Autoencoder to Identify Sex-Specific Sub-phenotypes in Alzheimer's Disease Progression Using Longitudinal Electronic Health Records

Weimin Meng[1], Jie Xu[1], Yu Huang[1], Cankun Wang[2], Qianqian Song[1], Anjun Ma[2], Lixin Song[3], Jiang Bian[1], Qin Ma[2], Rui Yin[1]

[1] Department of Health Outcomes &Biomedical Informatics, University of Florida, Gainesville, FL, 32610, USA
[2] Department of Biomedical Informatics, Ohio State University, Columbus, OH, 43210, USA
[3] School of Nursing, University of Texas Health Science Center at San Antonio, San Antonio, TX, 78229, USA

Abstract

Alzheimer's Disease (AD) is a complex neurodegenerative disorder significantly influenced by sex differences, with approximately two-thirds of AD patients being women. Characterizing the sex-specific AD progression and identifying its progression trajectory is a crucial step to developing effective risk stratification and prevention strategies. In this study, we developed an autoencoder to uncover sex-specific sub-phenotypes in AD progression leveraging longitudinal electronic health record (EHR) data from OneFlorida+ Clinical Research Consortium. Specifically, we first constructed temporal patient representation using longitudinal EHRs from a sex-stratified AD cohort. We used a long short-term memory (LSTM)-based autoencoder to extract and generate latent representation embeddings from sequential clinical records of patients. We then applied hierarchical agglomerative clustering to the learned representations, grouping patients based on their progression sub-phenotypes. The experimental results show we successfully identified five primary sex-based AD sub-phenotypes with corresponding progression pathways with high confidence. These sex-specific sub-phenotypes not only illustrated distinct AD progression patterns but also revealed differences in clinical characteristics and comorbidities between females and males in AD development. These findings could provide valuable insights for advancing personalized AD intervention and treatment strategies.

Introduction

Alzheimer's disease (AD), recognized as the predominant form of dementia, is a multifaceted and progressive neurodegenerative disorder that currently affects an estimated 6.9 million Americans as of 2024[1], with a potential rise to 13.85 million by 2060[2]. AD accounts for 60-80% of dementia cases as well as the 7th cause of death in the United States, which presents significant challenges in both diagnosis and treatment globally[1] and imposes substantial burdens on individuals affected, their families, healthcare systems, and the whole society. The progression of AD is hypothesized to include three main phases: preclinical AD, clinically significant mild cognitive impairment (MCI), and AD[3]. These phases represent a continuum where early pathological changes, such as the accumulation of beta-amyloid plaques and tau tangles, begin silently in the brain during the preclinical stage[4]. As the disease advances, individuals may experience noticeable cognitive decline, initially manifesting as MCI that significantly impacts daily functioning. Ultimately, this progression culminates in AD, marked by severe cognitive deficits, memory loss, and functional impairment[5]. However, the progression of AD is characterized by a complex array of longitudinally linked clinical features and outcomes, demonstrating a broad spectrum of manifestations across different groups of patients beyond three main phases[6–8]. There is considerable interindividual variability in AD progression[9], with some experiencing rapid loss of cognition while others progressing more slowly. Compelling evidence indicates that heterogeneity exists in AD progression through different intermediate stages with varied clinical presentations[10–13].

Many factors have been suggested to influence AD progression, such as age, sex[8,14], genetic variations[15–17], educational background[18], environmental exposures[19] and so forth. Basic and clinical research has indicated that sex difference is one of the most critical factors that contribute to its complexity[20–23]. Approximately two-thirds of all existing AD cases are in females[24], and recent studies suggest significant sex differences in clinical severity[25–27], neuropathological characteristics[28,29] and genetic factors[21,30] of AD. It is also reported that for women aged 65, the lifetime risk of developing AD is 21.2%, about twice the risk seen in men[1,31]. Additionally, female is a major risk factor for late-onset AD[32]. One of the reasons for the higher prevalence of AD in women might be their longer average life expectancy[33]. For three progressive pathological stages, they all present sex differences. For example, several studies reported a higher incidence of MCI in males[34–36], while females show a faster rate of cognitive decline when transitioning from normal cognition to MCI. The accelerated cognitive decline in females has partially been attributed to the effects of the APOE ε4 allele[37–41]. Among patients who have not yet been diagnosed with AD, female carriers exhibit more severe brain metabolic slowdown, hippocampal volume reduction, and cortical thinning compared to male carriers[38,42]. APOE ε4 significantly increases brain Aβ deposition and atrophy, and dramatically reduces brain connectivity in the default mode network of females[43–47]. Similarly, women are significantly at greater risk of the development of MCI to AD due to a more rapid speed of cognitive decline. Several biomarkers have been revealed (e.g., Cerebrospinal Fluid (CSF), neuroimage), providing clearer evidence of a negative correlation between the APOE ε4 genotype and female AD patients[48]. For instance, female AD patients with ε4 carriers greatly exhibited higher levels of CSF tau protein compared to male AD ε4 carriers, increasing the risk of cognitive decline leading to AD[38,48]. Another longitudinal study also showed that the ε4 allele caused a significantly higher risk of transition from MCI to AD in women than in men[38]. However, some other studies found women have greater cognitive resilience, though they have increased tau pathology[49,50]. It is suggested that women might be better able to preserve their brain structural properties after exposure to pathological tau[49]. Moreover, male patients have a higher death rate than female patients after the diagnosis of AD, with women having a mean age at death of 89.8 years versus men at an average of 87.3 years[28,51]. Additionally, women aged 75 years or older are more likely to be diagnosed with AD than men[52]. This is probably due to female carriers of APOE ε4 allele increased drastically and at a greater risk of cognitive decline than male carriers[50]. The mounting proof of sex differences in AD above-mentioned highlights the importance of understanding the underlying architecture and development in female and male AD progression.

In the last decade, the development of electronic health record (EHR) systems[53,54] has made it possible to collect large-scale longitudinal and diverse profiles of patients for AD research[55–57]. Typical EHRs consist of a wide variety of critical health events of patients collected through routine care, including diagnostic codes, comorbidities, medication use, laboratory measurements, and other relevant clinical information. Moreover, EHRs include longitudinal follow-up data inpatient or outpatient, offering long-term insights into AD development. These data not only can assist healthcare professionals in personalized treatment and long-term patient management but also support clinical research on disease progression and pathology. EHRs have been effectively utilized to predict patient outcomes and identify disease sub-phenotypes, including AD[58–61]. Xu *et al.*[56] proposed an outcome-oriented model using Long Short-Term Memory (LSTM)[62] to identify progression pathways from MCI to AD, deriving several AD progression subtypes related to comorbidities like cardiovascular diseases. Hinrichs *et al.* developed a multi-kernel learning framework that enables us to predict transitions from MCI to AD, revealing differences in disease progression between MCI converters and stable MCI patients[63]. In sex-specific analyses, Alice T. *et al.*[64] performed comprehensive phenotyping and network analyses, gaining insight into clinical characteristics and sex-specific clinical associations in AD. Additionally, the framework developed by Landi *et al.*[65] identified patient stratification at scale by leveraging deep representation learning

of EHR data to classify dementia subtypes, enhancing large-scale precise disease prediction. Furthermore, Tang *et al.* demonstrated how EHRs and knowledge networks can be leveraged for AD prediction and uncovering sex-specific biological insights[66]. Many previous studies have been hypothesis-driven, analyzing phenotypes of AD patients using clinical data such as neurocognitive tests[67] and neuroimaging[68]. They have focused on specific risk factors associated with AD, such as demographics (e.g., ages[28]), socioeconomics (e.g., exercise and occupation[69]) and comorbidities (e.g., hypertension[70] and vascular disease[71]). However, the role of sex in moderating the complexity and heterogeneity of AD progression remains largely unexplored. These approaches have not accounted for sex differences in AD progression, specifically how sex influences long-term pathological characteristics[72] of AD sub-phenotypes. Characterizing the sub-phenotypes of sex-specific AD progression and identifying the contributing factors is a crucial step for AD stratification and prevention. By revealing and understanding the differences, clinicians can tailor therapeutic strategies to maximize treatment benefits and minimize adverse effects in male and female AD patients.

In this study, we developed a long short-term memory (LSTM)-based autoencoder that can identify sex-specific sub-phenotypes in the progression of AD using large-scale longitudinal EHR data. Unlike traditional approaches that establish the models of clinical characteristics' correlation with patient outcomes, our proposed model learns temporal changes in patient conditions throughout the progression of AD, stratified by sex. Specifically, we identified and collected longitudinal EHRs of AD patients from the OneFlorida+ clinical research consortium. Then, we utilized an autoencoder model based on LSTM units that can automatically learn subsequence temporal features while accounting for sex stratification. We subsequently applied hierarchical agglomerative clustering to group patients into distinct clusters representing disease states or subtypes. We then used chi-square tests and visualization techniques to interpret the sex-based AD progression sub-phenotyping results. Finally, we validated the reproducibility and stability of the identified sub-phenotypes using the Silhouette score and Adjusted Rand Index (ARI). We demonstrated both differences and similarities in disease progression pathways between female and male AD patients. This study contributes to understanding the heterogeneity of AD progression on sex differences, potentially aiding in personalized care and treatment of AD. **Fig. 1** illustrates the overall framework of our study.

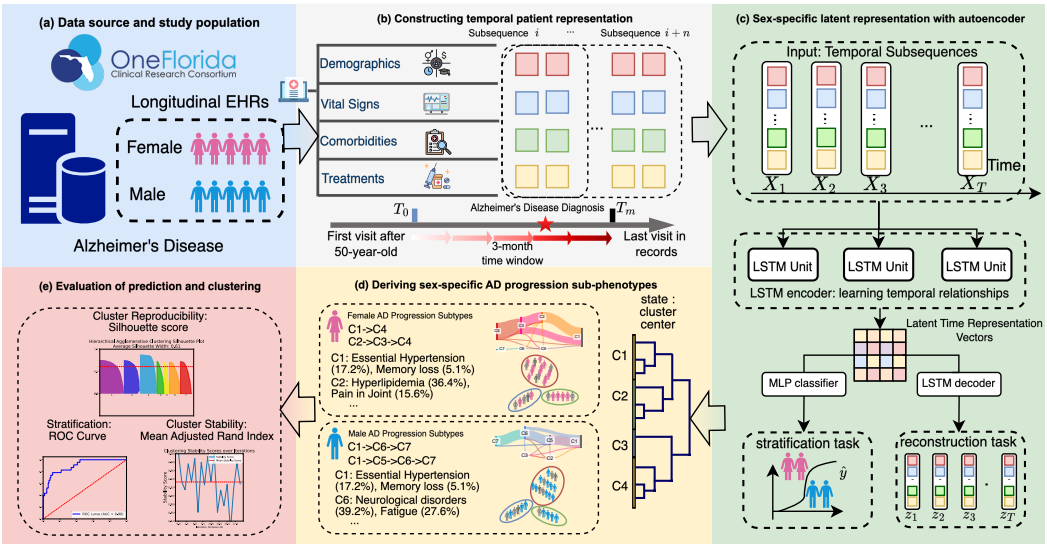


Fig. 1. The overview of the study: (a) Data source and study population; (b) Constructing temporal patient representation; (c) Sex-specific latent representation with autoencoder; (d) Deriving sex-specific AD progression sub-phenotypes; and (e) Evaluation of model and clustering.

Materials and methods

Data source and study population

The large amount of EHR data of AD patients used in this study comes from the OneFlorida+ Clinical Research Consortium[73,74], one of the clinical research networks in the Patient-Centered Outcomes Research Institute-funded National Patient-Centered Clinical Research Network (PCORnet). OneFlorida+, a collaborative clinical network of 14 health institutions, including community health systems, clinics, and academic health institutes, which covers over 20 million patients from Florida (~17 million), Georgia (~2.1 million), and Alabama (~1 million). It consists of diverse EHR-based information adhering to the PCORnet Common Data Model (CDM), including demographics, encounters, procedures, diagnosis, vital signs, and among others, covering massive and heterogeneous patient-centered clinical data. These structured EHRs were collected after January 2012. This study was approved by the University of Florida Institutional Review Board (protocol no. IRB202202820).

We identified the patient cohort by the following criteria: (1) The start of the observation time was defined as the earliest of January 2012; (2) The first visit should be at 50 years of age or older; (3) Patients have at least one diagnosis of AD, i.e., International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes: 331.0 and International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes G30.0, G30.1, G30.8, and G30.9. (4) The patients have at least two encounters. (5) patients have at least one visit record three years before AD diagnosis and over one year of disease progression after AD diagnosis. We excluded the patients with AD diagnosis dates that are not within the period between their earliest and latest visit times.

Variable selection and preprocess

The EHRs information from selected AD patients consists of thousands of different features. We categorize these features into four types, i.e., continuous features (e.g., body mass index and age), binary features (e.g., death), time-to-event features (e.g., diagnosis and treatment) and multi-class features (e.g., smoking status and medication code). We first transformed these features with different strategies based on categories to ensure they can be processed by the classifiers. For instance, we discretized age using bins of identical 10-year size and discretized BMI into four groups: underweight (≤18.5), normal weight (18.5–24.9), overweight (25–29.9), and obesity (≥30). We also transformed smoking status codes into non-smokers, current smokers, ex-smokers, and others. In addition, diagnosis and treatment codes were the main components of our variables. For diagnosis, the codes were mapped to Phecodes[75] for the sake of standardization, an EHR-specific codebase for supporting phenome-wide association studies (PheWAS). For treatment medication codes, using National Drug Codes (NDC) and RxNorm, we mapped them to the unified Anatomical Therapeutic Chemical (ATC) Classification codes[76]. Further, we used multiple imputations by chained equations (MICE) to tackle the missing values. Finally, we concatenated all the features to be a patient-centered binary matrix, excluding sex in the column which was used as the classification label.

We used visualization techniques to see the difference in the distribution of patients' features before representation learning. If there is a change in the sex stratification between the initial patient feature matrix and the latent representations learned by the model, where the later sex stratification becomes more significant, it can indicate the effectiveness of our model in sex stratification. Since we leveraged sub-sequence level latent embedding to unveil sub-phenotypes subsequently, we used sub-sequence level patient features for visualization. The dimension of the subsequence is generally high, so we first use principal component analysis (PCA) dimensionality reduction to visualize the distribution in the main components. In addition to

visualizing PCA components with scatter plots, we used the Mann-Whitney U test to quantitatively assess sex stratification by analyzing the significance of the correlation between sex and the principal components. The Mann-Whitney U test[77] is a non-parametric two-sample t-test used to test the null hypothesis. In this study, for example, we aim to test whether there is no statistically significant difference in the PCA components of subsequences between different sexes. Mann-Whitney U test takes the PCA component from two sexes as parameters, compares their ranks to assess whether they come from the same distribution, and returns the test statistic and p-value to indicate statistical significance. All analyses are conducted at a significance level of $p < 0.01$[78], as this is a commonly used p-value threshold for statistical significance in biomedical research. If $p < 0.01$, we reject the null hypothesis, indicating that the PCA components of subsequences differ significantly between sexes. Additionally, the primary feature in the patient feature matrix is the diagnosis features (1,711, 90.2%). To provide a more focused interpretation of sex differences in AD risk factors, we summarized three common AD comorbidity categories with sex differences based on previous studies[79–81]: neurological/mental disorders, cardiovascular diseases, and diabetes-related conditions (listed in Supplementary Table S1). These three comorbidity categories have been widely studied in previous research on sex differences in AD progression. We calculated the proportion of patients with these comorbidities.

Constructing temporal patient representation using longitudinal EHRs

To construct patients' longitudinal progression trajectory before (3 years) and after (1 year) the onset of AD, we first aggregated multiple measurements or event irregular time points of EHRs and converted them into 3-month blocks as a subsequence of patients. Each block formed a vector representing a specific event type (e.g., diagnoses, medications, etc.). To model the progression patterns, we split each patient into multiple subsequences. Fig. 2 presents the construction of the AD temporal trajectory using EHRs. For example, a diagnosis vector contained distinct diagnosis codes and their frequencies within a 6-month window. For the invariant features (e.g., race and gender), we treated them as static features and passed them at each time window. We selected the earliest patient encounter date 3 years before AD onset as the index date and the latest encounter date 1 year after AD diagnosis as the end date. Given N patients, the total number of time windows T, and the dimension of patient feature matrix D, we constructed the temporal patient representation. The representation is a matrix in which patients can be represented as a 3-tuple symbol and a sequence of vectors for patient i at visit t_j, where $t_j \in \{t_1, t_2, \ldots, t_T\}$ and each visit is denoted as $x_{i,t_j} \in \mathbb{R}^{T \times D}$. To increase the quantity and augment the data, the original data (where each patient corresponds to a temporal matrix) has been divided into multiple subsequences. Starting from the index date t_1 (the date of the first visit), every 2 windows (equivalent to 6 months) form one subsequence, until reaching the maximum length of the patient's EHR data. We then divided each patient into multiple subsequences with varying time lengths, starting from the index date, and new subsequences are created with 3-month increments (i.e., 6-month, 9-month, etc.) until the last encounter of patients within the end date. Each subsequence is treated as an independent temporal matrix sample and inputted into the model. Therefore, the l-th subsequence can be represented as $x_{i,t_j}^{(l_i)}$, in which $l_i \in \{l_1, l_2, \ldots, l_N\}$ is the index of each patient's subsequence. Finally, we got the whole temporal matrix with dimensions from (N, T, D) to $(\sum_{n=1}^{N} l_n, T, D)$. We concatenated these subsequences in different time lengths to construct the time-series AD progression trajectory for progression modeling.

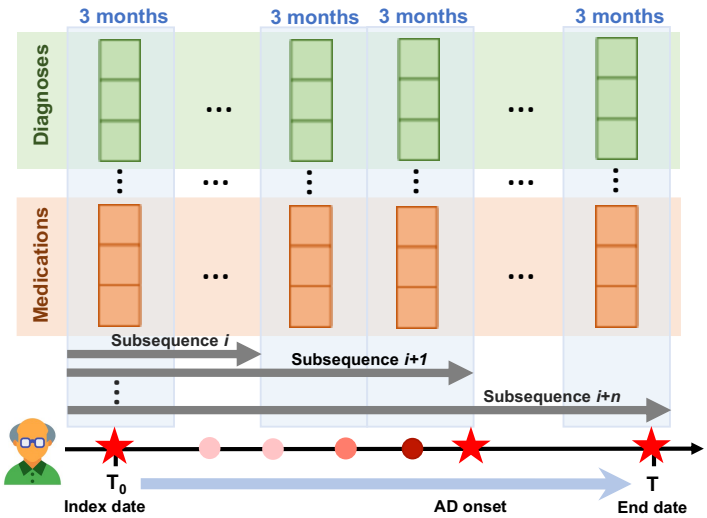


Fig. 2. The construction of AD temporal trajectory using EHRs.

Generating sex-specific AD latent representation with temporal autoencoder

We constructed an LSTM-based autoencoder model[82] to learn the representation of the subsequences of AD patients. This autoencoder consists of a three-layer LSTM encoder for latent temporal representation embedding learning, a two-layer LSTM decoder for reconstructing the subsequences, and a Multilayer Perceptron (MLP) predictor to stratify the patients into two sex groups (Fig. 1c). The multivariate time-series subsequences of each patient as the input were fed into the LSTM encoder cell, which creates a hidden state that represents the latent patient trajectory features. LSTM is particularly effective for handling longitudinal clinical patient data with long-term dependencies. Its internal memory units are capable of capturing features and correlations from the early stages to the later stages of patients. Then, the latent representation matrix from the last layer of the LSTM encoder was simultaneously fed into a two-layer LSTM decoder and an MLP. Firstly, it served the sequence reconstruction task handled by a two-layer LSTM decoder. The first layer of the LSTM decoder expanded the feature dimensions at each time step, while the second layer reconstructed the original sequence. This reconstruction enhanced the learning of effective temporal representations for patients. For the sequence reconstruction task, mean squared error (MSE) was used as the loss function. Secondly, the same latent representations were utilized for a sex-stratified AD classification task using an MLP. Through this classification task, we can stratify the latent representations learned by the LSTM encoder based on sex during training. For this task, we used the binary cross entropy (BCELoss) as the loss function and the area under the receiver operating characteristic curve (AUROC) as the predictive performance evaluation. The model was trained jointly using Adam optimizer, summing the MSE and BCELoss for optimization.

Deriving sex-specific AD progression sub-phenotypes by unsupervised clustering

To identify sex-specific AD progression sub-phenotypes, we first extracted the sex-specific AD progression embedding matrix learned by the LSTM encoder to clustering models. After obtaining temporal representations of each subsequence from the LSTM encoder's outputs involving sex-stratified clinical information of AD patients, we then used several clustering algorithms to identify them into different clusters (or states) that exhibit similar temporal properties. Here, we tested 4 types of clustering methods, including k-means, kernel k-means, density-based spatial clustering of applications with noise (DBSCAN), and hierarchical agglomerative clustering. We used hierarchical agglomerative clustering as an example to describe the clustering process. Initially, each subsequence can be regarded as a patient sample with varying-length clinical information.

Distances (or similarities) between all pairs of subsequences are computed, and the closest pair of subsequences is merged. The representation of clusters is updated, for instance, by using the average to represent the newly merged cluster. This process was repeated until all subsequences were merged into one cluster or reached the stopping criterion (e.g., a predetermined number of clusters was met). Different clusters represent groups of subsequences, with each cluster center representing a distinct patient state. To derive the AD progression sub-phenotypes, we assume that a patient has four subsequences containing clinical information with different time lengths (i.e., 3, 6, 9, and 12 months). Each subsequence can be grouped into a cluster (i.e., state) C_i (i = 1, 2, 3 …), assuming 3-month to C1, 6-month to C2, 9-month to C3 and 12-month to C4. The progression trajectory of this patient would be represented as "C1->C2->C3->C4", which will be categorized as one of the sex-specific AD progression sub-phenotypes.

Clustering reproducibility and stability evaluation

We evaluated the effectiveness of our identified clusters by analyzing their reproducibility and stability. Cluster reproducibility can evaluate the extent of separation and distinctiveness among the clusters, and stability can measure whether the clustering generated by the model is stable over multiple iterations. Regarding reproducibility, we employed the silhouette score[83], which can evaluate the quality of clustering by measuring how each data point compares to its own cluster versus other clusters. A higher silhouette score indicates better cluster separation, where data points are closer to their own cluster center than to the neighboring clusters. Notably, if the silhouette scores close to zero, it implies the data points are on the cluster boundaries, whereas negative scores suggest the data points are within another cluster, which is potentially incorrectly clustered. We trained the model with training data (i.e., from OneFlorida+ sites: source 1, source 3, source 11, N=1468 (88.2%)) and applied it to validation data (i.e., from OneFlorida+ sites: source 9, source 10, source 12, source 15, N=197 (11.8%)), and compared the average clustering silhouette scores. For the four clustering methods used, we also experimented on different cluster numbers, ranging from 1 to 20, for comparison. Further, we tested the stability of clustering models on the clustering model that has the best reproducibility performance. It is assessed if consistent cluster results can be reliably generated in iterative clustering. We utilized random subsets of data in each iteration and examined fluctuations in the Adjusted Rand Index (ARI) (a metric for measuring similarity between two clustering results)[84]. Higher ARI shows smaller fluctuations in changes, suggesting these clusters are more stable.

Results

Descriptive statistics and visualization of sex-stratified AD patients

Table 1 shows the statistics of the study AD cohorts stratified by sex. As shown in the table, females have a slightly larger proportion of AD patients than males (57.7% versus 42.3%) and are older at AD diagnosis (77.1 versus 76.1 years). Females also show a longer total disease duration than males (2,998 versus 2,903 days). This is probably because female patients have a longer life expectancy, which suggests a slightly higher risk of developing AD. Regarding mortality rates, males show a higher mortality rate (14.20%) compared to females (11.34%). In addition, we could observe racial differences, including a higher percentage of Hispanic patients among females (25.81%) compared to males (18.52%), and a higher proportion of Black or African American and White patients among males (16.3% and 79.12%) compared to females (14.5% and 77.11%), respectively. In terms of comorbidities, neurological/mental disorders are prevalent in a large majority of AD patients, with 89.1% of the total cohort. We observed that 91.2% of female AD patients have neurological/mental disorders, slightly higher than males of 86.2%. Meanwhile, cardiovascular diseases and diabetes-related conditions both present a higher prevalence in males (46.3% and 48.3%) compared to females (38.5% and 45.7%) in our cohorts. More details can be found in Supplementary Table S2.

Table 1. Descriptive statistics on the characteristics of the study cohort

	Total AD Patients (N = 1665)	Female AD Patients (N = 961)	Male AD Patients (N = 704)
Demographics and Vital Signs			
Age at AD diagnosis, mean (std)	76.7 (9.2)	77.1 (9.5)	76.1 (8.8)
Female, N (%)	961 (57.7%)	961 (100.0%)	0 (0.0%)
Male, N (%)	704 (42.3%)	0 (0.0%)	704 (100.0%)
Disease development, mean (std)			
Duration days	2958.1 (620.2)	2998.0 (613.7)	2903.2 (627)
Age at death	81.7 (8.9)	82.1 (9.4)	81.2 (8.4)
Mortality rate, N (%)	209 (12.55%)	109 (11.34%)	100 (14.20%)
Hispanic, N (%)			
Hispanic	426 (25.6%)	248 (25.8%)	178 (18.5%)
Not Hispanic	1230 (73.9%)	707 (73.6%)	523 (54.4%)
No Hispanic information	9 (0.5%)	6 (0.6%)	3 (0.3%)
Race, N (%)			
American Indian or Alaska Native	0 (0.0%)	0 (0.0%)	0 (0.0%)
Asian	10 (0.6%)	5 (0.5%)	5 (0.7%)
Black or African American	259 (15.6%)	157 (16.3%)	102 (14.5%)
White	1298 (77.96%)	741 (77.11%)	557 (79.12%)
Multiple races	15 (0.90%)	6 (0.62%)	9 (1.28%)
Unknown	83 (5.0%)	52 (5.4%)	31 (4.40%)
Comorbidity, **N (%)**			
Neurological/mental disorders	1483 (89.1%)	876 (91.2%)	607 (86.2%)
Cardiovascular diseases	696 (41.8%)	370 (38.5%)	326 (46.3%)
Diabetes-related conditions	779 (46.8%)	439 (45.7%)	340 (48.3%)

Due to the size of the patient subsequence-level features, we first utilized low-dimensional PCA visualization to plot patients' initial feature matrix colored by sex, shown in Fig. 3a. In this figure, each point represents one subsequence of the patient (1,897 features). From this figure, we can see that there is no clear separation between female and male AD patients. Then, we used violin plots and performed the Mann-Whitney U-test to check the distribution and significant differences between sex groups and two PCA components. For PCA component 1, it is noteworthy that the Mann-Whitney U-test p-value is 6.1e-23 (p-value < 0.01) in Fig. 3b, indicating a significant difference between sexes. This rejected our null hypothesis and suggested that there might be a significant difference in the initial low-dimensional patient subsequences between females and males. However, in the second component (Fig. 3c), the Mann-Whitney U-test p-value is 0.09 (p-value >= 0.01), in which we did not find significant differences which explain the overlap of sex groups before stratification.

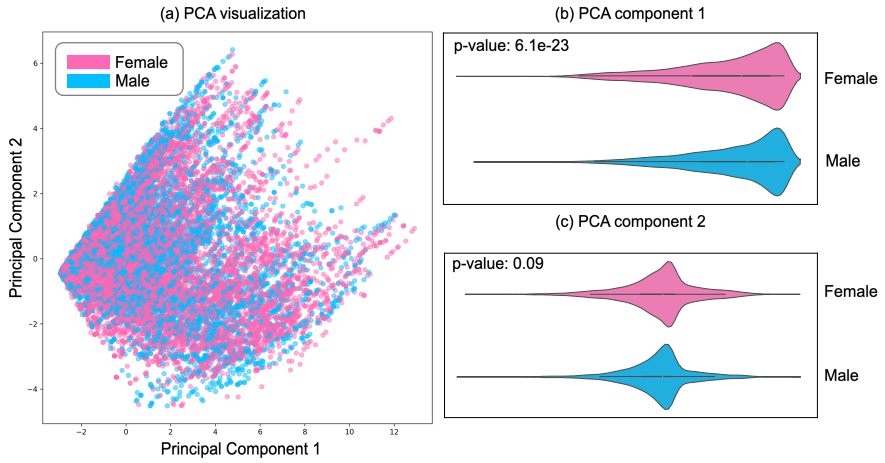


Fig. 3. Visualization of subsequence-level features. (a) PCA visualization of all patients' subsequence-level feature matrix with each dot representing a subsequence from one patient colored by sex (females and males); Violin plot shows the distribution of female and male patients along the PCA principal component 1 (b) (p-values 6.1e-23 from Mann-Whitney test); and principal component 2 (c) (p-values 0.09 from Mann-Whitney test).

Performance of prediction model, cluster reproducibility and stability

For the evaluation of the sex-stratified identification of AD patients, we achieved 0.903 of AUROC, showing a strong capability for prediction. Then, regarding the evaluation of the clustering's reproducibility and stability. As shown in Fig. 4a, we achieved a silhouette score of 0.61 for hierarchical agglomerative clustering, 0.53 for k-means clustering, 0.08 for kernel k-means clustering, and 0.45 for DBSCAN clustering, where the x-axis is the cluster index, and the y-axis is the silhouette score of each cluster. These plots indicate that the hierarchical agglomerative cluster analysis achieved the best performance, which can identify the clusters from the latent embeddings effectively. As for the cluster stability, we calculated and visualized hierarchical agglomerative clustering ARI results, in which the ARI slightly fluctuated around 0.93 over random sub-dataset sample sizes (Fig. 4b). The stability analysis indicates that consistent clustering results can be reliably generated, even when random subsets of data are used. These findings demonstrate that the hierarchical agglomerative clustering model is stable and can be used in the following experiments, and the identified clusters are not statistical artifacts. The evaluation of reproducibility and stability of clustering is crucial for ensuring that the identified clusters are meaningful and reproducible in different situations.

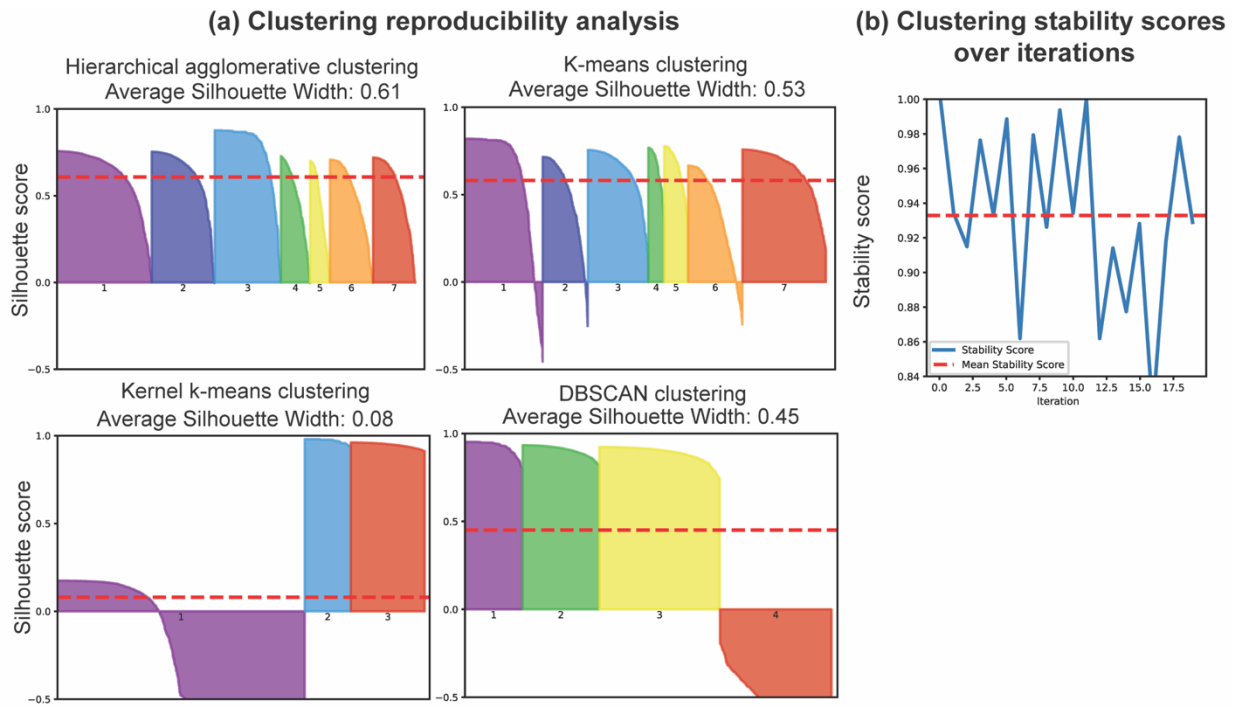


Fig. 4. Cluster Reproducibility and Stability Performance. (a) Clustering reproducibility analysis; (b) Clustering stability scores over iterations.

Clusters of sex-stratified subsequences from AD patients

The hierarchical agglomerative clustering generated seven different clusters from all subsequences of AD patients. We denoted these derived clusters (i.e., states) as C1 to C7. Fig. 5a (left) shows the visualization of the clusters using PCA, which presents the clusters of learned representation of AD patients by sex. We can find a clear sex stratification of AD patients through learned latent embeddings compared with Fig. 3a. Like Fig. 3b and Fig. 3c, we examined the correlation of PCA principal components and sex groups using the Mann-Whitney U test. We also used violin plots to visualize the distribution of latent embeddings' PCA principal components in Fig. 5a (left). The PCA principal component 1 exhibited a greater significant difference in sex groups (p-value = 4.3e-152) compared to Fig. 3b (p-value = 6.1e-23). PCA principal component 2 also showed significant differences in sex groups (p-value = 3.2e-19) compared to Figure 3c (p-value = 0.09). The right plot of Fig. 5a displays the identified clusters (C1-C7) of subsequences of these AD patients. We could observe that the cluster C4 (N=9,537, 34.9%) involves most subsequences of AD patients, followed by C1 (N=4,726, 17.3%), C7 (N=4,583, 16.8%), C6 (N=3,743, 13.7%), C3 (N=2,498, 9.1%), C5 (N=1,258, 4.6%), and C2 (N=983, 3.6%). To gain insights into the features that differentiate these clusters, the heatmap in Fig. 5b shows the patient percentages of the top features in each cluster (i.e., state). For better visualization and interpretation, we ordered the percentages by the values in C1 and colored the names of features by our summarized three main comorbidity categories (listed in Supplementary Table S1). We can see that essential hypertension is the most prevalent disease among all the clusters (e.g., 82.7% in C4, 89.6% in C7), with a particularly high occurrence in C4 and C7. Hyperlipidemia is the next most common feature in AD patients (e.g., 72.5% in C4, 79.5% in C7). Neurological disorders and memory loss also account for a significant proportion (e.g., neurological disorders: 62.8% in C7, memory loss: 51.3% in C7). Chronic pains, such as joint, back, and abdominal pain, are also common comorbidities in the states among AD patients. Moreover, compared with the phenotypic features in cluster C4, there is a larger proportion of severe comorbidities in C7 (e.g., 0.827 versus 0.896 in essential hypertension and 0.725 versus 0.795 in hyperlipidemia). The differences between these sex-specific cluster pairs (e.g., C4 and C7) could be crucial in understanding the sex-specific progression patterns of AD. Additionally, we could also observe the differences in the demographics of these clusters. For example, the ratio of female subsequences in clusters C3 and C4 is significantly higher than male, with 90.1% and 88.0%, respectively, whereas the male subsequences have a larger proportion in clusters in C6 (96.7%) and C7 (75.4%). This might indicate that C3 and C4 are the female-dominated states, while C6 and C7 are the male-dominated AD progression states. We also noticed that clusters C2 (2981.1 days) and C5 (2913.8 days) have shorter durations among all clusters, while C4 and C7 show longer durations in AD development. More details can be found in Supplementary Table S3.

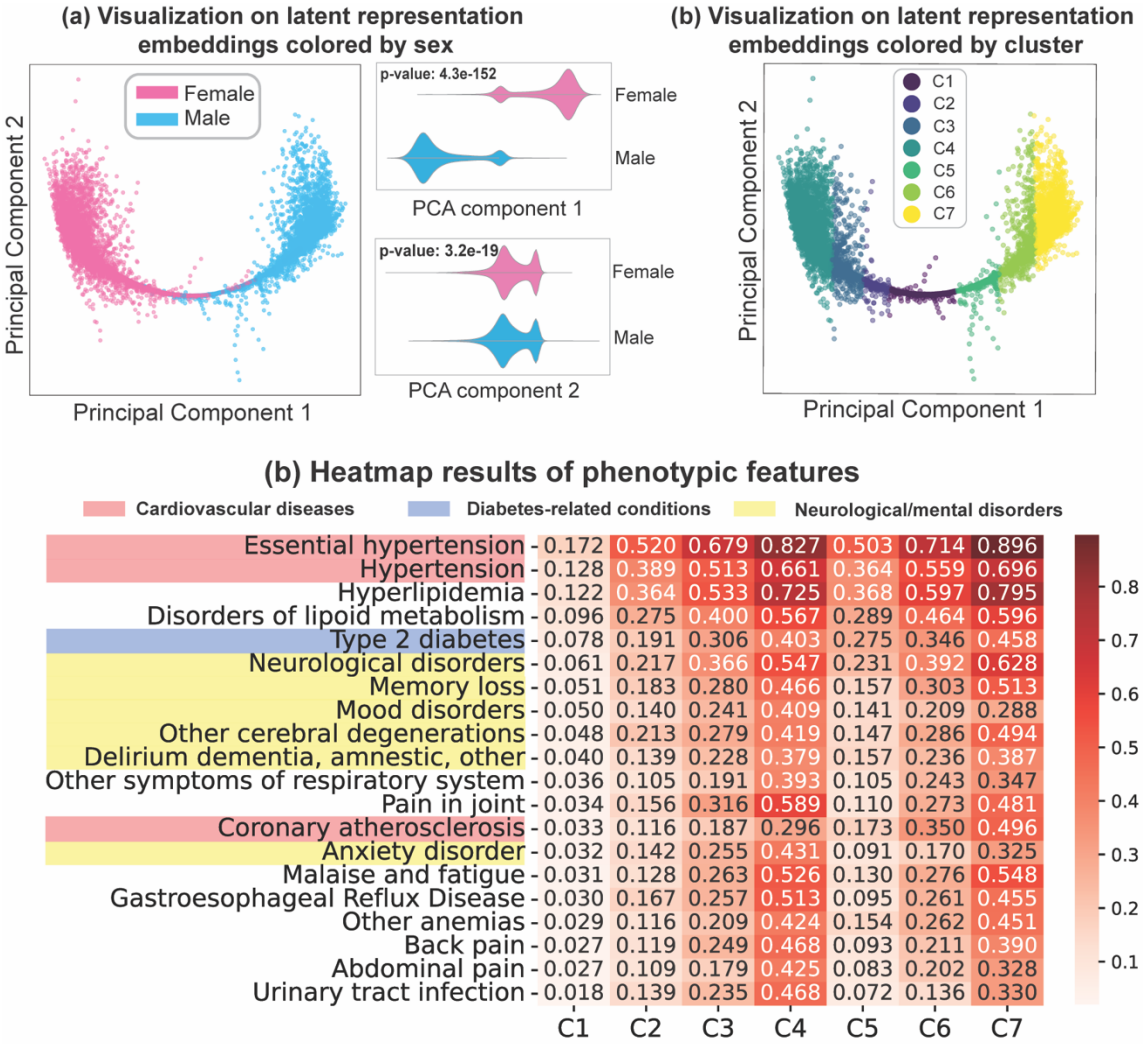


Fig. 5. The visualization and clustering results. (a) The visualization of latent representation embeddings by sex and violin plots show the distribution of female and male patients along the PCA principal component 1 (top) (p-values 4.3e-152 from Mann-Whitney test) and PCA principal component 2 (bottom) (p-values 3.2e-19 from Mann-Whitney test); (b) The visualization of latent representation embeddings by cluster; (c) Heatmap results of top 20 phenotypic features in each generated cluster (i.e., state) in AD patients.

Interpretability of identified sex-specific AD progression sub-phenotypes

We identified five primary AD progression sub-phenotypes by the split of subsequences' clusters derived from each patient illustrated in Fig. 6a, represented as S1 to S5 that involve distinct progression pathways: (1) S1: C1->C2->C3->C4, (2) S2: C1->C3->C4, (3) S3: C1->C5->C6->C7, (4) S4: C1->C6->C7, and (5) S5: C1->C5->C6. According to Fig. 6a, we could find that S1 and S2 are female-dominant sub-phenotypes, with 406 and 326 female patients, whereas there were only 7 and 3 male patients, respectively. On the opposite, the sub-phenotypes S3 to S5 are male-dominant. Further observations suggest (Fig. 5b and Supplementary Table S3), in the sub-phenotype S1, the state C4 had the larger proportion of comorbidity along the progression pathway C1->C2->C3->C4 (e.g., essential hypertension 82.7%) compared with other states. It also showed a longer duration of disease (i.e., 3,036.6 days) in AD development, followed by C3 (67.9%, 3,007.3 days), C2 (52.0%, 2,981.1 days), and C1 (17.2%, 2,958.3 days). This might suggest C4 is a progression state relatively associated with slow AD progression, while patients in C1 progressed more rapidly. Similarly, we found that C5 had the shortest disease duration days

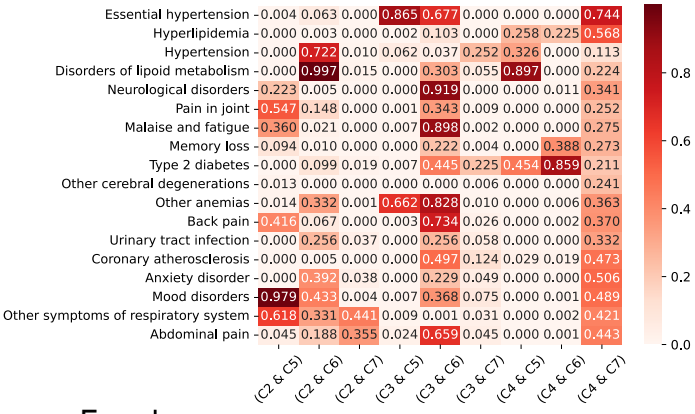
(2,913.8 days) in male-dominant sub-phenotype S3. Additionally, there are three other sub-phenotypes with progression pathways: C1->C3->C4, C1->C6->C7, and C1->C5->C6, respectively.

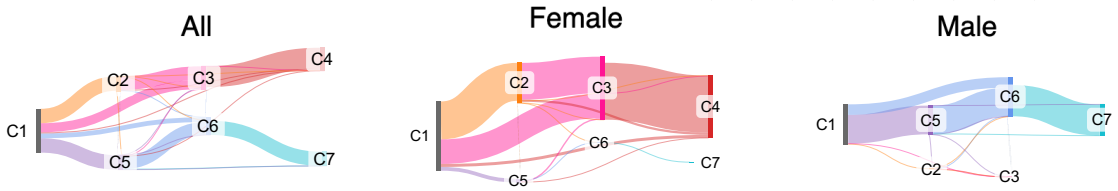
Interestingly, we noticed that the states C2, C3, and C4 were uniquely associated with female-dominant sub-phenotypes S1 and S2, whereas C5, C6, and C7 only existed in male-dominant sub-phenotypes. To understand the differences between these states in sex-specific AD sub-phenotypes, we conducted chi-square tests on paired states across clinical phenotypes. The p-values of these tests were shown in Fig. 6b, where we could see significant differences in most comorbidities among compared state pairs, like hyperlipidemia and other cerebral degenerations in C2 & C7, C3 & C5, and C3 & C7 with p-value < 0.05. These findings indicate these pairs of states could be crucial in understanding the AD progression sub-phenotypes between females and males. We also used Sankey diagrams to illustrate the progression pathways and transitions between sex-specific AD sub-phenotypes. In Fig. 5 (c), these diagrams show the main transitions from C1 to other states. The width of the Sankey diagram represents the size of patients transiting from one state to another, where pathways S1 and S3 are the two groups with the largest number of patients. The flows in the diagram further validated the differences in the progression in male and female AD patients with distinct sub-phenotypes.

(a) Patients by sub-phenotype patterns

Progression Pathway	# Female	# Male
[C1->C2->C3->C4]	406	7
[C1->C3->C4]	326	3
[C1->C5->C6->C7]	1	345
[C1->C6->C7]	6	172
[C1->C5->C6]	2	135

(b) Heatmap results of clusters





(c) Sankey diagrams of sub-phenotype pathways

Fig. 6. Sex-specific AD progression sub-phenotypes, comorbidity significant difference and pathway structures. (a) Patients by sub-phenotype patterns; (b) Heatmap results of cluster analysis; (c) Sankey diagrams of AD sub-phenotype progression pathways.

Discussion

In this study, we developed an LSTM-based auto-encoder model to identify sex-specific AD progression sub-phenotypes using longitudinal EHRs and extract associated clinical characteristics. Firstly, the OneFlorida+ Clinical Research Consortium enabled us to collect large-scale and heterogeneous clinical information from AD patients, including diagnosis, lab measurements, medication, etc. We then constructed temporal subsequences extracted from AD

patients to represent their progression trajectory. We achieved our goal of deriving sex-specific sub-phenotypes by the combination of an LSTM-based autoencoder with clustering techniques. The autoencoder generated sex-based embeddings of temporal subsequences representing various clinical information of AD patients. We then applied hierarchical agglomerative clustering to group these subsequences into different clusters (i.e., states), and then we obtained all the states of a patient's subsequences and concatenated these states as a progress trajectory (or sub-phenotype), from which we can extract and reveal the clinical characteristics for each sub-phenotype. Here, we uncovered five primary sex-specific AD sub-phenotypes (Fig. 6b). We evaluated our model's learned representation embeddings by AUROC and clustering reproducibility and stability using silhouette and ARI score, which demonstrated a decent performance with robustness for both. In this work, we emphasized the heterogeneity of AD progression and the importance of considering sex factors in disease modeling and treatment. The identified sub-phenotypes would not only present different trajectories of AD progression but also reveal differences in characteristics and comorbidities between male and female patients.

In our cohort, female patients showed a larger proportion of AD patients, with an older age of diagnosis and a longer time of disease development on average. The mortality rate for males (14.20%) was higher than that for females (11.34%). These characteristics are consistent with previous studies[1,33,85]. In our clustering results, we found 7 clusters (or states), in which cluster C4 included the largest number of patient subsequences. When we looked at the clinical characteristics of these clusters, we observed that essential hypertension was the most prevalent comorbidity among all the clusters, followed by hyperlipidemia (79.5%), hypertension (69.6%), neurological disorders (62.8%), and memory loss (51.3%). Moreover, in the identified five primary AD sub-phenotypes, two of them were female-dominant (S1 and S2), and the other three were male-dominant (S3, S4, and S5) sub-phenotypes, with distinct progression pathways for each sub-phenotype. (Fig. 6a). According to the results, the age of AD diagnosis in females was highest in cluster C4 within sub-phenotype S1, and C4 also had the largest proportion of comorbidities, such as essential hypertension and hyperlipidemia. We further observed that these sub-phenotypes encompassed varying states of rapid or slow disease progression. For example, compared with S3 (C1->C5->C6->C7), there is no relatively rapid progression state like C5 in S4 (C1->C6->C7). Based on the analysis of comorbidities among AD patients in different sex-specific sub-phenotypes, we found neurological/mental disorders, cardiovascular diseases, and diabetes-related conditions were more prevalent among female AD patients. Additionally, we compared the differences of paired clusters that uniquely existed in female or male dominant sub-phenotypes, such as C2 & C7, C3 & C5, and C3 & C7 (Fig. 6b). We discovered obvious distinctions in some comorbidities, including cardiovascular diseases (e.g., essential hypertension), diabetes-related conditions (e.g., type 2 diabetes), neurological/mental disorders (e.g., cerebral degeneration, memory loss), and chronic pain (e.g., pain of the joint). These findings suggest that AD progression varies significantly in terms of different sex groups, and our identified AD sub-phenotypes can provide an in-depth understanding of sex-specific disease progression pathways.

This study, while insightful in exploring and identifying sex differences of AD sub-phenotypes through advanced AI techniques, has several limitations that should be acknowledged. First, the reliance on EHR from one specific research network may introduce biases due to the demographic and geographic characteristics of the population. The patients in OneFlorida+ predominantly from Florida, Georgia, and Alabama, might not fully represent the broader, more diverse U.S. population or other global populations. The generalizability of the findings should be further validated with heterogeneous demographic profiles. Second, while the clustering model showed stability and reproducibility, the interpretation of the sex-specific AD sub-phenotypes remains challenging. The clinical significance of these clusters, especially their specific meaning

and potential utility in guiding personalized treatment plans, needs validation through prospective studies and clinical trials. The study's design limits the ability to identify detailed causal inferences about the progression pathways and their implications for disease development. Third, although we discovered sex-specific sub-phenotypes using data-driven methods, this study primarily relies on routine EHR data, lacking detailed cognitive assessment (e.g., the Mini-Mental State Exam) and biomarker data (e.g., brain imaging or cerebrospinal fluid analysis), which are crucial in AD progression and sex difference study. The incorporation of these data categories should be considered in future investigations, which can provide a more comprehensive and valuable insight into the potential mechanisms driving sex differences in AD progression and develop personalized therapeutic interventions.

Future research would focus on, firstly, validating these sub-phenotypes in larger and more diverse populations to enhance the generalizability of our findings. This underscores the need to adopt methods such as federated learning techniques[86] that can scale across large heterogeneous datasets. Secondly, exploring the underlying biological mechanisms driving these sex differences in AD progression, which is crucial for developing sex-stratified targeted therapies. This highlights the necessity of employing multimodal and multi-omics data and research approaches[87]. For example, incorporating the genetic data of male and female AD patients through whole genome or exome sequencing analyses can identify potential differential genes and variants associated with clinical manifestation in AD development. Thirdly, developing more robust and powerful computational models to precisely and quantitatively trace and identify phenotypic changes in the development of AD for patients, which can provide timely monitoring and feedback for disease management.

Supplementary Materials

The codes and supplementary materials are publicly available at https://github.com/UF-HOBI-Yin-Lab/AD_sex_subtype.

Acknowledgments

This study was partially supported by grants from Centers for Disease Control and Prevention (1U18DP006512), National Institute of Environmental Health Sciences (R21ES032762) and the NIH National Center for Advancing Translational Sciences (UL1TR001427).

Reference

1. 2024 Alzheimer's disease facts and figures. *Alzheimers. Dement.* **20**, 3708–3821 (2024).
2. Rajan, K. B. *et al.* Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimers. Dement.* **17**, 1966–1975 (2021).
3. Sperling, R. A. *et al.* Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers. Dement.* **7**, 280–292 (2011).
4. Albert, M. S. *et al.* The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Focus (Am. Psychiatr. Publ.)* **11**, 96–106 (2013).
5. Sun, B.-L. *et al.* Clinical research on Alzheimer's disease: Progress and perspectives. *Neurosci. Bull.* **34**, 1111–1118 (2018).
6. 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. Inform.* **107**, 103462 (2020).

7. Lee, C. & van der Schaar, M. Temporal phenotyping using deep predictive clustering of disease progression. *arXiv [physics.med-ph]* (2020).
8. Li, J.-Q. *et al.* Risk factors for predicting progression from mild cognitive impairment to Alzheimer's disease: a systematic review and meta-analysis of cohort studies. *J. Neurol. Neurosurg. Psychiatry* **87**, 476–484 (2016).
9. Jamalian, S. *et al.* Modeling Alzheimer's disease progression utilizing clinical trial and ADNI data to predict longitudinal trajectory of CDR-SB. *CPT Pharmacometrics Syst. Pharmacol.* **12**, 1029–1042 (2023).
10. Stallard, E. *et al.* Estimation and validation of a multiattribute model of Alzheimer disease progression. *Med. Decis. Making* **30**, 625–638 (2010).
11. Lam, B., Masellis, M., Freedman, M., Stuss, D. T. & Black, S. E. Clinical, imaging, and pathological heterogeneity of the Alzheimer's disease syndrome. *Alzheimers. Res. Ther.* **5**, 1 (2013).
12. Goyal, D. *et al.* Characterizing heterogeneity in the progression of Alzheimer's disease using longitudinal clinical and neuroimaging biomarkers. *Alzheimers Dement. (Amst.)* **10**, 629–637 (2018).
13. Young-Pearse, T. L., Lee, H., Hsieh, Y.-C., Chou, V. & Selkoe, D. J. Moving beyond amyloid and tau to capture the biological heterogeneity of Alzheimer's disease. *Trends Neurosci.* **46**, 426–444 (2023).
14. Hersi, M. *et al.* Risk factors associated with the onset and progression of Alzheimer's disease: A systematic review of the evidence. *Neurotoxicology* **61**, 143–187 (2017).
15. Cuyvers, E. & Sleegers, K. Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. *Lancet Neurol.* **15**, 857–868 (2016).
16. Naj, A. C., Schellenberg, G. D. & for the Alzheimer's Disease Genetics Consortium (ADGC). Genomic variants, genes, and pathways of Alzheimer's disease: An overview. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **174**, 5–26 (2017).
17. Reitz, C. & Mayeux, R. Use of genetic variation as biomarkers for mild cognitive impairment and progression of mild cognitive impairment to dementia. *J. Alzheimers. Dis.* **19**, 229–251 (2010).
18. Isaacson, R. S. *et al.* The clinical practice of risk reduction for Alzheimer's disease: A precision medicine approach. *Alzheimers. Dement.* **14**, 1663–1673 (2018).
19. Calderón-Garcidueñas, L. Smoking and cerebral oxidative stress and air pollution: A dreadful equation with particulate matter involved and one more powerful reason not to smoke anything! *Journal of Alzheimer's disease: JAD* vol. 54 109–112 (2016).
20. Dumitrescu, L., Mayeda, E. R., Sharman, K., Moore, A. M. & Hohman, T. J. Sex Differences in the Genetic Architecture of Alzheimer's Disease. *Curr. Genet. Med. Rep.* **7**, 13–21 (2019).
21. Dumitrescu, L. *et al.* Sex differences in the genetic predictors of Alzheimer's pathology. *Brain* **142**, 2581–2589 (2019).
22. Zhu, D., Montagne, A. & Zhao, Z. Alzheimer's pathogenic mechanisms and underlying sex difference. *Cell. Mol. Life Sci.* **78**, 4907–4920 (2021).
23. Guo, L., Zhong, M. B., Zhang, L., Zhang, B. & Cai, D. Sex Differences in Alzheimer's Disease: Insights From the Multiomics Landscape. *Biol. Psychiatry* **91**, 61–71 (2022).
24. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimers. Dement.* **15**, 321–387 (2019).
25. Barnes, L. L. *et al.* Sex differences in the clinical manifestations of Alzheimer disease pathology. *Arch. Gen. Psychiatry* **62**, 685–691 (2005).
26. Fan, C. C. *et al.* Sex-dependent autosomal effects on clinical progression of Alzheimer's disease. *Brain* **143**, 2272–2280 (2020).
27. Kamizato, C., Osawa, A., Maeshima, S. & Kagaya, H. Activity level by clinical severity and sex differences in patients with Alzheimer disease and mild cognitive impairment. (2023).

28. Oveisgharan, S. *et al.* Sex differences in Alzheimer's disease and common neuropathologies of aging. *Acta Neuropathol.* **136**, 887–900 (2018).
29. Ullah, M. F. *et al.* Impact of sex differences and gender specificity on behavioral characteristics and pathophysiology of neurodegenerative disorders. *Neurosci. Biobehav. Rev.* **102**, 95–105 (2019).
30. Eissman, J. M. *et al.* Sex differences in the genetic architecture of cognitive resilience to Alzheimer's disease. *Brain* **145**, 2541–2554 (2022).
31. *Gender and Incidence of Dementia in the Framingham Heart Study from Mid -Adult Life*.
32. Farrer, L. A. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. *JAMA* **278**, 1349 (1997).
33. *With Chartbook on Long-Term Trends in Health*. (United States; Hyattsville, MD).
34. Koivisto, K. *et al.* Prevalence of age-associated memory impairment in a randomly selected population from eastern Finland. *Neurology* **45**, 741–747 (1995).
35. Ganguli, M., Dodge, H. H., Shen, C. & DeKosky, S. T. Mild cognitive impairment, amnestic type: an epidemiologic study. *Neurology* **63**, 115–121 (2004).
36. *Prevalence of Mild Cognitive Impairment Is Higher in Men: The Mayo Clinic Study of Aging*.
37. *Effects of Age, Sex, and Ethnicity on the Association between Apolipoprotein E Genotype and Alzheimer Disease: A Meta-Analysis*.
38. Altmann, A., Tian, L., Henderson, V. W., Greicius, M. D. & Alzheimer's Disease Neuroimaging Initiative Investigators. Sex modifies the APOE-related risk of developing Alzheimer disease. *Ann. Neurol.* **75**, 563–573 (2014).
39. Kim, J., Basak, J. M. & Holtzman, D. M. The role of apolipoprotein E in Alzheimer's disease. *Neuron* **63**, 287–303 (2009).
40. Fatemi, F. *et al.* Sex differences in cerebrovascular pathologies on FLAIR in cognitively unimpaired elderly. *Neurology* **90**, e466–e473 (2018).
41. Ungar, L., Altmann, A. & Greicius, M. D. Apolipoprotein E, gender, and Alzheimer's disease: an overlooked, but potent and promising interaction. *Brain Imaging Behav.* **8**, 262–273 (2014).
42. Sampedro, F. *et al.* APOE-by-sex interactions on brain structure and metabolism in healthy elderly controls. *Oncotarget* **6**, 26663–26674 (2015).
43. Fleisher, A. Sex, apolipoprotein E ε4 status, and hippocampal volume in mild cognitive impairment. *Arch. Neurol.* **62**, 953 (2005).
44. Damoiseaux, J. S. *et al.* Gender modulates the APOE ε4 effect in healthy older adults: convergent evidence from functional brain connectivity and spinal fluid tau levels. *J. Neurosci.* **32**, 8254–8262 (2012).
45. Mosconi, L. *et al.* Sex differences in Alzheimer risk: Brain imaging of endocrine vs chronologic aging. *Neurology* **89**, 1382–1390 (2017).
46. Lopez-Lee, C., Torres, E. R. S., Carling, G. & Gan, L. Mechanisms of sex differences in Alzheimer's disease. *Neuron* **112**, 1208–1221 (2024).
47. Belloy, M. E. *et al.* APOE genotype and Alzheimer disease risk across age, sex, and population ancestry. *JAMA neurology* vol. 80 1284–1294 (2023).
48. Hohman, T. J. *et al.* Sex-specific association of apolipoprotein E with cerebrospinal fluid levels of tau. *JAMA Neurol.* **75**, 989 (2018).
49. Ossenkoppele, R. *et al.* Assessment of demographic, genetic, and imaging variables associated with brain resilience and cognitive resilience to pathological tau in patients with Alzheimer disease. *JAMA Neurol.* **77**, 632–642 (2020).
50. Digma, L. A. *et al.* Women can bear a bigger burden: ante- and post-mortem evidence for reserve in the face of tau. *Brain Commun.* **2**, fcaa025 (2020).
51. Dubal, D. B. Sex difference in Alzheimer's disease: An updated, balanced and emerging perspective on differing vulnerabilities. *Handb. Clin. Neurol.* **175**, 261–273 (2020).

52. Kovacs, G. G. *et al.* Mixed brain pathologies in dementia: the BrainNet Europe consortium experience. *Dement. Geriatr. Cogn. Disord.* **26**, 343–350 (2008).
53. Barbieri, C., Neri, L., Stuard, S., Mari, F. & Martín-Guerrero, J. D. From electronic health records to clinical management systems: how the digital transformation can support healthcare services. *Clin. Kidney J.* **16**, 1878–1884 (2023).
54. Woldemariam, M. T. & Jimma, W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. *BMJ Health Care Inform* **30**, (2023).
55. Meng, W. *et al.* An interpretable population graph network to identify rapid progression of Alzheimer's disease using UK Biobank. *medRxiv* (2024) doi:10.1101/2024.03.27.24304966.
56. Xu, J. *et al.* Identification of Outcome-Oriented Progression Subtypes from Mild Cognitive Impairment to Alzheimer's Disease Using Electronic Health Records. *medRxiv* (2023) doi:10.1101/2023.07.27.23293270.
57. Wu, P.-F. *et al.* Growth differentiation factor 15 is associated with Alzheimer's disease risk. *Front. Genet.* **12**, 700371 (2021).
58. Kumar, S., Abrams, Z., Schindler, S., Ghoshal, N. & Payne, P. Identifying Dementia Subtypes with Electronic Health Records. *arXiv [cs.LG]* (2022).
59. Xu, J. *et al.* Identification of outcome-oriented progression subtypes from mild cognitive impairment to Alzheimer's disease using electronic health records. *AMIA Annu. Symp. Proc.* **2023**, 764–773 (2023).
60. Xu, J. *et al.* Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn. Health Syst.* **4**, e10246 (2020).
61. Li, Q. *et al.* Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. *Alzheimers. Dement.* **19**, 3506–3518 (2023).
62. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
63. Hinrichs, C., Singh, V., Xu, G., Johnson, S. C. & Alzheimers Disease Neuroimaging Initiative. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* **55**, 574–589 (2011).
64. Tang, A. S. *et al.* Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. *Nat. Commun.* **13**, 675 (2022).
65. Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit. Med.* **3**, 96 (2020).
66. Tang, A. S. *et al.* Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. *Nat Aging* **4**, 379–395 (2024).
67. Engedal, K. *et al.* Sex differences on Montreal Cognitive Assessment and Mini-Mental State Examination scores and the value of self-report of memory problems among community dwelling people 70 years and above: The HUNT study. *Dement. Geriatr. Cogn. Disord.* **50**, 74–84 (2021).
68. *Alzheimer's Disease Neuroimaging Initiative et al. Multimodal Phenotyping of Alzheimer's Disease with Longitudinal Magnetic Resonance Imaging and Cognitive Function Data*.
69. Hasselgren, C. *et al.* Sex differences in dementia: on the potentially mediating effects of educational attainment and experiences of psychological distress. *BMC Psychiatry* **20**, 434 (2020).
70. Ou, Y.-N. *et al.* Blood pressure and risks of cognitive impairment and dementia: A systematic review and meta-analysis of 209 prospective studies. *Hypertension* **76**, 217–225 (2020).
71. Nucera, A. & Hachinski, V. Cerebrovascular and Alzheimer disease: fellow travelers or partners in crime? *J. Neurochem.* **144**, 513–516 (2018).
72. Murray, M. E. *et al.* Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol.* **10**, 785–796 (2011).

73. Shenkman, E. *et al.* OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute With a Community-Based Distributive Medical Education Model. *Acad. Med.* **93**, 451–455 (2018).
74. Hogan, W. R. *et al.* The OneFlorida Data Trust: a centralized, translational research data infrastructure of statewide scope. *J. Am. Med. Inform. Assoc.* **29**, 686–693 (2022).
75. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
76. WHOCC. ATCDDD - Home. https://atcddd.fhi.no/.
77. Nachar, N. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutor. Quant. Methods Psychol.* **4**, 13–20 (2008).
78. Gale, R. P., Hochhaus, A. & Zhang, M.-J. What is the (p-) value of the P-value? *Leukemia* **30**, 1965–1967 (2016).
79. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-wide association studies as a tool to advance precision medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).
80. Lumsden, A. L., Mulugeta, A., Zhou, A. & Hyppönen, E. Apolipoprotein E (APOE) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the UK Biobank. *EBioMedicine* **59**, 102954 (2020).
81. Korologou-Linden, R. *et al.* The causes and consequences of Alzheimer's disease: phenome-wide evidence from Mendelian randomization. *Nat. Commun.* **13**, 4726 (2022).
82. Sagheer, A. & Kotb, M. Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. *Sci. Rep.* **9**, 19038 (2019).
83. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
84. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846 (1971).
85. Chêne, G. *et al.* Gender and incidence of dementia in the Framingham Heart Study from mid-adult life. *Alzheimers. Dement.* **11**, 310–320 (2015).
86. Xu, J. *et al.* Federated Learning for Healthcare Informatics. *Int. J. Healthc. Inf. Syst. Inform.* **5**, 1–19 (2021).
87. *Integrative Analyses of Multimodal Clinical Neuroimaging,Genetic,and Data Identify Subtypes and Potential Treatments for Heterogeneous Parkinson Disease*.