Knowledge-guided Deep Temporal Clustering for Alzheimer's Disease Subtypes in Completed Clinical Trials

Dulin Wang, M.S.[1], Xiaotian Ma, M.S.[1], Paul E. Schulz, M.D.[2], Xiaoqian Jiang, Ph.D.[1], Yejin Kim, Ph.D.[1]
[1]McWilliams School of Biomedical Informatics, The University of Texas Health Center at Houston, Houston, TX, U.S.
[2] Department of Neurology, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, USA.

Abstract:

Alzheimer's disease (AD) is a multifaceted neurodegenerative disorder with varied patient progression. We aim to test the hypothesis that AD patients can be categorized into subgroups based on differences in progression. We leveraged data from three randomized clinical trials (RCTs) to develop a knowledge-guided, deep temporal clustering (KG-DTC) framework for AD subtyping. This model combined autoencoders for contextual information capture, k-means clustering for representation formation, and clinical outcome classification for clinical knowledge integration. The derived representations, encompassing demographics, APOE genotype, cognitive assessments, brain volumes, and biomarkers, were clustered using the Gaussian Mixture Model to identify AD subtypes. Our novel KG-DTC framework was developed using placebo data from 2,087 AD patients across three solanezumab clinical trials (EXPEDITION, EXPEDITION2, and EXPEDITION3), achieving high performance in outcome prediction and clustering. The KG-DTC model demonstrated superior clustering structures, especially when combined with k-means clustering loss. External validation with independent clinical trial data showed consistent clustering results, with a 0.33 silhouette score for three clusters. The model's stability was confirmed through a leave-one-out approach, with an average adjusted Rand Index around 0.945. Three distinct AD subtypes were identified, each exhibiting unique patterns of cognitive function, neurodegeneration, and amyloid beta levels. Notably, Subtype 3 (S3) showed rapid cognitive decline across multiple clinical measures (e.g., 0.64 in S1 vs. -1.06 in S2 vs. 15.09 in S3 of average ADAS total change score, $p<.001$). This innovative approach offers promising insights for understanding variability in treatment outcomes and personalizing AD treatment strategies.

Introduction

Alzheimer's disease (AD) is incurable and challenging to diagnose and treat due to its complexity, heterogeneity, and multifactor nature. Its progression patterns manifest through a broad spectrum of longitudinally linked clinical features and outcomes that vary across AD patients. Variability in AD has impeded many clinical trials for drug development[1,2]. Therefore, precision medicine for AD aims to understand why individuals respond differently to treatments, and temporal subtyping has become a crucial tool in identifying patient subgroups to answer

such questions. By transforming the raw data along disease progression into clinically relevant and interpretable information, temporal subtyping minimizes progression heterogeneity among individuals[3,4], guides clinicians to tailor treatment to AD subgroups, and ultimately develops successful AD subtype-specific drugs.

Recent studies have used electronic health records (EHRs) and Alzheimer's Disease Neuroimaging Initiative (ADNI) databases extensively to derive systematic and comprehensive AD subtypes[5–8]. However, these datasets are sparse or noisy. Randomized clinical trials (RCTs) are another rich but understudied multimodal data source. The placebo groups from RCTs allow us to investigate the disease progression without being confounded by the exposure to experimental drugs. A pooled group of placebo-treated patients from RCTs can increase the power and diversity of AD populations. RCT databases were used to identify phenotypes or subtypes in sepsis and respiratory disease[9,10]. Considering the successful utilization of RCT data in various diseases, using multimodal data from RCT to derive the AD subtypes is promising.

Temporal subtyping of AD is a data-driven, unsupervised learning task to group patients into similar disease progression patterns. Clustering is a widely used to discover subtype. Recent advances in deep representation learning overcome the limitations of traditional clustering methods that are hard to handle high-dimensional and multimodal data[6,11]. The deep learning-based clustering learns low-dimensional representation for multivariate longitudinal observations, which can be used in downstream tasks. However, the separate step learning approach has certain limitations. Firstly, it learns a low-dimensional representation for multivariate longitudinal observations, but this is not directly learned for clustering, which can hinder the overall clustering performance. Secondly, this method identifies temporal subtypes in an entirely unsupervised manner, ignoring any existing information about patients' clinical outcomes. This information, such as clinical trial outcomes, is crucial for understanding the progression of the disease and predicting future clinical outcomes. Thus, such an approach may not fully utilize all available resources for optimal results.

This study proposed knowledge-guided, deep temporal clustering, which is a unified framework to identify AD subtypes. We first developed a knowledge-guided deep clustering architecture to derive clustering-friendly vector-based representations. This architecture combined (1) temporal autoencoders (AEs) to capture contextual information and generate representations, (2) *k*-means clustering to encourage the representation to form clusters, and (3) clinical outcome classification to reflect clinical knowledge. Second, our model could generate informative representation through qualitative and quantitative analysis to derive meaningful subtypes. To this end, we applied the proposed framework to construct the representation learned from pooled RCTs with multimodal data, namely demographics, cognitive assessments, brain region volumes, biomarkers including amyloid-beta ($A\beta$), and genomic data on the apolipoprotein E (APOE) gene. We used the learned representation to cluster patients through the Gaussian mixture model (GMM) to derive AD subtypes and characterize the subtypes of AD to interpret potential disease progression patterns. From our extensive validation on the subtypes via reproducibility test, stability test, and significance test, we found that the subtypes were reproducible, stable, and clinically meaningful.

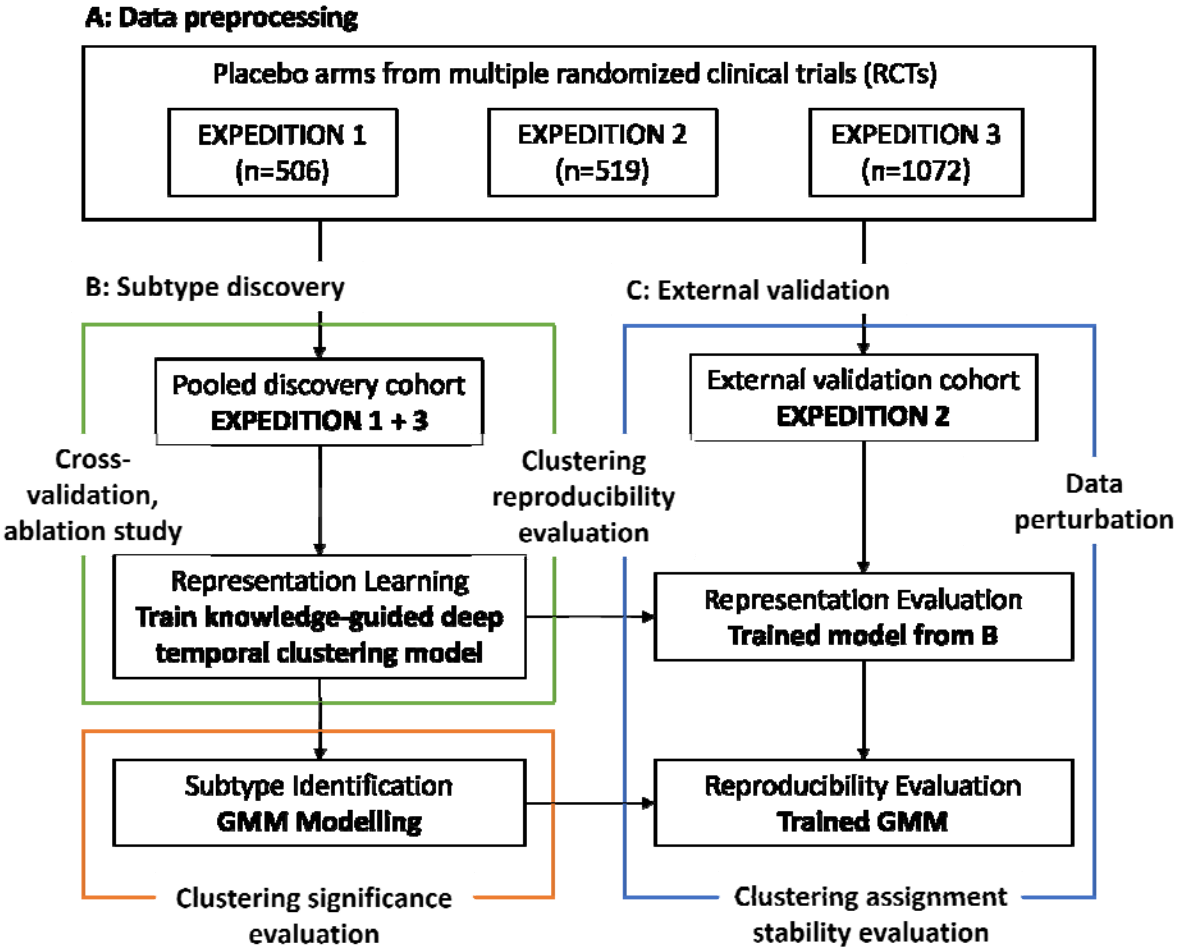


Fig. 1 Overall framework. Our framework to identify the subtypes of AD progression has three stages: 1) data preprocessing, 2) subtype discovery, and 3) external validation. After data preprocessing and pooling, we developed knowledge-guided deep temporal clustering (KG-DTC) methods and applied them to pooled clinical trial data. The KG-DTC method identified subtypes of AD patients using their longitudinal observations on efficacy measures. We internally validated the model using cross-validation and ablation study. We then externally validated the clustering result's reproducibility and stability with independent clinical trial data. After thoroughly evaluating the clustering model, we investigated the characteristics of clusters (i.e., subtypes) and common patterns within the clusters. This figure was adapted from Dinga.[12]

Results

Data summary

We developed our model using 2,087 AD patients from three RCT placebo arms (505 in EXPEDITION1[13], 518 from EXPEDITION2[13], and 1,064 from EXPEDITION3[14]). For each patient, we included visits with assessments conducted or biomarkers collected. As a result, we

selected visits 1, 2, 3, 4, 5, 6, 10, 13, 16, 19, and 23.0 for EXPEDITION 1 and 2, and visits 1, 2, 3, 5, 9, 12, 15, 18, and 22 for EXPEDITION 3. The total number of visits was 13,946. We set aside EXPEDITION2 as an external validation set and used EXPEDITION1 and EXPEDITION3 as a discovery set. The data comprises eight demographics, 28 baseline efficacy measurements, and 28 time-variant variables observed from baseline to the end of clinical trials. The efficacy measures include cognitive assessments, imaging biomarkers, fluid biomarkers, quality-of-life assessments, and a neuropsychiatric assessment (Supplementary Table S1. Demographics and longitudinal features summary.).

Knowledge-guided deep temporal clustering summary

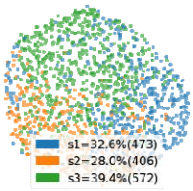
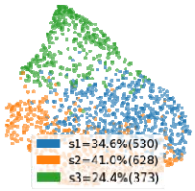
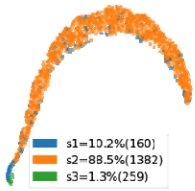
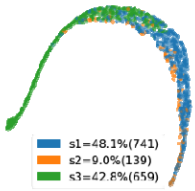
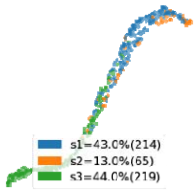
We developed a novel and unified framework (Fig. 1) for knowledge-guided deep temporal clustering (KG-DTC, Fig. 5) to identify subtypes from temporal multimodal data (Method: Subtype discovery: Knowledge-Guided Deep Temporal Clustering (KG-DTC) model).

Model cross-validation

We first evaluated the performance of KG-DTC in a cross-validation manner. We evaluated whether our model could learn a representation that can predict clinical outcomes when embedding the knowledge guidance. We also evaluated our model's clustering performance by silhouette scores and UMAP (uniform manifold approximation and projection for dimension reduction) visualization (Method: Model evaluation). For comparison, we ablated each component in KG-DTC (i.e., *k*-means clustering, knowledge guidance) and evaluated the three criterions above. Overall, we found that KG-DTC achieved high performance in both outcome prediction and clustering (Table 1). KG-DTC had *R*-squared (R^2) scores of 0.84 and 0.31 for two selected clinical outcome variables (i.e., ADAS and CDR) and a silhouette score of 0.26 with three clusters. The UMAP plot showed that three clusters separate the patients well. Other ablated models (M1, M2, and M3) failed to achieve a balance between outcome prediction and clustering. The UMAP showed that the clusters are not compact or not evenly separated.

Table 1 Evaluation of representation and clustering with ablated models, and reproducibility of clustering

Dataset	Pooled discovery data (EXPEDITION1+EXPEDITION3)				External data (EXPEDITION2)
Model	**M1:** Seq2seq reconstruction	**M2:** Seq2seq reconstruction + knowledge guidance	**M3:** Seq2seq reconstruction + clustering loss	**Ours:** Seq2seq clustering loss + knowledge guidance	**Trained DG-KDC**
R^2					
ADAS	-0.40	0.91	-0.41	0.84	0.84
CDR	-0.45	0.85	-0.15	0.31	0.32
Silhouette	0. 03 (3)	0.18 (3)	0. 37 (3)	0.26 (3)	0.33(3)

UMAP Clustering structure					

Silhouette: Highest silhouette scores (number of clusters). We use scores to examine outcome regression fit to evaluate representation quality; we use the highest silhouette scores with the number of clusters for clustering performance evaluation; we use UMAP visualization for clustering structure evaluation. We evaluated the model's representation, clustering, and effectiveness of components with three ablated models: 1) M1: simple model with only seq2seq reconstruction structure, 2) M2: model with seq2seq reconstruction and knowledge guidance, and 3) M3: model with seq2seq reconstruction and clustering loss.

We then investigated the contribution of each model component. We observed that joint optimization of *k*-means clustering significantly improved silhouette scores and made compact clustering structures in UMAP visualization. As shown in Table 1. UMAP clustering structure, it was evident that the representations learned by M3 and KG-DTC with *k*-means clustering loss have formed clearly separated and compressed clusters. In contrast, the representations learned by the ablated models M1 and M2 without *k*-means clustering loss fail to form distinct cluster shapes (highest silhouette scores of 0.03 and 0.18 at cluster counts 3, respectively), and the samples are scattered with a great amount of mixing. In addition, we observed that knowledge guidance helps preserve rich clinical context in representation. Knowledge guidance increased the outcome prediction accuracy (M1's of -0.40 and -0.45 vs. M3's of -0.41 and -0.15), and it also prevented the accuracy from extremely decreasing when it was combined with *k*-means clustering loss (M2's of 0.91 and 0.85 vs KG-DTC's of 0.84 and 0.31). These advances are also visualized in the UMAP plots. The representation learned from the ablated model M3 without knowledge guidance has formed separated clusters. Still, the samples are gathered to form one giant cluster, failing to generate informative and well-separated clusters.

Clustering reproducibility evaluation

After cross-validating our clustering model's performance, we investigated the reproducibility of the clusters on external data. We applied the KG-DTC models trained with the discovery data (EXPEDITION1 + EXPEDITION3) to the external data (EXPEDITION2) and compared the clusters from each set. As a result, we identified similar clustering results from the external set (Table 1). The proportions of clusters were 48.1, 9.0, and 42.8% in discovery data and 43.0, 13.0 and 44.0% in external data. In addition, we found that applying the trained KG-DTC to the external set achieved the outcome classification accuracy (the of 0.84 and 0.32) and 0.33 silhouette scores of 3 clusters, which is similar to the results with the discovery set. In the UMAP plots, clustering results from the external set also achieved a compact and distinct shape. Our models show great representation quality and clustering reproducibility in the external set.

Clustering stability evaluation

Now that we verified that the ORDTCR could reproduce similar clustering results with external data, we investigated whether the ORDTCR is stable enough to generate similar clustering results consistently. We evaluated our clustering model's assignment stability using the leave-one-out (LOO) approach[5] (Method: Clustering evaluation). We set leave-out sample size $n \in \{1:50\}$. Supplementary Fig. 1 showed that the average adjusted Rand Index (ARI) slightly fluctuated around 0.945 over different leave-out sample sizes. This finding demonstrated that our clustering model is stable, and the identified clusters are not statistical artifacts. Due to the stochastic nature of optimization, traditional clustering methods may produce inconsistent clustering results if the sample similarity is not well defined. The KG-DTC model projected the complicated multimodal longitudinal observation into a representation space where the sample similarity preserves the clinically meaningful patterns.

Clustering interpretation and its statistical significance

After thoroughly validating our proposed clustering models, we investigated common patterns of patients within the clusters. To illustrate the within-cluster distribution of cognitive scores, brain regional volumes, and amyloid beta deposition, we plotted violin plots to visualize data distribution across three clusters (Fig. 3 (A)). In addition, the main drawback of neural-network-based clustering is its inability to explain how static data and longitudinal sequences map to the latent dimensions. To address this and obtain interpretable patterns, we calculated the feature importance scores to determine the important baseline and longitudinal variables for cluster assignment. We built a cluster assignment (i.e., S1, S2, and S3) classification model and adapted it to feature permutation attribution algorithm1 (Method: Clustering evaluation). The AUROC for cluster assignment classification tasks were 0.90, 0.94, and 0.96. Using this approach, we calculated the importance scores of individual variables (Fig. 3 (B)).

To provide more detailed characterizations of each subtype, we statistically tested the important individual variables to determine whether they were distributed differently across clusters. As a result, we identified important features that distinguish the cluster assignment and generated profiles for each cluster. The profiles of clusters (or AD subtypes) have statistically distinctive characteristics. Also, the clusters were closely associated with the primary clinical outcomes of the trials. We identified three AD subtypes (Subtype 1 (S1), Subtype 2 (S2), and Subtype 3 (S3)) with distinct patterns in impaired cognitive function (C), neurodegeneration (N), and amyloid beta (A) (Table 2), as well as clinical outcomes in the trials (Fig. 4).

Table 2 Characteristics of Subtypes at Baseline and Endpoint Changes (The cells represent mean (median, std) unless specific illustration)

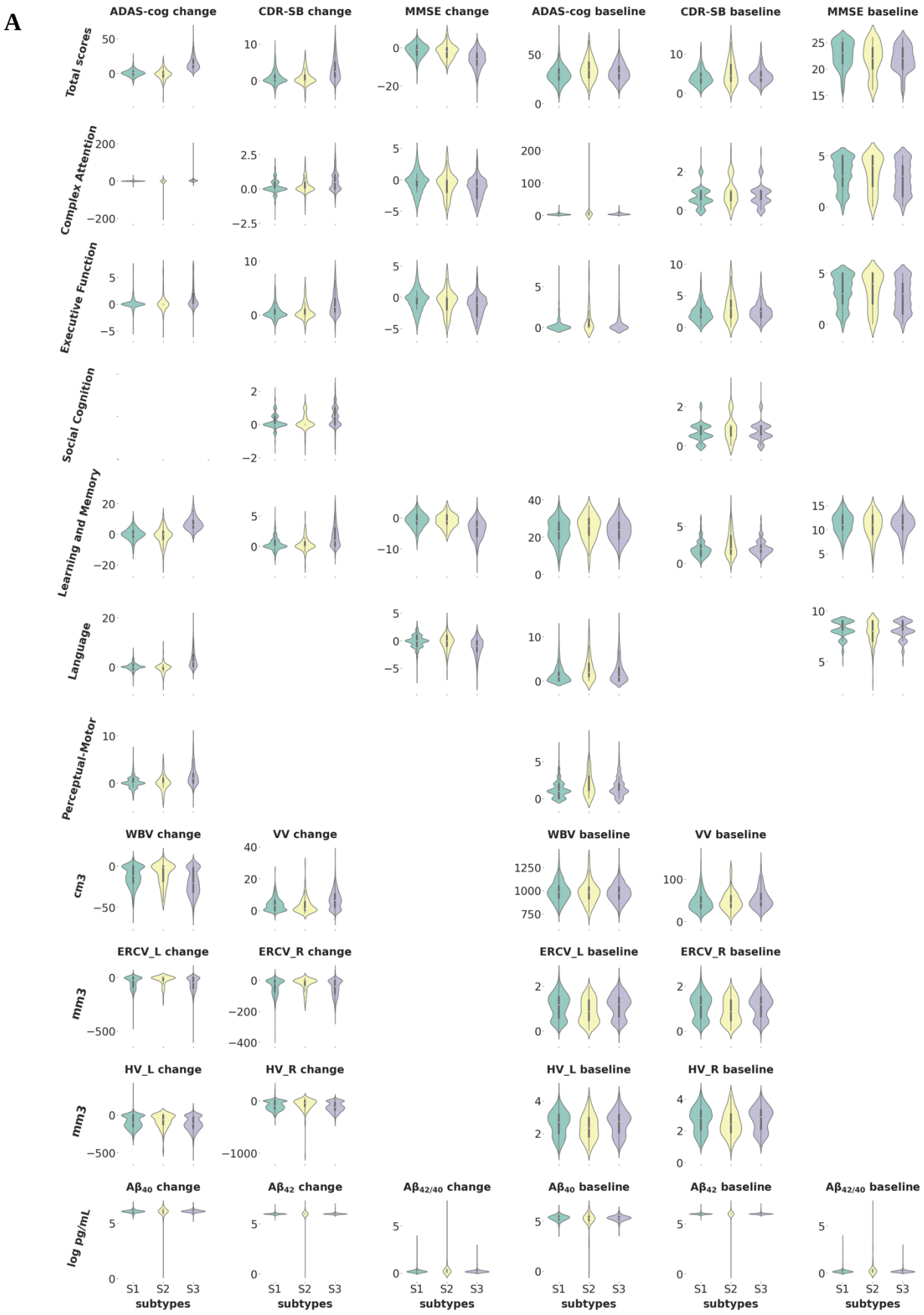
Subtypes	**Subtype 1 (S1)**	**Subtype 2 (S2)**	**Subtype 3 (S3)**	**P**
N of patients	741	139	659	
Demographics				
Age at first visit	74.15 (74.73, 7.82)	74.31 (75.28, 8.68)	72.83 (73.5, 7.98)	0.005
Females - no. (%)	419 (56.55)	83 (59.71)	396 (60.09)	0.383
Hispanic - no. (%)	66 (8.91)	19 (13.67)	44 (6.68)	0.02
Race				
Black or African American	22 (2.97)	9 (6.47)	12 (1.82)	0.009
White	611 (82.46)	109 (78.42)	572 (86.8)	0.015

APOE Gene				
No E4	252 (34.01)	55 (39.57)	216 (32.78)	0.307
One E4	370 (49.93)	59 (42.45)	311 (47.19)	0.224
Two E4	94 (12.69)	15 (10.79)	102 (15.48)	0.183
Baseline				
ADAS				
ADAS-cog14 total	29.32 (29.0, 8.91)	34.63 (34.0, 11.0)	31.76 (31.0, 9.3)	<.001
Complex Attention	3.71 (3.0, 3.2)	6.48 (4.0, 17.95)	4.71 (4.0, 3.24)	<.001
Executive Function	0.38 (0.0, 1.07)	0.91 (0.0, 1.63)	0.51 (0.0, 1.24)	<.001
Language	1.52 (1.0, 1.92)	2.71 (2.0, 2.71)	2.18 (1.0, 2.54)	<.001
Learning and Memory	23.24 (23.0, 6.54)	25.22 (26.0, 6.69)	23.75 (24.0, 5.93)	0.004
Perceptual-Motor Function	1.23 (1.0, 1.12)	1.87 (1.0, 1.62)	1.58 (1.0, 1.3)	<.001
CDR				
CDR total	4.07 (4.0, 2.06)	5.26 (4.5, 3.16)	4.41 (4.0, 2.01)	<.001
Complex Attention	0.7 (0.5, 0.52)	0.91 (1.0, 0.66)	0.76 (0.5, 0.54)	0.001
Executive Function	2.31 (2.0, 1.43)	3.13 (3.0, 2.16)	2.53 (2.5, 1.44)	<.001
Learning and Memory	1.97 (2.0, 1.1)	2.58 (2.0, 1.63)	2.09 (2.0, 1.01)	<.001
Social Cognition	0.66 (0.5, 0.45)	0.85 (1.0, 0.62)	0.72 (0.5, 0.47)	0.002
MMSE				
MMSE total	22.7 (23.0, 2.51)	21.94 (22.0, 3.09)	22.0 (22.0, 2.44)	<.001
Complex Attention	3.23 (3.0, 1.53)	3.22 (4.0, 1.72)	2.82 (3.0, 1.56)	<.001
Executive Function	3.23 (3.0, 1.53)	3.22 (4.0, 1.72)	2.82 (3.0, 1.56)	<.001
Language	8.16 (8.0, 0.84)	7.89 (8.0, 1.07)	8.08 (8.0, 0.86)	0.014
Learning and Memory	11.3 (11.0, 2.13)	10.83 (11.0, 2.51)	11.1 (11.0, 2.09)	0.154
NPI total	8.71 (4.0, 10.9)	8.96 (4.0, 11.23)	7.76 (4.0, 10.2)	0.201
ADL total	66.32 (68.0, 8.93)	58.62 (62.0, 15.34)	64.67 (67.0, 10.15)	<.001
EQ5D total	74.54 (80.0, 18.93)	69.71 (71.0, 23.56)	75.48 (80.0, 18.35)	0.041
QLADC total	36.34 (36.0, 5.98)	34.5 (35.0, 7.13)	36.3 (36.0, 5.83)	0.014
Aβ levels				
$A\beta_{40}$ (log pg/mL)	5.33 (5.36, 0.3)	5.14 (5.29, 0.87)	5.31 (5.32, 0.28)	0.046
$A\beta_{42}$ (log pg/mL)	1.61 (0.0, 1.87)	1.55 (0.0, 2.04)	1.52 (0.0, 1.82)	0.506
$A\beta_{42/40}$ ratio	0.19 (0.15, 0.16)	0.35 (0.16, 0.66)	0.18 (0.14, 0.13)	0.267
PET-SUVR	1.49 (1.49, 0.17)	1.48 (1.47, 0.19)	1.5 (1.5, 0.16)	0.107
vMRI (regional brain volumes are scaled by WBV)				
ERCV_L	1.1 (1.18, 0.55)	0.9 (0.87, 0.52)	1.11 (1.19, 0.53)	<.001
ERCV_R	1.04 (1.09, 0.53)	0.88 (0.79, 0.53)	1.06 (1.14, 0.5)	<.001
HV_L	2.61 (2.7, 0.75)	2.4 (2.34, 0.76)	2.65 (2.75, 0.73)	0.002
HV_R	2.7 (2.81, 0.78)	2.48 (2.42, 0.78)	2.74 (2.88, 0.76)	0.001
VV (cm^3)	47.41 (43.3, 22.26)	48.12 (43.7, 22.6)	53.54 (48.3, 23.54)	<.001
WBV (cm^3)	988.64 (981.71, 108.77)	980.34 (981.6, 117.18)	976.9 (973.18, 98.77)	0.223
Endpoint Changes				
ADAS				
ADAS-cog14 total	0.64 (1.0, 4.89)	-1.06 (0.0, 7.18)	15.09 (13.0, 9.88)	<.001
Complex Attention	0.17 (0.0, 3.04)	-1.34 (0.0, 17.14)	3.87 (3.0, 10.25)	<.001
Executive Function	0.18 (0.0, 1.14)	0.22 (0.0, 1.63)	1.15 (0.0, 2.05)	<.001
Language	0.07 (0.0, 1.42)	-0.4 (0.0, 1.93)	3.05 (2.0, 3.57)	<.001
Learning and Memory	0.06 (0.0, 3.76)	-1.0 (0.0, 5.28)	6.99 (6.0, 4.3)	<.001
Perceptual-Motor Function	0.16 (0.0, 1.18)	0.09 (0.0, 1.54)	1.57 (1.0, 2.02)	<.001
CDR				
CDR total	0.84 (0.5, 1.83)	0.83 (0.0, 1.88)	3.2 (2.5, 3.05)	<.001

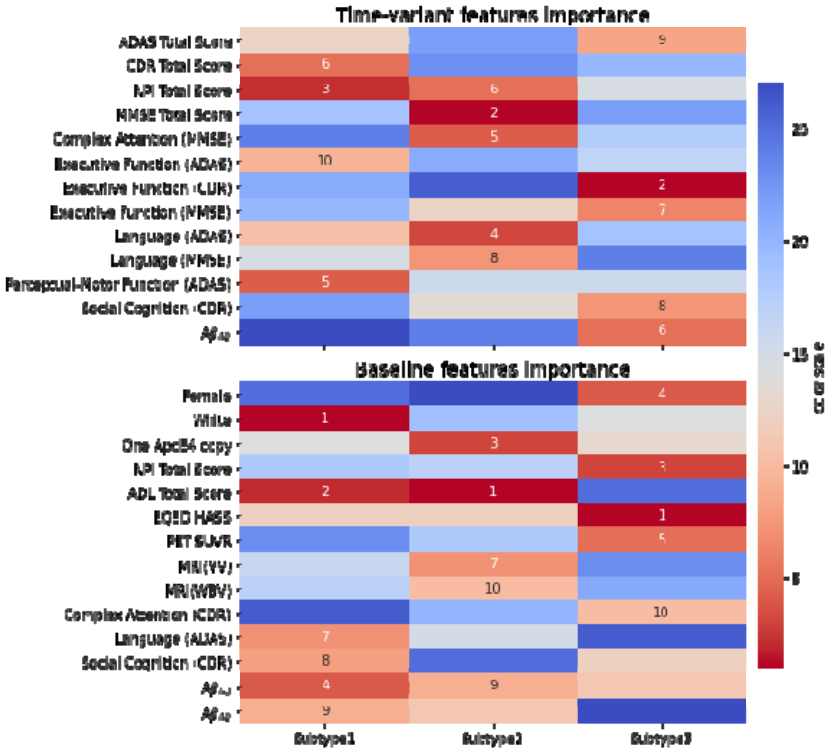
Complex Attention	0.15 (0.0, 0.48)	0.22 (0.0, 0.51)	0.55 (0.5, 0.68)	<.001
Executive Function	0.53 (0.0, 1.27)	0.63 (0.0, 1.43)	2.07 (1.5, 2.14)	<.001
Learning and Memory	0.41 (0.0, 1.02)	0.35 (0.0, 1.08)	1.67 (1.0, 1.68)	<.001
Social Cognition	0.15 (0.0, 0.43)	0.14 (0.0, 0.44)	0.47 (0.5, 0.56)	<.001
MMSE				
MMSE total	-1.42 (-1.0, 3.57)	-1.92 (-2.0, 3.77)	-6.09 (-6.0, 4.72)	<.001
Complex Attention	-0.44 (0.0, 1.7)	-0.71 (0.0, 1.86)	-1.4 (-1.0, 1.71)	<.001
Executive Function	-0.44 (0.0, 1.7)	-0.71 (0.0, 1.86)	-1.4 (-1.0, 1.71)	<.001
Language	-0.14 (0.0, 1.18)	-0.31 (0.0, 1.44)	-1.22 (-1.0, 1.72)	<.001
Learning and Memory	-0.86 (-1.0, 2.36)	-0.92 (-1.0, 2.18)	-3.59 (-3.0, 2.97)	<.001
NPI total	0.3 (0.0, 10.31)	0.18 (0.0, 9.31)	5.35 (2.0, 13.36)	<.001
ADL total	-4.08 (-3.0, 8.43)	-5.56 (-2.0, 10.78)	-13.04 (-10.0, 12.63)	<.001
EQ5D total	-0.99 (0.0, 18.96)	1.28 (0.0, 21.87)	-4.81 (0.0, 20.09)	0.001
QLADC total	-1.22 (0.0, 4.79)	-0.53 (0.0, 4.94)	-2.88 (-2.0, 5.03)	<.001
Aβ levels				
$A\beta_{40}$	6.08 (6.07, 0.15)	6.02 (6.07, 0.52)	6.1 (6.09, 0.14)	0.004
$A\beta_{42}$	6.02 (6.0, 0.12)	5.97 (6.0, 0.52)	6.02 (6.0, 0.11)	0.232
$A\beta_{42/40}$ ratio	0.23 (0.14, 0.34)	0.36 (0.19, 0.75)	0.21 (0.16, 0.25)	0.001
PET-SUVR	0.0 (0.0, 0.09)	0.0 (0.0, 0.06)	0.0 (0.0, 0.1)	0.222
vMRI				
ERCV_L	-46.94 (-25.23, 54.79)	-24.93 (0.0, 46.42)	-59.12 (-46.8, 61.76)	<.001
ERCV_R	-41.26 (-22.8, 48.03)	-23.78 (0.0, 40.5)	-50.16 (-38.4, 51.82)	<.001
HV_L	-94.12 (-93.07, 85.12)	-80.17 (-62.3, 92.19)	-115.58 (-116.46, 94.14)	<.001
HV_R	-95.63 (-98.0, 85.09)	-75.27 (-47.5, 118.06)	-118.92 (-122.14, 91.84)	<.001
VV	3.89 (3.15, 4.08)	3.17 (1.1, 4.65)	6.8 (6.2, 5.84)	<.001
WBV	-12.45 (-11.4, 12.26)	-9.9 (-1.8, 13.12)	-19.94 (-20.9, 15.31)	<.001

ADAS-cog 14: 14-item Alzheimer's Disease Assessment Scale – cognitive subscale (range 0 to 70, higher scores worse), CDR-SB: Clinical Dementia Rating scored by the sum of boxes method (range 0 to 18, higher scores worse), MMSE: Mini-Mental State Examination (range 0 to 30, lower scores worse), ADCS-ADL: Alzheimer's Disease Cooperative Study – Activities of Daily Living Scale (range 0 to 78, lower scores worse), NPI: Neuropsychiatric Inventory (NPI; range 0 to 144, higher scores worse), QoL-AD: Quality of Life in Alzheimer's Disease (range 13 to 52, lower scores worse), EQ-5D: 5-Dimensional EuroQol Quality of Life Scale Proxy Version (range 0 to 100, lower scores worse), $A\beta$: Amyloid-Beta, PET-SUVR: positron emission tomography with standardized uptake value ratio, vMRI: volumetric magnetic resonance imaging, ERCV: entorhinal cortex volume, HV: hippocampal volume, VV: ventricular volume, WBV: whole brain volume

Fig. 3 Feature distributions (A) and importance patterns (B) by clusters



B



Violin plots (A) illustrate the within-cluster distribution of cognitive scores, brain regional volumes, and amyloid beta deposition at baseline and change to the endpoint, respectively. Note the unequal sample size among the cluster types: S1 (N=741), S2 (N=139), and S3 (N=659). Heatmap (B) visualizes the top important baseline and time-invariant features.

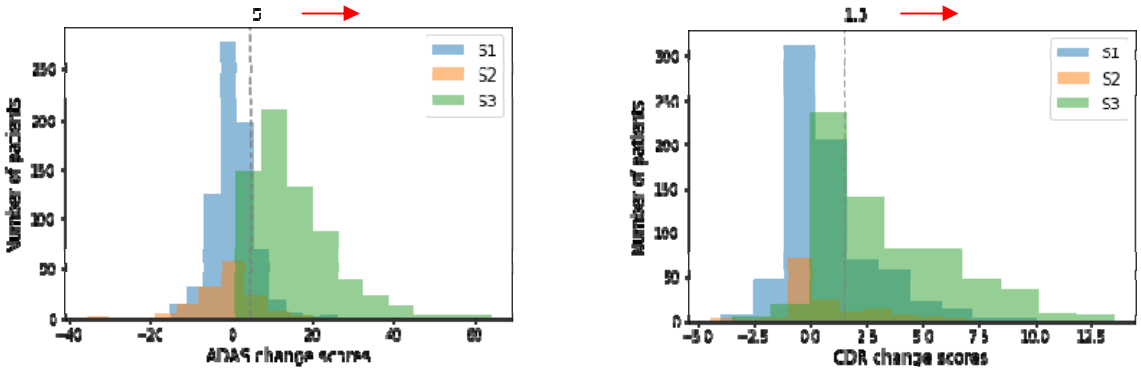
Specifically, S1 (48.1% in EXPEDITION1+EXPEDITION3; 43.0% in EXPEDITION2) had the largest cohort with 741 Alzheimer spectrum patients. S2 (9.0% in EXPEDITION1+EXPEDITION3; 13.0% in EXPEDITION2) had the smallest cohort with 139 patients containing the highest proportion of the Hispanic or Latino population (8.91% in S1 vs. 13.67% in S2 vs. 6.68% in S3, p=0.02; We omitted S1, S2, and S3 below). S2 also had the highest proportion of the Black or African American population (2.97% vs. 6.47% vs. 1.82%, p=0.009) and the lowest proportion of the White population (82.46% vs. 78.42% vs. 86.8%, p=0.015). S3 (42.8% in EXPEDITION1+EXPEDITION3; 44.0% in EXPEDITION2) had 659 slightly younger patients on average (74.15 vs. 74.31 vs. 72.83, p=0.005).

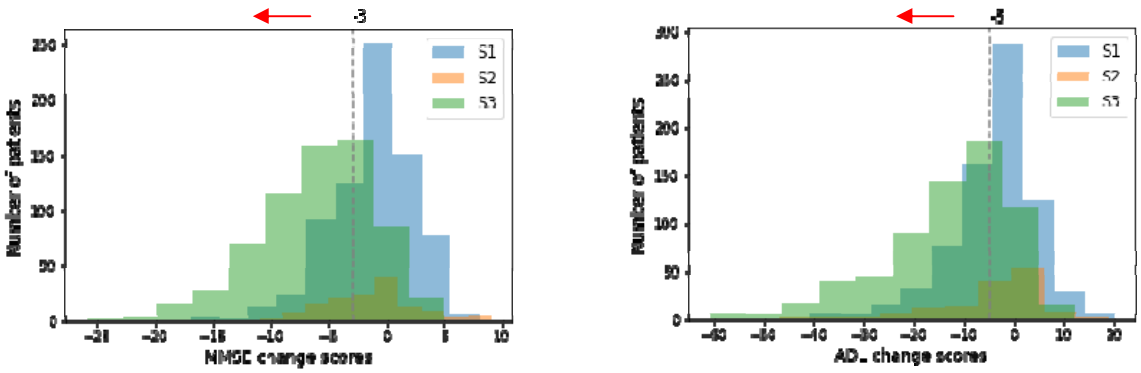
At baseline, S1 and S3 showed similar overall shapes of data distribution in the violin plots. The APOE genotype didn't show significant differences between subtypes. However, S2 had the highest portion of patients without the APOE 4 gene and the lowest portion of APOE 4 carriers. In terms of baseline cognitive assessments, S2 consistently exhibited higher scores across multiple domains and total scores of ADAS and CDR than S1 and S3 (e.g., 29.32 vs. 34.53 vs. 31.76 of average ADAS total score, p<.001; 4.07 vs. 5.26 vs. 4.41 of average CDR total scores, p<.001), which indicated worse cognitive performance or more severe disease symptoms. On the MMSE scale (a measure of cognitive function where lower scores indicated more severe cognitive impairment), S2 had a slightly lower mean score compared to S1 and S3 (e.g., 22.7 vs. 21.94 vs. 22.0 of average MMSE total scores, p<.001). S1 consistently exhibited

the most minor cognitive impairment (see details in Table 2). For levels, S2 had lower (5.33 vs. 5.14 vs. 5.31 log pg/mL, p=0.046), though the differences across the subtypes were significant only for , not for or the ratio. Furthermore, neuroimaging data suggested that patients in S3 generally showed greater entorhinal cortex and hippocampal volumes than S1 and S2, while S2 showed the lowest brain regional volumes (see details in Table 2). S3 also retained the most significant average ventricular volume (47.41 vs. 48.12 vs. 53.54 cm^3, p<.001). The PET-SUVR and whole brain volume didn't show significant differences across subtypes.

We then evaluated longitudinal trends using endpoint changes for time-variant variables. S1 and S2 showed similar shapes of cognitive data distribution in violin plots, while S1 and S3 showed similar shapes of neuroimaging data and levels. We observed that S3 significantly increased in scores of multiple domains and total scores of ADAS and CDR (e.g., 0.64 vs. -1.06 vs. 15.09 of average ADAS total change score, p<0.001; 0.84 vs. 0.83 vs. 3.2 of average CDR total change scores, p<.001), indicating a drastic cognitive decline. Interestingly, S2 showed preserved cognitive domains of complex attention (-1.34), language (-0.4), and learning and memory (-1.0) measured by ADAS. On the MMSE scale, S3 showed significantly decreased scores (e.g., -1.42 vs. -1.92 vs. -6.09 of average MMSE total change scores, p<.001), indicating deterioration. In levels, S3 showed the most increase in (6.08 vs. 6.02 vs. 6.1 log pg/mL, p=0.004) and the least increase in the ratio (0.23 vs. 0.36 vs. 0.21, p=0.001), while S2 was the opposite. In terms of brain imaging changes, S3 showed the most significant loss in regional brain volume (e.g., -46.94 vs. -24.93 vs. -59.12 mm^3 of average left entorhinal cortex change volume, p<.001; -94.12 vs. -90.17 vs. -115.58 mm^3 of average left hippocampal change volume, p<.001) and whole brain volume (-12.45 vs. -9.9 vs. -19.94 cm^3, p<.001), which indicated brain atrophy that might correlate with the cognitive decline observed in cognitive measurements. In addition, S3 showed the most significant increase in ventricular volume (3.89 vs. 3.17 vs. 6.8 cm^3, p<.001), suggesting greater brain atrophy. Here, S2 always showed the most minor changes compared to S3.

Fig 4 Distribution of different clinical outcome variables for identified clusters





Direction of fast decline

Finally, we plotted the distribution of four clinical outcome measures (i.e., ADAS, CDR, MMSE, and ADL) across clusters. We separated the population into different paths of progression using the middle value of each clinical outcome. We observed that the distribution of S3 lay mostly in the fast decline directions across all four measures.

Discussion and Conclusions

Our objective was to shed light on the heterogeneity of Alzheimer's Disease progression patterns by developing a cutting-edge, deep-learning framework to identify AD subtypes with static and longitudinal features. Specifically, we developed a representation model called the knowledge-guided deep temporal clustering representation (KG-DTC) to generate cluster-friendly and clinical outcome-related representations. The learned representation was examined with effectiveness in discriminating the heterogeneity of dynamic and complex disorders in the AD cohort. By applying this approach to the pooled randomized clinical trial data, we identified three distinct AD subtypes: S2 had the highest cognitive impairment at the baseline but preserved progression at the endpoint, while S3 had significant deterioration at the endpoint. S1 represented those who have milder symptoms at baseline and often demonstrate less decline over time. This finding provided a novel understanding of AD progression in combination with knowledge of neuropathological and clinical heterogeneity, which may pave the way for individualized AD progression forecasts and customized treatments for specific AD subtypes.

We employed a seq2seq architecture to encode multivaria longitudinal data and integrated *k*-means clustering loss and clinical outcome classification loss to form clinically meaningful cluster structures. Here, representation learning aimed to map the high-dimensional complex data to lower-dimensional space. However, traditional representation learning methods are not designed for clustering tasks. To enable the learned representations to favor clustering patients, we integrated the *k*-means clustering loss to encourage the representations to form clusters. Additionally, previous studies mostly focused on identifying AD subtypes in a purely unsupervised way[6,7,15]. While these studies obtained the AD subtypes without considering long-term clinical outcomes, the identified AD subtypes may not reflect our existing knowledge of outcomes of clinical interest. We filled this gap by introducing a clinical outcome supervision task; we employed a classification loss in the model to enable the learned representations to be informative with multiple clinical outcomes and unveil clinically meaningful and actionable insights.

We identified three distinct AD subtypes from pooled randomized clinical trials, which demonstrated heterogeneity in the presentation and progression of Alzheimer's disease. These subtypes were aligned with the NIA-AA's AT(N)(C) classification (based on fluid biomarkers, neurodegeneration, and cognitive symptoms) and with rapid/slow progressors (based on cognitive presentation) [16,17]. At the baseline, S2 has a higher representation of Hispanic and Black or African American populations. Race and ethnicity variations among the subtypes could hint at potential genetic or environmental risk factors specific to different populations.[18,19] Clinically, S2 manifests a more severe cognitive deficit at baseline, as evidenced by elevated ADAS and CDR scores, reduced MMSE scores, and lower levels of brain regional volumes. These suggested that S2 might represent a more severe or rapidly progressing form of the disease. However, when it comes to endpoint changes, S3 appears to deteriorate at a faster rate than S2 regarding cognitive measures and neurodegeneration, indicating that Subtype 3 might experience a delayed but rapid cognitive decline, which suggested an A(-) N(+) C(+) profile in the NIA-AA framework. Moreover, the significantly greater increase versus other subtypes of $A\beta_{42/40}$ in S2 suggests potential variations in amyloid pathology or deposition. Interestingly, certain cognitive domains in S2 showed lesser deterioration or even improvement, coupled with a reduced rate of regional brain atrophy, which suggested an A(+) N(-) C(-) profile in the NIA-AA framework.

Our data-driven subtyping can contribute to connecting the subtypes into the primary endpoint of clinical trials, which will facilitate patient specific therapeutic development. Traditional clinical trials often treat Alzheimer's patients as a homogenous group. However, the existence of distinct subtypes suggests that treatments could be more effective if they were more tailored. Targeted interventions can be designed by identifying which subtype an individual belongs to, which may increase the likelihood of therapeutic success. Moreover, aligning data-driven subtypes with established biomarker frameworks, such as the NIA-AA's AT(N)(C) classification, allows for a more integrated understanding of the disease. This can help in the identification of novel biomarkers and the development of therapies targeting specific pathways.

Our study has some limitations. We did not include the comorbidity risk factors when modeled in the context of AD heterogeneity. This reduced the interpretation ability of identified AD progression patterns that may be affected by other diseases or drugs. In the future, we envision combining and comparing comorbidities longitudinally, thus extending our current analyses to understand the contribution of comorbidities in AD subtypes. The inclusion of patients from two different trials is an advance since it increased the variability in the sample and therefore, represented the AD population better, but it is also a limitation due to data variability in multiple studies considering different efficacy measure standards and irregular visit intervals. Future work can focus on better visit alignment and missing imputations. Our model has to pre-specify the number of clusters in *k*-means clustering loss. However, we do not exactly know how many subtypes are in the datasets naturally, and it is sensitive to which datasets are used. Future studies should design a structure that can automatically optimize the selection of the best N of clusters during the model training process.

In conclusion, we discovered three longitudinal patterns of AD subtypes by a novel outcome-regularized deep temporal clustering approach. Our study is an important step towards solving an unmet need, i.e., uncovering the subtypes of AD disease progressions with observed heterogeneity in neurology and biology. Moreover, our proposed models unravel the heterogeneity in AD that can enable precision medicine and potentially lead to successful disease-modifying treatments in the future.

Methods

Study design and dataset

Figure 1 shows the overall study design, including data preprocessing, subtype discovery, model training, and external validation, and clustering significance and stability evaluation. In this retrospective study, we used patient data in placebo groups from three randomized AD trials: 505 in EXPEDITION1 (NCT00905372[13], 2009-2012), 518 in EXPEDITION2 (NCT00904683[13], 2009-2012), and 1064 in EXPEDITION3 (NCT01900665[14], 2013-2017). Patients in placebo groups allow us to investigate the disease progression without being confounded by the exposure to experimental drugs. The three trials tested Solanezumab, a humanized anti-amyloid monoclonal antibody, for its efficacy in slowing AD decline. They had similar entry criteria, including patients 55 years or older who met the criteria of the National Institute of Neurological and Communicative Diseases and Stroke–Alzheimer's Disease and Related Disorders Association for probable AD. EXPEDITION1 and EXPEDITION2 included mild to moderate patients with MMSE scores of 16 to 26, while EXPEDITION3 included mild patients with MMSE scores of 20 to 26. They then underwent 80-week observations. For each patient, the time interval between two adjacent longitudinal measurements ranges from 4 to 16 weeks.

Variables

We included 8 demographics, 28 baseline clinical conditions, and 28 longitudinal clinical conditions' changes in our model. The clinical conditions included cognitive assessments, imaging biomarkers, fluid biomarkers, quality-of-life assessments, and a neuropsychiatric assessment from three trials. The cognitive assessments included ADAS-cog14 (range 0 to 90, higher scores worse), ADCS-ADL (range 0 to 78, lower scores worse), CDR-SB (range 0 to 18, higher scores worse), and MMSE (range 0 to 30, lower scores worse). Imaging biomarkers contained v-MRI volumes of the whole brain and regional brains (e.g., hippocampi, entorhinal cortices, and ventricles) and amyloid PET imaging for the composite summary standard uptake value ratios (SUVRs). The fluid biomarkers are plasma $A\beta$ levels. The quality-of-life assessments are QoL-AD (range 13 to 52, lower scores worse) and EQ-5D Proxy (range 0 to 100, lower scores worse). The neuropsychiatric assessment included NPI (range 0 to 144, higher scores worse). Supplementary Table 1 provides a summary of all variables in each cohort.

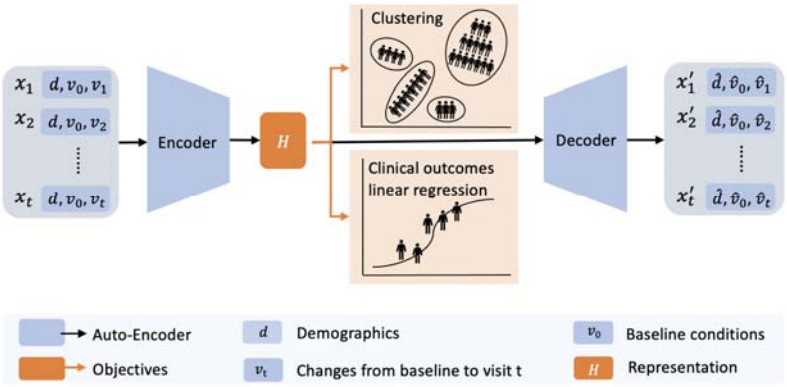
Data Preprocessing

We preprocessed the datasets by applying data harmonization, missing value imputation, and data transformation to all three trials. As each study can have variables with different names or units, careful harmonization was conducted. We matched the variable by reviewing the Case Report Form in each trial. We selected visits with assessments conducted or biomarkers collected (i.e., visits 1, 2, 3, 4, 5, 6, 10, 13, 16, 19, and 23.0 for EXPEDITION 1 and 2; visits 1, 2, 3, 5, 9, 12, 15, 18, and 22 for EXPEDITION 3). For features that have longitudinal changes, we calculated change values from the baseline to the visits, thus patients' longitudinal trends were represented as a set of change value sequences. To deal with missing values, we leveraged various imputation strategies. Within a sequence, a missing value indicated that a test was not conducted during that visit. Assuming a patient's condition stayed stable before and after a test visit, we filled in missing values by propagating the existing values forward and backward along a sequence. The remaining missing values were imputed by chained equations (MICE), which is

a multiple imputation method that builds a multivariate predictive model to infer the missing values using the remaining features in a round-robin fashion[20]. Data from different distributions were carefully standardized. We log-transformed skewed distributions (i.e., levels) into Gaussian and Gaussian into z-distributions to make variables comparable. To reduce cognitive variable redundancy, specific cognitive test items (e.g., test of Comprehension of Spoken Language, Word-Finding Difficulty, and Naming Objects and Fingers in ADAS) were grouped into cognitive functions as grouped variables (e.g., language in ADAS). Overall, the cognitive test items of ADAS, CDR, and MMSE were grouped into 6 cognitive functions (i.e., complex attention, executive function, language, learning and memory, perceptual-motor function, and social cognition), respectively.

Subtype discovery: Knowledge-Guided Deep Temporal Clustering (KG-DTC) model

Fig 5 KG-DTC model structure



[Caption] KG-DTC architecture. We used a temporal autoencoder to encapsulate the multivariate baseline (,) and temporal features ( ) into a hidden representation *H*. The hidden representation *H* is jointly derived to minimize the clustering loss and clinical outcome regression loss.

We developed a novel knowledge-guided deep temporal clustering model (see Fig. X). that identifies patient clusters by deriving cluster-friendly embedding of temporal observations in an end-to-end framework. Our model is built upon a sequence-to-sequence structure (seq2seq) that encapsulates time-invariant and time-variant observations into a representation. Here, a technical challenge is that the representation from the seq2seq does not necessarily form a cluster. Motivated by prior research that embeds clustering into representation learning[21,22], we encouraged the representation to form clusters by incorporating the *k*-means clustering loss during training. Clinical outcomes are the main results that are measured at the end of a study to see whether a given treatment worked. To leverage clinical outcomes as knowledge, we also guided the representation to be discriminative to clinical outcomes of interest, in order to identify highly responsive groups and less responsive groups. Details of each component are as follows.

Temporal autoencoder (Seq2seq). We used gated recurrent units (GRUs) autoencoders (AEs) for the seq2seq structure. We first instantiated the encoder as a batch normalization layer connected with a 2-layer stacked GRUs[23] to capture temporal and multiscale characteristics of input data. We then utilized a single-layer GRU as the decoder to reconstruct the input. GRU is a

variation of RNN with an additional relevance gate and updated gate, which can capture time dependencies over different periods and thus efficiently handles temporal patterns. AEs learn the embedded representations for high-dimensional data by reconstructing the input series and minimizing the reconstruction loss.

Given N patients, the patients data were divided into three parts: demographic information $g \in \mathbb{R}^{N\times d_0}$, baseline clinical conditions $v_0 \in \mathbb{R}^{N\times d_1}$, and longitudinal clinical condition changes over visit t $v_t \in \mathbb{R}^{N\times d_2\times T}$, where $t \in \{1,\dots,T\}$ and T was the total number of visits, and d_0, d_1, d_2 were the number of features in each part, respectively. The static components g and v_0 were repeated for each time step t. We then concatenated the $g,\ v_0, and\ v_t$ to form a sequence of feature vectors for patient i at visit t, denoted as $x_{i,t} = [g;\ v_0; v_t]$, $x_{i,t} \in \mathbb{R}^D$, where $D = d_0 + d_1 + d_2$, was the total number of features. Therefore, the sequence of concatenated vectors for patient i across all visits can be represented as: $x_i \in \mathbb{R}^{T\times D}$. We drop i for simplicity in the following notation.

We first applied batch normalization over input $x \in \mathbb{R}^{T\times D}$. The output $y \in \mathbb{R}^{T\times D}$ is

$$y = \gamma \times (x - E[x])/\sqrt{Var[x] + \varepsilon} + \beta,$$

where $E[x]$ is the expectation of x, $Var[x]$ is variance of x, γ and β are parameters to be learned, and ε is a constant added for numerical stability. For simplicity, we still use x as the output of batch normalization in the following notation. We then applied multi-layer GRU RNN to the x as encoder. Given input x_t^l at visit t, each layer l computed the following functions:

$$z_t^l = sigmoid\left(W_z^l \times x_t^l + U_z^l \times h_{t-1}^l + b_z^l\right),$$

$$r_t^l = sigmoid\left(W_r^l \times x_t^l + U_r^l \times h_{t-1}^l + b_r^l\right),$$

$$n_t^l = tanh\left(W_n^l \times x_t^l + U_n^l \times \left(r_t^l \odot h_{t-1}^l\right) + b_n^l\right),$$

$$h_t^l = \left(1 - z_t^l\right) \odot h_{t-1}^l + z_t^l \odot n_t^l,$$

where z_t^l, r_t^l, and n_t^l were update, reset, new gates, respectively. The update gate z_t^l helped the model to determine how much of the past information needs to be passed to the future. The reset gate r_t^l defined how much of the past information to forget. The new gate combined the new input with the past hidden state to create new candidates for the hidden state. The W and U were the weight matrices and b was the bias for each gate in layer l. The h_{t-1}^l was the past hidden state from the previous visit in the same layer. The h_t^l was the combination of the new gate and the past hidden state, controlled by the update gate. The output h_t^l was used as the input x_t^{l+1} for the next layer in the stacked GRU. Therefore, the final hidden representation from the encoder was $h_t^{l+1} \in \mathbb{R}^m$, where m was the hidden size.

The single-layer GRU decoder is essentially another RNN layer that takes hidden representation h_t^{l+1} generated by the encoder and aims to reconstruct the original input sequence. The computation process is the same as the encoder and the final hidden state returned by the decoder is x_t'. This sequence $x' = (x_1', x_2', \dots, x_T')$ is the reconstructed version of the original

input sequence. We used mean square error (MSE) to evaluate the quality of the reconstruction, which can be formulated as

$$L_{reconstruction} = 1/n \times \sum_{i=1}^{N} (\mathrm{x}' - \mathrm{x})^2 .$$

***K*-means clustering loss.** We denoted the hidden representation h_t^{l+1} for N patients as $H \in \mathbb{R}^{N \times m}$. It is noted that the H obtained from the temporal autoencoder does not guarantee distinctive clusters. Hence, following previous work[22], we encouraged H to form clusters while maintaining the reconstruction by regularizing the AEs through a soft *k*-means objective, defined as follows

$$L_{k-means} = Tr\,(H^T H) - Tr\,(F^T H^T H F)\; s.t. F^T F = I,$$

where $F \in \mathbb{R}^{N \times k}$ denoted the cluster membership matrix, $F^T F = I$, and k was the number of clusters. Minimizing the *k*-means clustering loss with H was equivalent to maximizing the trace $Tr\,(F^T H^T H F)$[23]. Since the learning of H was dynamic instead of static, the training process consisted of iteratively updating F and H. When fixing F, updating H can follow the standard stochastic gradient descent (SGD), encouraging the representations to form cluster structures. While fixing H, according to the Ky Fan theorem[24], we updated F using the closed-form solution to the trace maximization problem by computing the k-truncated singular value decomposition (SVD) of H. We fixed a cluster count of four in *k*-means clustering loss, which was commonly used in multiple previous AD subtyping studies[25–27].

Knowledge guidance. To enable the learned representation to be discriminative to the clinical outcomes, we introduced a multi-target multi-linear regression to encourage the learned representations to predict multiple clinical outcomes. We selected two primary clinical outcome assessments (i.e., total change of CDR-SB and ADAS-cog14 from baseline to end of observation) in the trials. We jointly trained the encoder that can detect multiple continuous clinical outcomes. Each patient had two clinical outcomes $y = \{y_{CDR}, y_{ADAS}\}$. We predicted the clinical outcomes $\hat{y}_j = W_{fc} \times H + b$, where $W_{fc} \in \mathbb{R}^{2 \times m}$ were the weights of the fully connected layers and b is the residual term. The loss between ground truth and predicted results are defined as

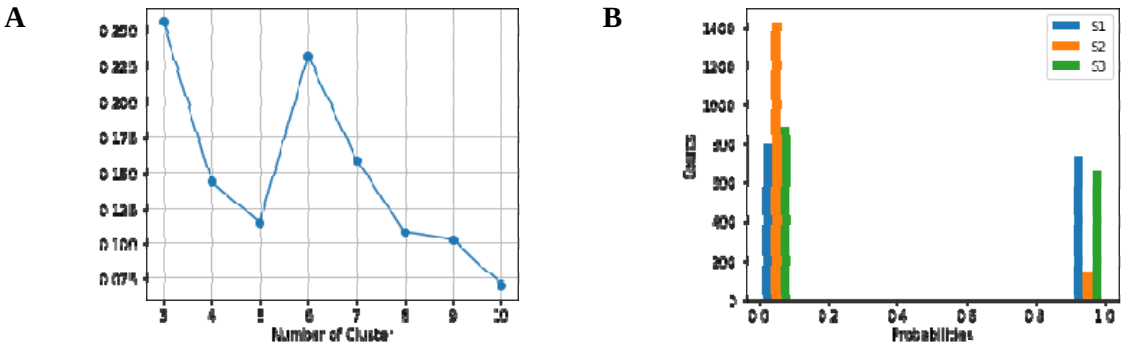
$$L_{regression} = \sum_{j=1}^{2} \sum_{i=1}^{N} (\hat{y}_{i,j} - y_{i,j})^2 .$$

We jointly optimized the three training losses $L_{reconstruction}$, $L_{k-means}$, and $L_{regression}$. To mitigate detrimental gradient interference in multi-objective learning, we use the gradient surgery to project conflicting gradients into the norm of other gradients[28]. Gradient surgery is a technique used to separate the gradients of different tasks, allowing the model to learn each task independently. This can help to prevent the gradients of one task from interfering with the gradients of another task, which can lead to suboptimal performance on one or more tasks. Throughout the training process, the batch size and hidden state feature size were set to 64 and 32, respectively.

After we trained the model, we obtained the deep temporal representation H for patients, which was used for downstream clustering tasks to determine distinct AD subtypes. Although the *k*-means clustering in our model provides the cluster membership (i.e., F), we trained a Gaussian

mixture model (GMM) and generated final clusters to account for soft membership. To determine the number of clusters k, we calculated the silhouette scores for k from 2 to 10 (Fig. 6 (A)). The silhouette score measures the similarity of an object to its own cluster compared to other clusters. It ranges from -1 to 1, with a higher value indicating more cluster separation. GMM is soft clustering, thus difficult to assign one sample to one cluster. We address this challenge by setting the probability threshold of membership. The distribution of cluster membership probabilities shows that subjects are concentrated in the probability interval of more than 0.9 (Fig. 6 (B)). Thus, after obtaining the clusters from GMM, we selected the representative "insider" patients for each cluster who had a cluster membership probability larger than 0.9.

Fig 6. Distribution of silhouette scores and predicted membership probabilities



Model evaluation

Cross-validation. We cross-validated the model on the discovery set to find the best model (as illustrated in Fig. X.). To this end, we randomly partitioned the discovery set into training, validation, and test sets in a ratio of 8:1:1 for training, best model selection, and final model performance testing. In the cross-validation we focused on balancing the tradeoff between multiple objectives: outcome prediction accuracy and clustering performance. We evaluated the model's capability to learn a representation that could effectively predict clinical outcomes. Despite incorporating clinical outcomes as a knowledge regularizer during training, the representation might not always predict these outcomes accurately due to the optimization of clustering.

The performance of the regression tasks was measured by the values. Higher values indicate that the model accounts for a good amount of variance of clinical outcome variables. We also evaluated our model's clustering performance by silhouette scores and UMAP visualization. A higher silhouette score indicates betterer clustering ability, and UMAP facilitates the mapping of learned representations into a 2D plot to visualize the shape and distribution of clusters.

Ablation study. To examine the contributions of individual components (i.e., clustering optimization and knowledge infusion) within our full model, we compared the full model against three ablated models to assess the importance of each module: M1 with only temporal encoder, M2 with temporal encoder and *k*-means clustering optimization, M3 with temporal encoder and

knowledge infusion regression tasks. Performance is also assessed by silhouette score, UMAP plots, and clinical outcome prediction accuracy.

Clustering evaluation

Reproducibility. To further demonstrate the generalizability of the model and reproducibility of the clustering patterns, we externally validated the trained model in an independent cohort. We applied the trained KG-DTC model using the discovery set to the external set and compared the clusters derived from both sets. We compared the distribution of clustering membership by comparing UMAP plots. To evaluate the reproducibility of clinical implication (i.e., clinical outcome prediction) and clustering performance, we calculated the prediction accuracy and silhouette scores on the external set.

Stability. We quantified the clustering stability using the ARI. ARI computes the similarity between two clustering results by counting pairs that are assigned to the same clusters in both clustering results[29]. It ranges from -1 to 1, with a score of 1 indicating a perfect match between two clustering results. To investigate the stability of clustering assignment against modest data perturbations, we adopted a LOO approach. In this approach, we randomly left out a small number of patients $n \in \{1:50\}$ in the external set, applied the trained KG-DTC model, and obtained perturbed clustering results. This process was repeated 200 times for each setting, which allowed us to compare the stability of clustering. We measured the similarity between the 200 different clustering results by calculating the ARI score.

Interpretation. We conducted a feature importance analysis using a permutation-based approach to identify variables that uniquely determine the clusters. We first built classification models to predict cluster membership for patients by adding two fully connected layers to the trained KG-DTC model. To measure the importance of each feature (including both time-variant and time-invariant features) in determining the cluster membership, we calculated feature importance scores by permuting the inputs of the model. Specifically, we randomly shuffled the values patient-wise for both static and time-varying features and then measured the changes in the performance of the model on this shuffled set. Features with a larger decrease in performance were more important in determining the clusters.

To gain a deeper understanding of clustering, we plotted the distribution of four clinical outcome measures (i.e., ADAS, CDR, MMSE, and ADL) across clusters. By separating the population into different paths of progression using the middle value of each clinical outcome, we were able to infer which cluster characteristics were informative regarding progression speed.

Statistical significance tests. We conducted post-hoc statistical tests to identify variables that showed significant differences across clusters. For non-parametric variables, we used the Kruskal-Wallis test. For variables that followed a Gaussian distribution, we employed the ANOVA test, and for categorical variables, we utilized the Chi-squared test. These statistical analyses enabled us to determine whether baseline clinical conditions and clinical condition changes at the final visit were significantly different across clusters. This provided valuable insights into the distinct factors contributing to the separate clusters identified by our model.

Author contributions

Concept and design: DW, XM, PS, XJ, and YK. Data analysis: DW, XM, XJ, and YK. Literature review: DW, PS, XJ, and YK. Model development: DW, XM, XJ, and YK. Verification of results: DW, XM, PS, XJ, and YK. Strategic guidance and oversight: PS, XJ, and YK. All authors had full access to all the data, contributed to data interpretation, drafted and edited the manuscript, approved the final manuscript, and accepted the responsibility to submit it for publication.

Acknowledgement

The authors would like to thank Christine M. Farrell, Ph.D. for her contribution to securing database access for this study, and Kristofer Harris for his contribution to clarifing the importance of identifying fast progressors without explicit definitions. This publication is based on research using data from data contributors Eli Lilly that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and Vivli and Eli Lilly are not in any way responsible for, the contents of this publication.

Competing interests

XJ is CPRIT Scholar in Cancer Research (RR180012), and he was supported in part by Christopher Sarofim Family Professorship, UT Stars award, UTHealth startup, the National Institute of Health (NIH) under award number R01AG066749, R01AG066749-03S1, R01LM013712, R01LM014520, R01AG082721, R01AG066749, U01AG079847, U01TR002062, U01CA274576 and the National Science Foundation (NSF) #2124789. YK is supported in part by UTHealth startup and the National Institute of Health (NIH) under award number R01AG082721 and R01AG066749. PS is funded by the McCord Family Professorship in Neurology, the Umphrey Family Professorship in Neurodegenerative Disorders, multiple NIH grants, several foundation grants, and contracts with multiple pharmaceutical companies related to the performance of clinical trials. He serves as a consultant and speaker for Eli Lilly, Biogen, and Acadia Pharmaceuticals. No other authors have declarations to disclose.

Data availability

Data access restrictions apply. Interested parties may contact Eli Lilly and Company for dataset licensing.

References

1. Devi, G. & Scheltens, P. Heterogeneity of Alzheimer's disease: consequence for drug trials? *Alzheimers Res. Ther.* **10**, 122 (2018).

2. Gabler, N. B. *et al.* Dealing with heterogeneity of treatment effects: is the literature up to the challenge? *Trials* vol. 10 (2009).

3. Kim, Y., Lhatoo, S., Zhang, G.-Q., Chen, L. & Jiang, X. Temporal phenotyping for transitional disease progress: An application to epilepsy and Alzheimer's disease. *J Biomed Inf.* **107**, 103462 (2020).
4. Lee, C. & Schaar, M. V. D. Temporal Phenotyping using Deep Predictive Clustering of Disease Progression. in *Proceedings of the 37th International Conference on Machine Learning* 5767–5777 (PMLR, 2020).
5. Mitelpunkt, A. *et al.* Novel Alzheimer's disease subtypes identified using a data and knowledge driven strategy. *Sci. Rep.* **10**, 1327 (2020).
6. Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *Npj Digit. Med.* **3**, 1–11 (2020).
7. Kim, Y. *et al.* Multimodal phenotyping of alzheimer's disease with longitudinal magnetic resonance imaging and cognitive function data. *Sci. Rep.* **10**, 1–10 (2020).
8. Kikuchi, M. *et al.* Identification of mild cognitive impairment subtypes predicting conversion to Alzheimer's disease using multimodal data. *Comput. Struct. Biotechnol. J.* **20**, 5296–5308 (2022).
9. Aldewereld, Z. T. *et al.* Identification of Clinical Phenotypes in Septic Patients Presenting With Hypotension or Elevated Lactate. *Front. Med.* **9**, 794423 (2022).
10. Liu, X. *et al.* Identification of distinct clinical phenotypes of acute respiratory distress syndrome with differential responses to treatment. *Crit. Care* **25**, 320 (2021).
11. Ma, Q., Zheng, J., Li, S. & Cottrell, G. Learning Representations for Time Series Clustering. in (2019).
12. Dinga, R. *et al.* Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage Clin.* **22**, 101796 (2019).

13. Doody, R. S. *et al.* Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N. Engl. J. Med.* **370**, 311–321 (2014).

14. Honig, L. S. *et al.* Trial of solanezumab for mild dementia due to Alzheimer's disease. *N. Engl. J. Med.* **378**, 321–330 (2018).

15. Feng, Y. *et al.* Deep multiview learning to identify imaging-driven subtypes in mild cognitive impairment. *BMC Bioinformatics* **23**, 402 (2022).

16. Jack, C. R. *et al.* NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* **14**, 535–562 (2018).

17. Craig J. Thalhauser, N. L. K. Alzheimer's disease: rapid and slow progression. *J R Soc Interface* **9**, 119 (2012).

18. Anderson, N. B., Bulatao, R. A., Cohen, B., & National Research Council (US) Panel on Race, Ethnicity, and Health in Later Life. *Ethnic Differences in Dementia and Alzheimer's Disease*. (National Academies Press (US), 2004).

19. Wang, H., Yang, F., Zhang, S., Xin, R. & Sun, Y. Genetic and environmental factors in Alzheimer's and Parkinson's diseases and promising therapeutic intervention via fecal microbiota transplantation. *Npj Park. Dis.* **7**, 1–10 (2021).

20. Van Buuren, S. *Flexible imputation of missing data*. (CRC press, 2018).

21. Xie, J., Girshick, R. & Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. in *Proceedings of The 33rd International Conference on Machine Learning* 478–487 (PMLR, 2016).

22. Ma, Q., Zheng, J., Li, S. & Cottrell, G. W. Learning Representations for Time Series Clustering. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).

23. Cho, K. *et al.* Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Preprint at http://arxiv.org/abs/1406.1078 (2014).

24. Fan, K. Maximum Properties and Inequalities for the Eigenvalues of Completely Continuous Operators. *Proc. Natl. Acad. Sci. U. S. A.* **37**, 760–766 (1951).

25. Ferreira, D., Pereira, J. B., Volpe, G. & Westman, E. Subtypes of Alzheimer's Disease Display Distinct Network Abnormalities Extending Beyond Their Pattern of Brain Atrophy. *Front. Neurol.* **10**, (2019).

26. Zhang, B., Lin, L., Wu, S. & Al-Masqari, Z. H. M. A. Multiple Subtypes of Alzheimer's Disease Base on Brain Atrophy Pattern. *Brain Sci.* **11**, 278 (2021).

27. Chen, P. *et al.* Four Distinct Subtypes of Alzheimer's Disease Based on Resting-State Connectivity Biomarkers. *Biol. Psychiatry* (2022) doi:10.1016/j.biopsych.2022.06.019.

28. Yu, T. *et al.* Gradient Surgery for Multi-Task Learning. in *Advances in Neural Information Processing Systems* vol. 33 5824–5836 (Curran Associates, Inc., 2020).

29. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).

Appendix Supplementary Materials

Supplementary Figure 1.