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], Alice S. Tang[1,10]

[1]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA

[2]Health Data Science Graduate Program, University of California, San Francisco, San Francisco, CA

[3]Biological and Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, CA

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

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

[6]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA

[7]Global Brain Health Institute, University of California, San Francisco, San Francisco, CA

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

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

[10]School of Medicine, University of California, San Francisco, San Francisco, CA

Authorship note: Marina Sirota and Alice S. Tang share co-corresponding authors.

Corresponding author: Marina Sirota, Bakar Computational Health Sciences Institute, University of California, San Francisco, 490 Illinois St, 2nd Floor, PO Box 2933, San Francisco, CA 94143 (marina.sirota@ucsf.edu)

Alice S. Tang, Bakar Computational Health Sciences Institute, University of California, San Francisco, 490 Illinois St, 2nd Floor, PO Box 2933, San Francisco, CA 94143 (alice.tang@ucsf.edu)

Conflict-of-interest statement:

Yukari Katsuhara was affiliated with Takeda Pharmaceutical Company Limited during the manuscript preparation but did not perform any work for the company and did not receive any compensation during this period. The authors declare that this does not influence the content of this manuscript.

ORCID information:

Yukari Katsuhara (0000-0003-2335-5413), Umair Khan (0000-0002-6361-4996), Zachary A. Miller (0000-0002-5991-3053), Tomiko T. Oskotsky (0000-0001-7393-5120), Marina Sirota (0000-0002-7246-6083), Alice S. Tang (0000-0003-4745-0714)

Abstract

Background: Alzheimer's disease is a progressive neurodegenerative disorder with no curative treatment. Identifying distinct subphenotypes and understanding potential personalized modifications remain critical unmet needs.

Methods: We applied unsupervised learning techniques to electronic medical records from UCSF to identify distinct Alzheimer's disease subphenotypes based on comorbidity profiles. We conducted enrichment analyses to determine cluster-specific comorbidities. Based on the observed sex-based differences, we subsequently conducted sex-stratified analyses to assess differences in disease manifestations between males and females. Findings were validated using an independent UC-Wide dataset.

Results: 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. Sex-stratified analyses revealed significant differences in comorbidity distributions across clusters. Notably, in Cluster 2, circulatory diseases were more prevalent among males, whereas in Cluster 3, bladder stones were more common among females. Key results were consistent across the UCSF and UC-Wide datasets.

Conclusions: Our study identifies clinically meaningful Alzheimer's disease subphenotypes and highlights sex-specific variations, suggesting potential underlying biological factors such as Apolipoprotein E and gut microbiome alterations contributing to Alzheimer's disease heterogeneity. These findings underscore the need for further research into the biological mechanisms driving these differences and may inform the development of individualized therapeutic regimens.

Funding: This study was supported by grants from the National Institute on Aging (R01AG060393 and R01AG057683).

Introduction

Alzheimer's disease (AD) is the most common cause of dementia, characterized by progressive cognitive decline, neurodegeneration, and increasing burdens on patients, caregivers, and healthcare systems. With rising prevalence and no curative treatments, AD represents a major public health crisis worldwide. Despite decades of research, its biological mechanism remains incompletely understood, making early diagnosis and effective management challenging. 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). Clinically, AD has traditionally been conceptualized as an amnestic disorder predominantly affecting memory, in those 65 and older, but growing evidence suggests that it can also manifest at an earlier age and with non-amnestic behavioral, executive, language, visuospatial, and asymmetric motor coordination phenotypes.

The greatest risk factor for developing AD is age followed by Apolipoprotein E (APOE) genetic status. 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 (3). Increasingly, sex has also emerged as a critical factor in understanding AD risk, progression, and clinical manifestations. Women account for approximately two-thirds of AD cases in the United States, and their lifetime risk of developing AD is significantly higher than that of men (4). However, once diagnosed, men with AD tend to have a shorter survival time, while women experience greater memory impairment (5-7). Recent studies indicate that sex may modify multiple aspects of AD, including susceptibility of AD, progression, and molecular pathology including APOE (8-13). Despite these insights, the impact of sex on AD presentations is not fully understood, highlighting the need for further research into sex-specific disease subphenotypes.

Recent advances in electronic medical records (EMRs) and machine learning have enabled large-scale, data-driven approaches to studying AD heterogeneity. 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 (14). Data-driven phenotyping approaches have revealed distinct AD subtypes influenced by factors such as sex and race (15-18), but comprehensive and unbiased analyses for sex-specific patterns among subphenotypes of AD remain largely unexplored. Deeper characterizing AD-related comorbidities and their subphenotypes may uncover novel risks for the disease and might facilitate the development of individualized therapeutic regimens that could serve as adjuvants to the current anti-amyloid- and anti-tau-based treatments, maximizing therapeutic benefits for patients through a truly personalized therapeutic approach.

In this study, we apply unsupervised clustering to a large EMR dataset of mainly LOAD presentations to identify if subphenotypes within this group exist. By leveraging real-world clinical data, we aim to uncover novel insights into LOAD heterogeneity, which may inform precision medicine approaches and improve comprehensive, sex-specific care for individuals with AD.

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. 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.

Low-dimensional embeddings derived from principal components of diagnostic data reveal five sub-phenotypes 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 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 sub-phenotypes (Supplemental Table 1).

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. Other demographic characteristics of each cluster are presented in Table 1.

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).

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, and benign neoplasm of the conjunctiva. Interestingly, Cluster 3 (682 significant comorbidities) showed negative associations with 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 2).

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).

Additionally, cluster-specific comorbidities with high odds ratios were identified (Figure 6A). Cluster 1 showed the strongest association with complications of kidney transplant (OR = 227, Bonferroni-corrected p-value < 0.05). In Cluster 2, phlebitis and thrombophlebitis 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).

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 3). 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).

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 4). 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).

Discussion

In this study, we analyzed multi-site EMR by applying dimensionality reduction, followed by clustering analysis, and in the process we identified five sub-phenotypes of AD based on comorbidities. Enrichment analysis comparing a specific cluster with the other clusters revealed clinical heterogeneity and highlighted sex-specific differences. Furthermore, our study demonstrated the robustness and generalizability of the identified heterogeneity by validating findings in the independent UC-Wide cohort. In the following sections, we present our findings and discuss their implications.

Cluster 1: high comorbidity burden with non-AD-specific conditions

We performed comorbidity enrichment analysis for overall populations. Cluster 1 was characterized by a high comorbidity burden, suggesting the presence of numerous comorbidities that are not clearly related to AD, such as complications of kidney transplant. The high comorbidity burden trend in this cluster remains consistent between males and females.

Cluster 2: older age and cardiovascular conditions

Cluster 2 has a relatively high proportion of Black or African American individuals. Given that AD prevalence is about twice as high in this population compared to non-Hispanic Whites, this suggests the role of APOE-independent risk factors, such as social determinants of health (4, 19, 20).

Additionally, patients in Cluster 2 were generally older and exhibited enriched comorbidities such as phlebitis and thrombophlebitis. This condition is known as age-related and is associated with functional decline, such as reduced mobility (21). Furthermore, 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 (22, 23). 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 (24, 25). These findings highlight fibrin as a potential therapeutic target for immunotherapy designed to reduce fibrin-induced neurotoxicity (26). 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 (27). This suggests a possibility that increased fibrinogen aggregation and Aβ in AD patients may contribute to vascular abnormalities such as phlebitis.

Cluster 3: younger age and gastrointestinal dysfunction

Cluster 3 includes relatively younger individuals and, interestingly, exhibited significant negative associations with mental disorders. Given that mental disorders are commonly reported comorbidities in AD patients (28), 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 (29). 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 (3, 30-32). Our findings indicate that the broad disease classification such as "Other specified mental disorders" constrains our ability to precisely determine which mental disorders were less prevalent in this cluster. Moreover, as the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) code for EOAD was not used as criteria for cohort selection in this study, we cannot ascertain whether this relatively younger population exhibits the same degree of heterogeneity as EOAD.

Additionally, we identified infection and inflammatory reactions due to a urinary catheter, as well as intestinal obstruction, as positively associated comorbidities in Cluster 3. Some evidence shows the association between AD and infection or inflammatory reactions. For example, recent evidence suggests that infections and inflammatory diseases are interconnected, and the association between AD and inflammatory diseases exhibits sex-specific differences (33, 34). Furthermore, AD is a well-established contributor to urinary bladder and urethral dysfunction, often necessitating catheterization, which in turn increases the risk of urinary tract infections, cross-infections, and other related complications (35). 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 (36-38). 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. The central nervous system has been demonstrated to be bidirectionally connected to the gastrointestinal tract, the enteric nervous system, and the mycobiome via the sympathetic and parasympathetic nervous systems, forming the brain–gut–microbiota axis (39, 40). Research has shown that AD patients exhibit reduced microbial diversity compared to cognitively healthy individuals (41). 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: advanced age and frailty-associated conditions

Cluster 4 also has a relatively high proportion of Black or African American individuals and, along with Cluster 2, is suggested to represent a population with APOE-independent risk factors (4, 19, 20). Additionally, 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 (42).

Cluster 5: low comorbidity burden with potential overlap with frontotemporal dementia

Cluster 5 has a significantly higher proportion of males. Given the rarity of male-dominant groups in AD, this finding highlights the need for further investigation into the underlying factors that contributed to the composition of this cluster.

Cluster 5 showed the fewest positively associated comorbidities, with one notable exception, Pick's disease. Pick's disease can refer to both the clinical syndrome of behavioral variant frontotemporal dementia (bvFTD) and an underlying pathology. In the context of the medical record, it is most likely referring to the clinical syndrome, as the pathology diagnosis requires an autopsy. Of note, while the underlying causes of Pick's disease are most non-AD frontotemporal lobar degeneration (FTLD) pathologies, up to 10% of the time bvFTD is due to underlying AD (43, 44). Additionally, the clinical syndrome of bvFTD is typically younger than that of amnestic AD, and this is the case regardless of the

FTD underlying pathology (45). Furthermore, bvFTD has been shown to have a higher percentage of male presentation (46, 47). Considering the demographics of Cluster 5—where the lower quartile of age is younger and the proportion of males is higher— there is a possibility that this cluster includes patients who reflect bvFTD due to underlying AD (sometimes referred to as bvAD or frontal variant AD) or were individuals where non-AD FTLD conditions were on the differential diagnosis.

Sex specific analysis:

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 (48) and cardiovascular disease (CVD) (49), are well-established risk factors for AD. Previous studies have demonstrated strong associations between these conditions and cerebrovascular pathology, as well as cognitive decline (28, 50).

Our findings suggest that Cluster 2 represents a specific subgroup of AD males with pronounced vascular and metabolic dysfunction, which contrasts with previous reports indicating that these conditions are more common in AD females generally. For instance, hypertension has been reported to be more common in AD women (11, 51). Additionally, a meta-analysis found that women with type 2 diabetes mellitus have a 19% higher risk of developing vascular dementia compared to men, although no significant sex differences were observed in AD risk (52). Despite this, the sex-specific association between diabetes, CVD, and AD remains understudied, making it difficult to draw definitive conclusions.

A potential explanation for these sex differences lies in differential susceptibility to APOE-related mechanisms. While the interplay between APOE, AD, sex, and comorbidities remains incompletely understood, genetic studies have demonstrated that variations in the APOE genotype are associated with an increased risk of hypertension, CVD and diabetes (53-55). Furthermore, existing evidence suggests that the impact of APOE varies by sex, even when accounting for dosage (heterozygous or homozygous genotypes), with AD risk considerably higher for women than men (13, 56). Although our study lacks direct APOE genotype data, the observed patterns indicate that variations in APOE genotype distribution may contribute to the sex differences seen in Cluster 2. Notably, recent research utilizing machine-learning techniques has proposed that the progression pathways of AD differ by sex and that multiple pathways exist within each sex. Among men, a subgroup has been reported to exhibit a distinct, rapid-progression pathway linked to such as CVD and diabetes (18). While APOE genotype data was not included in that research, the reported findings align with the possibility that diseases influenced by APOE genotype, such as CVD and diabetes, contribute to disease heterogeneity in males. This supports the hypothesis that male AD patients in Cluster 2 may represent a distinct subset characterized by more pronounced vascular and metabolic dysfunction, potentially influenced by the effects of APOE genotype.

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, 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 (57-59). 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 (60). 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 (41, 61). While the relationship between microbiome alteration, AD progression, and oxalate metabolism remains unclear, understanding these mechanisms may provide new therapeutic opportunities targeting the gut microbiome to improve gastrointestinal and neurological health in AD patients (62, 63).

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 (64), 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.

Validation analysis:

In our validation analysis, many of these key associations from overall analysis and sex-stratified analysis were replicated in the UC-Wide dataset, supporting the robustness and generalizability of our findings. However, some Pearson correlation coefficients were not particularly high, suggesting the presence of unexplored heterogeneity within both datasets. This discrepancy may partly result from demographic differences between the UCSF and UC-Wide cohorts, which could have influenced the distribution of AD severity and comorbidity patterns.

Limitations:

There are several limitations in this study. First, the presence of imbalanced data resulted in some odds ratios being disproportionately inflated. While key findings were consistent across both datasets, suggesting a certain level of reliability, potential biases due to sample size variation cannot be ruled out. Second, the potential for misdiagnosis of AD remains a concern, as our study did not incorporate biomarkers for confirmation, which may have introduced diagnostic inaccuracies. Third, data quality poses another limitation, as EMRs are primarily maintained for clinical rather than research purposes, leading to missing values and potential inconsistencies.

Additionally, age data was not consistently recorded across datasets, with UCSF capping reported ages at 91 and UC-Wide data capping at 90, which may have impacted age-related analyses. Furthermore, the case-control study design precludes establishing causal relationships or determining the temporal sequence of disease progression. Regarding generalizability, our analysis was limited to medical institutions in California, necessitating caution when extrapolating these findings to other regions or populations.

Finally, while our study primarily focused on LOAD, further research that focuses on EOAD is needed to fully elucidate the heterogeneity of AD. Despite these limitations, our research provides a comprehensive analysis of AD comorbidities, identifying distinct sub-phenotypes that highlight the heterogeneity in disease manifestation.

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. Our findings suggest the presence of sub-phenotypes related to APOE genotypes and microbiome alterations related to AD, which may contribute to the observed differences in comorbidity patterns across sexes. Future studies incorporating genomic and microbiome data 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.

Methods

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 using K-means clustering after dimensionality reduction.

Comprehensive comorbidity enrichment analyses, including sex-stratified analysis, were conducted to characterize these clusters. Validation was performed using independent data from UCDDP to ensure the robustness of the findings.

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.

Patient cohort identification

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 (30). Additionally, only patients with at least one documented comorbidity in addition to AD were included in the analysis. 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.

Dimensionality reduction and K-means clustering

To represent de-identified AD patients, one-hot encoding was performed for all diagnoses, excluding those explicitly related to Alzheimer's disease. Patients with no recorded comorbidities other than AD were excluded from the clustering analysis. PCA was used to reduce the dimensionality of the dataset while preserving diagnostic information, with the number of components selected to cover at least 80% of the cumulative variance. K-means clustering was then applied to these PCA components, and the optimal number of clusters was determined based on a combination of 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, enrichment analyses were conducted by comparing the frequency of diagnoses using ICD-10-CM codes between each cluster and all other clusters combined. 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 (65). 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.

Statistics

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.

Study approval

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.

Data availability

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.

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.

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). Additionally, ChatGPT-4o (released on May 13, 2024, by OpenAI) was utilized in the manuscript preparation process, specifically for refining language, and enhancing readability.

References

1. Jack CR, Jr., Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* 2018;14(4):535-62.

2. Uddin MS, and Lim LW. Glial cells in Alzheimer's disease: From neuropathological changes to therapeutic implications. *Ageing Research Reviews.* 2022;78:101622.
3. Miller ZA, Ossenkoppele R, Graff-Radford NR, Allen IE, Shwe W, Rosenberg L, et al. Left-handedness, learning disability, autoimmune disease, and seizure history influence age at onset and phenotypical targeting of Alzheimer's disease. *medRxiv.* 2022:2022.12.17.22283307.
4. 2024 Alzheimer's disease facts and figures. *Alzheimers Dement.* 2024;20(5):3708-821.
5. Larson EB, Shadlen MF, Wang L, McCormick WC, Bowen JD, Teri L, et al. Survival after initial diagnosis of Alzheimer disease. *Ann Intern Med.* 2004;140(7):501-9.
6. Xie J, Brayne C, and Matthews FE. Survival times in people with dementia: analysis from population based cohort study with 14 year follow-up. *BMJ.* 2008;336(7638):258-62.
7. Santiago JA, and Potashkin JA. Biological and Clinical Implications of Sex-Specific Differences in Alzheimer's Disease. *Handb Exp Pharmacol.* 2023;282:181-97.
8. Irvine K, Laws KR, Gale TM, and Kondel TK. Greater cognitive deterioration in women than men with Alzheimer's disease: a meta analysis. *J Clin Exp Neuropsychol.* 2012;34(9):989-98.
9. Ferretti MT, Iulita MF, Cavedo E, Chiesa PA, Schumacher Dimech A, Santuccione Chadha A, et al. Sex differences in Alzheimer disease — the gateway to precision medicine. *Nature Reviews Neurology.* 2018;14(8):457-69.
10. Aggarwal NT, and Mielke MM. Sex Differences in Alzheimer's Disease. *Neurol Clin.* 2023;41(2):343-58.
11. Gilsanz P, Mayeda ER, Glymour MM, Quesenberry CP, Mungas DM, DeCarli C, et al. Female sex, early-onset hypertension, and risk of dementia. *Neurology.* 2017;89(18):1886-93.
12. Rogalski EJ, Rademaker A, Harrison TM, Helenowski I, Johnson N, Bigio E, et al. ApoE E4 is a susceptibility factor in amnestic but not aphasic dementias. *Alzheimer Dis Assoc Disord.* 2011;25(2):159-63.
13. Arnold M, Nho K, Kueider-Paisley A, Massaro T, Huynh K, Brauner B, et al. Sex and APOE ε4 genotype modify the Alzheimer's disease serum metabolome. *Nat Commun.* 2020;11(1):1148.
14. Geifman N, Kennedy RE, Schneider LS, Buchan I, and Brinton RD. Data-driven identification of endophenotypes of Alzheimer's disease progression: implications for clinical trials and therapeutic interventions. *Alzheimer's Research & Therapy.* 2018;10(1):4.
15. Xu J, Wang F, Xu Z, Adekkanattu P, Brandt P, Jiang G, et al. Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn Health Syst.* 2020;4(4):e10246.
16. Tang AS, Oskotsky T, Havaldar S, Mantyh WG, Bicak M, Solsberg CW, et al. Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. *Nature Communications.* 2022;13(1):675.
17. Woldemariam SR, Tang AS, Oskotsky TT, Yaffe K, and Sirota M. Similarities and differences in Alzheimer's dementia comorbidities in racialized populations identified from electronic medical records. *Communications Medicine.* 2023;3(1):50.
18. Meng W, Xu J, Huang Y, Wang C, Song Q, Ma A, et al. Autoencoder to Identify Sex-Specific Sub-phenotypes in Alzheimer's Disease Progression Using Longitudinal Electronic Health Records. *medRxiv.* 2024.
19. Clinic M. Alzheimer's disease research center community outreach and engagement.
20. Association As. Alzheimer's impact movement. race, ethnicity, and Alzheimer's.
21. Simões AMN, Vendramim P, and Pedreira MLG. Risk factors for peripheral intravenous catheter-related phlebitis in adult patients. *Rev Esc Enferm USP.* 2022;56:e20210398.

22. Cortes-Canteli M, Zamolodchikov D, Ahn HJ, Strickland S, and Norris EH. Fibrinogen and altered hemostasis in Alzheimer's disease. *J Alzheimers Dis.* 2012;32(3):599-608.
23. Cortes-Canteli M, Paul J, Norris EH, Bronstein R, Ahn HJ, Zamolodchikov D, et al. Fibrinogen and beta-amyloid association alters thrombosis and fibrinolysis: a possible contributing factor to Alzheimer's disease. *Neuron.* 2010;66(5):695-709.
24. Fibrin induces neurotoxic microglia gene programs in neurodegeneration. *Nature Immunology.* 2023;24(7):1062-3.
25. Merlini M, Rafalski VA, Rios Coronado PE, Gill TM, Ellisman M, Muthukumar G, et al. Fibrinogen Induces Microglia-Mediated Spine Elimination and Cognitive Impairment in an Alzheimer's Disease Model. *Neuron.* 2019;101(6):1099-108.e6.
26. Ryu JK, Rafalski VA, Meyer-Franke A, Adams RA, Poda SB, Rios Coronado PE, et al. Fibrin-targeting immunotherapy protects against neuroinflammation and neurodegeneration. *Nat Immunol.* 2018;19(11):1212-23.
27. Bergamaschini L, Rossi E, Storini C, Pizzimenti S, Distaso M, Perego C, et al. 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.* 2004;24(17):4181-6.
28. Avitan I, Halperin Y, Saha T, Bloch N, Atrahimovich D, Polis B, et al. Towards a Consensus on Alzheimer's Disease Comorbidity? *J Clin Med.* 2021;10(19).
29. Nedelec T, Couvy-Duchesne B, Monnet F, Daly T, Ansart M, Gantzer L, et al. 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.* 2022;4(3):e169-e78.
30. Wingo TS, Lah JJ, Levey AI, and Cutler DJ. Autosomal recessive causes likely in early-onset Alzheimer disease. *Arch Neurol.* 2012;69(1):59-64.
31. Sirkis DW, Bonham LW, Johnson TP, La Joie R, and Yokoyama JS. Dissecting the clinical heterogeneity of early-onset Alzheimer's disease. *Mol Psychiatry.* 2022;27(6):2674-88.
32. Mantyh WG, Cochran JN, Taylor JW, Broce IJ, Geier EG, Bonham LW, et al. Early-onset Alzheimer's disease explained by polygenic risk of late-onset disease? *Alzheimers Dement (Amst).* 2023;15(4):e12482.
33. Ramey GD, Tang A, Phongpreecha T, Yang MM, Woldemariam SR, Oskotsky TT, et al. Exposure to autoimmune disorders is associated with increased Alzheimer's disease risk in a multi-site electronic health record analysis. *Cell Rep Med.* 2025:101980.
34. Sundaresan B, Shirafkan F, Ripperger K, and Rattay K. The Role of Viral Infections in the Onset of Autoimmune Diseases. *Viruses.* 2023;15(3).
35. Liao L. Evaluation and Management of Neurogenic Bladder: What Is New in China? *Int J Mol Sci.* 2015;16(8):18580-600.
36. Gómez-Gómez ME, and Zapico SC. Frailty, Cognitive Decline, Neurodegenerative Diseases and Nutrition Interventions. *Int J Mol Sci.* 2019;20(11).
37. Makhnevich A, Perrin A, Talukder D, Liu Y, Izard S, Chiuzan C, et al. Thick Liquids and Clinical Outcomes in Hospitalized Patients With Alzheimer Disease and Related Dementias and Dysphagia. *JAMA Internal Medicine.* 2024;184(7):778-85.
38. Yurtdaş Depboylu G, Acar Tek N, Akbulut G, Günel Z, and Kamanlı B. Functional Constipation in Elderly and Related Determinant Risk Factors: Malnutrition and Dietary Intake. *J Am Nutr Assoc.* 2023;42(6):541-7.
39. Seo DO, and Holtzman DM. Current understanding of the Alzheimer's disease-associated microbiome and therapeutic strategies. *Exp Mol Med.* 2024;56(1):86-94.
40. Kuźniar J, Kozubek P, Czaja M, and Leszek J. Correlation between Alzheimer's Disease and Gastrointestinal Tract Disorders. *Nutrients.* 2024;16(14).

41. Vogt NM, Kerby RL, Dill-McFarland KA, Harding SJ, Merluzzi AP, Johnson SC, et al. Gut microbiome alterations in Alzheimer's disease. *Sci Rep.* 2017;7(1):13537.
42. Cunha BA. Pneumonia in the elderly. *Clin Microbiol Infect.* 2001;7(11):581-8.
43. Perry DC, Brown JA, Possin KL, Datta S, Trujillo A, Radke A, et al. Clinicopathological correlations in behavioural variant frontotemporal dementia. *Brain.* 2017;140(12):3329-45.
44. Chare L, Hodges JR, Leyton CE, McGinley C, Tan RH, Kril JJ, et al. New criteria for frontotemporal dementia syndromes: clinical and pathological diagnostic implications. *J Neurol Neurosurg Psychiatry.* 2014;85(8):865-70.
45. Knopman DS, and Roberts RO. Estimating the number of persons with frontotemporal lobar degeneration in the US population. *J Mol Neurosci.* 2011;45(3):330-5.
46. Wagner M, Lorenz G, Volk AE, Brunet T, Edbauer D, Berutti R, et al. Clinico-genetic findings in 509 frontotemporal dementia patients. *Molecular Psychiatry.* 2021;26(10):5824-32.
47. Ranasinghe KG, Rankin KP, Pressman PS, Perry DC, Lobach IV, Seeley WW, et al. Distinct Subtypes of Behavioral Variant Frontotemporal Dementia Based on Patterns of Network Degeneration. *JAMA Neurol.* 2016;73(9):1078-88.
48. Wang Y, Yang P, Yan Z, Liu Z, Ma Q, Zhang Z, et al. The Relationship between Erythrocytes and Diabetes Mellitus. *J Diabetes Res.* 2021;2021:6656062.
49. Lowe GD. The relationship between infection, inflammation, and cardiovascular disease: an overview. *Ann Periodontol.* 2001;6(1):1-8.
50. Pankratz VS, Roberts RO, Mielke MM, Knopman DS, Jack CR, Geda YE, et al. Predicting the risk of mild cognitive impairment in the Mayo Clinic Study of Aging. *Neurology.* 2015;84(14):1433-42.
51. Gong J, Harris K, Peters SAE, and Woodward M. 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.* 2021;19(1):110.
52. Chatterjee S, Peters SAE, Woodward M, Mejia Arango S, Batty GD, Beckett N, et al. 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.* 2015;39(2):300-7.
53. Anthopoulos PG, Hamodrakas SJ, and Bagos PG. Apolipoprotein E polymorphisms and type 2 diabetes: a meta-analysis of 30 studies including 5423 cases and 8197 controls. *Mol Genet Metab.* 2010;100(3):283-91.
54. Shi J, Liu Y, Liu Y, Li Y, Qiu S, Bai Y, et al. Association between ApoE polymorphism and hypertension: A meta-analysis of 28 studies including 5898 cases and 7518 controls. *Gene.* 2018;675:197-207.
55. Bennet AM, Di Angelantonio E, Ye Z, Wensley F, Dahlin A, Ahlbom A, et al. Association of apolipoprotein E genotypes with lipid levels and coronary risk. *Jama.* 2007;298(11):1300-11.
56. Riedel BC, Thompson PM, and Brinton RD. Age, APOE and sex: Triad of risk of Alzheimer's disease. *The Journal of Steroid Biochemistry and Molecular Biology.* 2016;160:134-47.
57. Li ML, Song SC, Yang F, Gao C, Zhou B, and Wang Q. Risk assessment and prevention of urolithiasis in urban areas of Baoding, China. *Medicine (Baltimore).* 2024;103(2):e35880.
58. Stamatelou K, and Goldfarb DS. Epidemiology of Kidney Stones. *Healthcare (Basel).* 2023;11(3).
59. Siener R, Herwig H, Rüdy J, Schaefer RM, Lossin P, and Hesse A. Urinary stone composition in Germany: results from 45,783 stone analyses. *World J Urol.* 2022;40(7):1813-20.

60. Miller AW, Penniston KL, Fitzpatrick K, Agudelo J, Tasian G, and Lange D. Mechanisms of the intestinal and urinary microbiome in kidney stone disease. *Nat Rev Urol.* 2022;19(12):695-707.

61. Dodiya HB, Lutz HL, Weigle IQ, Patel P, Michalkiewicz J, Roman-Santiago CJ, et al. Gut microbiota-driven brain Aβ amyloidosis in mice requires microglia. *J Exp Med.* 2022;219(1).

62. Bashir B, Gulati M, Vishwas S, Gupta G, Dhanasekaran M, Paudel KR, et al. Bridging gap in the treatment of Alzheimer's disease via postbiotics: Current practices and future prospects. *Ageing Res Rev.* 2025;105:102689.

63. Jiang X, Zheng Y, Sun H, Dang Y, Yin M, Xiao M, et al. Fecal Microbiota Transplantation Improves Cognitive Function of a Mouse Model of Alzheimer's Disease. *CNS Neurosci Ther.* 2025;31(2):e70259.

64. Mevcha A, and Drake MJ. Etiology and management of urinary retention in women. *Indian J Urol.* 2010;26(2):230-5.

65. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, and Pfister H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph.* 2014;20(12):1983-92.

	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* (%)							
Female	217 (66%)	358 (67%)	514 (63%)	1,083 (71%)	3,143 (61%)	5,315 (64%)	< 0.05
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* (%)							
White or Caucasian	148 (45%)	248 (46%)	439 (54%)	897 (59%)	3,431 (66%)	5,163 (62%)	< 0.05
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* (%)							
Alive	218 (66%)	258 (48%)	550 (67%)	912 (60%)	4,342 (84%)	6,280 (75%)	< 0.05
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* (%)							
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%)	< 0.05
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* (%)							
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%)	< 0.05
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* (%)							
Alive	1,221 (69.2%)	6,265 (51.7%)	717 (34.5%)	1,617 (39.0%)	3,367 (58.1%)	13,187 (50.9%)	< 0.05
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* (%)							
Site 1	320 (18.1%)	2019 (16.7%)	315 (15.1%)	590 (14.2%)	1,112 (19.2%)	4,356 (16.8%)	< 0.05
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.

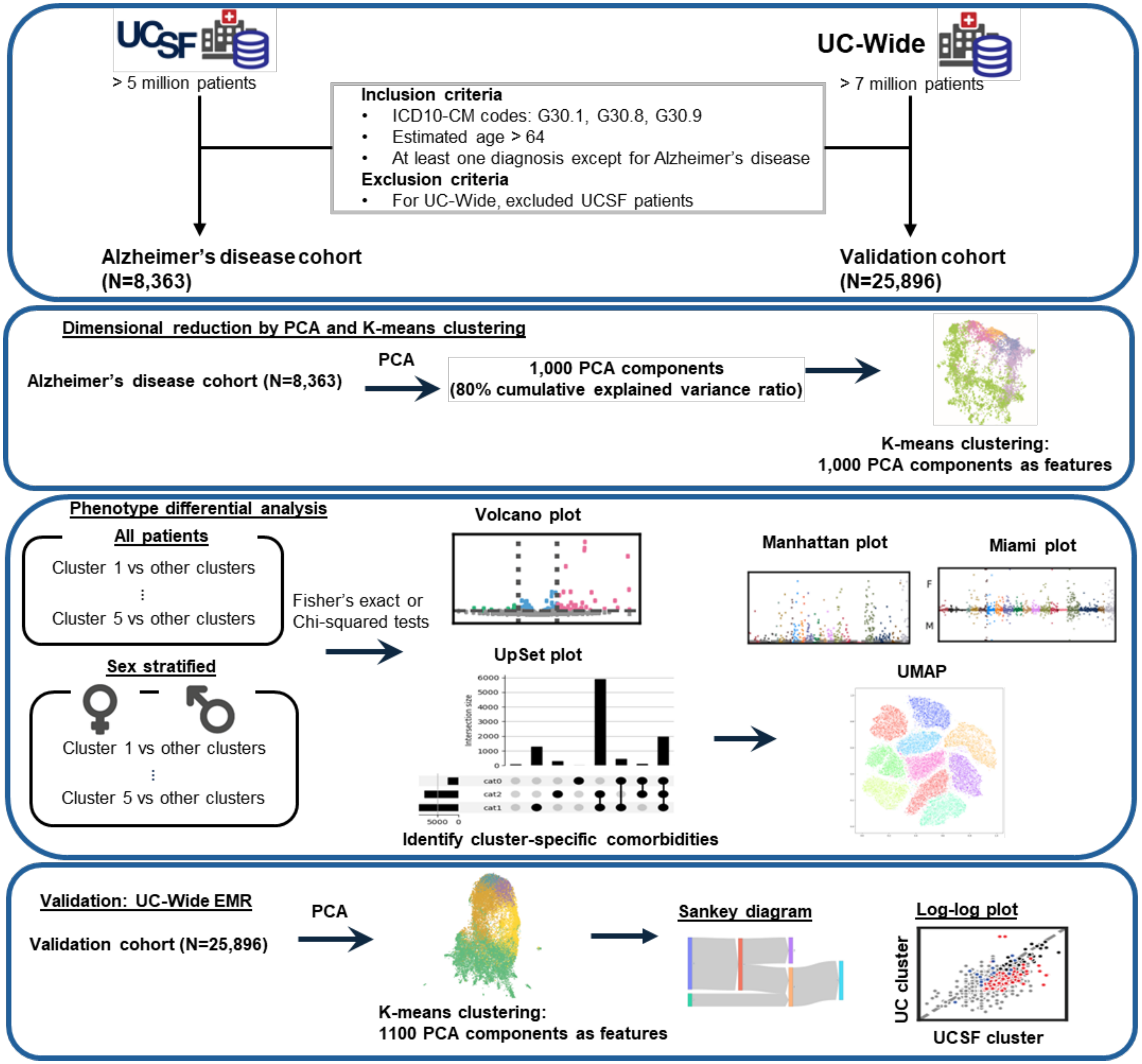
UCSF
> 5 million patients
UC-Wide
> 7 million patients
Inclusion criteria
• ICD10-CM codes: G30.1, G30.8, G30.9
• Estimated age > 64
• At least one diagnosis except for Alzheimer's disease
Exclusion criteria
• For UC-Wide, excluded UCSF patients
Alzheimer's disease cohort (N=8,363)
Validation cohort (N=25,896)
Dimensional reduction by PCA and K-means clustering
Alzheimer's disease cohort (N=8,363)
PCA
1,000 PCA components (80% cumulative explained variance ratio)
K-means clustering: 1,000 PCA components as features
Phenotype differential analysis
All patients
Cluster 1 vs other clusters
⋮
Cluster 5 vs other clusters
Sex stratified
Cluster 1 vs other clusters
⋮
Cluster 5 vs other clusters
Fisher's exact or Chi-squared tests
Volcano plot
UpSet plot
Intersection size
6000
5000
4000
3000
2000
1000
0
cat0
cat2
cat1
5000
0
Identify cluster-specific comorbidities
Manhattan plot
Miami plot
F
M
UMAP
Validation: UC-Wide EMR
Validation cohort (N=25,896)
PCA
K-means clustering: 1100 PCA components as features
Sankey diagram
Log-log plot
UC cluster
UCSF cluster

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

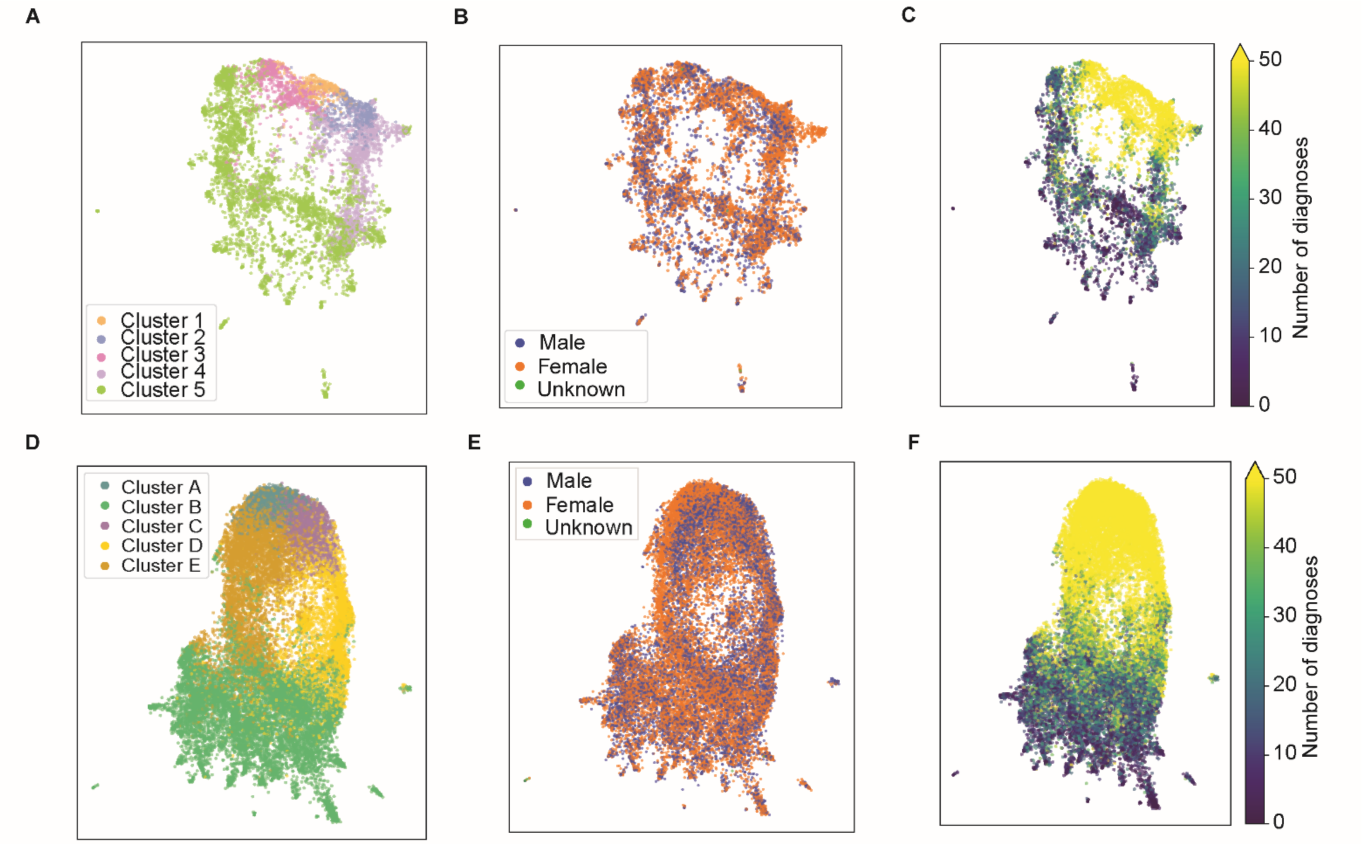

Figure 2. UMAP visualizations using PCA components as features. UMAP plots are colored by cluster (**A** and **D**), sex (**B** and **E**), and number of comorbidities (**C** and **F**). Each dot represents a patient. **A–C** correspond to the UCSF cohort, while **D–G** represent the validation cohort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

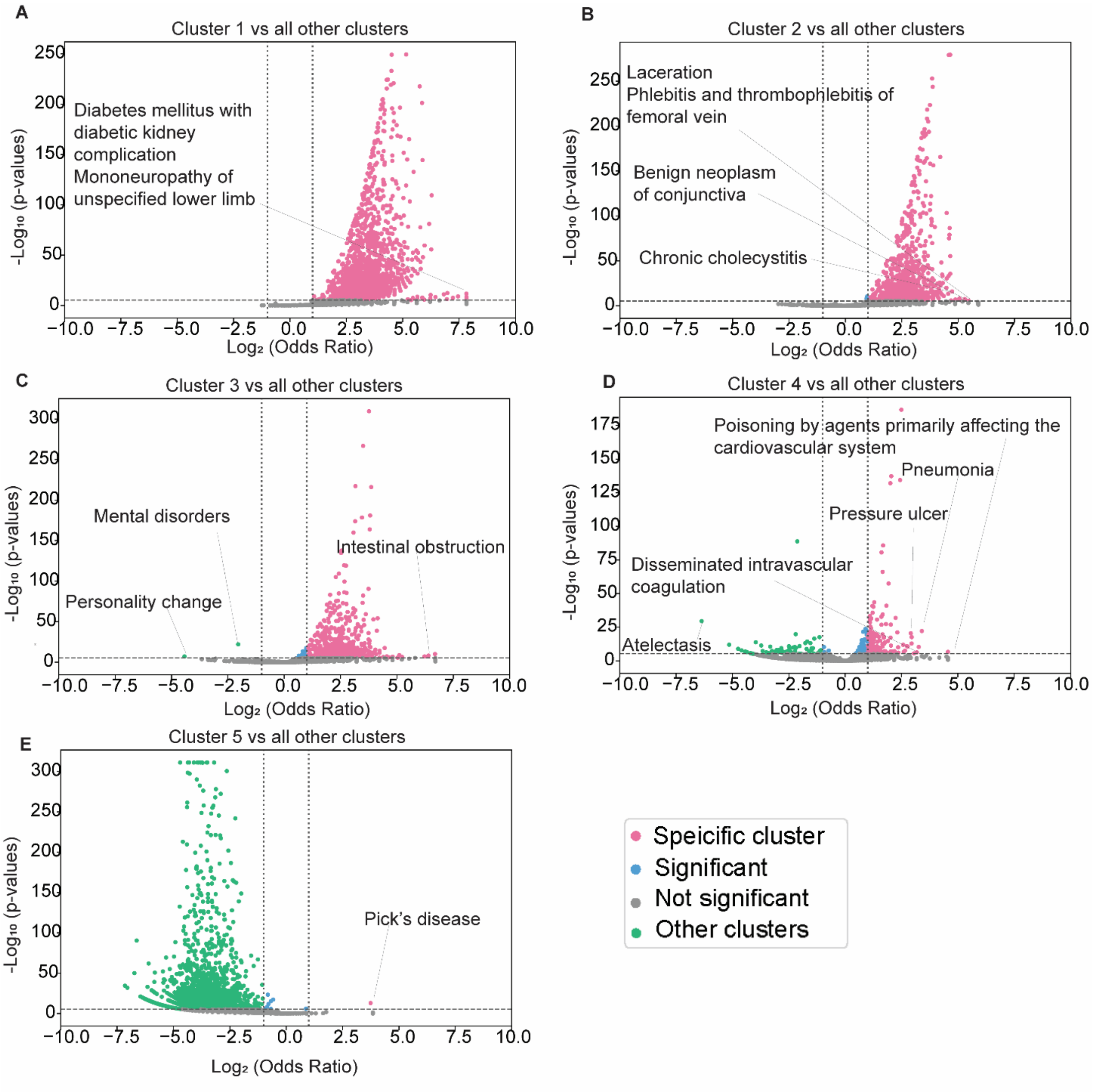

Figure 3. Volcano plot identifies overall trends of each cluster

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.

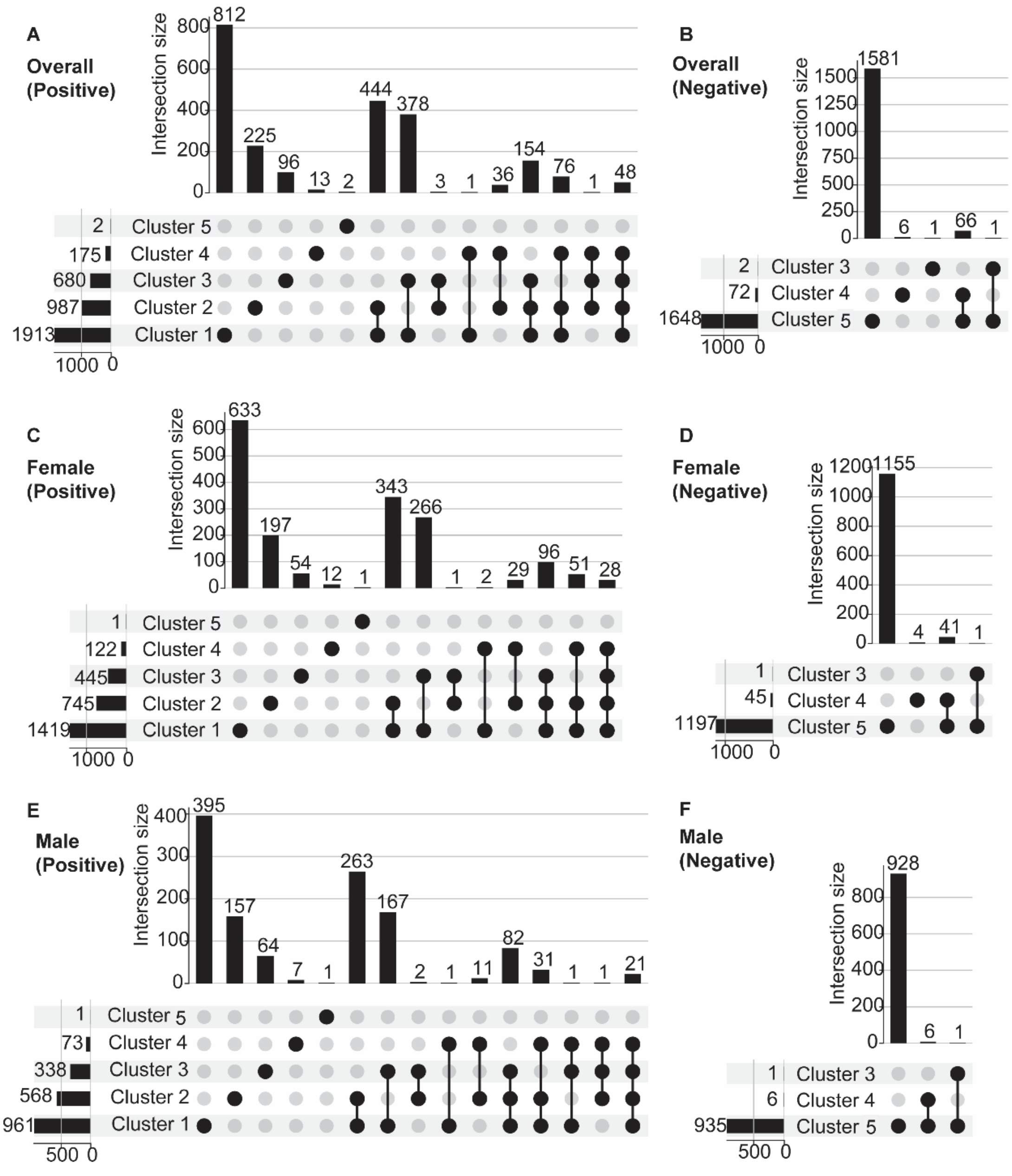

Figure 4. UpSet plot identifies significant cluster-specific comorbidities.
UpSet plot illustrating significant cluster-specific comorbidities. Panels represent the overall population (**A:** Positive associations, **B:** Negative associations), females (**C:** Positive associations, **D:** Negative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

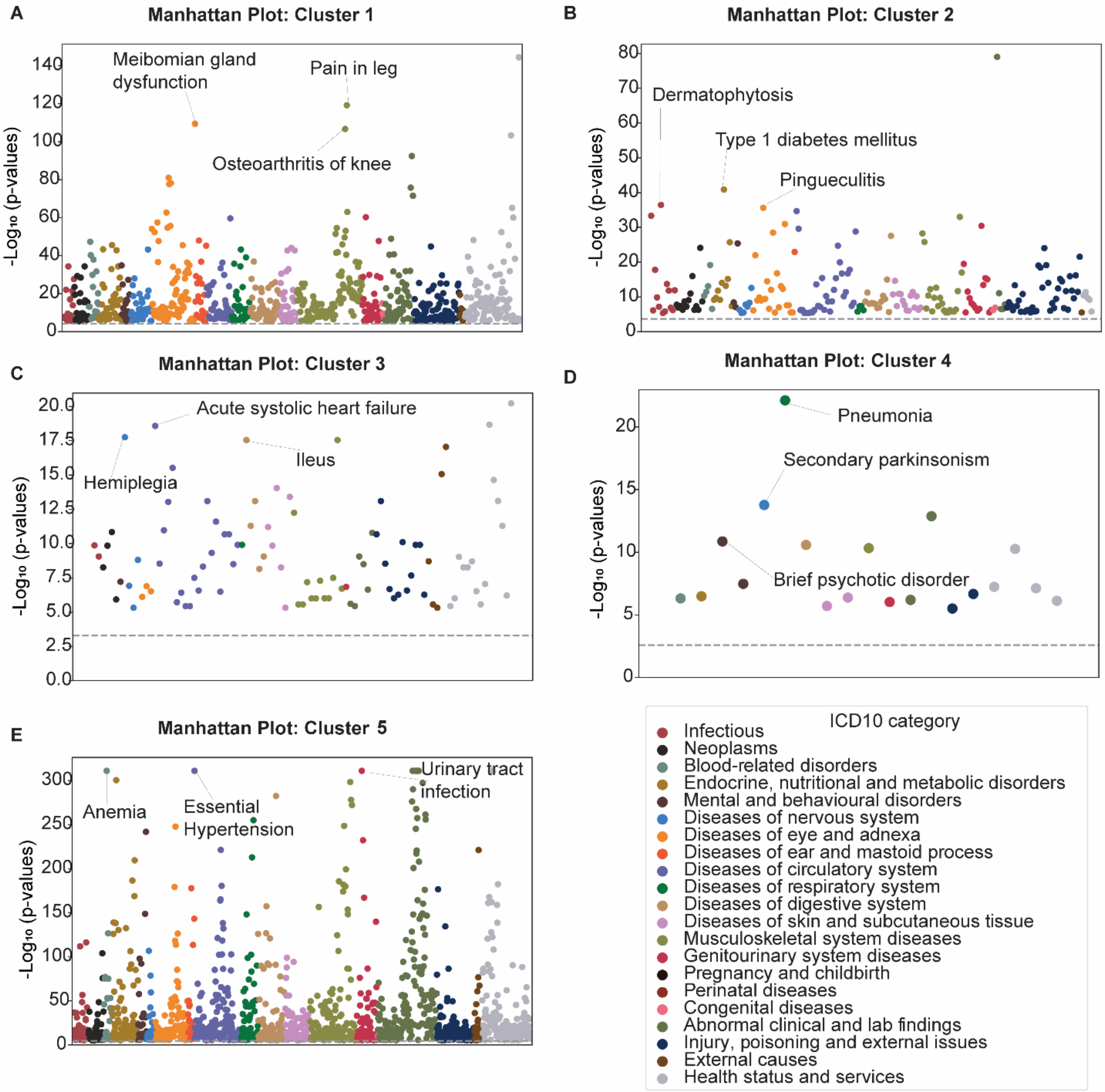

Figure 5. Manhattan plot for cluster-specific significant comorbidities

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.

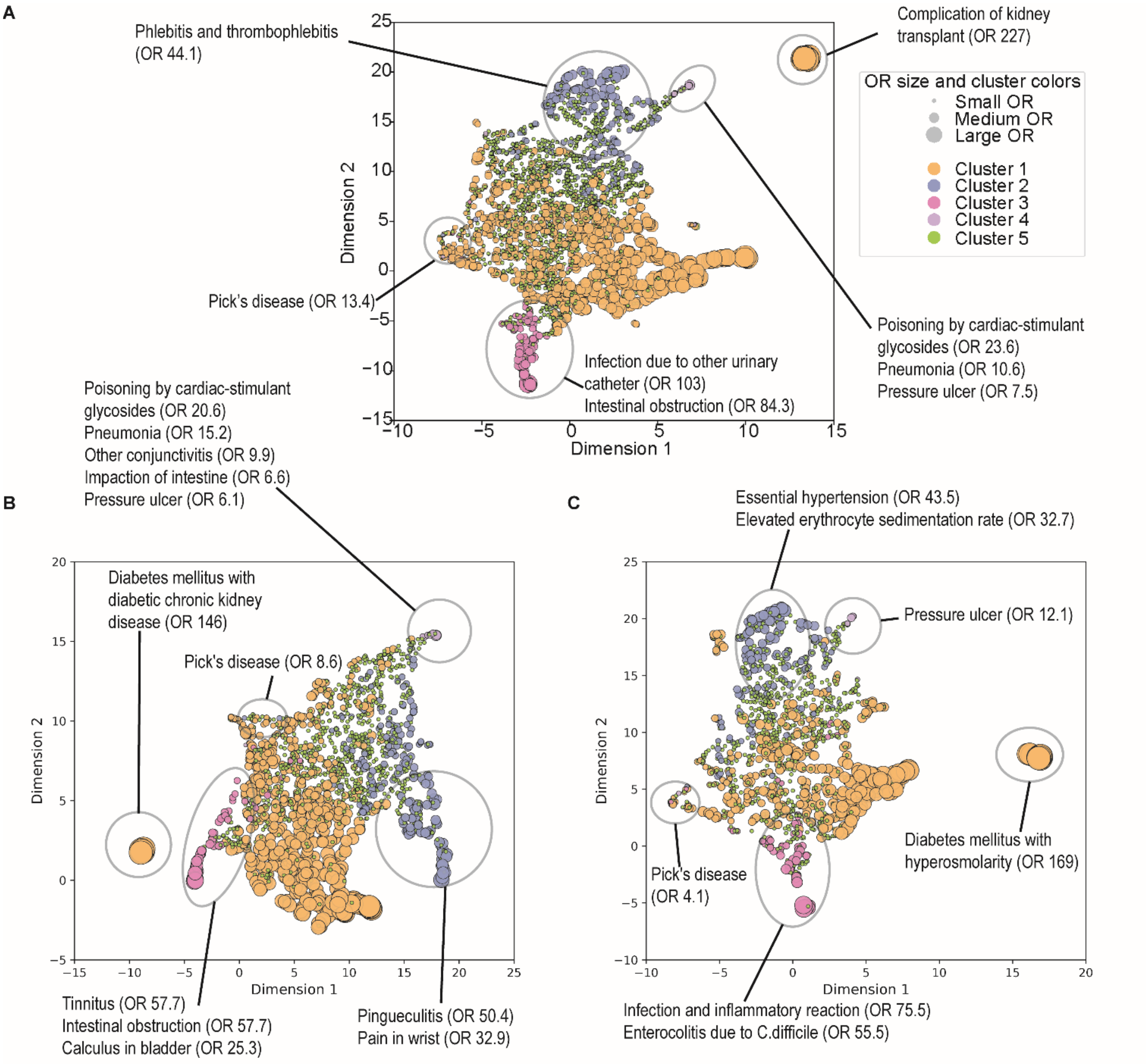

Figure 6. UMAP visualizations using odds ratios as features for each cluster. This figure presents a UMAP visualization of ICD-10-CM codes based on their standardized ORs across different clusters. Each dot represents a specific ICD-10-CM code, with the dot size corresponding to the OR of the significantly 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.

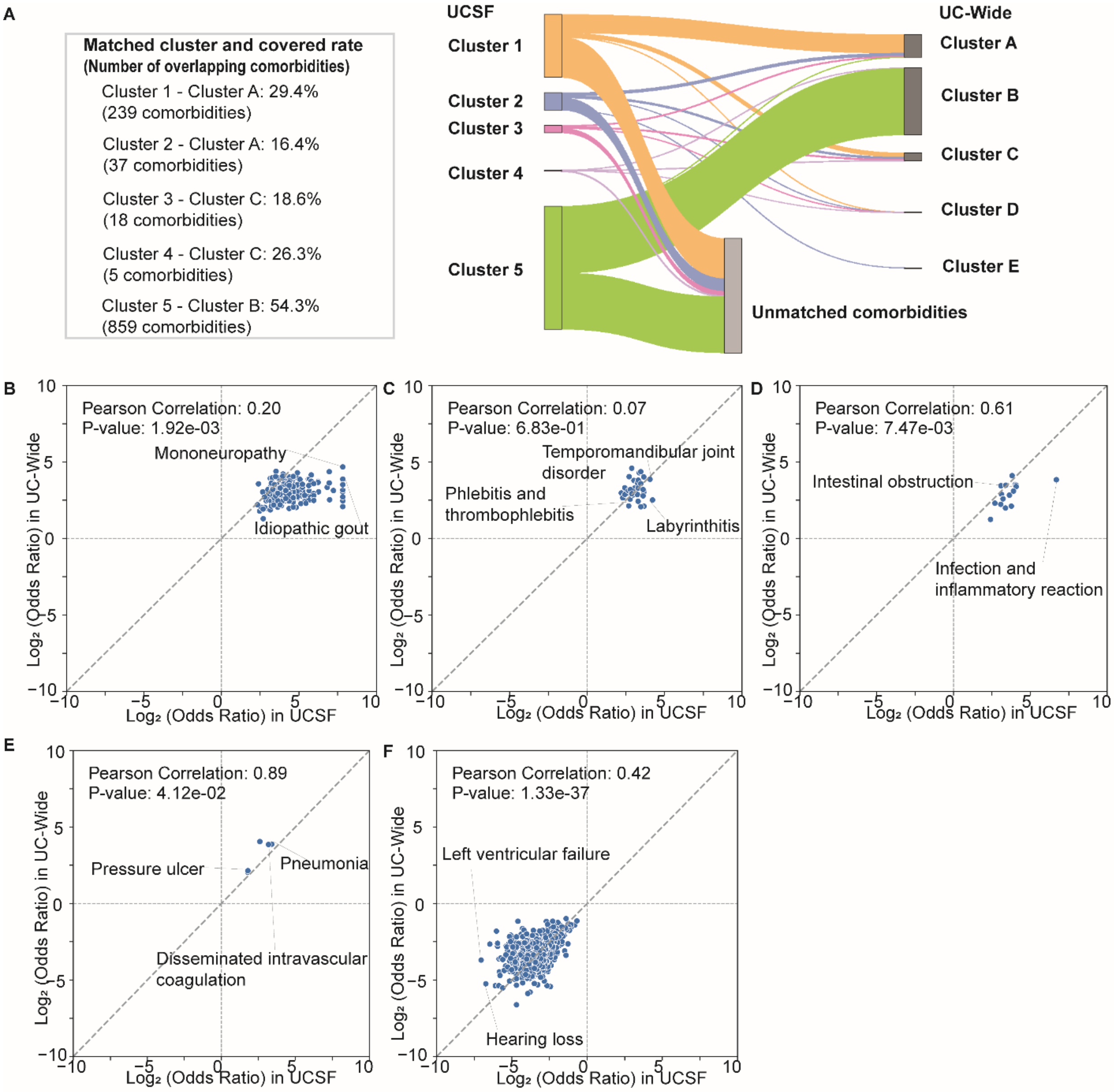

Figure 7. Validation analysis of cluster-specific significant comorbidities.

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.

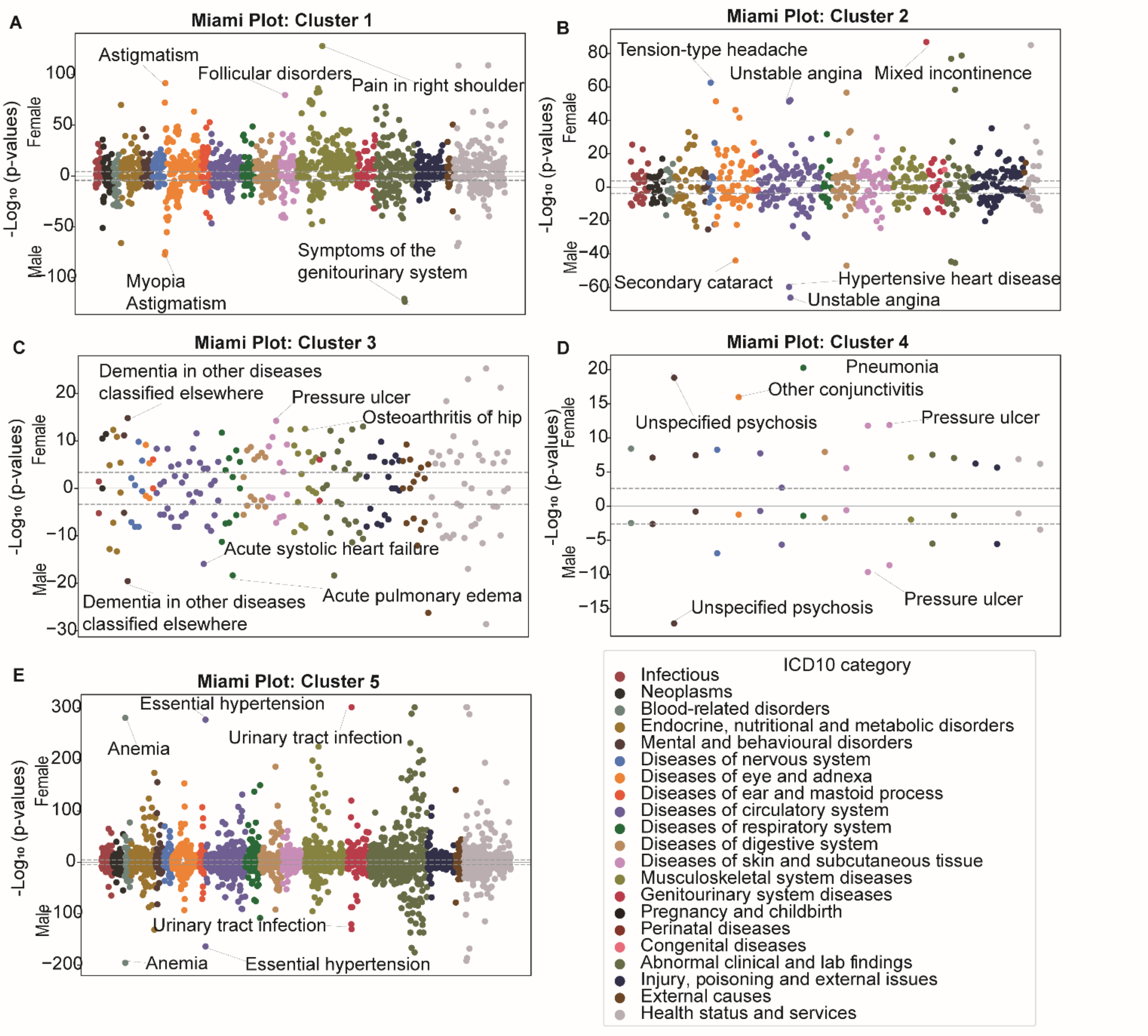

Figure 8. Miami plot for cluster-specific significant comorbidities. Miami Plot displaying enriched ICD10-CM codes identified using the two-sided Fisher's exact test (for cases < 5) or the Chi-squared test (for cases ≥ 5). Enrichment was determined based on a Bonferroni-corrected p-value < 0.05. ICD10-CM,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

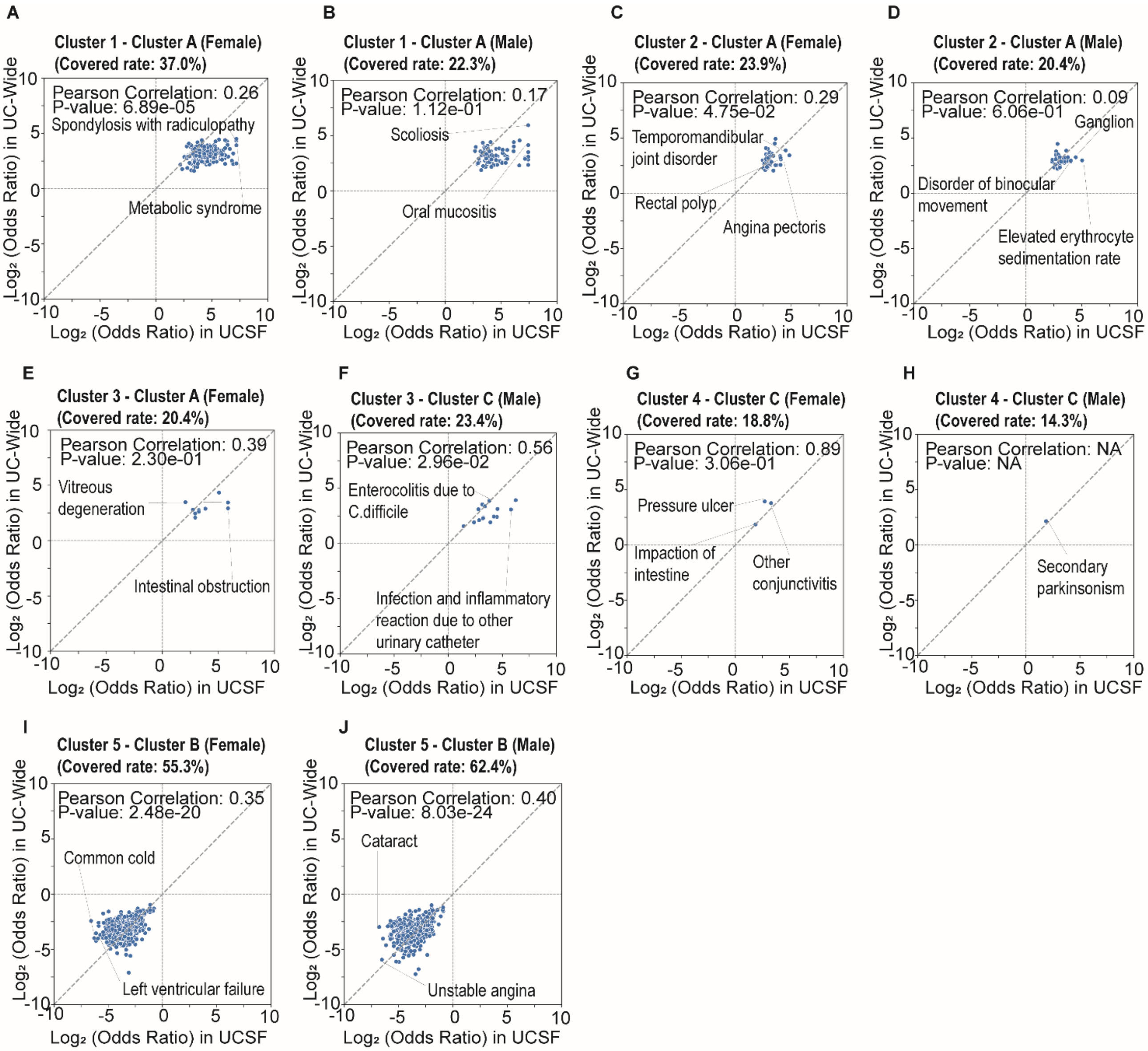

Figure 9. Validation analysis of cluster-specific significant comorbidities across sex. Log-log plot illustrating the significant overlapping comorbidities and the concordance of odds ratios for these conditions in the sex-stratified population. The covered rate represents the proportion of UCSF cluster-sex-specific 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.