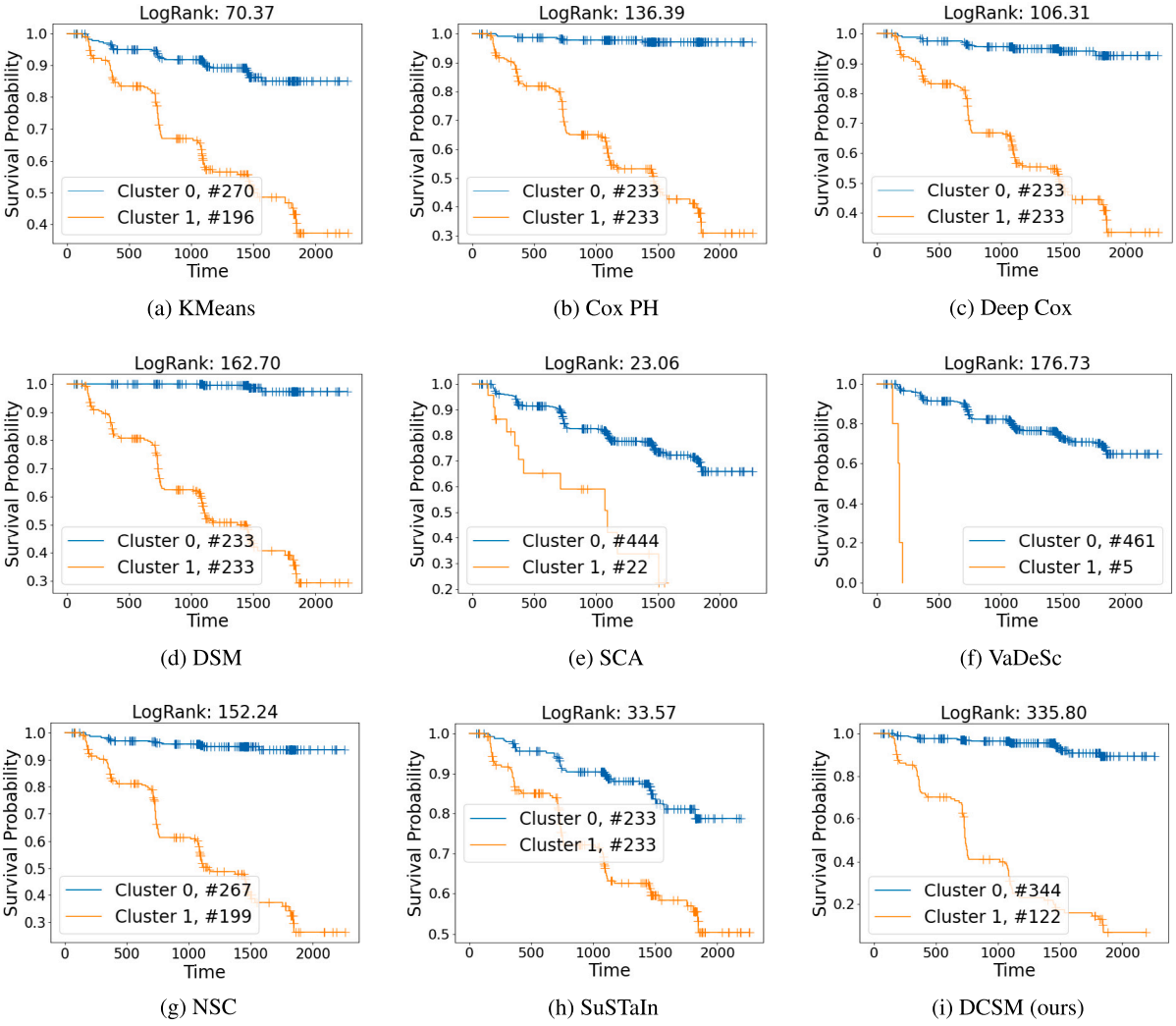


Fig. A.10. The Kaplan–Meier plots of KMeans, Cox PH, Deep Cox, DSM, SCA, Vadesc, NSC, SuSTaIn, and DCSM on dataset AV45 of one run. The cross mark on the curve means censoring. Cluster 0 means low-risk group while Cluster 1 means high-risk group.

Appendix C. Statistics regarding age, education and sex

We provide the statistics of high- and low-risk groups with respect to Age, Education and Gender and use histogram to illustrate the difference between low- and high-risk groups. Table C.6 summarizes the subject characteristics of high- and low- risk group clustered by all methods. Figs. C.13–C.15 shows the histogram of all methods. We find that the average age of subjects in high-risk group are higher than that in low-risk group. This is reasonable since older people are more risky to have cognition impairment. Besides, the cluster results indicates that more education time could reduce the risk to be diagnosed to be AD. Finally, for AV45 and FDG datasets, the ratio of high risk and low risk in female group are smaller than that in male group, while for VBM dataset, the ratio of high risk and low risk in female group are larger than that in male group.

Appendix D. C-index and LogRank results of three risk groups and different event rates

To further demonstrate the superiority of our model, we also add the experiment for $K = 3$, where the patients are clustered into three risk groups. The performance of LogRank results are presented in Table D.7 where we can see that our DCSM outperforms the seven baseline methods. Besides, we also present the C-Index results in Table D.8.

To investigate the benefits of incorporating censored data, we conducted extensive experiments with different event rates, ranging from 1/8 to 1, by varying the number of censored samples. A smaller event rate corresponds to a higher number of censored samples. When the event rate is 1, it indicates that no censored data is included in the dataset. For the time-to-event prediction task, the modified data with specific event rates were divided into a 70% training set and a 30% testing set. Fig. 8a shows that as the event rate increases (less censored data included), the C-Index performance worsens, highlighting the importance of censored data for accurate time-to-event prediction. For the clustering task, the training data consists of modified data with specific event rates, while the testing data comprises the original entire dataset to ensure a fair comparison across different cases. Fig. 8b shows that a higher event rate will cause a lower LogRank performance, demonstrating the benefit of incorporating censored data. The LogRank performance outperforms the others when the event rate is 2/8 (1/4) because the event rate of the original dataset is exactly 1/4. Maintaining the same distribution between the training and testing results in the best performance. The more detailed quantitative values are presented in Table D.9. In this table, to obtain the mean and standard deviation

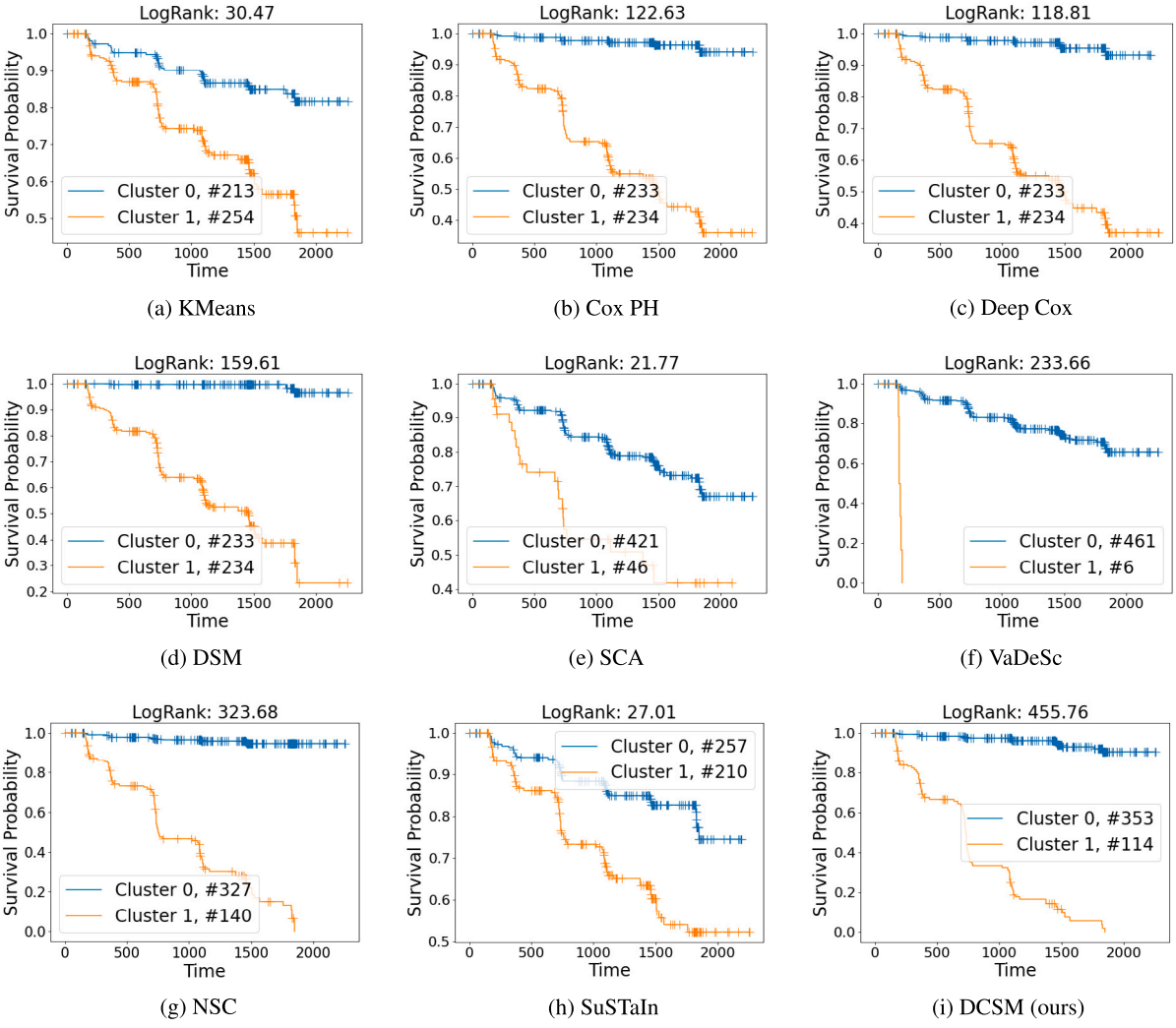
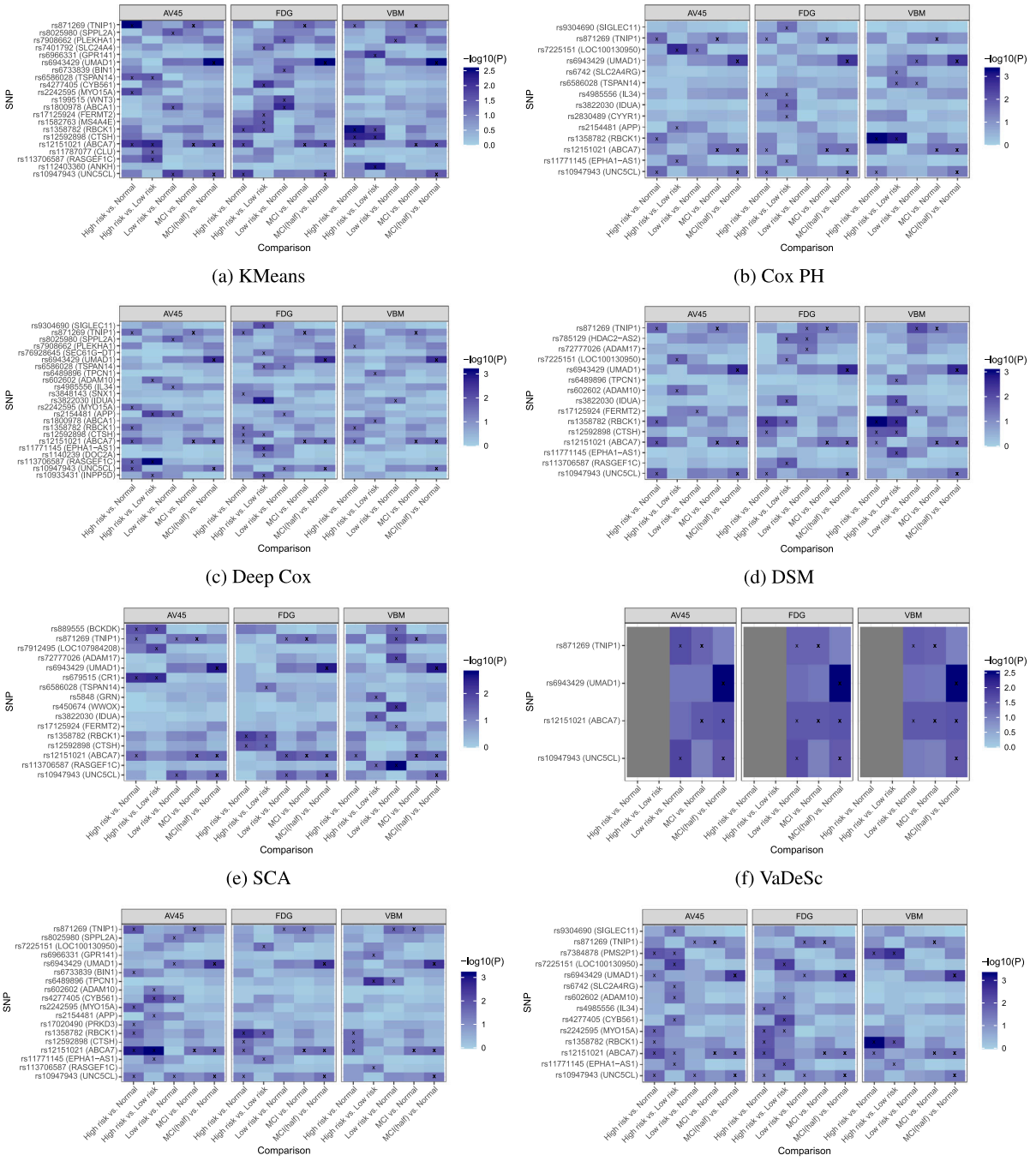


Fig. A.11. The Kaplan–Meier plots of KMeans, Cox PH, Deep Cox, DSM, SCA, Vadesc, NSC, SuSTaIn, and DCSM on dataset FDG of one run. The cross mark on the curve means censoring. Cluster 0 means low-risk group while Cluster 1 means high-risk group.

Table C.6
Subject characteristics of high- and low-risk group clustered by all methods.

Dataset	Method	Age		Education		Gender (M/F)	
	Risk Group	High risk	Low risk	High risk	Low risk	High risk	Low risk
AV45	KMeans	73.67 ± 6.78	70.82 ± 7.71	16.03 ± 2.84	16.28 ± 2.63	126/69	140/129
	Cox PH	73.03 ± 7.15	70.99 ± 7.64	16.06 ± 2.86	16.29 ± 2.57	140/93	126/105
	Deep Cox	73.32 ± 6.92	70.73 ± 7.76	16.04 ± 2.88	16.31 ± 2.55	136/95	130/103
	DSM	73.18 ± 7.15	70.85 ± 7.59	15.99 ± 2.81	16.36 ± 2.62	133/99	133/99
	SCA	73.56 ± 4.69	71.94 ± 7.57	15.59 ± 2.84	16.20 ± 2.71	13/9	253/189
	VaDeSC	73.96 ± 5.53	71.99 ± 7.48	16.80 ± 1.10	16.16 ± 2.74	4/1	262/197
	NSC	72.98 ± 7.42	71.31 ± 7.43	16.04 ± 2.82	16.27 ± 2.65	114/83	152/115
	DCSM	73.67 ± 6.99	71.43 ± 7.54	15.93 ± 2.73	16.26 ± 2.71	69/52	197/146
FDG	KMeans	74.67 ± 6.69	68.85 ± 7.09	15.86 ± 2.83	16.50 ± 2.55	154/99	111/101
	Cox PH	73.67 ± 7.25	70.35 ± 7.30	15.97 ± 2.89	16.33 ± 2.54	134/99	131/101
	Deep Cox	73.42 ± 7.22	70.60 ± 7.44	15.86 ± 2.86	16.44 ± 2.55	137/96	128/104
	DSM	72.82 ± 7.19	71.20 ± 7.64	16.16 ± 2.80	16.14 ± 2.65	141/92	124/108
	SCA	75.79 ± 7.04	71.60 ± 7.39	15.48 ± 2.76	16.22 ± 2.71	25/21	240/179
	VaDeSC	73.82 ± 7.87	71.99 ± 7.46	16.67 ± 1.03	16.14 ± 2.74	5/1	260/199
	NSC	71.51 ± 7.41	72.23 ± 7.48	16.31 ± 2.68	16.08 ± 2.74	78/61	187/139
	DCSM	73.40 ± 7.09	71.56 ± 7.52	15.99 ± 2.67	16.20 ± 2.74	62/51	203/149

(*continued on next page*)



(a) KMeans (b) Cox PH

(c) Deep Cox (d) DSM

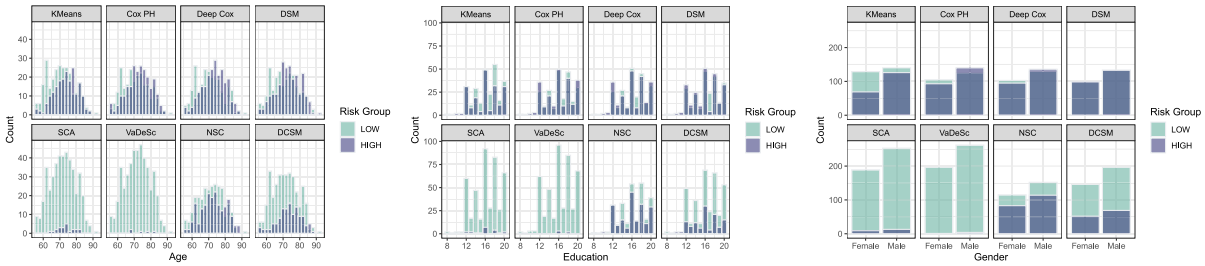
(e) SCA (f) VaDeSc

(g) NSC (h) DCSM

Fig. B.12. Targeted genetic association results of all the baseline methods and our DCSM. In the parentheses of the vertical label is the name of the gene closest to the corresponding SNP. The darkness of each blue patch represents the negative logarithm of *p-value* regarding each SNP. The darker the patch is, the more significant the SNP is. "x" marker on the patch means the SNP is statistically significant after FDR correction.

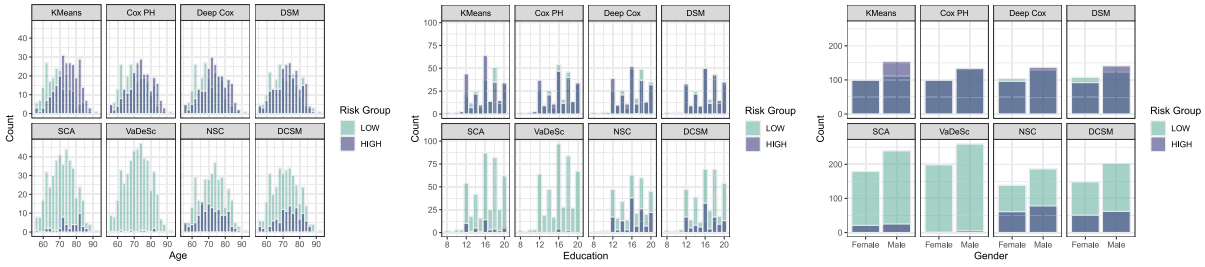
Table C.6 (*continued*).

Dataset	Method	Age		Education		Gender (M/F)	
	Risk Group	High risk	Low risk	High risk	Low risk	High risk	Low risk
VBM	KMeans	75.33 ± 6.60	69.01 ± 6.90	15.91 ± 2.85	16.37 ± 2.66	136/78	128/118
	Cox PH	73.46 ± 7.29	70.44 ± 7.32	16.19 ± 2.76	16.12 ± 2.77	138/92	126/104
	Deep Cox	73.46 ± 7.38	70.45 ± 7.23	16.20 ± 2.79	16.11 ± 2.74	131/99	133/97
	DSM	73.44 ± 7.40	70.47 ± 7.23	16.13 ± 2.84	16.18 ± 2.68	138/92	126/104
	SCA	73.25 ± 7.07	64.82 ± 5.15	16.12 ± 2.80	16.34 ± 2.53	242/147	22/49
	VaDeSC	74.45 ± 9.38	71.93 ± 7.45	17.50 ± 1.91	16.14 ± 2.76	1/3	263/193
	NSC	71.24 ± 7.02	72.41 ± 7.70	16.07 ± 2.70	16.21 ± 2.80	103/76	161/120
	DCSM	74.24 ± 6.70	71.17 ± 7.55	16.30 ± 2.63	16.10 ± 2.80	72/45	192/151



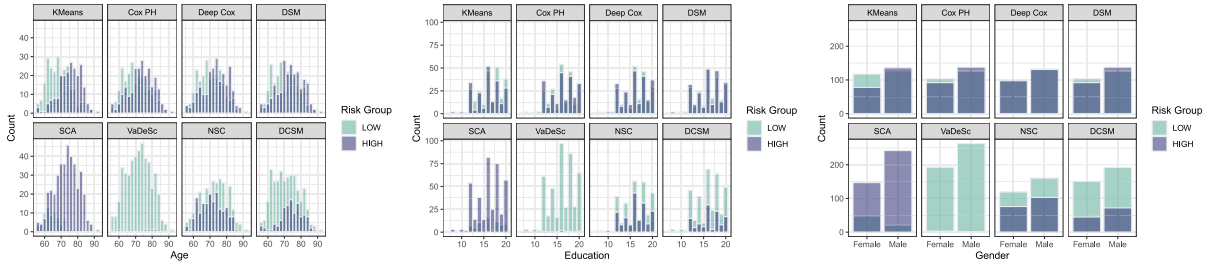
(a) Age (b) Education (c) Gender

Fig. C.13. Distribution of (a) age, (b) education and (c) gender of high and low risk group. Dataset is AV45.



(a) Age (b) Education (c) Gender

Fig. C.14. Distribution of (a) age, (b) education and (c) gender of high and low risk group. Dataset is FDG.



(a) Age (b) Education (c) Gender

Fig. C.15. Distribution of (a) age, (b) education and (c) gender of high and low risk group. Dataset is VBM.

Table D.7
LogRank comparison for three risk groups. The best one in each block is in bold.

Methods	AV45	FDG	VBM
KMeans	64.33 ± 2.64	30.31 ± 1.58	15.03 ± 1.08
Cox PH	272.35 ± 6.26	211.68 ± 4.61	176.56 ± 9.55
Deep Cox	225.99 ± 10.51	273.15 ± 30.74	232.60 ± 8.97
DSM	320.61 ± 10.87	296.99 ± 14.27	257.73 ± 15.08
SCA	45.04 ± 28.08	22.25 ± 10.50	10.84 ± 5.53
VaDeSC	196.54 ± 142.24	158.62 ± 133.80	237.66 ± 182.73
NSC	209.17 ± 42.35	313.82 ± 79.63	241.32 ± 121.48
SuSTaIn	41.25 ± 5.09	23.36 ± 8.05	19.76 ± 9.10
DCSM	**337.21 ± 14.18**	**371.52 ± 29.06**	**298.90 ± 10.32**

Table D.8
C-Index comparison for three risk groups. The best one in each block is in bold.

Methods	AV45	FDG	VBM
Cox PH	0.6958 ± 0.0259	0.6473 ± 0.0698	0.6311 ± 0.0327
Deep Cox	0.5775 ± 0.2511	**0.7531 ± 0.0276**	0.6703 ± 0.0257
DSM	**0.7637 ± 0.0247**	0.7429 ± 0.0295	0.6704 ± 0.0198
SCA	0.5208 ± 0.0801	0.5948 ± 0.1001	0.5382 ± 0.1011
VaDeSC	0.4187 ± 0.0535	0.4109 ± 0.0883	0.5149 ± 0.0641
NSC	0.7357 ± 0.0266	0.7535 ± 0.0360	0.6529 ± 0.0222
DCSM	0.7617 ± 0.0182	0.7426 ± 0.0250	**0.7050 ± 0.0236**

(std) values of C-Index, we use five different seeds to split the data and take the mean and std values over the five sets of results. For the LogRank results, we use bootstrap to get 95% of the entire data as the testing set for five times and get the mean and std values.

Appendix E. Different threshold of cluster results of Cox PH, Deep Cox and DSM

To cluster MCI patients, we first obtain their predicted risk scores, and then we set a threshold (usually the *median* or the *mean* of the whole risk scores) to get the subtypes. The patients whose predicted risk scores lower than the threshold are clustered as low risk group, while the others are clustered as high risk group. We compared the performance of LogRank using both median and mean as thresholds for the Cox PH, Deep Cox, and DSM models, as presented in Table E.10. The results show that using the median as a threshold produces significantly better outcomes for these baselines. This presents a challenge for our method, but it still outperforms the improved versions of the baselines. The reason *mode* was not used as the threshold is that the mode can appear anywhere in an ordered sequence only if that number repeats more than other numbers, and thus may not provide a reasonable stratification.

Table D.9
C-Index and LogRank results of three dataset based on different event rate of our DCSM model.

Performance	C-Index			LogRank		
Event rate	AV45	FDG	VBM	AV45	FDG	VBM
1/8	0.8317 ± 0.0316	0.8062 ± 0.0334	0.7466 ± 0.0401	196.17 ± 35.23	195.26 ± 11.61	163.57 ± 3.67
2/8	0.7878 ± 0.0396	0.7690 ± 0.0370	0.6820 ± 0.0382	298.90 ± 36.76	390.80 ± 21.27	321.02 ± 18.48
3/8	0.7225 ± 0.0087	0.7326 ± 0.0243	0.6989 ± 0.0308	262.49 ± 14.12	235.33 ± 8.60	186.62 ± 9.57
4/8	0.7017 ± 0.0332	0.6880 ± 0.0412	0.6623 ± 0.0331	106.00 ± 4.01	127.82 ± 5.65	97.44 ± 5.85
5/8	0.6726 ± 0.0531	0.6350 ± 0.0617	0.6462 ± 0.0570	116.53 ± 2.81	70.39 ± 6.58	72.68 ± 4.60
6/8	0.6859 ± 0.0353	0.6412 ± 0.0443	0.5791 ± 0.0469	108.67 ± 4.21	58.05 ± 2.44	55.30 ± 1.03
7/8	0.6433 ± 0.0297	0.6084 ± 0.0688	0.5788 ± 0.0535	65.94 ± 4.22	27.12 ± 1.21	15.94 ± 1.03
8/8	0.5201 ± 0.0352	0.5771 ± 0.0783	0.5267 ± 0.0418	70.61 ± 4.88	38.64 ± 3.62s	2.02 ± 0.61

Table E.10
LogRank comparison of Cox PH, Deep Cox, and DSM between using median and mean as thresholds.

Cluster threshold		Median (Ours)	Mean
AV45	Cox PH	133.60 ± 3.65	37.89 ± 0.57
	Deep Cox	121.49 ± 10.99	37.93 ± 5.84
	DSM	160.62 ± 3.79	41.35 ± 4.09
FDG	Cox PH	117.80 ± 1.88	35.99 ± 1.15
	Deep Cox	95.39 ± 16.07	21.42 ± 3.46
	DSM	124.26 ± 1.74	46.55 ± 2.05
VBM	Cox PH	89.07 ± 3.97	40.96 ± 0.52
	Deep Cox	63.49 ± 2.76	17.23 ± 3.98
	DSM	120.41 ± 2.47	37.51 ± 1.42

References

1000 Genomes Project Consortium, et al., 2015. A global reference for human genetic variation. Nature 526 (7571), 68.

Akram, S., Ann, Q.U., 2015. Newton raphson method. Int. J. Sci. Eng. Res. 6 (7), 1748–1752.

Aksman, L.M., Wijeratne, P.A., Oxtoby, N.P., Eshaghi, A., Shand, C., Altmann, A., Alexander, D.C., Young, A.L., 2021. pySuStaIn: a python implementation of the subtype and stage inference algorithm. SoftwareX 16, 100811.

Alashwal, H., El Halaby, M., Crouse, J.J., Abdalla, A., Moustafa, A.A., 2019. The application of unsupervised clustering methods to Alzheimer's disease. Front. Comput. Neurosci. 13, 31.

Ali, D.G., Bahrani, A.A., Barber, J.M., El Khouli, R.H., Gold, B.T., Harp, J.P., Jiang, Y., Wilcock, D.M., Jicha, G.A., 2022. Amyloid-PET levels in the precuneus and posterior cingulate cortices are associated with executive function scores in preclinical Alzheimer's disease prior to overt global amyloid positivity. J. Alzheimer's Dis. 88 (3), 1127–1135.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. Neuroimage 11 (6), 805–821.

Bailly, M., Destrieux, C., Hommet, C., Mondon, K., Cottier, J.P., Beaufils, E., Vierron, E., Vercouillie, J., Ibazizene, M., Voisin, T., et al., 2015. Precuneus and cingulate cortex atrophy and hypometabolism in patients with Alzheimer's disease and mild cognitive impairment: MRI and 18 F-FDG PET quantitative analysis using freesurfer. BioMed Res. Int. 2015.

Bland, J.M., Altman, D.G., 1998. Survival probabilities (the Kaplan-Meier method). Bmj 317 (7172), 1572–1580.

Chang, Y.T., Huang, C.W., Chen, N.C., Lin, K.J., Huang, S.H., Chang, W.N., Hsu, S.W., Hsu, C.W., Chen, H.H., Chang, C.C., 2016. Hippocampal amyloid burden with downstream fusiform gyrus atrophy correlate with face matching task scores in early stage Alzheimer's disease. Front. Aging Neurosci. 8, 145.

Chang, C., et al., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4.

Chapfuwa, P., Li, C., Mehta, N., Carin, L., Henao, R., 2020. Survival cluster analysis. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 60–68.

Collij, L.E., Heeman, F., Salvadó, G., Ingala, S., Altomare, D., de Wilde, A., Konijnenberg, E., van Buchem, M., Yaqub, M., Markiewicz, P., et al., 2020. Multitracer model for staging cortical amyloid deposition using PET imaging. Neurology 95 (11), e1538–e1553.

Cox, D.R., 1972. Regression models and life-tables. J. R. Stat. Soc. Ser. B Stat. Methodol. 34 (2), 187–202.

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., Schlessinger, D., et al., 2016. Next-generation genotype imputation service and methods. Nat. Genet. 48 (10), 1284–1287.

Dawber, T.R., Meadors, G.F., Moore, Jr., F.E., 1951. Epidemiological approaches to heart disease: the Framingham study. Am. J. Public Health Nations Health 41 (3), 279–286.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, Vol. 96, No. 34. pp. 226–231.

Feng, Y., Kim, M., Yao, X., Liu, K., Long, Q., Shen, L., Alzheimer's Disease Neuroimaging, Initiative, 2022. Deep multiview learning to identify imaging-driven subtypes in mild cognitive impairment. BMC Bioinform. 23 (Suppl 3), 402.

Fjell, A.M., McEvoy, L., Holland, D., Dale, A.M., Walhovd, K.B., Alzheimer's Disease Neuroimaging Initiative, et al., 2014. What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus. Progress Neurobiol. 117, 20–40.

Fleming, T.R., Harrington, D.P., 2013. Counting Processes and Survival Analysis, vol. 625, John Wiley & Sons.

Flynn, R., 2012. Survival analysis. J. Clin. Nurs. 21 (19pt20), 2789–2797.

Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., et al., 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage 60 (3), 1880–1889.

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A., 1982. Evaluating the yield of medical tests. Jama 247 (18), 2543–2546.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. J. R. Stat. Soc. Ser. C (Appl. Stat.) 28 (1), 100–108.

Hou, B., Li, H., Jiao, Z., Zhou, Z., Zheng, H., Fan, Y., 2023. Deep clustering survival machines with interpretable expert distributions. In: 2023 IEEE 20th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–4.

Hou, B.J., Zhang, L., Zhou, Z.H., 2017. Storage fit learning with unlabeled data. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 1844–1850.

Hou, B.J., Zhou, Z.H., 2018. Learning with interpretable structure from RNN. arXiv preprint arXiv:1810.10708.

Hou, B.J., Zhou, Z.H., 2020. Learning with interpretable structure from gated RNN. IEEE Trans. Neural Netw. Learn. Syst. 31 (7), 2267–2279.

Huang, K.L., Lin, K.J., Hsiao, I.T., Kuo, H.C., Hsu, W.C., Chuang, W.L., Kung, M.P., Wey, S.P., Hsieh, C.J., Wai, Y.Y., et al., 2013. Regional amyloid deposition in amnestic mild cognitive impairment and Alzheimer's disease evaluated by [18F] AV-45 positron emission tomography in Chinese population. PLoS One 8 (3), e58974.

Jack, Jr., C.R., Barnes, J., Bernstein, M.A., Borowski, B.J., Brewer, J., Clegg, S., Dale, A.M., Carmichael, O., Ching, C., DeCarli, C., Desikan, R.S., Fennema-Notestine, C., Fjell, A.M., Fletcher, E., Fox, N.C., Gunter, J., Gutman, B.A., Holland, D., Hua, X., Insel, P., Kantarci, K., Killiany, R.J., Krueger, G., Leung, K.K., Mackin, S., Maillard, P., Malone, I.B., Mattsson, N., McEvoy, L., Modat, M., Mueller, S., Nosheny, R., Ourselin, S., Schuff, N., Senjem, M.L., Simonson, A., Thompson, P.M., Rettmann, D., Vemuri, P., Walhovd, K., Zhao, Y., Zuk, S., Weiner, M., 2015. Magnetic resonance imaging in Alzheimer's disease neuroimaging initiative 2. Alzheimers Dement 11 (7), 740–756.

Jagust, W.J., Bandy, D., Chen, K., Foster, N.L., Landau, S.M., Mathis, C.A., Price, J.C., Reiman, E.M., Skovronsky, D., Koeppe, R.A., et al., 2010. The Alzheimer's disease neuroimaging initiative positron emission tomography core. Alzheimer's Dementia 6 (3), 221–229.

Jagust, W.J., Landau, S.M., Koeppe, R.A., Reiman, E.M., Chen, K., Mathis, C.A., Price, J.C., Foster, N.L., Wang, A.Y., 2015. The Alzheimer's disease neuroimaging initiative 2 PET core: 2015. Alzheimer's & Dementia 11 (7), 757–771.

Jeanselme, V., Tom, B., Barrett, J., 2022. Neural survival clustering: Non-parametric mixture of neural networks for survival clustering. In: Conference on Health, Inference, and Learning. PMLR, pp. 92–102.

Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y., 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med. Res. Methodol. 18 (1), 1–12.

Klein, J.P., 1991. Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators. Scand. J. Stat. 333–340.

Knaus, W.A., Harrell, F.E., Lynn, J., Goldman, L., Phillips, R.S., Connors, A.F., Dawson, N.V., Fulkerson, W.J., Califf, R.M., Desbiens, N., et al., 1995. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. Ann. Intern. Med. 122 (3), 191–203.

Kvamme, H., Borgan, Ø., Scheel, I., 2019. Time-to-event prediction with neural networks and Cox regression. arXiv preprint arXiv:1907.00825.

Kyle, R.A., Therneau, T.M., Rajkumar, S.V., Larson, D.R., Plevak, M.F., Offord, J.R., Dispenzieri, A., Katzmann, J.A., Melton, III, L.J., 2006. Prevalence of monoclonal gammopathy of undetermined significance. N. Engl. J. Med. 354 (13), 1362–1369.

Manduchi, L., Marcinkevičs, R., Massi, M.C., Weikert, T., Sauter, A., Gotta, V., Müller, T., Vasella, F., Neidert, M.C., Pfister, M., et al., 2021. A deep variational approach to clustering survival data. arXiv preprint arXiv:2106.05763.

Mantel, N., et al., 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother. Rep. 50 (3), 163–170.

Mosconi, L., 2005. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease: FDG-PET studies in MCI and AD. Eur. J. Nucl. Med. Mol. Imaging 32, 486–510.

Nagpal, C., Li, X., Dubrawski, A., 2021. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. IEEE J. Biomed. Health Inf. 25 (8), 3163–3175.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32. pp. 8024–8035.

Petrovic, P., Kalisch, R., Singer, T., Dolan, R.J., 2008. Oxytocin attenuates affective evaluations of conditioned faces and amygdala activity. J. Neurosci. 28 (26), 6607–6615.

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C., Alzheimer's Disease Neuroimaging Initiative, et al., 2011. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. Psychiatry Res. Neuroimaging 194 (1), 7–13.

Reynolds, D.A., et al., 2009. Gaussian mixture models. Encycl. Biom. 741 (659–663).

Saykin, A.J., Shen, L., Yao, X., Kim, S., Nho, K., Risacher, S.L., Ramanan, V.K., Foroud, T.M., Faber, K.M., Sarwar, N., Munsie, L.M., Hu, X., Soares, H.D., Potkin, S.G., Thompson, P.M., Kauwe, J.S., Kaddurah-Daouk, R., Green, R.C., Toga, A.W., Weiner, M.W., Alzheimer's Disease Neuroimaging, Initiative, 2015. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. Alzheimers Dement 11 (7), 792–814.

Shen, L., Thompson, P.M., 2020. Brain imaging genomics: Integrated analysis and machine learning. Proc. IEEE 108 (1), 125–162.

Shen, L., Thompson, P.M., Potkin, S.G., Bertram, L., Farrer, L.A., Foroud, T.M., Green, R.C., Hu, X., Huentelman, M.J., Kim, S., Kauwe, J.S., Li, Q., Liu, E., Macciardi, F., Moore, J.H., Munsie, L., Nho, K., Ramanan, V.K., Risacher, S.L., Stone, D.J., Swaminathan, S., Toga, A.W., Weiner, M.W., Saykin, A.J., Alzheimer's Disease Neuroimaging, Initiative, 2014. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. Brain Imaging Behav. 8 (2), 183–207.

Tarone, R.E., 1975. Tests for trend in life table analysis. Biometrika 62 (3), 679–690.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage (ISSN: 1053-8119) 15 (1), 273–289. http://dx.doi.org/10.1006/nimg.2001.0978, URL https://www.sciencedirect.com/science/article/pii/S1053811901909784.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38 (16), e164.

Weiner, M.W., et al., 2017. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. Alzheimer's Dementia 13 (4), e1–e85.

Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., et al., 2018. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. Nature Commun. 9 (1), 4273.

Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C., 2014. A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain 137 (9), 2564–2577.

Zhou, Z.H., 2018. A brief introduction to weakly supervised learning. Natl. Sci. Rev. 5 (1), 44–53.

Zhou, Z., Ataee Tarzanagh, D., Hou, B., Tong, B., Xu, J., Feng, Y., Long, Q., Shen, L., 2024. Fair canonical correlation analysis. Adv. Neural Inf. Process. Syst. 36.