Decoding subphenotypes in electronic medical records within late-onset Alzheimer's disease reveals heterogeneity and sex-specific differences

Yukari Katsuhara[1,2], Umair Khan[1,3], Zachary A. Miller[4,5], Isabel E. Allen[6,7,8], Tomiko T. Oskotsky[1,9], Marina Sirota[1,11,12,*], Alice S. Tang[1,10,11,*]

[1]Bakar Computational Health Sciences Institute, UCSF, San Francisco, CA, USA

[2]Health Data Science Graduate Program, UCSF, San Francisco, CA, USA

[3]Biological and Medical Informatics Graduate Program, UCSF, San Francisco, CA, USA

[4]Memory and Aging Center, Department of Neurology, UCSF Weill Institute for Neurosciences, UCSF, San Francisco, CA, USA

[5]Dyslexia Center, Department of Neurology and Psychiatry, UCSF Weill Institute for Neurosciences, UCSF, San Francisco, CA, USA

[6]Department of Epidemiology and Biostatistics, UCSF, San Francisco, CA, USA

[7]Global Brain Health Institute, UCSF, San Francisco, CA, USA

[8]Global Brain Health Institute, Trinity College Dublin, Dublin, Ireland

[9]Division of Clinical Informatics and Digital Transformation, Department of Medicine, UCSF, San Francisco, CA, USA

[10]School of Medicine, UCSF, San Francisco, CA, USA

[11]Senior author

[12]Lead contact

[*]Correspondence: marina.sirota@ucsf.edu (M.S.) and alice.tang@ucsf.edu (A.S.T.) share co-corresponding authorship.

Summary

We applied unsupervised learning techniques to electronic medical records from UCSF to identify distinct Alzheimer's disease subgroups based on comorbidity profiles. Given the well-known female sex predominance in Alzheimer's disease prevalence, we performed sex-stratified analyses to evaluate differences in disease manifestations based on sex. Findings were validated using an independent UC-Wide dataset. Among 8,363 patients, we identified five Alzheimer's disease subphenotypes, characterized by comorbidities related to cardiovascular conditions, gastrointestinal disorders, and frailty-related conditions such as pneumonia and pressure ulcers. We further refined significant comorbidity variations across clusters through sex-stratified analyses, observing a higher prevalence of circulatory diseases in males in Cluster 2 and bladder stones in females in Cluster 3. Key results were consistent across the UCSF and UC-Wide datasets. Our study identifies clinically meaningful Alzheimer's disease subgroups, along with sex-specific variations, suggesting underlying biological factors, and indicates the potential utility of these findings in informing individualized therapeutic regimens.

Keywords

Alzheimer's, electronic medical records, sex differences, unsupervised machine learning, phenotype, cluster analysis, individualized therapy

Introduction

Alzheimer's disease (AD) is the most common cause of dementia. With rising prevalence and no curative treatments, AD represents a major public health crisis worldwide. Clinically, AD has traditionally been conceptualized as an amnestic disorder predominantly affecting memory in those 65 and older, but it can also manifest at an earlier age and with non-amnestic behavioral, executive, language, visuospatial, and asymmetric motor coordination phenotypes. Pathologically, AD is defined by the accumulation of β-amyloid (Aβ) plaques, hyperphosphorylated tau neurofibrillary tangles, glial changes (microglia and astrocytes), and subsequent neurodegeneration [1,2]. Despite decades of research, its biological mechanism remains incompletely understood, making early diagnosis and effective management challenging.

The greatest known risk factor for developing AD is age followed by Apolipoprotein E (*APOE*) genetic status [3,4]. Cardiovascular and metabolic risk factors have also been shown to affect risk in typical amnestic late-onset AD (LOAD) presentations, while novel risk factors may apply in earlier onset and non-amnestic presentations [5], highlighting our incomplete understanding of AD vulnerability. Increasingly, sex has emerged as a critical factor in understanding AD risk, progression, and clinical manifestations, as women account for approximately two-thirds of AD cases in the United States, with a higher lifetime risk of developing AD than men [6]. Recent studies indicate that sex may modify multiple aspects of AD, including susceptibility of AD, progression, and molecular pathology including *APOE* [7-12]. Men with AD tend to have a shorter survival time, while women experience greater memory impairment [13-15]. Yet, here too, the mechanisms underlying these sex differences remain poorly understood, underscoring the need for more research into sex-specific risk factors associated with AD.

Recent advances in various types of big data and machine learning have enabled large-scale, data-driven investigations of AD heterogeneity. Transcriptomic approaches have identified three molecular subtypes of AD, each characterized by distinct combinations of dysregulated pathways, including susceptibility to tau-mediated neurodegeneration, amyloid-β–related neuroinflammation, synaptic signaling, immune activity, mitochondrial organization, and myelination[16]. Similarly, neuroimaging studies using MRI have

revealed three neuroanatomical subtypes, each defined by different initial regions of atrophy onset [17]. Cognitive subtypes have also been identified based on performance in memory, visuospatial ability, language, and executive function. However, these studies are largely restricted to carefully selected research cohorts, and it remains unclear whether their findings generalize to larger, more diverse populations.

In contrast, electronic medical records (EMRs) capture a wide range of real-world health data, including multimorbidity patterns, medication histories, and disease trajectories, making them valuable resources for identifying novel AD subtypes and potential therapeutic targets [18]. Prior EMR-based studies have identified AD subtypes associated with cardiovascular disease, psychiatric disorders, age of onset, and sensory impairments [19]. Another study identified five distinct patient groups, characterized by mental health conditions, nontypical AD, typical AD, cardiovascular disease, and combined memory problems with cancer [20]. From a sex-specific perspective, evidence suggests that men and women differ in comorbidity profiles [21] and may follow distinct AD progression trajectories [19,22]. Similar heterogeneity in comorbidity patterns has also been reported across racial and ethnic groups [23]. Nevertheless, subtyping methodologies are not yet standardized, with substantial variability in study populations (e.g., case definitions including early- vs. late-onset AD, and inclusion or exclusion of healthy controls) and analytic approaches. Comprehensive and unbiased analyses focusing specifically on sex-specific patterns among subphenotypes of AD remain largely unexplored.

In this study, we apply unsupervised clustering to a large EMR dataset of AD patients, consisting largely of LOAD presentations, to identify if subphenotypes within this group exist. By leveraging real-world clinical data, we aim to uncover novel insights into AD heterogeneity, which may inform precision medicine approaches and improve comprehensive, sex-specific care for patients suffering from neurodegenerative disease. Deeper characterization of AD-related comorbidities and their subphenotypes may uncover novel risks for the disease and facilitate the development of individualized therapeutic

regimens that could serve as adjuvants to the current anti-amyloid- and upcoming anti-tau-based treatments, maximizing therapeutic benefits for patients through a truly personalized therapeutic approach.

Results

Patient characteristics

AD cohorts were identified from University of California (UC) San Francisco (UCSF) data and UC-Wide data, with UCSF patients excluded from the latter, respectively. This study was designed as a cross-sectional analysis at the patient level, where diagnoses across all visits were aggregated into a single binary profile for each patient to ensure independence of observations. We identified 8,363 AD patients from the UCSF EMRs (5,315 female (64%); median age: 90.0 [IQR: 84.0–91.0]) (Figure 1). From the UC-Wide validation dataset, we identified 25,896 AD patients (16,036 female (61.9%); median age: 89.0 [IQR: 82.0–90.0]). In both datasets, AD patients were predominantly “White” (62.0% at UCSF, 65.7% in the UC-Wide dataset), followed by those classified as "Other" and “Asian”. At UCSF, most AD patients (75.0%) were recorded as "Alive," whereas in the UC-Wide dataset, only 50.9% were alive. Additionally, the median number of comorbidities recorded at UCSF was lower compared to the UC-Wide dataset (UCSF: median = 23.0 [IQR: 9.0–63.0]; UC-Wide: median = 40.0 [IQR: 16.0–88.0]). Other demographic characteristics of AD patients are presented in Table 1 and 2.

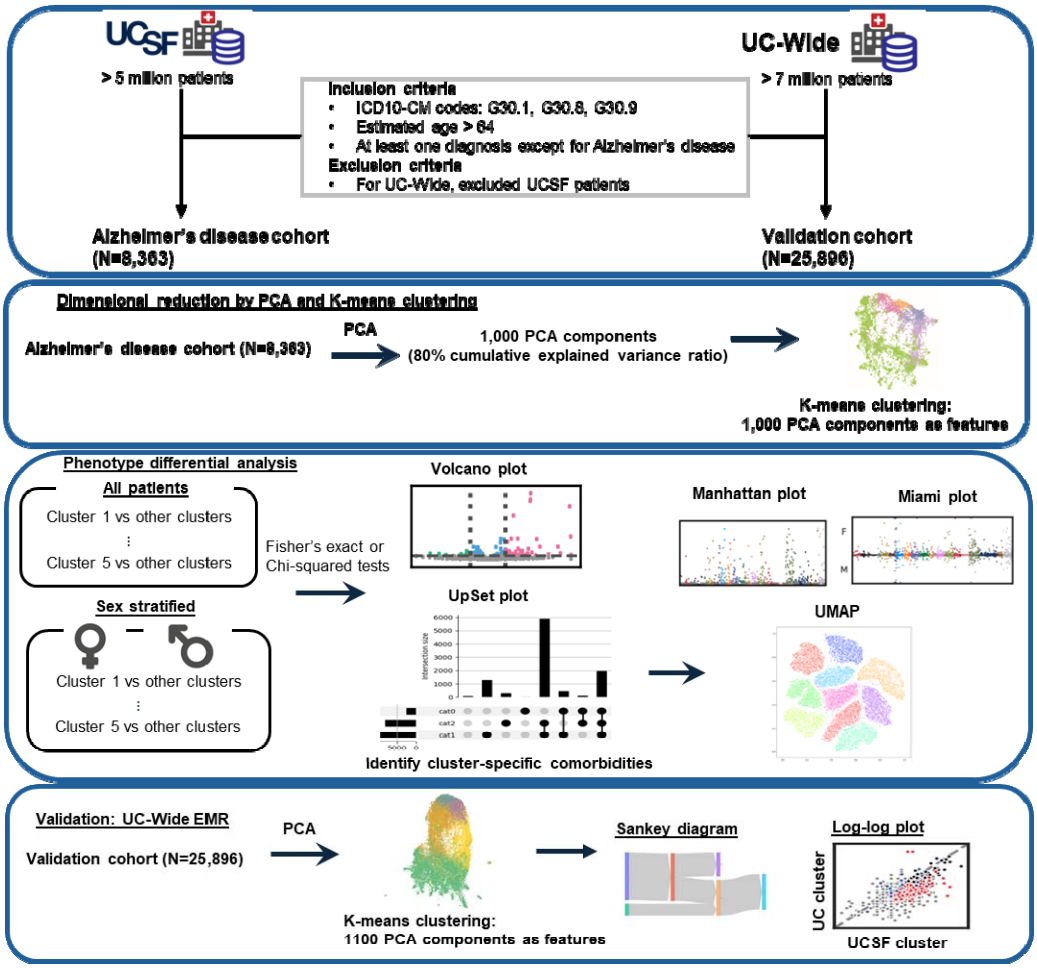

Figure 1. Cohort selection and overview of study design. Alzheimer's disease patients > 64 years with at least one comorbidity were included in this study. (1) K-means clustering was performed following Principal Component Analysis (PCA). (2) Each cluster was characterized based on cluster-specific comorbidities. For validation, (3) the same approach was applied to the UC-Wide cohort, excluding UCSF patients. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification; PCA, Principal Component Analysis; UC, University of California

Low-dimensional embeddings derived from principal components of diagnostic data reveal five subphenotypes of AD

The analysis data consisted of patient-level diagnostic records, where each row represented an individual patient and each column corresponded to one of 33,031 unique diagnosis names across 8,363 AD patients. Diagnosis names were one-hot encoded, with a binary value indicating the presence or absence of each diagnosis for a given patient. To mitigate the impact of redundant or noisy features, Principal Component

Analysis (PCA) was applied to diagnosis names (33,031 diagnosis name features) prior to K-means clustering. For the UCSF dataset, we selected 1,000 principal components, capturing approximately 80% of the cumulative explained variance, as input features for clustering.

Cluster determination was based on the analysis of within-cluster sum of squares (WSS) and silhouette scores. The WSS analysis indicated that after four clusters, the rate of WSS reduction diminished substantially, suggesting that increasing the number of clusters beyond four did not provide meaningful improvements in partitioning. Conversely, a sharp decline in the silhouette score was observed between five and six clusters (88.1%), indicating a considerable loss in clustering quality. These trends suggest that either a four- or five-cluster (Cluster 1–Cluster 5) solution could be appropriate. To balance cluster compactness with the representation of heterogeneity, the five-cluster solution was selected, as the four-cluster solution may not fully capture the diversity of AD subphenotypes (Supplemental Table 1).

We characterized each cluster based on comorbidity profiles. In this step, comorbidities were defined using International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes to facilitate clinical interpretation and enhance the interpretability of cluster characteristics. Within the UCSF clusters, Cluster 5 had significantly fewer comorbidities (median = 12.0 [IQR: 5.0–21.0], Bonferroni-corrected p-values < 0.05) compared to the other clusters, whereas Cluster 1 had the highest number of comorbidities (median = 313.5 [IQR: 258.0–379.0], Bonferroni-corrected p-values < 0.05). In terms of age distribution, Cluster 4 (median = 91.0 [IQR: 90.0–91.0], Bonferroni-corrected p-values < 0.05) was significantly older than Cluster 1, 3, and 5, followed by Cluster 2. Conversely, Cluster 3 (median = 88.0 [IQR: 82.0–90.0], Bonferroni-corrected p-values < 0.05) was significantly younger than all other clusters (Supplemental Table 2). Black or African American individuals comprised 13% of Cluster 2 and 12% of Cluster 4, representing a higher proportion compared to other clusters (Bonferroni-corrected p-values < 0.05, Supplemental Table 2). Regarding sex, Cluster 5 included a significantly higher proportion of males than Cluster 4, with these males displaying a younger age distribution. Other

demographic characteristics and sex-stratified distributions for each cluster are presented in Table 1 and Supplemental Table 3.

For the UC-Wide dataset, 19,835 unique diagnosis names from 25,896 AD patients were used as features for PCA. 1,100 principal components covering approximately 80% of the cumulative explained variance were selected as input features for K-means clustering. Consistent with the UCSF dataset, five clusters (Cluster A–Cluster E) were identified based on silhouette scores and WSS (Supplemental Table 1). The WSS reduction analysis indicated that the decrease from five to six clusters was relatively small, suggesting that increasing the number of clusters beyond five does not substantially improve clustering compactness. In contrast, the silhouette score analysis revealed a significant decline in clustering structure when increasing from four to five clusters (50.0%), indicating a reduction in cluster separation. Based on these findings, a four- or five-cluster solution appeared to provide the most stable clustering structure. Given that this study aims to capture the heterogeneity of comorbidity patterns in AD, the five-cluster solution was selected.

Significant differences in the number of comorbidities were observed between Cluster A (median = 201.0 [IQR: 166.0–253.0], Bonferroni-corrected p-values < 0.05) and Cluster B (median = 15.0 [IQR: 7.0–25.0], Bonferroni-corrected p-values < 0.05) (Supplemental Table 2). Other demographic characteristics of each cluster are presented in Table 2. A low-dimensional Uniform Manifold Approximation and Projection (UMAP) visualization of the PCA-transformed components illustrates the distribution of AD patient clusters, sex differences, the number of comorbidities, and locations (Figure 2).

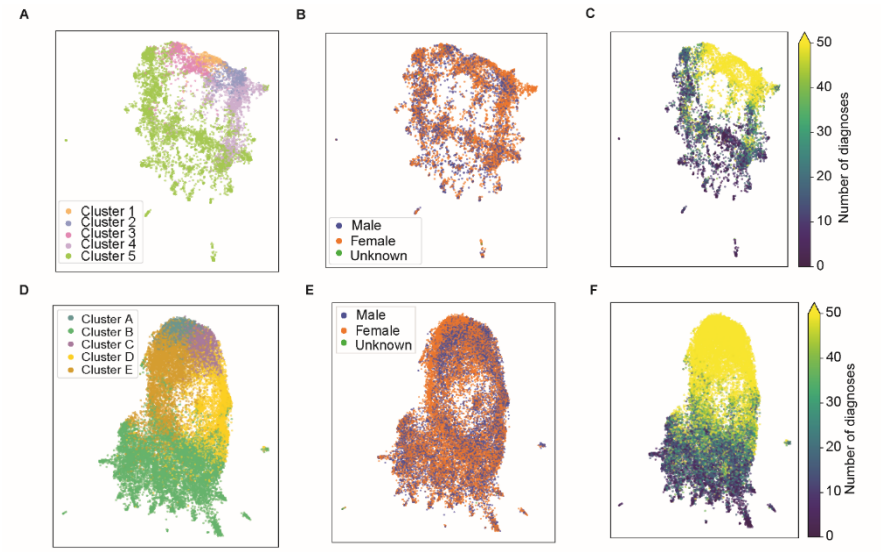

Figure 2. UMAP visualizations using PCA components as features. UMAP plots are colored by cluster (**A** and **D**), sex (**B** and **E**), and number of comorbidities (**C** and **F**). Each dot represents a patient. **A–C** correspond to the UCSF cohort, while **D–G** represent the validation cohort in the UC-Wide dataset. UC, University of California

Comorbidity enrichment analysis shows significant cluster-related comorbidities at UCSF

Our analysis identified significant diagnoses in each cluster within the UCSF dataset (two-sided Fisher's exact or Chi-squared test, Bonferroni-corrected p-value < 0.05) (Figure 3). Cluster 1 had 1,911 significant comorbidities, most of which were positively associated. In contrast, 1,650 significant comorbidities were identified in Cluster 5, predominantly showing negative associations. However, a few conditions exhibited positive associations, including Pick's disease. Cluster 2 – 4 displayed unique comorbidity enrichment patterns. For example, in Cluster 2 (985 significant comorbidities), the top positively associated diagnoses included laceration, phlebitis and thrombophlebitis of femoral vein, and benign neoplasm of the conjunctiva. Interestingly, Cluster 3 (682 significant comorbidities) showed relatively lower prevalence of mental disorders and personality changes. Finally, Cluster 4 (245 significant comorbidities) exhibited positive associations with pneumonia and disseminated intravascular coagulation (Supporting Data). Manhattan plots illustrated the distribution of statistical significance (Supplemental Figure 1).

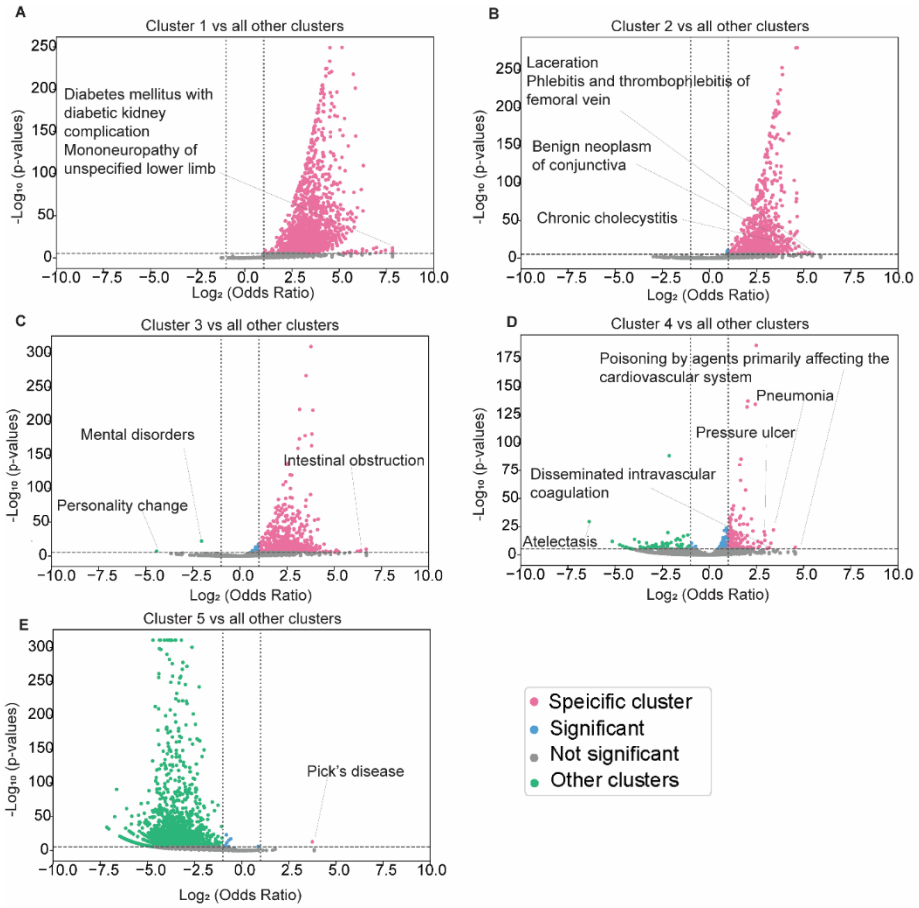

Figure 3. Volcano plot identifies overall trends of each cluster

The volcano plot displays enriched ICD10-CM codes identified using the two-sided Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). Enrichment was determined based on a Bonferroni-corrected p-value < 0.05. Odds ratios were calculated by comparing each specific cluster to all other clusters. **A:** Cluster 1, **B:** Cluster 2, **C:** Cluster 3, **D:** Cluster 4, **E:** Cluster 5. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification.

Significant cluster-specific comorbidities at UCSF are subsequently validated in the UC-Wide validation cohort

We identified cluster-specific comorbidities using UpSet plots (Figure 4). Cluster 1 exhibited the highest number of positively associated comorbidities (812 comorbidities), followed by Cluster 2 (225 comorbidities). Cluster 3 had 96 positively associated comorbidities and one negatively associated comorbidity (personality change), while Cluster 4 had 13 positively and 6 negatively associated comorbidities. In contrast, Cluster 5 had only two positively associated comorbidities, whereas 1,581

comorbidities were negatively associated. Examining the most significantly associated comorbidities specific to each cluster, we found that Cluster 3 was characterized by acute systolic heart failure and ileus, whereas pneumonia and brief psychotic disorder were observed in Cluster 4. In Cluster 5, we observed essential hypertension, anemia, and urinary tract infection as the most significantly associated comorbidities (Figure 5).

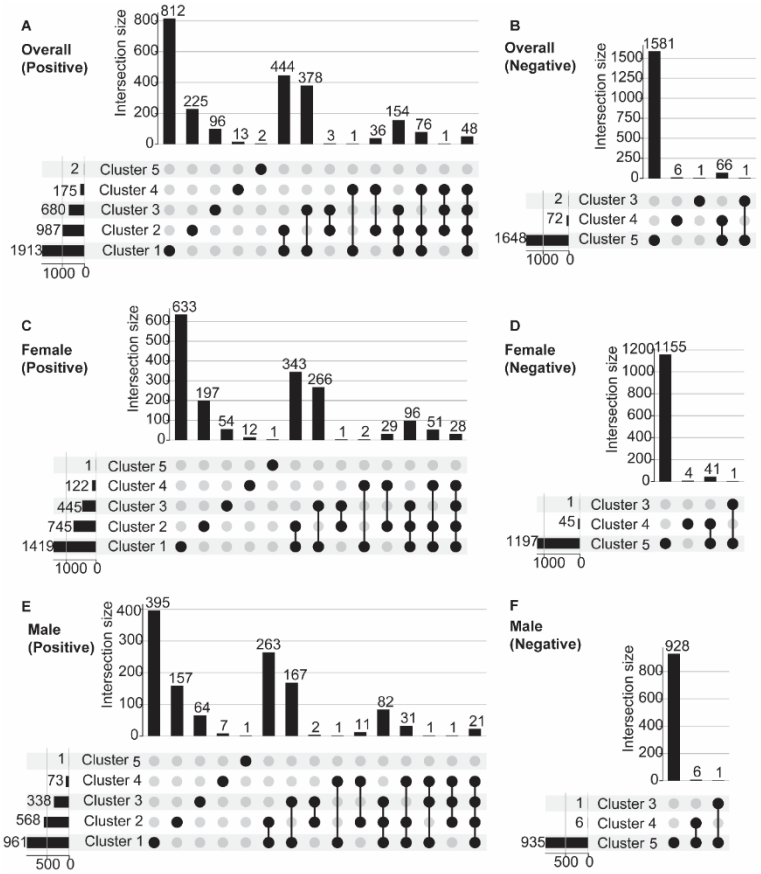

Figure 4. UpSet plot identifies significant cluster-specific comorbidities.
UpSet plot illustrating significant cluster-specific comorbidities. Panels represent the overall population (**A:** Positive associations, **B:** Negative associations), females (**C:** Positive associations, **D:** Negative associations), and males (**E:** Positive associations, **F:** Negative associations).

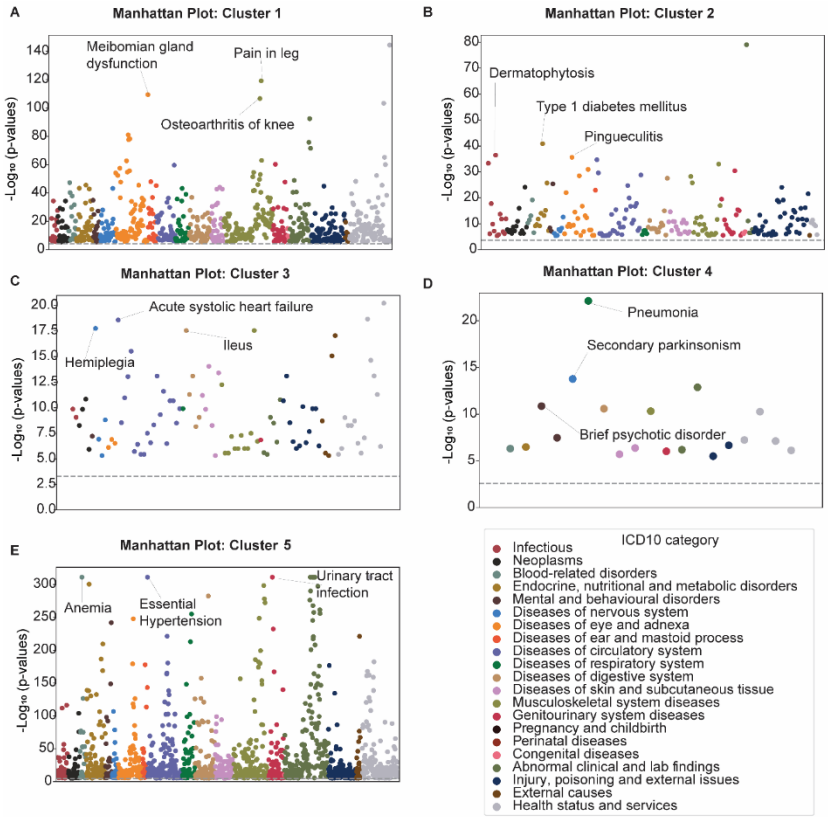

Figure 5. Manhattan plot for cluster-specific significant comorbidities
Manhattan Plot displaying enriched ICD10-CM codes identified using the two-sided Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). Enrichment was determined based on a Bonferroni-corrected p-value < 0.05. **A:** Cluster 1, **B:** Cluster 2, **C:** Cluster 3, **D:** Cluster 4, **E:** Cluster 5. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification.

Additionally, cluster-specific comorbidities with high odds ratios were identified (Figure 6A). Cluster 1 showed numerous significant associations with a variety of conditions, including diabetes mellitus with other diabetic kidney complication, contusion of the right upper arm, displaced fracture of shaft of right clavicle, complication of kidney transplant, encounter for aftercare following kidney transplant, and mononeuropathy of right lower (all ORs = 227, Bonferroni-corrected p-values < 0.05). In Cluster 2, phlebitis and thrombophlebitis of femoral vein had the highest odds ratio (OR = 44.1, Bonferroni-corrected p-value < 0.05). Cluster 3 was characterized by infection and inflammatory reaction due to a urinary catheter (OR = 103, Bonferroni-corrected p-value < 0.05), followed by intestinal obstruction (OR = 84.3, Bonferroni-corrected p-value < 0.05). Additionally, dysphagia was significantly associated with

Cluster 3 (OR = 5.3, Bonferroni-corrected p-value < 0.05, Supporting Data). Cluster 4 was significantly associated with poisoning by cardiac-stimulant glycosides (OR = 23.6, Bonferroni-corrected p-value < 0.05), pneumonia (OR = 10.6, Bonferroni-corrected p-value < 0.05), and pressure ulcer (OR = 7.5, Bonferroni-corrected p-value < 0.05). Lastly, in Cluster 5, Pick's disease was significantly enriched (OR = 13.4, Bonferroni-corrected p-value < 0.05). However, ORs in our study should be interpreted with caution, as the frequency of certain diagnoses in each cluster was relatively small, which may have led to inflated odds ratios (Supporting Data).

In the UC-Wide validation dataset, we identified corresponding UC-Wide clusters for each UCSF cluster based on the overlap in significant comorbidities, as visualized by Sankey plots (Figure 7A). The proportion of significant comorbidities in each UCSF cluster that were also captured in the corresponding UC-Wide clusters ranged from 20% to 55%. For instance, 239 of the 812 (29.4%) significant comorbidities identified in UCSF Cluster 1 were also observed in the matched UC-Wide Cluster A, which itself contained 1,976 cluster-specific significant comorbidities.

To further evaluate the consistency of comorbidity enrichment across datasets, we examined the correlation of shared comorbidities using Log-Log plots (Figure 7B). UCSF cluster-specific comorbidities with high odds ratios including phlebitis and thrombophlebitis in Cluster 2, intestinal obstruction and infection and inflammatory reaction due to a urinary catheter in Cluster 3, and pneumonia and pressure ulcer in Cluster 4, were largely replicated in their corresponding UC-Wide clusters. However, Cluster 2 did not show a significant correlation with its matched UC-Wide Cluster A (Pearson r = 0.07, p = 0.68).

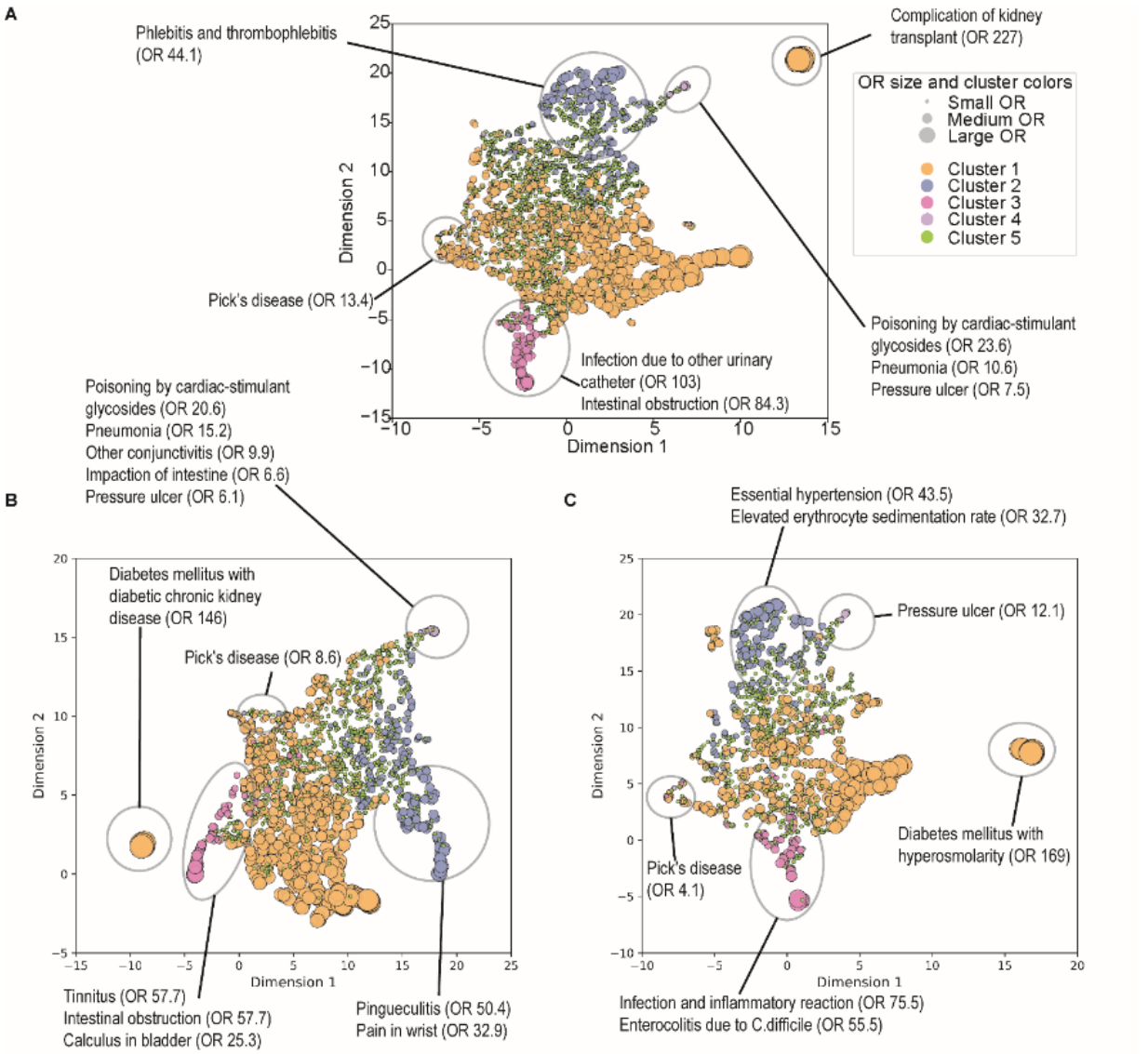

Figure 6. UMAP visualizations using odds ratios as features for each cluster. This figure presents a UMAP visualization of ICD-10-CM codes based on their standardized ORs across different clusters. Each dot represents a specific ICD-10-CM code, with the dot size corresponding to the OR of the significantly associated cluster and the color indicating the corresponding cluster. The input features for the UMAP projection were derived from a matrix of standardized ORs of cluster-specific comorbidities. These standardized ORs were calculated by comparing the prevalence of each diagnosis within a specific cluster against all other clusters using Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). A Bonferroni-corrected p-value threshold of 0.05 was applied to identify significantly enriched diagnoses. **A:** Overall population, **B:** Female, **C:** Male. ICD-10-CM, International Classification of Diseases, 10th Revision, Clinical Modification; UMAP, Uniform Manifold Approximation and Projection.

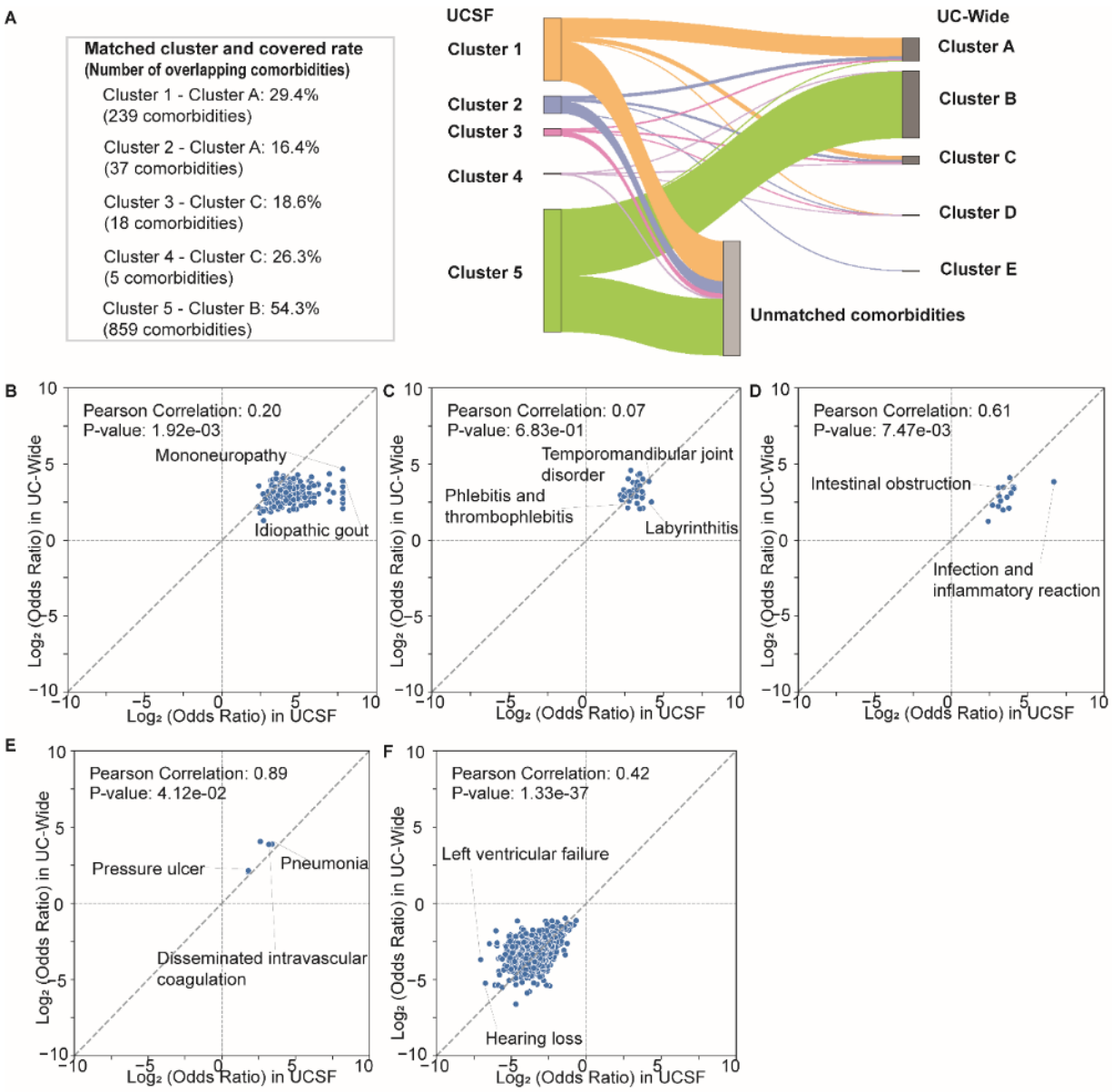

Figure 7. Validation analysis of cluster-specific significant comorbidities.
A Sankey plot illustrating the overlap of significant comorbidities across UCSF and UC-Wide populations within each cluster. The coverage rate represents the proportion of UCSF cluster-specific comorbidities that were statistically significant in UCSF and remained significant in UC-Wide clusters, identifying the most closely matched UC cluster. Number of cluster-specific comorbidities: Cluster 1 (812), Cluster 2 (225), Cluster 3 (97), Cluster 4 (19), Cluster 5 (1,583), Cluster A (1,976), Cluster B (2,731), Cluster C (819), Cluster D (40), Cluster E (19)
B–F Log-log plot showing significant overlapping comorbidities and the concordance of odds ratios for overlapping conditions in the overall population. P-values were calculated from Pearson correlation analysis. UC, University of California.

Sex-stratified analysis identifies significantly associated cluster-sex-specific comorbidities at UCSF, which are subsequently validated in the UC-Wide validation cohort

After illustrating the distribution of statistical significance in Miami plots, we identified cluster-sex-specific comorbidities using UpSet plots (Figure 4, Supplemental Figure 2). When we focus on the significant comorbidities exclusive to each cluster, in Cluster 1, females showed a strong association with pain-related conditions, whereas males were primarily associated with symptoms of the genitourinary system. In Cluster 2, females exhibit a strong association with tension-type headaches, whereas males are more frequently associated with circulatory diseases, including unstable angina and hypertensive heart disease. In Cluster 3, females show significant associations with pressure ulcers and osteoarthritis, while males are predominantly linked to respiratory diseases, such as acute pulmonary edema, and circulatory conditions, including acute systolic heart failure. In Cluster 4, females are significantly associated with pneumonia. In Cluster 5, we observed essential hypertension, anemia, and urinary tract infection across both sexes (Figure 8).

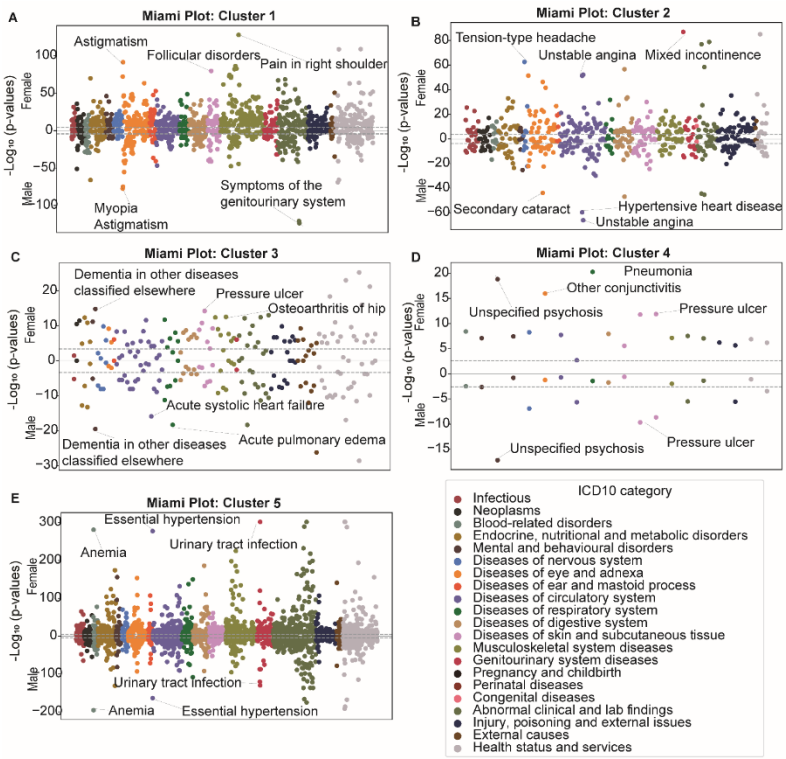

Figure 8. Miami plot for cluster-specific significant comorbidities. Miami Plot displaying enriched ICD10-CM codes identified using the two-sided Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). Enrichment was determined based on a Bonferroni-corrected p-value < 0.05. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification. **A:** Cluster 1, **B:** Cluster 2, **C:** Cluster 3, **D:** Cluster 4, **E:** Cluster 5.

Considering odds ratios, diabetes-related conditions were significantly associated with Cluster 1 in both sexes, with diabetes mellitus due to an underlying condition with diabetic chronic kidney disease being predominant in females (OR = 146) and diabetes mellitus with hyperosmolarity in males (OR = 169). Similarly, Pick's disease remained highly associated with Cluster 5 across both sexes (Female OR = 8.6, Male OR = 4.1). In Cluster 2, females exhibit stronger associations with eye diseases, such as pingueculitis (OR=50.4, Bonferroni-corrected p-value < 0.05), whereas males are more associated with circulatory diseases, including hypertension (OR=43.5, Bonferroni-corrected p-value < 0.05) and elevated erythrocyte sedimentation rate (ESR) (OR=32.7, Bonferroni-corrected p-value < 0.05). In Cluster 3, females demonstrate high odds ratios for tinnitus (OR=57.7, Bonferroni-corrected p-value < 0.05), intestinal obstruction (OR=57.7, Bonferroni-corrected p-value < 0.05), and bladder calculus (OR=25.3, Bonferroni-corrected p-value < 0.05), while males are significantly associated with infection and inflammatory reaction (OR=75.5, Bonferroni-corrected p-value < 0.05) and enterocolitis (OR=55.5, Bonferroni-corrected p-value < 0.05). Lastly, in Cluster 4, females have strong associations with poisoning by cardiac-stimulant glycosides (OR=20.6, Bonferroni-corrected p-value < 0.05) and pneumonia (OR=15.2, Bonferroni-corrected p-value < 0.05), whereas males are associated with pressure ulcers (OR=12.1, Bonferroni-corrected p-value < 0.05) (Figure 6B and 6C, Supporting Data). However, sex-stratified odds ratios should be interpreted with caution due to the limited number of cases for certain diagnoses, which may lead to variability in estimates.

In the UC-Wide dataset, we identified corresponding UC-Wide clusters based on the coverage rate of UCSF cluster-specific comorbidities in each sex (Supplemental Figure 3). The matched UC-Wide clusters remained consistent with those identified in the overall population, except for UCSF Cluster 3 among

females. For UCSF Cluster 4 among males, two UC-Wide clusters had the same coverage rate. In this case, the cluster that was consistently matched in the overall analysis and had one of the highest coverage rates was selected as the matched cluster. Between 15% and 60% of significant comorbidities in each UCSF cluster were recaptured by the matched UC-Wide clusters. Overall, UCSF cluster-specific comorbidities with high odds ratios —such as elevated ESR in Cluster 2 among males, intestinal obstruction in Cluster 3 among females, and infection and inflammation reaction due to urinary catheter, and enterocolitis in Cluster 3 among males—were largely replicated in the UC-Wide clusters. However, some Pearson correlation coefficients were neither strong nor statistically significant (Figure 9).

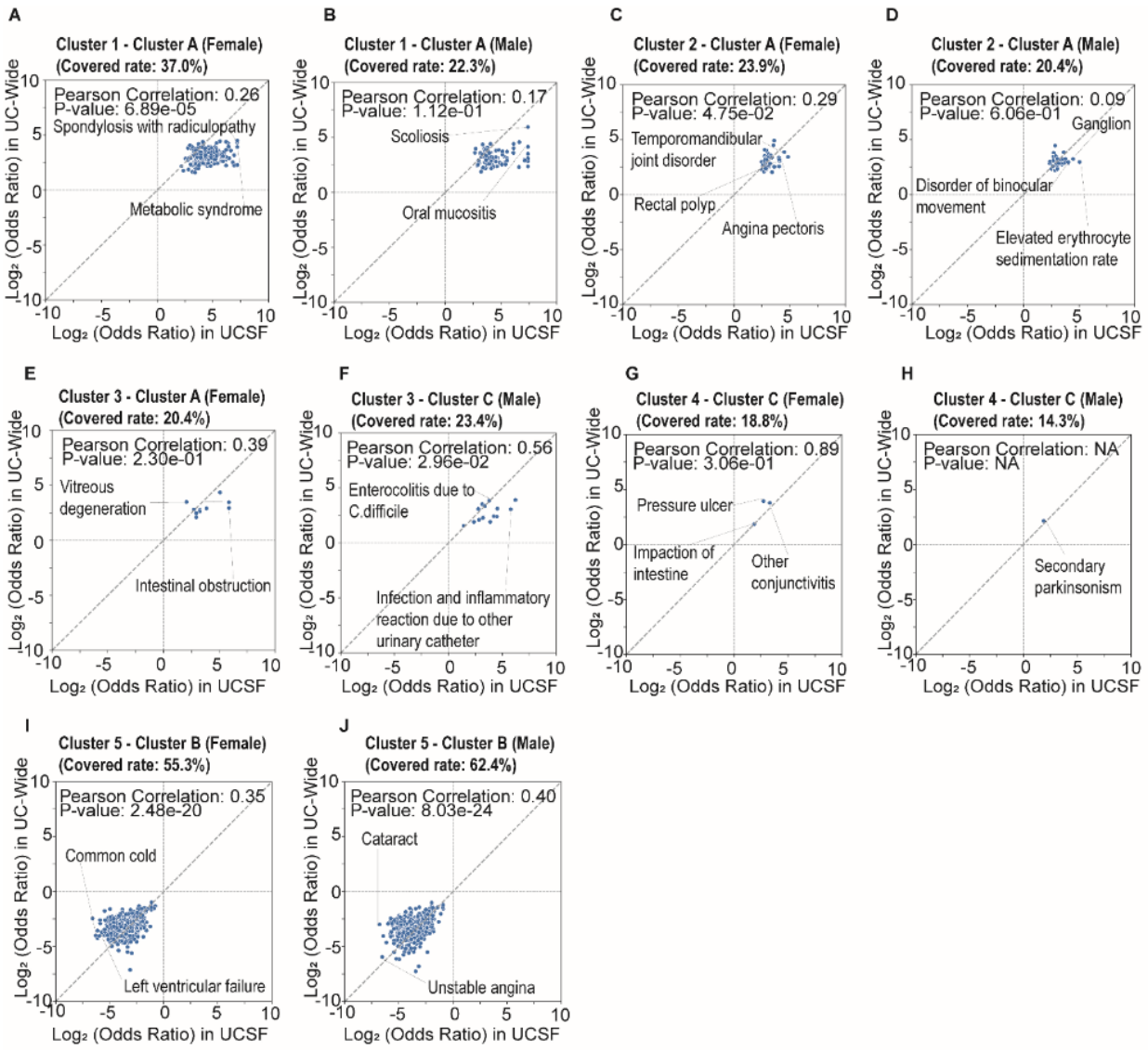

Figure 9. Validation analysis of cluster-specific significant comorbidities across sex. Log-log plot illustrating the significant overlapping comorbidities and the concordance of odds ratios for these conditions in the sex-stratified population. The covered rate represents the proportion of UCSF cluster-sex-specific comorbidities that remain significant in both the UCSF and UC-Wide datasets. **A, C, E, G, I**: Cluster 1–Cluster 5 (Female) vs. Matched UC-Wide Cluster (Female), **B, D, F, H, J:** Cluster 1–Cluster 5 (Male) vs. Matched UC-Wide Cluster (Male). P-values were calculated from Pearson correlation analysis. UC, University of California.

Discussion

In this study of 8,363 AD patients from UCSF and 25,896 from an independent UC-Wide cohort, we analyzed multi-site EMR by applying dimensionality reduction, followed by clustering analysis, and in the process, we identified five subphenotypes of AD based on comorbidities.

A key strength of our approach is its focus on patients with LOAD, enabling the characterization of heterogeneity within the most common clinical presentations and minimizing potential noise from other dementia types. Importantly, whereas most prior studies have compared AD patients with cognitively normal controls[19,21,23], our study advances the field by performing comparisons within the AD population, thereby revealing clinically meaningful heterogeneity among patients. Furthermore, we comprehensively explored comorbidity profiles across the entire cohort using an unbiased enrichment analysis and confirmed the reproducibility of our findings in a completely independent validation dataset.

Enrichment analysis comparing a specific cluster with the other clusters revealed clinical heterogeneity. Sex-stratified analysis further highlighted sex-specific differences within the identified clusters. In the following sections, we present our findings for each cluster, followed by the identification of notable sex-specific differences, and discuss their implications.

Cluster 1: high comorbidity burden with complex clinical profiles

We performed comorbidity enrichment analysis for overall populations. Cluster 1 was characterized by a high comorbidity burden, which may indicate the presence of patients with complex clinical profiles. In particular, diseases with the highest odds ratios and lowest p-values included several kidney-related

conditions—diabetes mellitus with other diabetic kidney complications, complications of kidney transplant, and encounter for aftercare following kidney transplant. These findings highlight the prominence of renal comorbidities in this cluster. However, additional enriched conditions such as contusion of the right upper arm, displaced fracture of shaft of right clavicle, and mononeuropathy of right lower were also identified, reflecting a broader range of systemic complications beyond kidney disease alone. Kidney transplantation is a recognized risk factor for AD and is often associated with multiple comorbidities due to long-term immunosuppressive therapy and underlying chronic conditions[24,25], therefore, kidney-related conditions are suggested as a key component of this cluster, although they do not solely define it. Further stratification of this cluster is warranted to identify subgroups with specific comorbidity patterns, which may facilitate the development of more targeted therapeutic treatment.

Cluster 2: African American individuals, phlebitis and cardiovascular conditions

Cluster 2 presented with a relatively higher proportion of Black or African American individuals. AD prevalence is about twice as high in this population compared to non-Hispanic Whites. Studies have suggested that *APOE* ε4-based mechanisms of AD may be less relevant to Black or African American than non-Hispanic Whites, and thus this cluster may be highlighting *APOE* ε4-independent risks, such as social determinants of health [6,26-28]. Patients in Cluster 2 also exhibited enriched comorbidities such as phlebitis and thrombophlebitis of femoral vein. Superficial thrombophlebitis, the most common form of thrombophlebitis, has been increasingly recognized as a condition associated with deep vein thrombosis and pulmonary embolism, both of which can lead to serious clinical outcomes [29-31]. Established risk factors for these thrombotic conditions include a hypercoagulable state and the presence of abnormal fibrin clot structures [32-34]. Previous literature suggests that increased fibrinogen aggregation and Aβ fibrillization occur in AD brain parenchyma and vessels and induce the formation of abnormal fibrin clots that are more resistant to degradation [35,36]. Additionally, recent studies have shown that blood-derived fibrin deposits in the central nervous system are a common feature of autoimmune and neurodegenerative

diseases, including AD [37,38]. These findings highlight fibrin as a potential therapeutic target for immunotherapy designed to reduce fibrin-induced neurotoxicity [39]. There is no strong evidence for the impact on peripheral blood vessels caused by the abnormal fibrin clots in the brain; however, previous evidence shows that peripheral treatment with a low molecular weight heparin reduces plaques and Aβ accumulation in a mouse model of Alzheimer's disease [40]. This suggests a possibility that increased fibrinogen aggregation and Aβ in AD patients may contribute to vascular abnormalities such as phlebitis.

Cluster 3: lower rates of mental illness and increased infections and gastrointestinal dysfunction

This group exhibited significant negative associations with mental disorders. Given that mental disorders are commonly reported comorbidities in AD patients [41], this finding is particularly noteworthy. One possible explanation for the lower prevalence of mental disorders in Cluster 3 is the significantly younger age of patients in this cluster compared to others. This aligns with prior research indicating that AD is a progressive disease in which the risk of comorbidities, including mental disorders, increases over time [42]. On the other hand, previous evidence has demonstrated that early-onset AD (EOAD) is associated with a more aggressive clinical course and distinct biological mechanisms compared to LOAD and in some reports actually possesses a higher amount of psychiatric symptoms burden [5,43-46]. However, since this study focused on LOAD population and the ICD-10-CM's broad disease classification, such as "Other specified mental disorders," we cannot ascertain whether this relatively younger population reflects an early onset presentation of AD or is instead at the younger age range of LOAD.

Additionally, we identified infection and inflammatory reactions due to a urinary catheter, as well as intestinal obstruction, as positively associated comorbidities in Cluster 3. Increasing evidence shows reciprocal associations between AD, infection, and inflammatory reactions [47-49,]. Therefore, the combination of inflammation and reduced mobility associated with AD may contribute to the comorbidity profile observed in Cluster 3.

Regarding intestinal obstruction, several possible mechanisms can be considered. First, reduced mobility, including oropharyngeal dysphagia, and altered nutritional intake associated with AD progression may contribute to gastrointestinal dysfunction such as constipation [50-52]. Notably, dysphagia was identified as a Cluster 3-specific comorbidity (OR = 5.3, Bonferroni-corrected p-value < 0.05, Supporting Data), suggesting that individuals in this cluster may be at higher risk of malnutrition and associated gastrointestinal deficiency. Secondly, recent studies have explored the associations between AD and gut microbiota [53,54]. Research has shown that AD patients exhibit reduced microbial diversity compared to cognitively healthy individuals [55]. It is hypothesized that AD-related alterations in gut microbiota may contribute to gastrointestinal dysfunction, potentially increasing the risk of intestinal obstruction. However, further research is required to elucidate the interplay between AD progression, gut microbiota alterations, and gastrointestinal complications, including intestinal obstruction.

Cluster 4: oldest age and strong association with pneumonia

Patients in Cluster 4 constituted the oldest subgroup and exhibited a higher prevalence of pneumonia, a condition commonly associated with aging and characterized by functional decline, including impaired swallowing function [56]. This observation is particularly noteworthy given the longstanding hypothesis that infectious respiratory diseases may precipitate the onset of AD. This hypothesis was originally proposed over a century ago and has since gained support from a growing body of scientific evidence [57-59], including recent studies that demonstrate lower risks of developing AD in those who were vaccinated against respiratory pathogens, including pneumococcal and influenza[60-63]. Our findings underscore the clinical relevance of implementing preventive strategies, such as vaccination, to reduce modifiable risk factors for neurodegeneration in older populations.

Cluster 5: male sex, low comorbidity burden, potential overlap with frontotemporal dementia

Cluster 5 includes a significantly higher proportion of males with a relatively younger age distribution. Given that AD possesses a female predominance, the enrichment of younger males is particularly

intriguing. While this showed the fewest positively associated comorbidities, there was one notable exception, Pick's disease. Pick's disease, also referred to as behavioral variant frontotemporal dementia (bvFTD), is second to AD in being the most common neurodegenerative disorder in individuals with first symptoms younger than 65 [64,65]. Pick's disease is also notable for its higher prevalence of male sex [66,67]. Considering the demographics of Cluster 5, we wonder if this cluster might be identifying groups where conditions like bvFTD were on the differential diagnosis. Given the inherent limitations of EHR-based research, we were unable to incorporate neuroimaging data such as MRI scans to further investigate this hypothesis. Future studies integrating imaging or biomarker data would be valuable to determine whether this cluster captures patients with bvFTD or other diagnostic entities.

Sex specific analysis:

While sex related results only came out in our first pass analysis in cluster 5, given the significant female bias to AD, we performed follow up sex specific analyses. Here, our findings underscored the presence of sex-specific disease manifestations in Clusters 2 and 3. In Cluster 2, hypertension and elevated ESR were strongly associated with males. Alongside hypertension, diseases known to increase ESR, such as diabetes [68] and cardiovascular disease (CVD) [69], are well-established risk factors for AD. Previous studies have demonstrated strong associations between these conditions and cerebrovascular pathology, as well as cognitive decline [41,70].

Our findings suggest that Cluster 2 represents a specific subgroup of AD males with pronounced vascular and metabolic dysfunction. A similar trend has been reported by Alexander et al.[20], who identified a subtype enriched for cardiovascular comorbidities predominantly among male AD patients. Unlike prior work that reported these patterns only as percentages, our study formally evaluated these differences using statistical tests, providing stronger evidence for sex-specific enrichment. This observation is noteworthy because it contrasts with several previous reports indicating that these conditions are more common in AD females [10,71,72]. A potential explanation for these sex differences may involve *APOE* genotype, which is known as a risk factor for hypertension, CVD, and diabetes [73-75]. Furthermore, existing

evidence suggests that the impact of *APOE* varies by sex and genotypes [12,74-76]. Given this, it is plausible that male sex and *APOE* genotype could contribute to the observed vascular-metabolic phenotype in Cluster 2. Because our dataset lacks *APOE* information, we were not able to directly assess the potential combined effects of male sex and *APOE* genotype. Further studies incorporating genetic data are needed to test this hypothesis.

Another potential explanation for the observed comorbidity patterns in Cluster 2 involves the interaction between racial composition and susceptibility to APOE ε4-related mechanisms. Although Cluster 2 includes a relatively higher proportion of Black or African American individuals overall, sex-stratified demographics revealed that the male subgroup had a higher proportion of White individuals and a lower proportion of Black individuals compared to the female subgroup.

Previous studies suggest that White populations are more susceptible to *APOE* ε4-associated AD than Black populations, and *APOE* ε4 has been linked to an increased risk of vascular dysfunction [26,27,74,75,77]. Therefore, this racial stratification may partially explain the higher burden of *APOE* ε4-related cardiovascular conditions observed in males within Cluster 2, who predominantly belong to the White population.

Our findings hypothesized that male AD patients in Cluster 2 may represent a distinct subset characterized by more pronounced vascular and metabolic dysfunction, potentially influenced by the complex interplay of *APOE* genotypes and ethnic background. However, due to the absence of *APOE* genomic information in our dataset, this hypothesis remains speculative and requires further investigation to confirm its validity. Future research should integrate genetic data to further explore the role of *APOE* and other genetic factors in shaping comorbidity patterns. Additionally, understanding the interplay between sex, race, *APOE* genotype, and vascular risk factors could provide valuable insights for personalized risk assessment and targeted interventions in AD patients.

In Cluster 3, our study found that calculus in the bladder was more strongly associated with females. Generally, urolithiasis, including bladder calculi, is more prevalent in males, and the most common kidney stone is calcium oxalate [78-80]. One possible explanation for this finding is the influence of sex-specific differences in microbiome alterations. Dysbiosis may lead to increased oxalate accumulation, which plays a central role in bladder stone formation [81]. Recent studies have highlighted the relationship between gut microbiome alterations, AD, and sex differences. These findings suggest that microbial diversity changes observed in AD patients may be more relevant to females, potentially contributing to sex-specific differences in disease susceptibility and progression [55,82]. While the relationship between microbiome alteration, AD progression, and oxalate metabolism remains unclear, future research integrating microbiome data may provide new therapeutic opportunities targeting the gut microbiome to improve gastrointestinal and neurological health in AD patients [83,84].

If AD and bladder calculus are linked through microbiome dysregulation, Cluster 3 may overrepresent females in a later stage post-diagnosis. Since urinary retention is rare in females [85], symptoms may go unnoticed, leading to delayed diagnosis. Consequently, bladder calculus may only be detected once both urological and AD symptoms have advanced. However, as our study lacks data on diagnosis timing, further research incorporating disease onset and progression timelines is needed to validate this hypothesis. Such investigations may also facilitate the development of personalized therapeutic regimens tailored to disease stage and progression.

While comorbidity profiles provided meaningful insights into sex-specific patterns, we acknowledge that administrative codes alone cannot fully capture the biological complexity underlying sex differences in AD. Future studies integrating biomarkers, genetic data, and neuroimaging will be necessary to comprehensively elucidate sex-related mechanisms and their contribution to clinical heterogeneity. Additionally, psychosocial and lifestyle factors that differ by sex — such as social isolation, smoking, and physical activity[86-88] — may also contribute to the observed clustering patterns and represent important avenues for future investigation.

Final thoughts:

Through unsupervised learning, we identified five patient subgroups and revealed that AD-related comorbidities can be categorized into clinically meaningful subgroups: Cluster 2 was primarily associated with cardiovascular conditions, Cluster 3 with gastrointestinal disorders, and Cluster 4 with frailty-related conditions. Although these comorbidities correspond to well-recognized clinical syndromes, their enrichment within specific clusters indicates that they are not uniformly distributed among AD patients and may represent distinct subtypes with unique risk profiles. Our findings suggest the presence of subphenotypes related to *APOE* genotypes and microbiome alterations related to AD, which may contribute to the observed differences in comorbidity patterns across sexes.

Future studies should investigate whether the uneven distribution of these clinical syndromes influences the trajectory of cognitive decline, the timing of AD diagnosis, or treatment decisions, and should further explore underlying risk factors and biological mechanisms. Incorporating genomic and microbiome data into such studies could provide deeper insights into the complex interplay between genetic susceptibility, microbiome composition, and disease progression in AD.

Currently, anti-beta amyloid therapies are nearly universal and FDA-approved, with the next step being the addition of anti-tau therapies. Our findings suggest that, following these approaches, individuals may benefit from targeted therapeutics tailored to their specific cluster profiles, such as interventions to reduce systemic inflammation, protect against urinary tract infections, or prevent gastrointestinal diseases. Recognizing distinct subphenotypes may also inform how patients differ in their clinical trajectories and care needs. This insight could guide earlier referrals, proactive care planning, and the inclusion of multidisciplinary support — such as nephrologists, gastroenterologists, and physical therapists — to better manage comorbid conditions and potentially slow functional decline.

Limitations of the study

There are several limitations in this study. First, the presence of imbalanced data resulted in some odds ratios being disproportionately inflated. Although we applied Bonferroni correction to control for multiple testing, extremely rare events may still represent random variation or coding anomalies rather than true biological differences or clinically meaningful associations. Our findings should therefore be interpreted as hypothesis-generating rather than definitive evidence. While key findings were consistent in validation analysis, suggesting a certain level of reliability, potential biases due to sample size variation cannot be ruled out.

Second, some Pearson correlation coefficients in validation analyses were not particularly high. This finding suggests the presence of unexplored heterogeneity within both datasets. This discrepancy may partly result from demographic differences or coding practices between the UCSF and UC-Wide cohorts, which could have influenced the distribution of AD severity and comorbidity patterns. Nevertheless, several key comorbidity patterns were consistently replicated across datasets, indicating that these findings represent reproducible signals within the complexity of real-world data and provide a valuable foundation for future hypothesis-driven studies. Future multi-institutional studies with harmonized coding practices and larger, more diverse cohorts will be essential to further validate and refine these clusters.

Third, the use of EMRs, which are primarily maintained for clinical care and billing purposes rather than research, introduces potential misclassification, missing data, and variability in coding practices. Importantly, our study did not incorporate biomarkers, genetic testing, or neuroimaging data, which may have introduced diagnostic inaccuracies of AD. In addition, age data were not consistently recorded across datasets, with UCSF capping reported ages at 91 and UC-Wide data capping at 90. To minimize the impact of this limitation, age was used only for descriptive statistics and not for stratified analyses in this study. Nevertheless, this truncation may bias the representation of the oldest-old in descriptive summaries, and future studies with access to unmasked age data will be important to validate age-stratified findings and explore heterogeneity among patients over 90 years of age. Regarding

generalizability, our analysis was limited to medical institutions in California, necessitating caution when extrapolating these findings to other regions or populations.

Fourth, this study employed a cross-sectional design, precluding causal inference or determination of temporal relationships. Future longitudinal studies are warranted to explore the timing of comorbidity onset relative to AD diagnosis, identify early risk indicators, and link comorbidity patterns with outcomes such as survival or cognitive decline.

Finally, while our study primarily focused on LOAD, the lack of EOAD analysis constrains the understanding of AD heterogeneity. Future research should extend the investigation to include EOAD populations to provide a more comprehensive understanding of AD subphenotypes.

Despite these limitations, our research provides a comprehensive, robust, analysis of AD comorbidities, identifying distinct subphenotypes that highlight the heterogeneity in disease manifestation.

Resource availability

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Marina Sirota (marina.sirota@ucsf.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Ethical review and approval were waived for this study because it did not meet the definition of human participants research, and obtaining informed consent would not be possible or necessary. The datasets from UCHDW and UCDDP were fully de-identified prior to analysis, and the requirement for written informed consent was waived by the respective institutions in compliance with applicable ethical

guidelines. The data supporting the findings of this study are not publicly available due to their sensitive nature. Access is restricted to UCSF-affiliated individuals and approved collaborators. Individuals not affiliated with UCSF may establish an official collaboration with a UCSF-affiliated investigator by contacting the principal investigator, Marina Sirota (marina.sirota@ucsf.edu). Requests for collaboration are typically processed within a few weeks. UCSF-affiliated individuals seeking access to the UCSF EMR database may contact UCSF's Clinical and Translational Science Institute (ctsi@ucsf.edu) or the UCSF Information Commons team (Info.Commons@ucsf.edu) for more details. The UCDDP database is accessible only to UC researchers who have completed analyses within their respective UC institutions and have provided a justified rationale for scaling their analyses across multiple UC health centers. Censored source data for phenotype comorbidity enrichment analysis used to create Figure 3 and 6 are provided in Supporting Data.

The code corresponding to this research is available at https://github.com/yukari-katsuhara/SexSubphenotypes_AD.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request

Acknowledgements

This research was primarily supported by grants from the National Institute on Aging (NIA) (R01AG060393 and R01AG057683). The authors sincerely thank John Kornak, Thomas Hoffmann, Christopher Seaman, and all members of the Sirota Lab for their insightful discussions and valuable guidance throughout the preparation of this manuscript. The authors also thank the Center for Data-driven Insights and Innovation at UC Health (CDI2; https://www.ucop.edu/uc-health/functions/center-for-data-driven-insights-and-innovations-cdi2.html), for its analytical and technical support related to use of the UC Health Data Warehouse and related data assets, including the UC Data Discovery Platform (UCDDP).

Author contributions

Y.K., U.K., T.T.O., A.S.T., and M.S. designed research studies, conducted experiments, and acquired data. Y.K. analyzed data and wrote the manuscript. A.S.T. mentored Y.K. All authors interpreted results, edited and reviewed the manuscript.

Declaration of interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT-4o (released on May 13, 2024, by OpenAI) in order to refine language and enhance readability. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplemental information

Document S1. Figures S1–S3 and Tables S1 and S2

Table S3. Excel file containing additional data too large to fit in a PDF, related to Figures 3 and 6

STAR Methods

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Original Code	This paper	https://github.com/yukari-katsuhara/SexSubphenotypes_AD/tree/main
Software and algorithms		
Python version 3.7.9	Python Software Foundation	https://www.python.org/
sklearn.decomposition.PCA	scikit-learn	https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

sklearn.cluster.KMeans	scikit-learn	https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
sklearn.preprocessing.OneHotEncoder	scikit-learn	https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
sklearn.preprocessing.StandardScaler	scikit-learn	https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
umap	umap	https://umap-learn.readthedocs.io/en/latest/index.html
scipy	Python Package Index (PyPI)	https://scipy.org/
statsmodels.api	Python Package Index (PyPI)	https://www.statsmodels.org/stable/index.html
upsetplot.UpSet	Python Package Index (PyPI)	https://upsetplot.readthedocs.io/
R version 4.3.1	R Development Core Team	https://www.r-project.org/
tidyverse	R Development Core Team	https://www.tidyverse.org/
gt	R Development Core Team	https://gt.rstudio.com/
gtsummary	R Development Core Team	https://www.danieldsjoberg.com/gtsummary/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cohort identification

All clinical data used in this study were obtained from the University of California's Health Data Warehouse (UCHDW) and the University of California Data Discovery Platform (UCDDP). This study

identified subgroups of AD patients from UCSF EMR and validated the results using independent UC-Wide data from UCDDP to ensure the robustness of the findings.

AD patients were identified from over five million records in the UCSF EMR dataset, which contains data from 1982 to 2020. Due to the de-identification process, all dates were shifted by up to one year while preserving relative dates. Additionally, birth dates before 1930 were adjusted to 1930, effectively treating all individuals born before 1930 as having an estimated age of 90 years. Patients were included in the study if they met the inclusion criteria of an estimated age above 64 years and a diagnosis of AD based on the presence of the ICD-10-CM codes G30.1, G30.8, or G30.9. We focused on LOAD because it is a highly polygenic disorder, whereas EOAD is thought to have a distinct etiology and a stronger genetic predisposition [43]. F00.1 was intentionally excluded because, in our dataset, it was frequently applied in the context of psychiatric or behavioral diagnoses and often encompassed vascular or mixed dementia cases. Including F00.1 could therefore introduce heterogeneity from non-AD etiologies, and its exclusion allowed us to define a more clinically homogeneous AD cohort. This study was conducted as a cross-sectional analysis at the patient level. For each patient, diagnoses were aggregated across all available visits to create a single binary profile indicating the presence or absence of each diagnosis name. This approach ensured that each row in the dataset corresponded to an independent individual and mitigated the issue of correlated repeated measures across visits. Additionally, only patients with at least one documented comorbidity in addition to AD were included in the analysis, after excluding diagnosis records explicitly containing the term “Alzheimer” in diagnosis names from the comorbidity feature set. This decision was made to ensure the stability of the clustering algorithm, as a binary feature matrix with no positive entries would result in an empty vector. Moreover, because the primary objective of this study was to characterize comorbidity patterns, including patients without any comorbidities could dilute the signal within clusters and hinder the detection of meaningful patterns. Sex was determined based on the most recent sex assignment recorded in the electronic medical record. UCDDP includes the de-identified records from UCSF, UC Los Angeles (UCLA), UC San Diego (UCSD), UC Davis (UCD), UC Irvine

(UCI), and UC Riverside (UCR). As in the UCSF UCHDW dataset, the de-identification process adjusted the estimated age of all individuals 91 years or older to 91 years. The validation cohort was identified using the same criteria from UCDDP, excluding UCSF records, and consisted of UC-Wide clinical data from 2012 to 2024.

Sex as a biological variable

Our study examined both males and females using stratified analyses to identify sex differences in AD subphenotypes. By conducting these analyses, we aimed to clarify potential sex-specific variations in disease characteristics.

METHOD DETAILS

Dimensionality reduction and K-means clustering

After excluding diagnosis records explicitly containing the term “Alzheimer” in diagnosis names, one-hot encoding was performed for all diagnosis names. PCA was applied to diagnosis names to reduce the dimensionality of the dataset while preserving diagnostic information. The number of components was selected to cover at least 80% of the cumulative variance. K-means clustering was then applied to these PCA components created from diagnosis names, and the optimal number of clusters was determined based on a combination of silhouette scores and WSS reduction analysis. For robustness, we additionally calculated the Calinski–Harabasz and Davies–Bouldin indices, which yielded results consistent with the silhouette scores and WSS reduction analysis. To visualize the clustering results, patient distributions were projected onto a lower-dimensional space using UMAP.

Comorbidity enrichment analysis for overall trends of each cluster

To examine overall trends within each cluster, we performed enrichment analyses by comparing the frequency of each comorbidity in a given cluster with that in all other clusters combined. In this step, ICD-10-CM codes were used to define comorbidities, thereby facilitating clinical interpretation and

improving the interpretability of cluster characteristics. For each diagnosis, the proportion of patients in the target cluster was compared to the proportion in the other clusters using Fisher's exact test when the sample size was below five or the chi-squared test when the sample size was larger. Statistical significance was determined using a Bonferroni-corrected threshold of p-value < 0.05, and the directionality of associations was determined based on odds ratios. The statistical significance of associations was visualized using volcano plots, with effect sizes and magnitudes, while Manhattan plots (for the overall population) and Miami plots (for sex-stratified populations) illustrated the distribution of p-values across diagnoses. Sex-stratified analyses were performed to assess whether comorbidity patterns differed between male and female patients.

Comorbidity enrichment analysis for cluster- or cluster-sex-specific comorbidities

Cluster-specific comorbidities were identified by visualizing significant comorbidities using UpSet plots [89]. The statistical significance of these comorbidities was displayed in Manhattan plots for the overall population and Miami plots for sex-stratified populations. To further explore the phenotypic characteristics of each cluster, a UMAP visualization was generated using standardized odds ratios of cluster-specific comorbidities. Each dot represents a specific ICD-10-CM code, with its size corresponding to the odds ratio of the significantly associated cluster and its color indicating the respective cluster. The input features for the UMAP projection were derived from a matrix of standardized odds ratios in each cluster, where each row represents an ICD-10-CM code and each column corresponds to a cluster. Sex-stratified analyses were conducted to further investigate differences in comorbidity patterns between males and females.

Sensitivity analysis

To validate the robustness of the clustering and comorbidity enrichment results, the same approach used for the UCSF dataset was applied to the UC-Wide dataset. AD patients were identified using the same inclusion criteria, and dimensionality reduction and K-means clustering were performed as described

previously. Due to the de-identification process, all estimated ages above 91 were recorded as 91 years. Cluster-specific comorbidities in the UC-Wide dataset were analyzed using the same statistical methods and visualized with UpSet plots. The relationships between clusters in the UCSF and UC-Wide datasets were examined by comparing cluster-specific comorbidities, and the degree of overlap was assessed using Sankey plots. The coverage rate for UCSF clusters was calculated as the proportion of UCSF cluster-specific comorbidities that were statistically significant in UCSF and remained significant in UC-Wide clusters. The UC-Wide clusters with the highest coverage rates were identified as the best matches for each UCSF cluster. Log-log plots were used to compare the odds ratios of matched clusters between the UCSF and UC-Wide datasets for the overall population and sex-stratified subgroups. Pearson correlation analysis was performed to assess the linear relationship between the odds ratios, and the correlation coefficient and p-value were obtained to evaluate the statistical significance of the association.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical and computational analyses were conducted using Python, except for demographic table generation, which was performed in R. In the demographic table analysis, categorical variables such as sex, race, and death status were compared across clusters using the Chi-squared test, while continuous variables such as age and the number of comorbidities were analyzed using the Kruskal-Wallis rank sum test. When the Kruskal-Wallis test indicated a significant difference among clusters, post-hoc pairwise comparisons were conducted using Dunn's test with Bonferroni correction to identify which specific groups differed significantly.

For comorbidity enrichment analysis, the proportions of patients in each ICD-10-CM code were compared using either Fisher's exact test, if fewer than five patients were present in a given ICD-10-CM code, or the Chi-squared test otherwise. Diagnoses were considered significant if the p-value was below 0.05 after Bonferroni correction, and the directionality of the association was determined using the OR.

References

1. Jack, C.R., Jr., Bennett, D.A., Blennow, K., Carrillo, M.C., Dunn, B., Haeberlein, S.B., Holtzman, D.M., Jagust, W., Jessen, F., Karlawish, J., et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dement *14*, 535-562. 10.1016/j.jalz.2018.02.018.
2. Uddin, M.S., and Lim, L.W. (2022). Glial cells in Alzheimer's disease: From neuropathological changes to therapeutic implications. Ageing Research Reviews *78*, 101622. https://doi.org/10.1016/j.arr.2022.101622.
3. Hersi, M., Irvine, B., Gupta, P., Gomes, J., Birkett, N., and Krewski, D. (2017). Risk factors associated with the onset and progression of Alzheimer's disease: A systematic review of the evidence. Neurotoxicology *61*, 143-187. 10.1016/j.neuro.2017.03.006.
4. Liu, C.C., Liu, C.C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol *9*, 106-118. 10.1038/nrneurol.2012.263.
5. Miller, Z.A., Ossenkoppele, R., Graff-Radford, N.R., Allen, I.E., Shwe, W., Rosenberg, L., Olguin, D.J., Erkkinen, M.G., Butler, P.M., Spina, S., et al. (2022). Left-handedness, learning disability, autoimmune disease, and seizure history influence age at onset and phenotypical targeting of Alzheimer's disease. medRxiv, 2022.2012.2017.22283307. 10.1101/2022.12.17.22283307.
6. 2024 Alzheimer's disease facts and figures. (2024). Alzheimers Dement *20*, 3708-3821. 10.1002/alz.13809.
7. Irvine, K., Laws, K.R., Gale, T.M., and Kondel, T.K. (2012). Greater cognitive deterioration in women than men with Alzheimer's disease: a meta analysis. J Clin Exp Neuropsychol *34*, 989-998. 10.1080/13803395.2012.712676.
8. Ferretti, M.T., Iulita, M.F., Cavedo, E., Chiesa, P.A., Schumacher Dimech, A., Santuccione Chadha, A., Baracchi, F., Girouard, H., Misoch, S., Giacobini, E., et al. (2018). Sex differences in Alzheimer disease — the gateway to precision medicine. Nature Reviews Neurology *14*, 457-469. 10.1038/s41582-018-0032-9.
9. Aggarwal, N.T., and Mielke, M.M. (2023). Sex Differences in Alzheimer's Disease. Neurol Clin *41*, 343-358. 10.1016/j.ncl.2023.01.001.
10. Gilsanz, P., Mayeda, E.R., Glymour, M.M., Quesenberry, C.P., Mungas, D.M., DeCarli, C., Dean, A., and Whitmer, R.A. (2017). Female sex, early-onset hypertension, and risk of dementia. Neurology *89*, 1886-1893. doi:10.1212/WNL.0000000000004602.
11. Rogalski, E.J., Rademaker, A., Harrison, T.M., Helenowski, I., Johnson, N., Bigio, E., Mishra, M., Weintraub, S., and Mesulam, M.M. (2011). ApoE E4 is a susceptibility factor in amnestic but not aphasic dementias. Alzheimer Dis Assoc Disord *25*, 159-163. 10.1097/WAD.0b013e318201f249.
12. Arnold, M., Nho, K., Kueider-Paisley, A., Massaro, T., Huynh, K., Brauner, B., MahmoudianDehkordi, S., Louie, G., Moseley, M.A., Thompson, J.W., et al. (2020). Sex and APOE ε4 genotype modify the Alzheimer's disease serum metabolome. Nat Commun *11*, 1148. 10.1038/s41467-020-14959-w.
13. Larson, E.B., Shadlen, M.F., Wang, L., McCormick, W.C., Bowen, J.D., Teri, L., and Kukull, W.A. (2004). Survival after initial diagnosis of Alzheimer disease. Ann Intern Med *140*, 501-509. 10.7326/0003-4819-140-7-200404060-00008.
14. Xie, J., Brayne, C., and Matthews, F.E. (2008). Survival times in people with dementia: analysis from population based cohort study with 14 year follow-up. BMJ *336*, 258-262. 10.1136/bmj.39433.616678.25.
15. Santiago, J.A., and Potashkin, J.A. (2023). Biological and Clinical Implications of Sex-Specific Differences in Alzheimer's Disease. Handb Exp Pharmacol *282*, 181-197. 10.1007/164_2023_672.
16. Neff, R.A., Wang, M., Vatansever, S., Guo, L., Ming, C., Wang, Q., Wang, E., Horgusluoglu-Moloch, E., Song, W.-m., Li, A., et al. (2021). Molecular subtyping of

Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. Science Advances *7*, eabb5398. doi:10.1126/sciadv.abb5398.

17. Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., et al. (2018). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nature Communications *9*, 4273. 10.1038/s41467-018-05892-0.
18. Geifman, N., Kennedy, R.E., Schneider, L.S., Buchan, I., and Brinton, R.D. (2018). Data-driven identification of endophenotypes of Alzheimer's disease progression: implications for clinical trials and therapeutic interventions. Alzheimer's Research & Therapy *10*, 4. 10.1186/s13195-017-0332-0.
19. Xu, J., Wang, F., Xu, Z., Adekkanattu, P., Brandt, P., Jiang, G., Kiefer, R.C., Luo, Y., Mao, C., Pacheco, J.A., et al. (2020). Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. Learn Health Syst *4*, e10246. 10.1002/lrh2.10246.
20. Alexander, N., Alexander, D.C., Barkhof, F., and Denaxas, S. (2020). Using Unsupervised Learning to Identify Clinical Subtypes of Alzheimer's Disease in Electronic Health Records. Stud Health Technol Inform *270*, 499-503. 10.3233/shti200210.
21. Tang, A.S., Oskotsky, T., Havaldar, S., Mantyh, W.G., Bicak, M., Solsberg, C.W., Woldemariam, S., Zeng, B., Hu, Z., Oskotsky, B., et al. (2022). Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. Nature Communications *13*, 675. 10.1038/s41467-022-28273-0.
22. Meng, W., Xu, J., Huang, Y., Wang, C., Song, Q., Ma, A., Song, L., Bian, J., Ma, Q., and Yin, R. (2024). Autoencoder to Identify Sex-Specific Sub-phenotypes in Alzheimer's Disease Progression Using Longitudinal Electronic Health Records. medRxiv. 10.1101/2024.07.07.24310055.
23. Woldemariam, S.R., Tang, A.S., Oskotsky, T.T., Yaffe, K., and Sirota, M. (2023). Similarities and differences in Alzheimer's dementia comorbidities in racialized populations identified from electronic medical records. Communications Medicine *3*, 50. 10.1038/s43856-023-00280-2.
24. McAdams-DeMarco, M.A., Bae, S., Chu, N., Gross, A.L., Brown, C.H.t., Oh, E., Rosenberg, P., Neufeld, K.J., Varadhan, R., Albert, M., et al. (2017). Dementia and Alzheimer's Disease among Older Kidney Transplant Recipients. J Am Soc Nephrol *28*, 1575-1583. 10.1681/asn.2016080816.
25. Ruck, J.M., Chu, N.M., Liu, Y., Li, Y., Chen, Y., Mathur, A., Carlson, M.C., Crews, D.C., Chodosh, J., Segev, D.L., and McAdams-DeMarco, M. (2024). Association of Postoperative Delirium With Incident Dementia and Graft Outcomes Among Kidney Transplant Recipients. Transplantation *108*, 530-538. 10.1097/tp.0000000000004779.
26. Evans, D.A., Bennett, D.A., Wilson, R.S., Bienias, J.L., Morris, M.C., Scherr, P.A., Hebert, L.E., Aggarwal, N., Beckett, L.A., Joglekar, R., et al. (2003). Incidence of Alzheimer disease in a biracial urban community: relation to apolipoprotein E allele status. Arch Neurol *60*, 185-189. 10.1001/archneur.60.2.185.
27. Tang, M.X., Stern, Y., Marder, K., Bell, K., Gurland, B., Lantigua, R., Andrews, H., Feng, L., Tycko, B., and Mayeux, R. (1998). The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. Jama *279*, 751-755. 10.1001/jama.279.10.751.
28. Weuve, J., Barnes, L.L., Mendes de Leon, C.F., Rajan, K.B., Beck, T., Aggarwal, N.T., Hebert, L.E., Bennett, D.A., Wilson, R.S., and Evans, D.A. (2018). Cognitive Aging in Black and White Americans: Cognition, Cognitive Decline, and Incidence of Alzheimer Disease Dementia. Epidemiology *29*, 151-159. 10.1097/ede.0000000000000747.
29. Nasr, H., and Scriven, J.M. (2015). Superficial thrombophlebitis (superficial venous thrombosis). Bmj *350*, h2039. 10.1136/bmj.h2039.

30. Décousus, H., Bertoletti, L., and Frappé, P. (2015). Spontaneous acute superficial vein thrombosis of the legs: do we really need to treat? J Thromb Haemost *13 Suppl 1*, S230-237. 10.1111/jth.12925.
31. Di Minno, M.N., Ambrosino, P., Ambrosini, F., Tremoli, E., Di Minno, G., and Dentali, F. (2016). Prevalence of deep vein thrombosis and pulmonary embolism in patients with superficial vein thrombosis: a systematic review and meta-analysis. J Thromb Haemost *14*, 964-972. 10.1111/jth.13279.
32. Simões, A.M.N., Vendramim, P., and Pedreira, M.L.G. (2022). Risk factors for peripheral intravenous catheter-related phlebitis in adult patients. Rev Esc Enferm USP *56*, e20210398. 10.1590/1980-220X-REEUSP-2021-0398en.
33. Undas, A. (2017). Prothrombotic Fibrin Clot Phenotype in Patients with Deep Vein Thrombosis and Pulmonary Embolism: A New Risk Factor for Recurrence. Biomed Res Int *2017*, 8196256. 10.1155/2017/8196256.
34. Czysz A, H.S. (2023). Superficial Thrombophlebitis. StatPearls [Internet].
35. Cortes-Canteli, M., Zamolodchikov, D., Ahn, H.J., Strickland, S., and Norris, E.H. (2012). Fibrinogen and altered hemostasis in Alzheimer's disease. J Alzheimers Dis *32*, 599-608. 10.3233/jad-2012-120820.
36. Cortes-Canteli, M., Paul, J., Norris, E.H., Bronstein, R., Ahn, H.J., Zamolodchikov, D., Bhuvanendran, S., Fenz, K.M., and Strickland, S. (2010). Fibrinogen and beta-amyloid association alters thrombosis and fibrinolysis: a possible contributing factor to Alzheimer's disease. Neuron *66*, 695-709. 10.1016/j.neuron.2010.05.014.
37. Fibrin induces neurotoxic microglia gene programs in neurodegeneration. (2023). Nature Immunology *24*, 1062-1063. 10.1038/s41590-023-01542-w.
38. Merlini, M., Rafalski, V.A., Rios Coronado, P.E., Gill, T.M., Ellisman, M., Muthukumar, G., Subramanian, K.S., Ryu, J.K., Syme, C.A., Davalos, D., et al. (2019). Fibrinogen Induces Microglia-Mediated Spine Elimination and Cognitive Impairment in an Alzheimer's Disease Model. Neuron *101*, 1099-1108.e1096. 10.1016/j.neuron.2019.01.014.
39. Ryu, J.K., Rafalski, V.A., Meyer-Franke, A., Adams, R.A., Poda, S.B., Rios Coronado, P.E., Pedersen, L., Menon, V., Baeten, K.M., Sikorski, S.L., et al. (2018). Fibrin-targeting immunotherapy protects against neuroinflammation and neurodegeneration. Nat Immunol *19*, 1212-1223. 10.1038/s41590-018-0232-x.
40. Bergamaschini, L., Rossi, E., Storini, C., Pizzimenti, S., Distaso, M., Perego, C., De Luigi, A., Vergani, C., and De Simoni, M.G. (2004). Peripheral treatment with enoxaparin, a low molecular weight heparin, reduces plaques and beta-amyloid accumulation in a mouse model of Alzheimer's disease. J Neurosci *24*, 4181-4186. 10.1523/jneurosci.0550-04.2004.
41. Avitan, I., Halperin, Y., Saha, T., Bloch, N., Atrahimovich, D., Polis, B., Samson, A.O., and Braitbard, O. (2021). Towards a Consensus on Alzheimer's Disease Comorbidity? J Clin Med *10*. 10.3390/jcm10194360.
42. Nedelec, T., Couvy-Duchesne, B., Monnet, F., Daly, T., Ansart, M., Gantzer, L., Lekens, B., Epelbaum, S., Dufouil, C., and Durrleman, S. (2022). Identifying health conditions associated with Alzheimer's disease up to 15 years before diagnosis: an agnostic study of French and British health records. The Lancet Digital Health *4*, e169-e178. https://doi.org/10.1016/S2589-7500(21)00275-2.
43. Wingo, T.S., Lah, J.J., Levey, A.I., and Cutler, D.J. (2012). Autosomal recessive causes likely in early-onset Alzheimer disease. Arch Neurol *69*, 59-64. 10.1001/archneurol.2011.221.
44. Sirkis, D.W., Bonham, L.W., Johnson, T.P., La Joie, R., and Yokoyama, J.S. (2022). Dissecting the clinical heterogeneity of early-onset Alzheimer's disease. Mol Psychiatry *27*, 2674-2688. 10.1038/s41380-022-01531-9.

45. Mantyh, W.G., Cochran, J.N., Taylor, J.W., Broce, I.J., Geier, E.G., Bonham, L.W., Anderson, A.G., Sirkis, D.W., Joie, R., Iaccarino, L., et al. (2023). Early-onset Alzheimer's disease explained by polygenic risk of late-onset disease? Alzheimers Dement (Amst) *15*, e12482. 10.1002/dad2.12482.
46. Falgàs, N., Allen, I.E., Spina, S., Grant, H., Piña Escudero, S.D., Merrilees, J., Gearhart, R., Rosen, H.J., Kramer, J.H., Seeley, W.W., et al. (2022). The severity of neuropsychiatric symptoms is higher in early-onset than late-onset Alzheimer's disease. Eur J Neurol *29*, 957-967. 10.1111/ene.15203.
47. Ramey, G.D., Tang, A., Phongpreecha, T., Yang, M.M., Woldemariam, S.R., Oskotsky, T.T., Montine, T.J., Allen, I., Miller, Z.A., Aghaeepour, N., et al. (2025). Exposure to autoimmune disorders is associated with increased Alzheimer's disease risk in a multi-site electronic health record analysis. Cell Rep Med, 101980. 10.1016/j.xcrm.2025.101980.
48. Sundaresan, B., Shirafkan, F., Ripperger, K., and Rattay, K. (2023). The Role of Viral Infections in the Onset of Autoimmune Diseases. Viruses *15*. 10.3390/v15030782.
49. Liao, L. (2015). Evaluation and Management of Neurogenic Bladder: What Is New in China? Int J Mol Sci *16*, 18580-18600. 10.3390/ijms160818580.
50. Gómez-Gómez, M.E., and Zapico, S.C. (2019). Frailty, Cognitive Decline, Neurodegenerative Diseases and Nutrition Interventions. Int J Mol Sci *20*. 10.3390/ijms20112842.
51. Makhnevich, A., Perrin, A., Talukder, D., Liu, Y., Izard, S., Chiuzan, C., D'Angelo, S., Affoo, R., Rogus-Pulia, N., and Sinvani, L. (2024). Thick Liquids and Clinical Outcomes in Hospitalized Patients With Alzheimer Disease and Related Dementias and Dysphagia. JAMA Internal Medicine *184*, 778-785. 10.1001/jamainternmed.2024.0736.
52. Yurtdaş Depboylu, G., Acar Tek, N., Akbulut, G., Günel, Z., and Kamanlı, B. (2023). Functional Constipation in Elderly and Related Determinant Risk Factors: Malnutrition and Dietary Intake. J Am Nutr Assoc *42*, 541-547. 10.1080/27697061.2022.2096150.
53. Seo, D.O., and Holtzman, D.M. (2024). Current understanding of the Alzheimer's disease-associated microbiome and therapeutic strategies. Exp Mol Med *56*, 86-94. 10.1038/s12276-023-01146-2.
54. Kuźniar, J., Kozubek, P., Czaja, M., and Leszek, J. (2024). Correlation between Alzheimer's Disease and Gastrointestinal Tract Disorders. Nutrients *16*. 10.3390/nu16142366.
55. Vogt, N.M., Kerby, R.L., Dill-McFarland, K.A., Harding, S.J., Merluzzi, A.P., Johnson, S.C., Carlsson, C.M., Asthana, S., Zetterberg, H., Blennow, K., et al. (2017). Gut microbiome alterations in Alzheimer's disease. Sci Rep *7*, 13537. 10.1038/s41598-017-13601-y.
56. Cunha, B.A. (2001). Pneumonia in the elderly. Clin Microbiol Infect *7*, 581-588. 10.1046/j.1198-743x.2001.00328.x.
57. Whitson, H.E., Colton, C., El Khoury, J., Gate, D., Goate, A., Heneka, M.T., Kaddurah-Daouk, R., Klein, R.S., Shinohara, M.L., Sisodia, S., et al. (2022). Infection and inflammation: New perspectives on Alzheimer's disease. Brain Behav Immun Health *22*, 100462. 10.1016/j.bbih.2022.100462.
58. Lehrer, S., and Rheinstein, P.H. (2022). Circulating Chlamydia Trachomatis Antigens in Subjects With Alzheimer's Disease. In Vivo *36*, 2650-2653. 10.21873/invivo.12999.
59. Lehrer, S., and Rheinstein, P.H. (2024). RPTOR Is an Alzheimer's Disease Susceptibility Gene Associated with the Risk Factors Body Mass Index and Infectious Encephalitis. J Alzheimers Dis Rep *8*, 715-721. 10.3233/adr-230185.
60. Ukraintseva, S., Duan, M., Simanek, A.M., Holmes, R., Bagley, O., Rajendrakumar, A.L., Yashkin, A.P., Akushevich, I., Tropsha, A., Whitson, H., et al. (2023). Vaccination

Against Pneumonia May Provide Genotype-Specific Protection Against Alzheimer's Disease. J Alzheimers Dis *96*, 499-505. 10.3233/jad-230088.

61. Lehrer, S., and Rheinstein, P.H. (2022). Vaccination Reduces Risk of Alzheimer's Disease, Parkinson's Disease and Other Neurodegenerative Disorders. Discov Med *34*, 97-101.

62. Iwai-Saito, K., Sato, K., Fujii, M., and Kondo, K. (2024). Pneumococcal vaccination, but not influenza vaccination, is negatively associated with incident dementia among Japanese older adults: The JAGES 2013–2022 prospective cohort study. Brain, Behavior, and Immunity *120*, 452-463. https://doi.org/10.1016/j.bbi.2024.06.020.

63. Huo, X., and Finkelstein, J. (2024). Pneumococcal Vaccination Lowers the Risk of Alzheimer's Disease: A Study Utilizing Data from the IBM® MarketScan® Database. Stud Health Technol Inform *310*, 961-965. 10.3233/shti231107.

64. Perry, D.C., Brown, J.A., Possin, K.L., Datta, S., Trujillo, A., Radke, A., Karydas, A., Kornak, J., Sias, A.C., Rabinovici, G.D., et al. (2017). Clinicopathological correlations in behavioural variant frontotemporal dementia. Brain *140*, 3329-3345. 10.1093/brain/awx254.

65. Musa, G., Slachevsky, A., Muñoz-Neira, C., Méndez-Orellana, C., Villagra, R., González-Billault, C., Ibáñez, A., Hornberger, M., and Lillo, P. (2020). Alzheimer's Disease or Behavioral Variant Frontotemporal Dementia? Review of Key Points Toward an Accurate Clinical and Neuropsychological Diagnosis. J Alzheimers Dis *73*, 833-848. 10.3233/jad-190924.

66. Wagner, M., Lorenz, G., Volk, A.E., Brunet, T., Edbauer, D., Berutti, R., Zhao, C., Anderl-Straub, S., Bertram, L., Danek, A., et al. (2021). Clinico-genetic findings in 509 frontotemporal dementia patients. Molecular Psychiatry *26*, 5824-5832. 10.1038/s41380-021-01271-2.

67. Ranasinghe, K.G., Rankin, K.P., Pressman, P.S., Perry, D.C., Lobach, I.V., Seeley, W.W., Coppola, G., Karydas, A.M., Grinberg, L.T., Shany-Ur, T., et al. (2016). Distinct Subtypes of Behavioral Variant Frontotemporal Dementia Based on Patterns of Network Degeneration. JAMA Neurol *73*, 1078-1088. 10.1001/jamaneurol.2016.2016.

68. Wang, Y., Yang, P., Yan, Z., Liu, Z., Ma, Q., Zhang, Z., Wang, Y., and Su, Y. (2021). The Relationship between Erythrocytes and Diabetes Mellitus. J Diabetes Res *2021*, 6656062. 10.1155/2021/6656062.

69. Lowe, G.D. (2001). The relationship between infection, inflammation, and cardiovascular disease: an overview. Ann Periodontol *6*, 1-8. 10.1902/annals.2001.6.1.1.

70. Pankratz, V.S., Roberts, R.O., Mielke, M.M., Knopman, D.S., Jack, C.R., Geda, Y.E., Rocca, W.A., and Petersen, R.C. (2015). Predicting the risk of mild cognitive impairment in the Mayo Clinic Study of Aging. Neurology *84*, 1433-1442. doi:10.1212/WNL.0000000000001437.

71. Gong, J., Harris, K., Peters, S.A.E., and Woodward, M. (2021). Sex differences in the association between major cardiovascular risk factors in midlife and dementia: a cohort study using data from the UK Biobank. BMC Med *19*, 110. 10.1186/s12916-021-01980-z.

72. Chatterjee, S., Peters, S.A.E., Woodward, M., Mejia Arango, S., Batty, G.D., Beckett, N., Beiser, A., Borenstein, A.R., Crane, P.K., Haan, M., et al. (2015). Type 2 Diabetes as a Risk Factor for Dementia in Women Compared With Men: A Pooled Analysis of 2.3 Million People Comprising More Than 100,000 Cases of Dementia. Diabetes Care *39*, 300-307. 10.2337/dc15-1588.

73. Anthopoulos, P.G., Hamodrakas, S.J., and Bagos, P.G. (2010). Apolipoprotein E polymorphisms and type 2 diabetes: a meta-analysis of 30 studies including 5423 cases and 8197 controls. Mol Genet Metab *100*, 283-291. 10.1016/j.ymgme.2010.03.008.

74. Shi, J., Liu, Y., Liu, Y., Li, Y., Qiu, S., Bai, Y., Gu, Y., Luo, J., Cui, H., Li, Y., et al. (2018). Association between ApoE polymorphism and hypertension: A meta-analysis of 28

studies including 5898 cases and 7518 controls. Gene *675*, 197-207. 10.1016/j.gene.2018.06.097.

75. Bennet, A.M., Di Angelantonio, E., Ye, Z., Wensley, F., Dahlin, A., Ahlbom, A., Keavney, B., Collins, R., Wiman, B., de Faire, U., and Danesh, J. (2007). Association of apolipoprotein E genotypes with lipid levels and coronary risk. Jama *298*, 1300-1311. 10.1001/jama.298.11.1300.
76. Riedel, B.C., Thompson, P.M., and Brinton, R.D. (2016). Age, APOE and sex: Triad of risk of Alzheimer's disease. The Journal of Steroid Biochemistry and Molecular Biology *160*, 134-147. https://doi.org/10.1016/j.jsbmb.2016.03.012.
77. Maestre, G., Ottman, R., Stern, Y., Gurland, B., Chun, M., Tang, M.X., Shelanski, M., Tycko, B., and Mayeux, R. (1995). Apolipoprotein E and Alzheimer's disease: ethnic variation in genotypic risks. Ann Neurol *37*, 254-259. 10.1002/ana.410370217.
78. Li, M.L., Song, S.C., Yang, F., Gao, C., Zhou, B., and Wang, Q. (2024). Risk assessment and prevention of urolithiasis in urban areas of Baoding, China. Medicine (Baltimore) *103*, e35880. 10.1097/md.0000000000035880.
79. Stamatelou, K., and Goldfarb, D.S. (2023). Epidemiology of Kidney Stones. Healthcare (Basel) *11*. 10.3390/healthcare11030424.
80. Siener, R., Herwig, H., Rüdy, J., Schaefer, R.M., Lossin, P., and Hesse, A. (2022). Urinary stone composition in Germany: results from 45,783 stone analyses. World J Urol *40*, 1813-1820. 10.1007/s00345-022-04060-w.
81. Miller, A.W., Penniston, K.L., Fitzpatrick, K., Agudelo, J., Tasian, G., and Lange, D. (2022). Mechanisms of the intestinal and urinary microbiome in kidney stone disease. Nat Rev Urol *19*, 695-707. 10.1038/s41585-022-00647-5.
82. Dodiya, H.B., Lutz, H.L., Weigle, I.Q., Patel, P., Michalkiewicz, J., Roman-Santiago, C.J., Zhang, C.M., Liang, Y., Srinath, A., Zhang, X., et al. (2022). Gut microbiota-driven brain Aβ amyloidosis in mice requires microglia. J Exp Med *219*. 10.1084/jem.20200895.
83. Bashir, B., Gulati, M., Vishwas, S., Gupta, G., Dhanasekaran, M., Paudel, K.R., Chellappan, D.K., Anand, K., Negi, P., Singh, P.K., et al. (2025). Bridging gap in the treatment of Alzheimer's disease via postbiotics: Current practices and future prospects. Ageing Res Rev *105*, 102689. 10.1016/j.arr.2025.102689.
84. Jiang, X., Zheng, Y., Sun, H., Dang, Y., Yin, M., Xiao, M., and Wu, T. (2025). Fecal Microbiota Transplantation Improves Cognitive Function of a Mouse Model of Alzheimer's Disease. CNS Neurosci Ther *31*, e70259. 10.1111/cns.70259.
85. Mevcha, A., and Drake, M.J. (2010). Etiology and management of urinary retention in women. Indian J Urol *26*, 230-235. 10.4103/0970-1591.65396.
86. Ren, Y., Savadlou, A., Park, S., Siska, P., Epp, J.R., and Sargin, D. (2023). The impact of loneliness and social isolation on the development of cognitive decline and Alzheimer's Disease. Front Neuroendocrinol *69*, 101061. 10.1016/j.yfrne.2023.101061.
87. Qiu, C., Kivipelto, M., and von Strauss, E. (2009). Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention. Dialogues Clin Neurosci *11*, 111-128. 10.31887/DCNS.2009.11.2/cqiu.
88. Su, S., Shi, L., Zheng, Y., Sun, Y., Huang, X., Zhang, A., Que, J., Sun, X., Shi, J., Bao, Y., et al. (2022). Leisure Activities and the Risk of Dementia: A Systematic Review and Meta-analysis. Neurology *99*, e1651-e1663. 10.1212/wnl.0000000000200929.
89. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. IEEE Trans Vis Comput Graph *20*, 1983-1992. 10.1109/tvcg.2014.2346248.

	Cluster 1 *N* = 328		Cluster 2 *N* = 538		Cluster 3 *N* = 815		Cluster 4 *N* = 1,521		Cluster 5 *N* = 5,161		Overall *N* = 8,363		p-val
Sex, *n* (%)													< 0.05
Female	217	(66%)	358	(67%)	514	(63%)	1,083	(71%)	3,143	(61%)	5,315	(64%)	
Male	110	(34%)	180	(33%)	301	(37%)	438	(29%)	2,000	(39%)	3,029	(36%)	
Unknown	<10	(-)	<10	(-)	<10	(-)	<10	(-)	18	(0.3%)	19	(0.2%)	
Race, *n* (%)													< 0.05
White or Caucasian	148	(45%)	248	(46%)	439	(54%)	897	(59%)	3,431	(66%)	5,163	(62%)	
Other/Unknown	45	(14%)	98	(18%)	75	(9.2%)	205	(13%)	886	(17%)	1,309	(16%)	
Asian	108	(33%)	90	(17%)	222	(27%)	109	(7.2%)	347	(6.7%)	876	(10%)	
Black or African American	27	(8.2%)	70	(13%)	65	(8.0%)	180	(12%)	240	(4.7%)	582	(7.0%)	
Native Hawaiian or Other Pacific Islander	<10	(-)	32	(5.9%)	14	(1.7%)	130	(8.5%)	257	(5.0%)	433	(5.2%)	
Death Status, *n* (%)													< 0.05
Alive	218	(66%)	258	(48%)	550	(67%)	912	(60%)	4,342	(84%)	6,280	(75%)	
Deceased	110	(34%)	280	(52%)	265	(33%)	609	(40%)	819	(16%)	2,083	(25%)	
Age (yr), Median (IQR)	89.5	(84.0-91.0)	90.0	(90.0-91.0)	88.0	(82.0-90.0)	91.0	(90.0-91.0)	90.0	(81.0-91.0)	90.0	(84.0-91.0)	< 0.05
Comorbidities, Median (IQR)	313.5	(258.0-379.0)	132.0	(106.0-175.5)	114.0	(80.0-156.0)	47.0	(33.0-65.0)	12.0	(5.0-21.0)	23.0	(9.0-63.0)	< 0.05

Table 1. Patient demographics in UCSF. Medians and interquartile ranges are presented as median (25th percentile-75th percentile). "Comorbidities" refers to "Number of comorbidities," which represents the total count of diagnoses per patient. The category "Unknown" in Sex includes records classified as “Unknown.” The category "Other/Unknown" in Race includes records classified as “American Indian or Alaska Native,” “Other,” “0,” “Unknown/Declined,” “Unknown,” and “Declined.” P-values were calculated using Pearson's Chi-squared test for Sex, Race, Death Status, and Location, and Kruskal-Wallis rank sum test for Age and Comorbidities. p-val, p-value.

	Cluster A *N* = 1,765		Cluster B *N* = 12,112		Cluster C *N* = 2,080		Cluster D *N* = 4,146		Cluster E *N* = 5,793		Overall *N* = 25,896		p-val
Sex, *n* (%)													< 0.05
Female	1,228	(69.6%)	7,414	(61.2%)	1,170	(56.2%)	2,370	(57.2%)	3,854	(66.5%)	16,036	(61.9%)	
Male	537	(30.4%)	4,672	(38.6%)	910	(43.8%)	1,776	(42.8%)	1,938	(33.5%)	9,833	(38.0%)	
Unknown	<10	(-)	26	(0.2%)	<10	(-)	<10	(-)	<10	(-)	27	(0.1%)	
Race, *n* (%)													< 0.05
White or Caucasian	1,234	(69.9%)	7,828	(64.6%)	1,322	(63.6%)	2,549	(61.5%)	4,085	(70.5%)	17,018	(65.7%)	
Other/Unknown	248	(14.1%)	2,851	(23.5%)	290	(13.9%)	728	(17.6%)	775	(13.4%)	4,892	(18.9%)	
Asian	147	(8.3%)	993	(8.2%)	271	(13.0%)	556	(13.4%)	625	(10.8%)	2,592	(10.0%)	
Black or African American	131	(7.4%)	419	(3.5%)	187	(9.0%)	287	(6.9%)	295	(5.1%)	1,319	(5.1%)	
Native Hawaiian or Other Pacific Islander	<10	(-)	21	(0.2%)	10	(0.5%)	26	(0.6%)	13	(0.2%)	785	(0.3%)	
Death Status, *n* (%)													< 0.05
Alive	1,221	(69.2%)	6,265	(51.7%)	717	(34.5%)	1,617	(39.0%)	3,367	(58.1%)	13,187	(50.9%)	
Deceased	544	(30.8%)	5,847	(48.3%)	1,363	(65.6%)	2,529	(61.0%)	2,426	(41.9%)	12,709	(49.1%)	
Location, *n* (%)													< 0.05
Site 1	320	(18.1%)	2019	(16.7%)	315	(15.1%)	590	(14.2%)	1,112	(19.2%)	4,356	(16.8%)	
Site 2	123	(7.0%)	2,629	(21.7%)	185	(8.9%)	1,032	(24.9%)	566	(9.8%)	4,535	(17.5%)	
Site 3	765	(43.3%)	4,211	(34.8%)	1,185	(57.0%)	1,395	(33.6%)	2,845	(49.1%)	10,401	(40.2%)	
Site 4	<10	(-)	85	(0.7%)	<10	(-)	<10	(-)	<10	(-)	92	(0.4%)	
Site 5	557	(31.6%)	3,168	(26.2%)	395	(19.0%)	1,128	(27.2%)	1,264	(21.8%)	6,512	(25.1%)	
Age (yr), Median (IQR)	87.0	(82.0-90.0)	88.0	(81.0-89.0)	90.0	(87.0-90.0)	90.0	(84.0-90.0)	88.0	(82.0-90.0)	89.0	(82.0-90.0)	< 0.05
Comorbidities, Median (IQR)	201.0	(166.0-253.0)	15.0	(7.0-25.0)	153.0	(127.0-185.0)	52.0	(39.0-71.0)	76.0	(57.0-100.0)	40.0	(16.0-88.0)	< 0.05

Table 2. Patient demographics in UC-Wide. Medians and interquartile ranges are presented as median (25th percentile-75th percentile). "Comorbidities" refers to "Number of comorbidities," which represents the total count of diagnoses per patient. The category "Unknown" in Sex includes records classified as “Unknown” and “*Unspecified.” The category "Other/Unknown" in Race includes records classified as “American Indian or Alaska Native,” “Other Race,” and “Unknown.” Site1 to Site5 are obfuscated representations of the following institutions: UCLA, UCSD, UCD, University of California, Irvine (UCI), and University of California, Riverside (UCR). P-values were calculated using Pearson's Chi-squared test for Sex, Race, Death Status, and Location, and Kruskal-Wallis rank sum test for Age and Comorbidities. p-val, p-value.

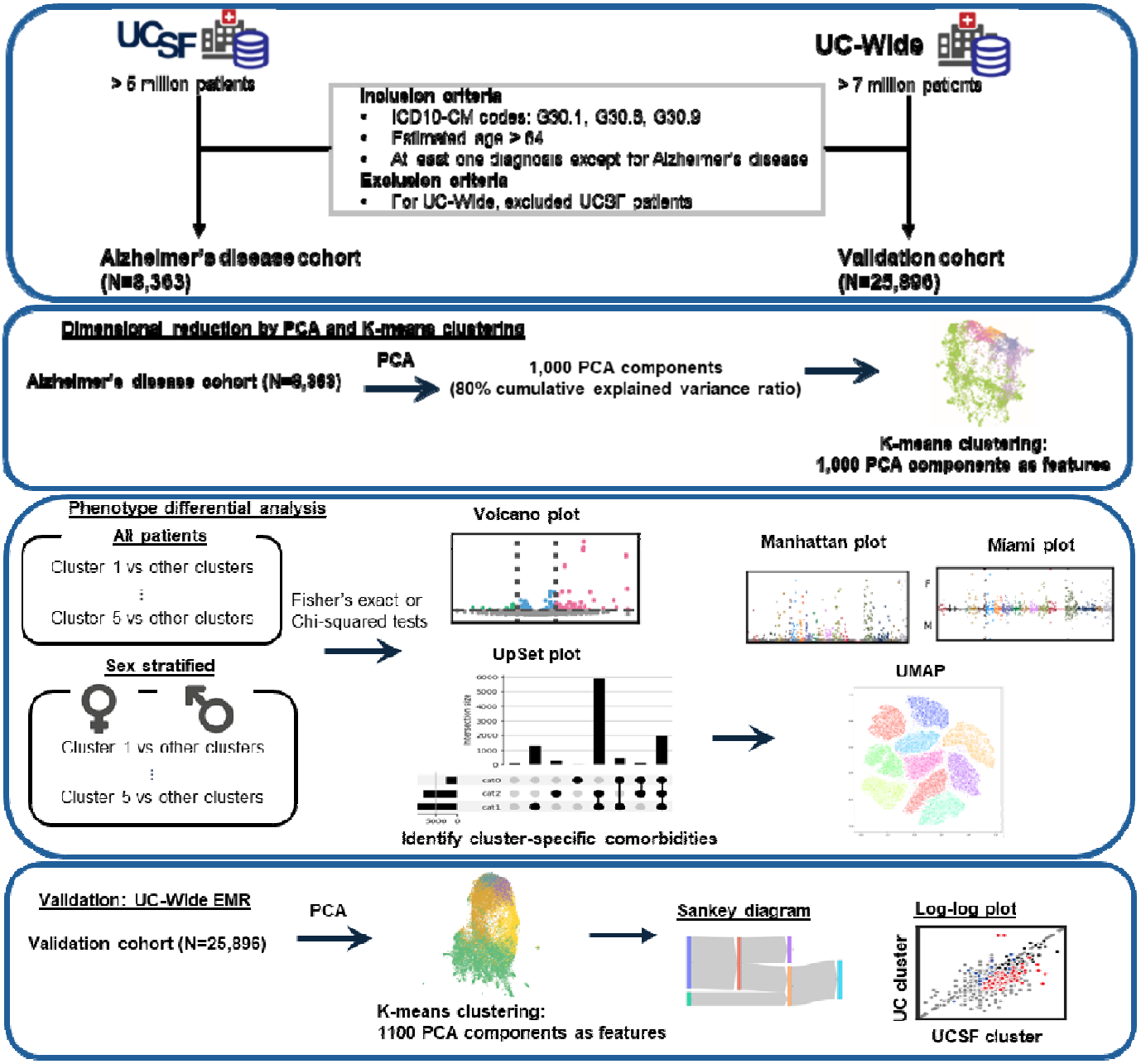

Figure 1. Cohort selection and overview of study design. Alzheimer's disease patients > 64 years with at least one comorbidity were included in this study. (1) K-means clustering was performed following Principal Component Analysis (PCA). (2) Each cluster was characterized based on cluster-specific comorbidities. For validation, (3) the same approach was applied to the UC-Wide cohort, excluding UCSF patients. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification; PCA, Principal Component Analysis; UC, University of California

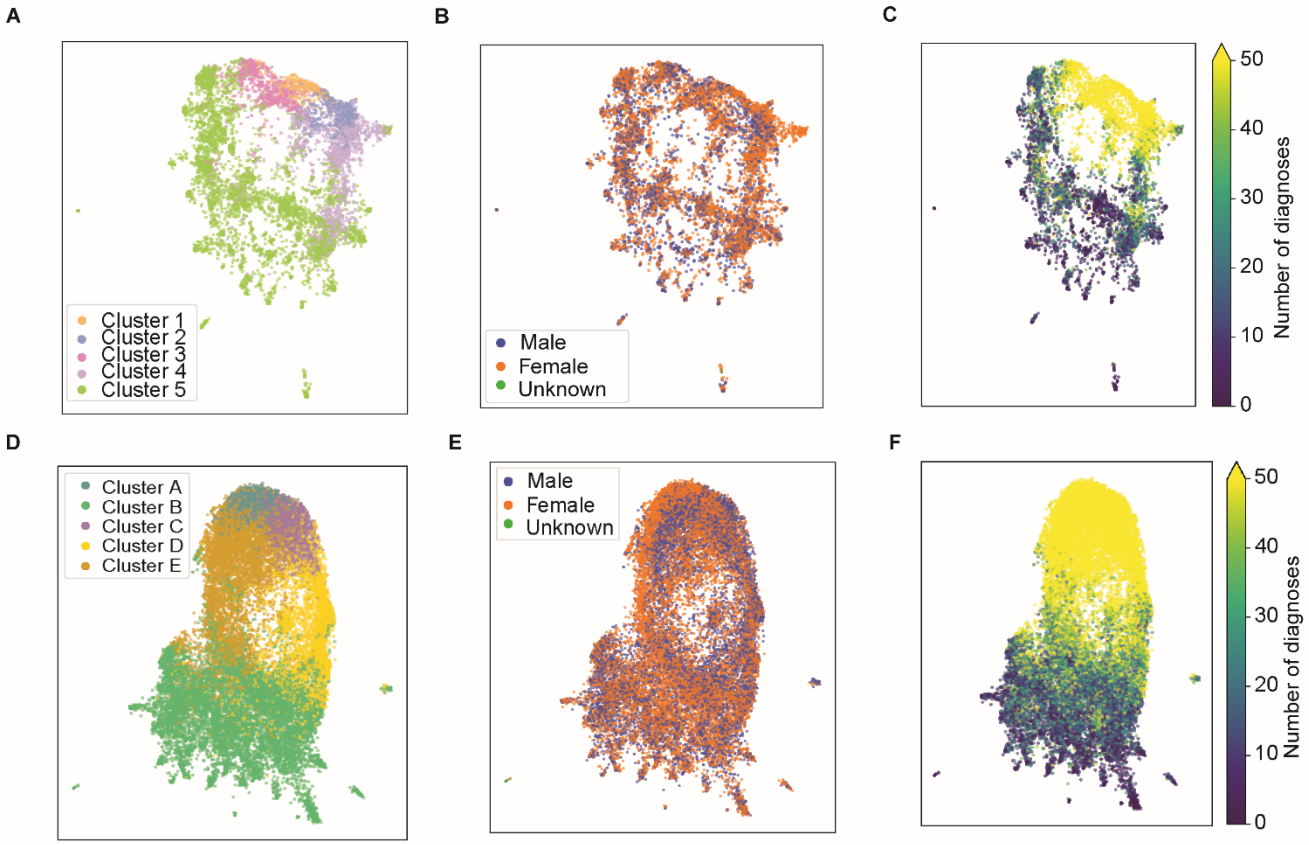

Figure 2. UMAP visualizations using PCA components as features. UMAP plots are colored by cluster (**A** and **D**), sex (**B** and **E**), and number of comorbidities (**C** and **F**). Each dot represents a patient. **A–C** correspond to the UCSF cohort, while **D–G** represent the validation cohort in the UC-Wide dataset. UC, University of California

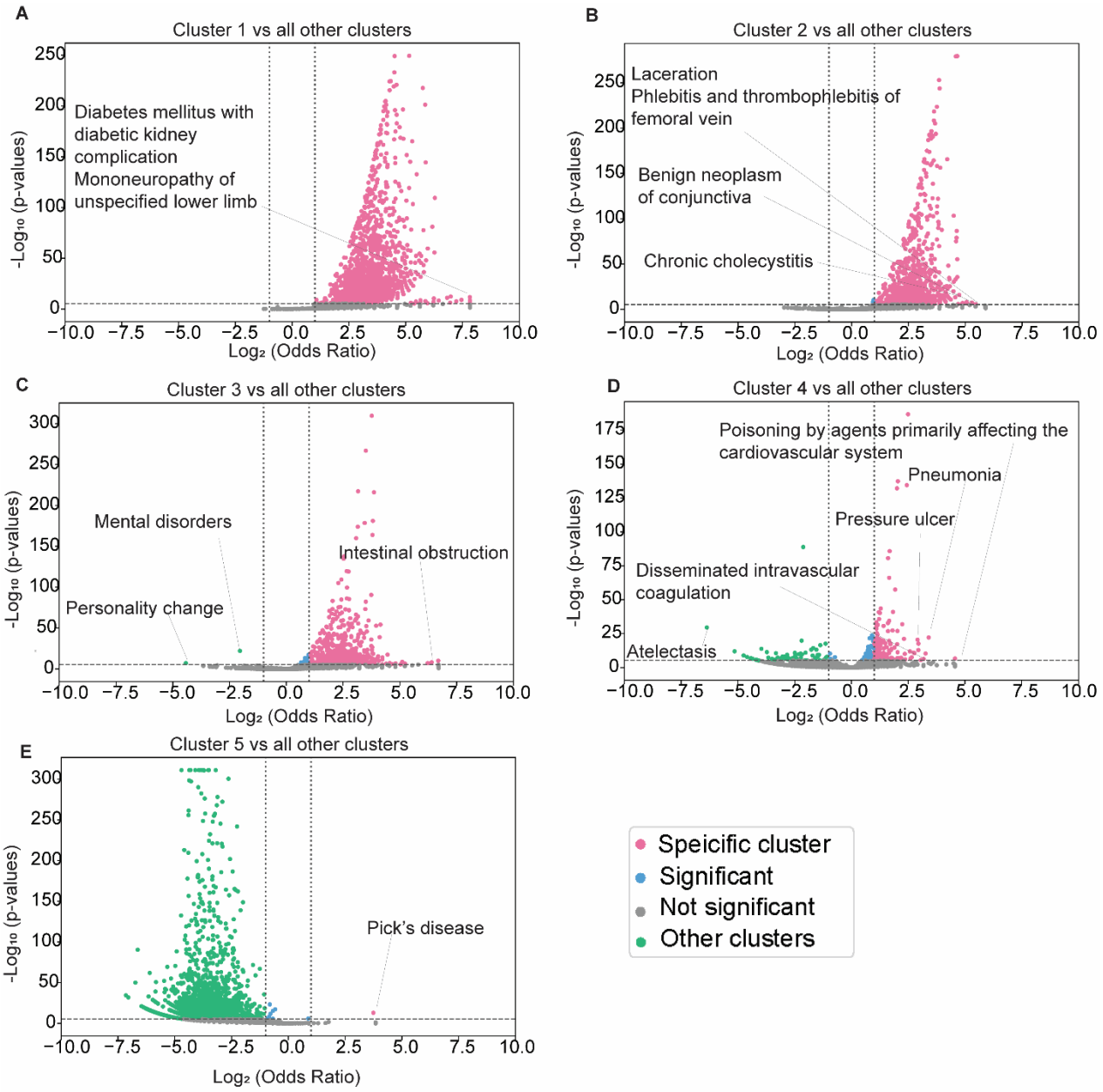

Figure 3. Volcano plot identifies overall trends of each cluster

The volcano plot displays enriched ICD10-CM codes identified using the two-sided Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). Enrichment was determined based on a Bonferroni-corrected p-value < 0.05. Odds ratios were calculated by comparing each specific cluster to all other clusters. **A:** Cluster 1, **B:** Cluster 2, **C:** Cluster 3, **D:** Cluster 4, **E:** Cluster 5. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification.

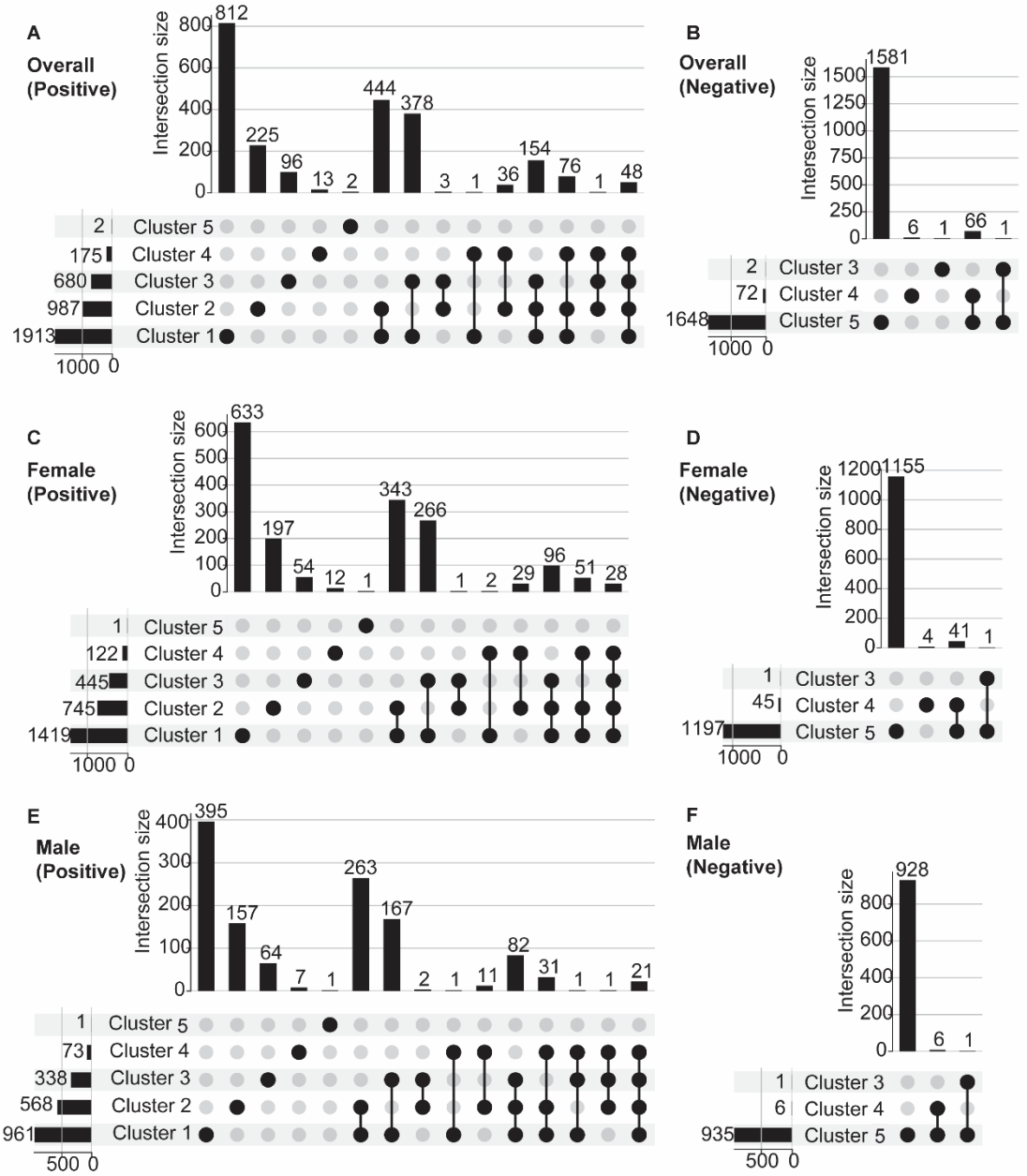

Figure 4. UpSet plot identifies significant cluster-specific comorbidities.

UpSet plot illustrating significant cluster-specific comorbidities. Panels represent the overall population (**A:** Positive associations, **B:** Negative associations), females (**C:** Positive associations, **D:** Negative associations), and males (**E:** Positive associations, **F:** Negative associations).

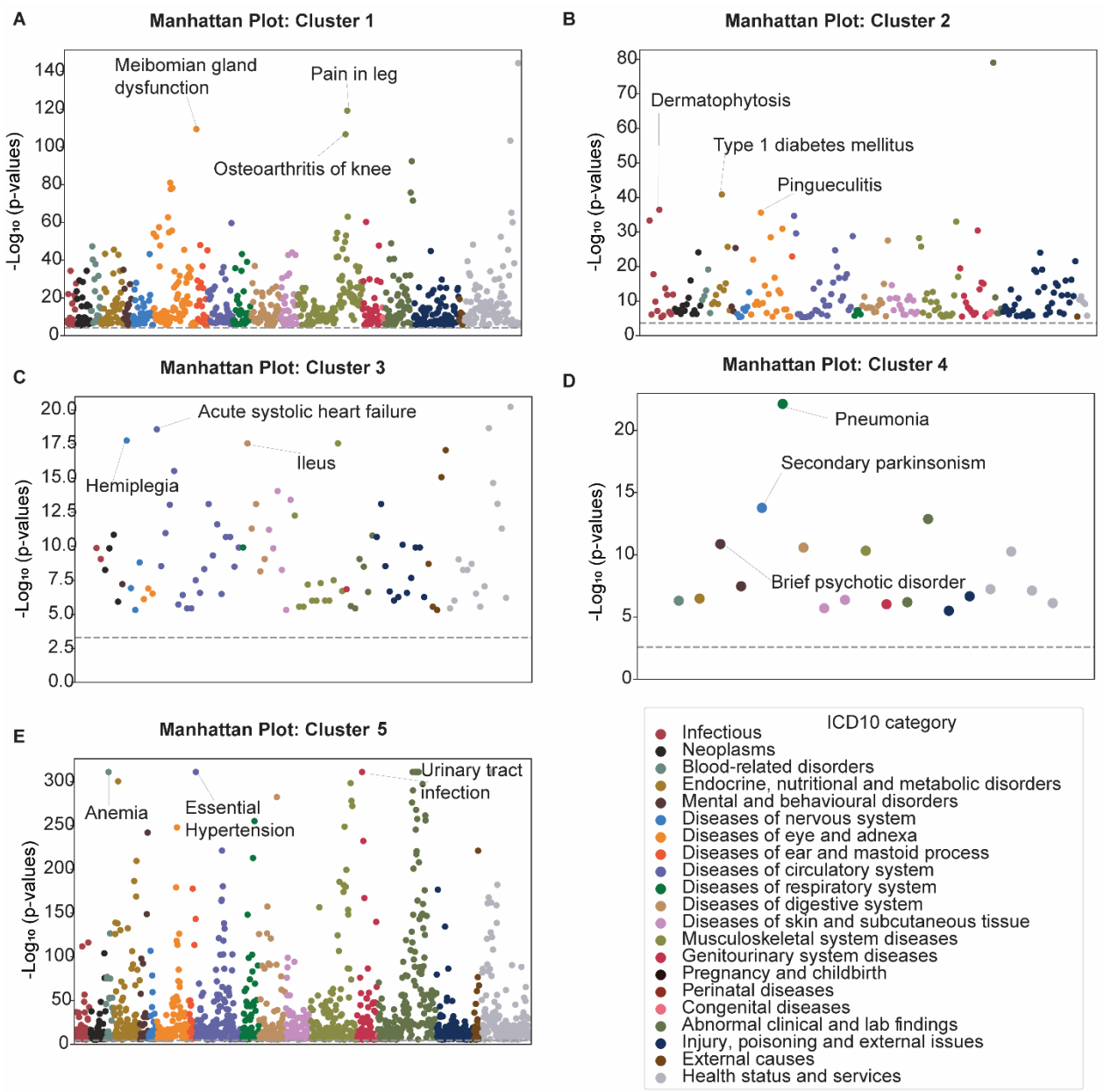

Figure 5. Manhattan plot for cluster-specific significant comorbidities

Manhattan Plot displaying enriched ICD10-CM codes identified using the two-sided Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). Enrichment was determined based on a Bonferroni-corrected p-value < 0.05. **A:** Cluster 1, **B:** Cluster 2, **C:** Cluster 3, **D:** Cluster 4, **E:** Cluster 5. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification.

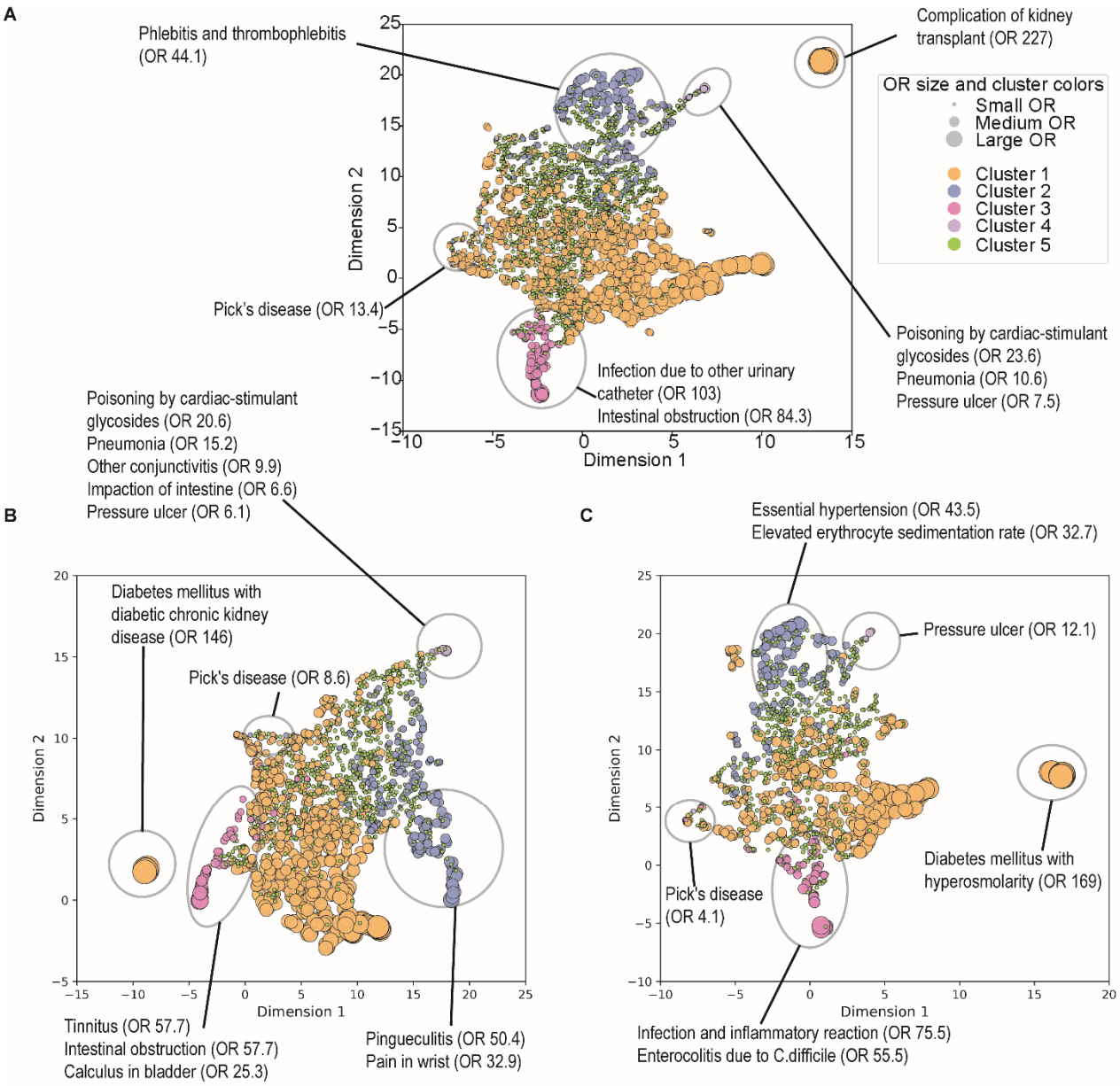

Figure 6. UMAP visualizations using odds ratios as features for each cluster. This figure presents a UMAP visualization of ICD-10-CM codes based on their standardized ORs across different clusters. Each dot represents a specific ICD-10-CM code, with the dot size corresponding to the OR of the significantly associated cluster and the color indicating the corresponding cluster. The input features for the UMAP projection were derived from a matrix of standardized ORs of cluster-specific comorbidities. These standardized ORs were calculated by comparing the prevalence of each diagnosis within a specific cluster against all other clusters using Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). A Bonferroni-corrected p-value threshold of 0.05 was applied to identify significantly enriched diagnoses. **A:** Overall population, **B:** Female, **C:** Male. ICD-10-CM, International Classification of Diseases, 10th Revision, Clinical Modification; UMAP, Uniform Manifold Approximation and Projection.

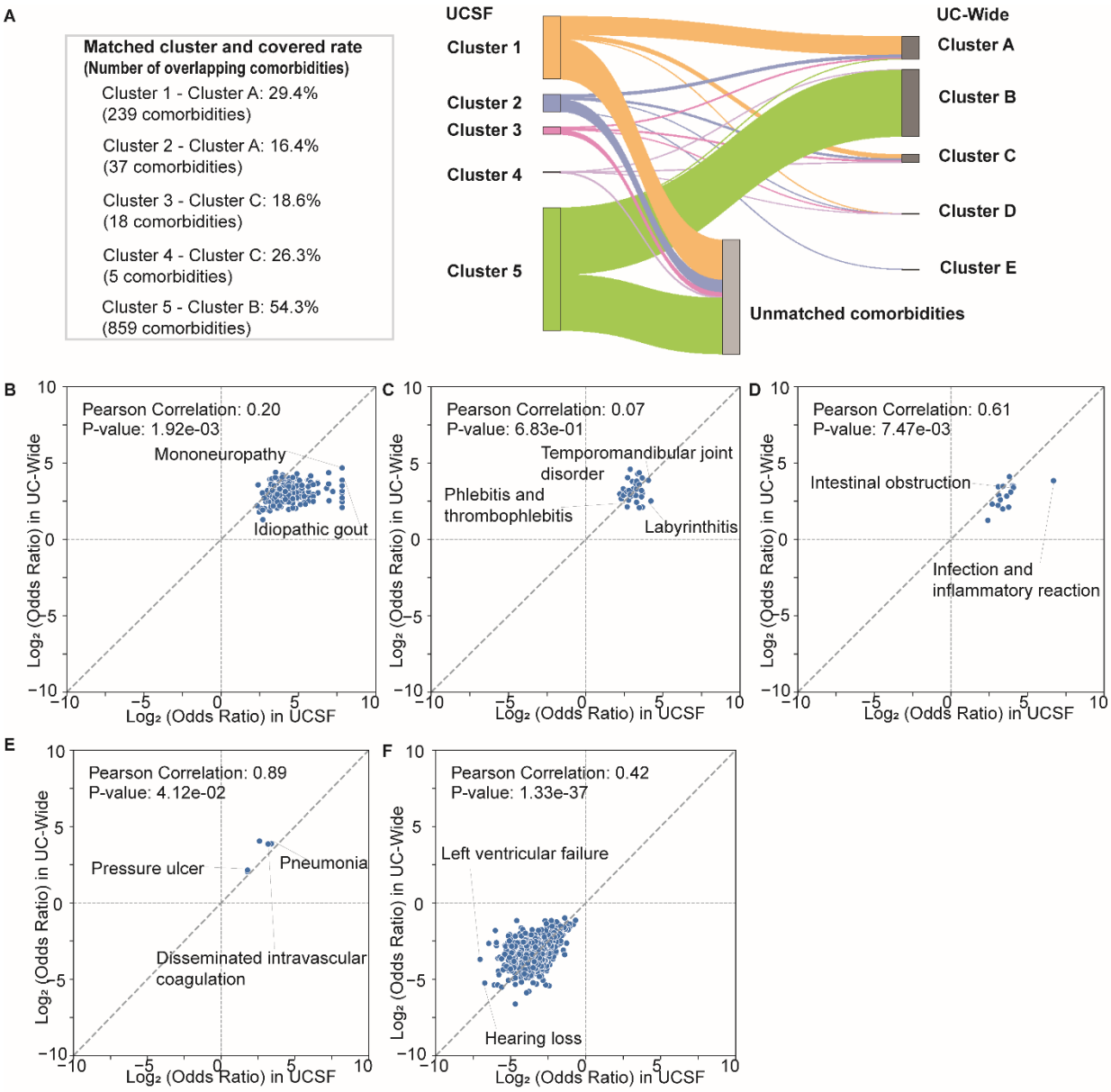

Figure 7. Validation analysis of cluster-specific significant comorbidities.

A Sankey plot illustrating the overlap of significant comorbidities across UCSF and UC-Wide populations within each cluster. The coverage rate represents the proportion of UCSF cluster-specific comorbidities that were statistically significant in UCSF and remained significant in UC-Wide clusters, identifying the most closely matched UC cluster. Number of cluster-specific comorbidities: Cluster 1 (812), Cluster 2 (225), Cluster 3 (97), Cluster 4 (19), Cluster 5 (1,583), Cluster A (1,976), Cluster B (2,731), Cluster C (819), Cluster D (40), Cluster E (19)

B–F Log-log plot showing significant overlapping comorbidities and the concordance of odds ratios for overlapping conditions in the overall population. P-values were calculated from Pearson correlation analysis. UC, University of California.

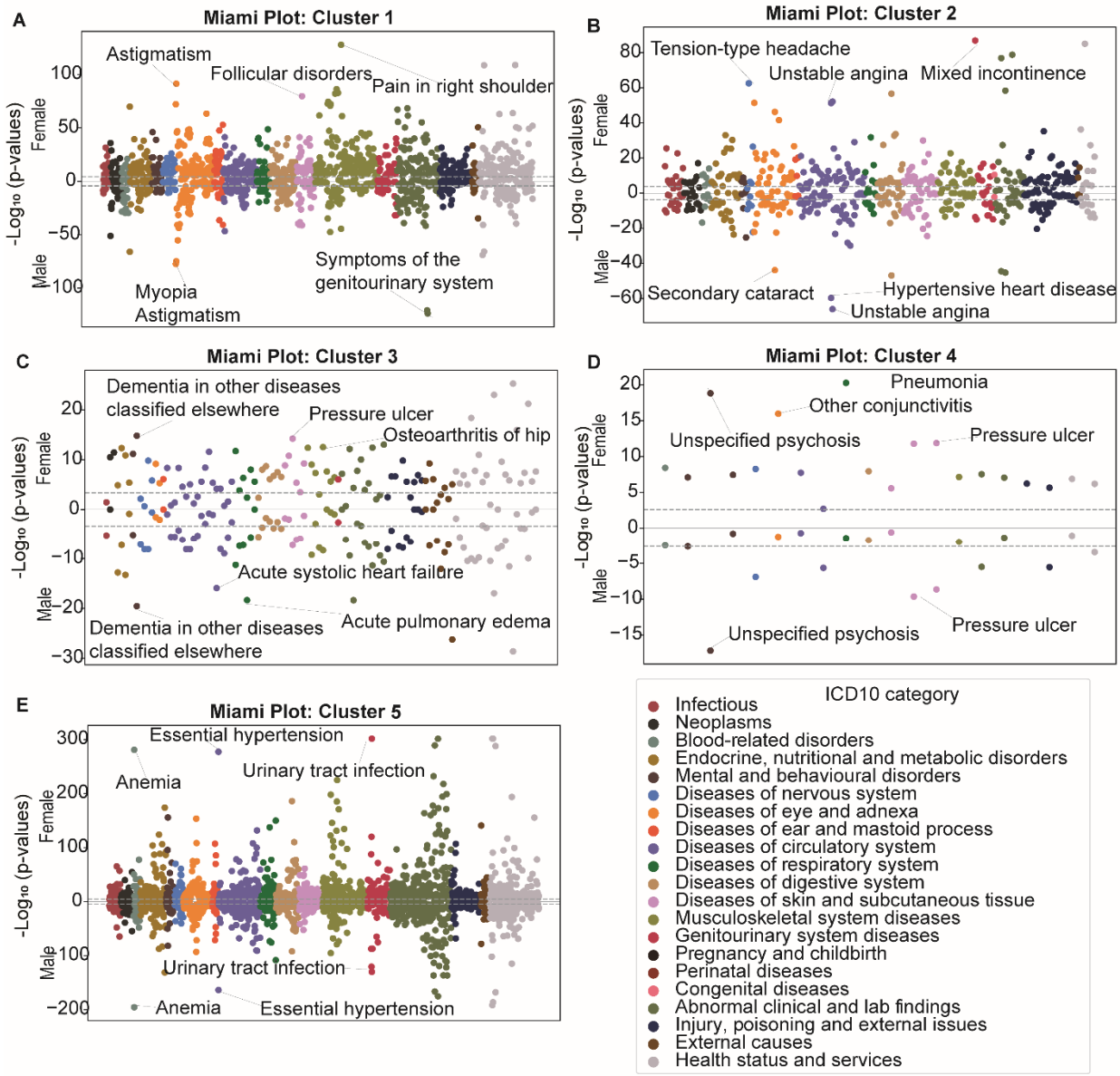

Figure 8. Miami plot for cluster-specific significant comorbidities. Miami Plot displaying enriched ICD10-CM codes identified using the two-sided Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). Enrichment was determined based on a Bonferroni-corrected p-value < 0.05. ICD10-CM, the International Classification of Diseases, 10th Revision, Clinical Modification. **A:** Cluster 1, **B:** Cluster 2, **C:** Cluster 3, **D:** Cluster 4, **E:** Cluster 5.

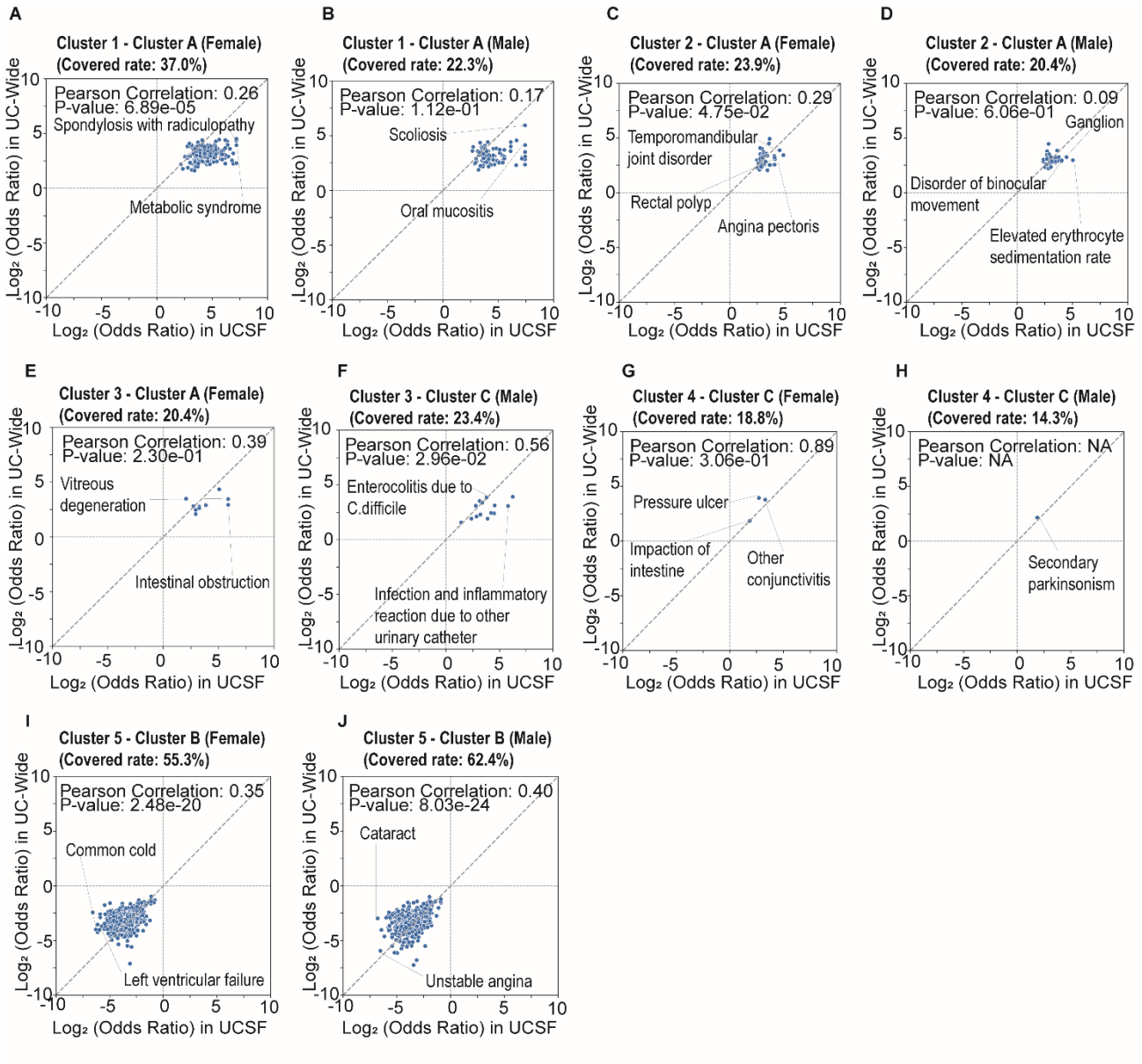

Figure 9. Validation analysis of cluster-specific significant comorbidities across sex. Log-log plot illustrating the significant overlapping comorbidities and the concordance of odds ratios for these conditions in the sex-stratified population. The covered rate represents the proportion of UCSF cluster-sex-specific comorbidities that remain significant in both the UCSF and UC-Wide datasets. **A, C, E, G, I**: Cluster 1–Cluster 5 (Female) vs. Matched UC-Wide Cluster (Female), **B, D, F, H, J:** Cluster 1–Cluster 5 (Male) vs. Matched UC-Wide Cluster (Male). P-values were calculated from Pearson correlation analysis. UC, University of California.