










Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations

Alice S. Tang ^{1,2,3✉}, Tomiko Oskotsky ^{1,4}, Shreyas Havaladar⁵, William G. Mantyh⁶, Mesude Bicak^{5,7}, Caroline Warly Solsberg ^{8,9,10}, Sarah Woldemariam ¹, Billy Zeng³, Zicheng Hu ¹, Boris Oskotsky ¹, Dena Dubal ⁹, Isabel E. Allen¹¹, Benjamin S. Glicksberg ^{5,7} & Marina Sirota ^{1,4✉}

Alzheimer's Disease (AD) is a neurodegenerative disorder that is still not fully understood. Sex modifies AD vulnerability, but the reasons for this are largely unknown. We utilize two independent electronic medical record (EMR) systems across 44,288 patients to perform deep clinical phenotyping and network analysis to gain insight into clinical characteristics and sex-specific clinical associations in AD. Embeddings and network representation of patient diagnoses demonstrate greater comorbidity interactions in AD in comparison to matched controls. Enrichment analysis identifies multiple known and new diagnostic, medication, and lab result associations across the whole cohort and in a sex-stratified analysis. With this data-driven method of phenotyping, we can represent AD complexity and generate hypotheses of clinical factors that can be followed-up for further diagnostic and predictive analyses, mechanistic understanding, or drug repurposing and therapeutic approaches.

¹Bakar Computational Health Sciences Institute, UCSF, San Francisco, CA, USA. ²Graduate Program in Bioengineering, UCSF, San Francisco, CA, USA. ³School of Medicine, UCSF, San Francisco, CA, USA. ⁴Department of Pediatrics, UCSF, San Francisco, CA, USA. ⁵Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁶Department of Neurology, University of Minnesota School of Medicine, Minneapolis, MN, USA. ⁷Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁸Pharmaceutical Sciences and Pharmacogenomics, UCSF, San Francisco, CA, USA. ⁹Department of Neurology and Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA. ¹⁰Memory and Aging Center, UCSF, San Francisco, CA, USA. ¹¹Department of Epidemiology and Biostatistics, UCSF, San Francisco, CA, USA. ✉email: alice.tang@ucsf.edu; marina.sirota@ucsf.edu

Alzheimer's disease (AD) is the most common cause of dementia, making up 60–80% of cases, with a large and increasing burden on patients, caregivers, and society¹. AD is characterized by brain atrophy and accumulation of beta-amyloid plaques and tau tangles seen on brain pathology after death. The disease erodes memory and cognitive functions, causing interference with daily activities and contributing to emotional, social, and economic burden on patients and their families. AD is incurable and challenging to understand and diagnose. One reason AD is difficult to study is because it is a complex, heterogeneous, and multifactorial disease that takes many years to manifest². This complexity, along with the slow insidious progression of the disease, makes it difficult to fully characterize disease phenotypes and associations.

Sex is one factor that has been shown to be important in AD, with a higher prevalence in women afflicted by the disease at a 2:1 ratio compared to men¹. While women have an increased estimated lifetime risk of AD, there is mixed evidence of risk between men and women of the same age^{3,4}. Recent findings show that sex contributes to differing vulnerabilities or resilience to AD, as men with AD progress to death quicker^{5,6} while women with this disease show higher cognitive resilience despite increased tau pathology^{5,7,8}. How sex contributes to these differences in prevalence and vulnerability is a question of fervent interest among researchers in the AD field⁹. Recent studies in mice demonstrate that a second X chromosome may contribute to AD resilience⁶. Further sex-specific human studies in Alzheimer's disease also show sex modification of AD risk¹⁰, progression¹¹, and molecular phenotype^{11–15}. As such, sex is an important factor to consider in studying and phenotyping AD.

While many efforts have evaluated the association of individual risk factors with AD, unbiased approaches to these associations are limited. Prior work, largely hypothesis-driven, focused on select comorbidities associated with AD, such as hypertension¹⁶, vascular disorders¹⁷, diabetes¹⁸, obesity¹⁹, and others^{20–22}. However, how sex modulates AD complexity and heterogeneity has still not been fully explored. Prior big data approaches to AD have examined genotype-phenotype associations^{23,24} and molecular analyses^{14,25–27} to characterize AD and sex differences^{12,13}. Other work on phenotyping patients with AD using clinical data has examined neuroimaging²⁸, neuropsychiatric phenotype²⁹, chart reviews³⁰, and billing records independently. Thus, an unbiased comprehensive approach to phenotype AD and identify sex associations using full clinical records is needed.

With the rise in electronic medical record (EMR) use over the past decade³¹, there is abundant underutilized clinical data on patients covering comorbidities, medications, and lab values. This type of data set provides a great opportunity to deeply investigate diseases and identify associations to facilitate understanding disease prevention and progression. Recently, EMR has been utilized for other diseases for creating comorbidity networks³², identifying disease subtypes³³ and predicting disease outcomes^{34,35} highlighting the potential of utilizing EMR data to extract insight and utility for complex and heterogeneous diseases³⁶, but a big data integrative analysis with EMR data has not yet been applied to characterize AD.

Deep phenotyping is a data-driven approach that has been used to provide more detailed stratification and representation of a disease in the era of precision medicine^{37,38}. Here, we take an integrative approach through deep clinical phenotyping and network analysis to provide insight into AD clinical characteristics with a focus on sex differences. For the first time to our awareness, integrative phenotyping and association analysis is used to identify, in an unbiased manner, unique clinical features associated with AD itself—and reveals potential previously unknown sex-specific associations in the context of diagnoses, medications, and lab test results.

Results

From the UCSF EMR database (~5 million patients), we identified 8804 patients with AD (5558 female, 86.5 mean age (6.4 standard deviation)) and 17,608 propensity score (PS)-matched control patients (11,117 females, 86.5 mean age (6.4 standard deviation)). From the Mount Sinai EMR (~4 million patients), 5958 patients with AD (4138 females, 88.3 mean age (8.7 standard deviation)) and 11,916 PS-matched controls (8446 females, 88.7 mean age (11.4 standard deviation)) were identified (Fig. 1). Male and female groups were identified by the most recent sex assignment in the EMR, and race/ethnicity information was extracted from the EMR as reported by the patient. Post-matching analysis demonstrated the adequate balance in covariates with standardized mean differences in age and categorical distributions below 0.1 (or below 0.2 between matched sex groups). Demographic characteristics of patients with AD and matched control patients are shown in Table 1 and Supplementary Table 1.

Embedding with diagnosis shows separation between AD and controls. Due to the size of our cohort, we first performed low-dimensional visualizations using diagnoses as features to visualize patient separation. Low-dimensional UMAP visualizations of non-AD diagnoses (47,439 features, ICD-10-CM codes) show that distributions for patients with AD and controls are significantly different among the first two UMAP components (two-sided Mann–Whitney *U*-test, *p*-value < 1e–5, Fig. 2a, b) at both UCSF and Mount Sinai, with a progressive separation between groups. For the UCSF data, sex, and death status show significant correlations with the first component, while age is significantly correlated with both components (two-sided Mann–Whitney *U*-test *p*-value < 0.01, Fig. 2a, Supplementary Fig. 1). Sex, death status, and age are significantly correlated with both components at Mount Sinai (two-sided Mann–Whitney *U*-test *p*-value < 0.01, Fig. 2b, Supplementary Fig. 1).

Association analysis identifies associated comorbidities in AD. Among each diagnostic hierarchical level (Level 2 categories, Level 3 categories, and full diagnosis names), the majority of AD disease networks contain more nodes and edges compared with control networks (Supplementary Table 3). In UCSF Level 3 diagnosis networks, more nodes and edges occur in AD vs control networks. As shown in Fig. 3a, when thresholding Level 3 diagnosis categories by >10% of patients, there are 144 diagnosis pairs among patients with AD compared to one pair in controls. When comparing node-level network metrics between groups, thresholded by >1% of patients within a group, AD and control networks are significantly different when compared on closeness centrality, degree, neighborhood connectivity, and stress centrality indicating a higher degree of connectivity among AD networks across all levels (two-sided Mann–Whitney *U*-test, *p*-value < 0.01, Fig. 3c). In Mount Sinai Level 3 diagnostic networks, more edges occur in AD networks compared to control networks, with significantly different distributions across AD and control networks on degree, neighborhood connectivity, and stress centrality (two-sided Mann–Whitney *U*-test, *p*-value < 0.01, Supplementary Table 3). Across the board, network metrics normalized by the metric are significantly correlated between UCSF and Mount Sinai (Spearman's $\rho = 0.44$, *p*-value < 1e–4, Fig. 3e).

Within Level 2 diagnosis categories, there are 166 significant diagnosis categories (two-sided Fisher's exact or Chi-squared test, Bonferroni-corrected *p*-value < 0.05), with 120 diagnosis categories significantly enriched (odds ratio (OR) > 2) uniquely in the AD group and no significantly enriched diagnosis categories uniquely in the control group (Fig. 4a, top). Within Level 3 diagnosis categories,

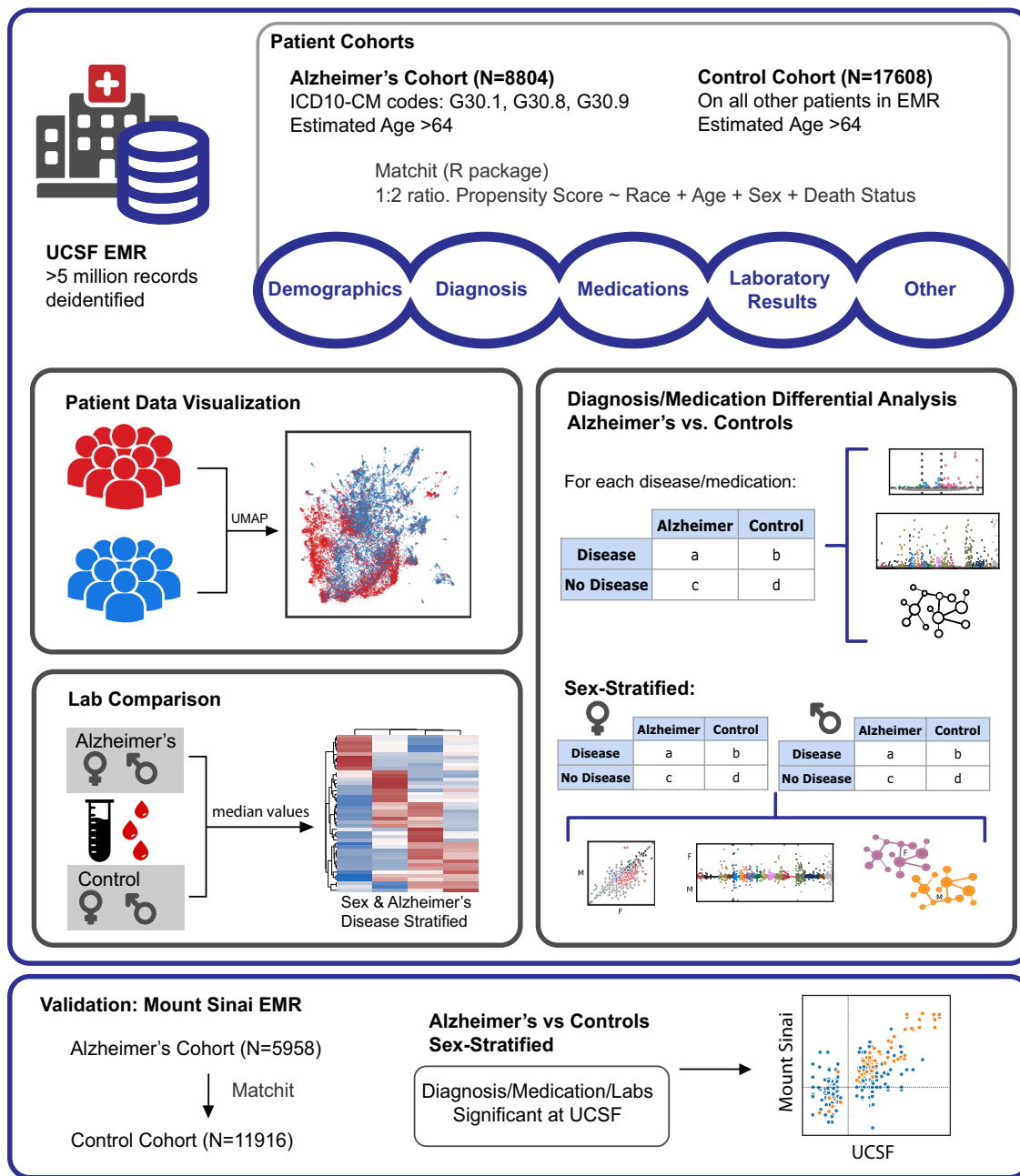


Fig. 1 Workflow visualization. Visualization of patient cohort identification from the UCSF EMR and methods for deep phenotyping and enrichment analysis. Validation analysis is done with Mount Sinai EMR to assess correlations.

there are 501 significant categories, with 391 and 4 categories significantly enriched in AD and control groups, respectively (two-sided Fisher exact or Chi-squared test, Bonferroni-corrected p -value < 0.05, Supplementary Table 2). Within full diagnosis names, there are 1627 significant diagnoses, with 1491 and 7 diagnoses enriched uniquely in AD and control groups, respectively. Top significant diagnoses in AD include vascular dementia, hypertension, hyperlipidemia, urinary tract infection, syncope, hypothyroidism, and osteoporosis, while top significant diagnoses in controls include neoplasms of liver and brain (two-sided Fisher exact or Chi-squared test, Bonferroni-corrected p -value < 0.05, Fig. 4a, bottom, Supplementary Data 1). Top ICD diagnostic blocks in AD include mental health and behavioral diseases, genitourinary diseases, endocrine and metabolic diseases, and circulatory system diseases (Fig. 4b). In the validation cohort, 1495 of 1627 significant UCSF diagnoses mapped to Mount Sinai EMR codes, of which 889

(60.13%) are significant (two-sided Fisher's exact or Chi-squared test, Bonferroni p -value < 0.05). Overall comorbidity odds ratios at UCSF are significantly correlated with those of the validation cohort at Mount Sinai (Spearman ρ = 0.65, p -value < $1e-5$, Fig. 4c).

Sex-stratified AD vs. control association analysis identifies vascular and musculoskeletal disorders in females with AD and behavioral/neurological disorders in male AD. When stratifying diagnoses by sex (see "Methods" section), AD disease networks are significantly different on metrics of degree and neighborhood connectivity in both males and females compared to their respective controls among all diagnostic hierarchical levels (p -value < 0.001). Comparison of sex-specific AD network for diagnosis name shows significantly greater neighborhood connectivity, and lower eccentricity in female networks (two-sided

Table 1 Patient demographics.

	UCSF				Mount Sinai			
	Overall	AD	Control	SMD	Overall	AD	Control	SMD
<i>n</i>	26,412	8804	17,608		17,874	5958	11,916	
Sex, <i>n</i> (%)								
Female	16,675 (63.1)	5558 (63.1)	11,117 (63.1)	<0.001	12,584 (70.4)	4138 (69.5)	8446 (70.9)	0.031
Male	9659 (36.6)	3220 (36.6)	6439 (36.6)		5290 (29.6)	1820 (30.5)	3470 (29.1)	
Unknown	78 (0.3)	26 (0.3)	52 (0.3)					
Estimated age, mean (SD)	86.5 (6.4)	86.5 (6.4)	86.5 (6.4)	<0.001	88.6 (10.6)	88.3 (8.7)	88.7 (11.4)	−0.039
Race, <i>n</i> (%)								
American Indian/ Alaska Native	27 (0.1)	9 (0.1)	18 (0.1)	<0.001	20 (0.1)	8 (0.1)	12 (0.1)	0.129
Asian	2638 (10.3)	879 (10.3)	1759 (10.3)		177 (1.0)	78 (1.3)	99 (0.8)	
Black/African American	1758 (6.9)	586 (6.9)	1172 (6.9)		3732 (20.9)	1214 (20.4)	2518 (21.1)	
Native Hawaiian/ Pacific Islander	1356 (5.3)	452 (5.3)	904 (5.3)		9 (0.1)	5 (0.1)	4 (0.0)	
Other	2230 (8.7)	743 (8.7)	1487 (8.7)		3922 (21.9)	1496 (25.1)	2426 (20.4)	
Unknown	2017 (7.6)	673 (7.6)	1344 (7.6)		786 (4.4)	253 (4.2)	533 (4.5)	
White/Caucasian	16,386 (64.0)	5462 (64.0)	10,924 (64.0)		9228 (51.6)	2904 (48.7)	6324 (53.1)	
Death status, <i>n</i> (%)								
Alive	20,146 (76.3)	6714 (76.3)	13,432 (76.3)	0.001	9371 (52.4)	3264 (54.8)	6107 (51.3)	0.078
Deceased	6266 (23.7)	2090 (23.7)	4176 (23.7)		882 (4.9)	306 (5.1)	576 (4.8)	
Unknown					7621 (42.6)	2388 (40.1)	5233 (43.9)	

Summary table of sex, estimated age, death status, and first race among Alzheimer's and control cohorts at UCSF and Mount Sinai. Patients are propensity score-matched at a 1:2 Alzheimer to control ratio with the demographics shown in the table. SD standard deviation, SMD standardized mean difference.

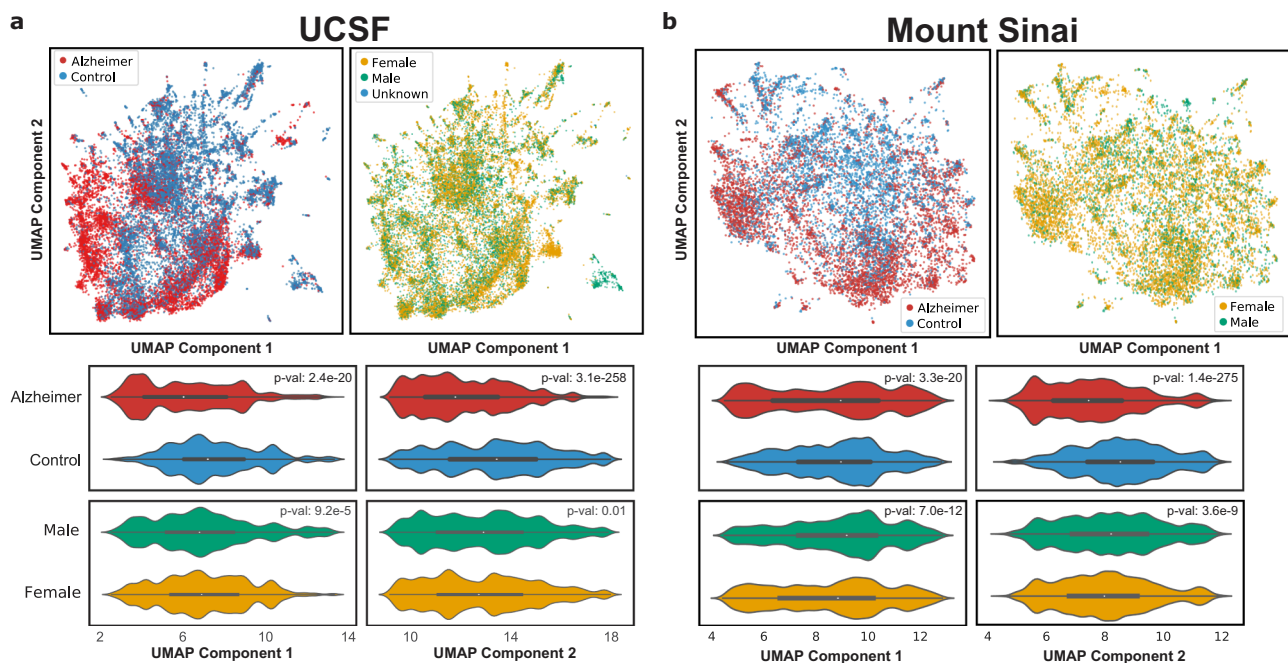


Fig. 2 UMAPs using comorbidities as features provide a topographical view of the distribution of patients. Top row: UMAP of all patients (AD and controls), with each dot representing a patient, colored by AD status (**a** top left, **b** top left) or by sex (**a** top right, **b** top right). Middle and bottom rows: violin plots show the distribution of patients with AD and controls along the UMAP principal components for UCSF (**a**) and Mount Sinai (**b**), and *p*-values determined from comparing distributions with a two-sided Mann-Whitney *U*-test. Alzheimer vs Control: UCSF *p*-value 2.4e−20 (component 1) and 3.1e−258 (component 2). Mount Sinai *p*-value 3.3e−20 and 1.4e−275. Male vs female: UCSF *p*-value 9.2e−5 and 0.01. Mount Sinai *p*-value 7.0e−12 and 3.6e−9.

Mann-Whitney *U*-test, *p*-value < 0.01 both metrics, Fig. 3d, Supplementary Table 3). Within the validation cohort, similarly, female AD networks show significantly greater neighborhood connectivity compared to male AD networks (two-sided Mann-Whitney *U*-test, *p*-value < 0.01, Supplementary Table 3).

When thresholding full diagnosis names by >10% of patients within a sex group, female patients with AD have 58 shared co-diagnosis pairs compared to 38 in male patients with AD (Fig. 3b and Supplementary Table 3), and 3 shared co-diagnosis pairs were identified for both control sex groups.

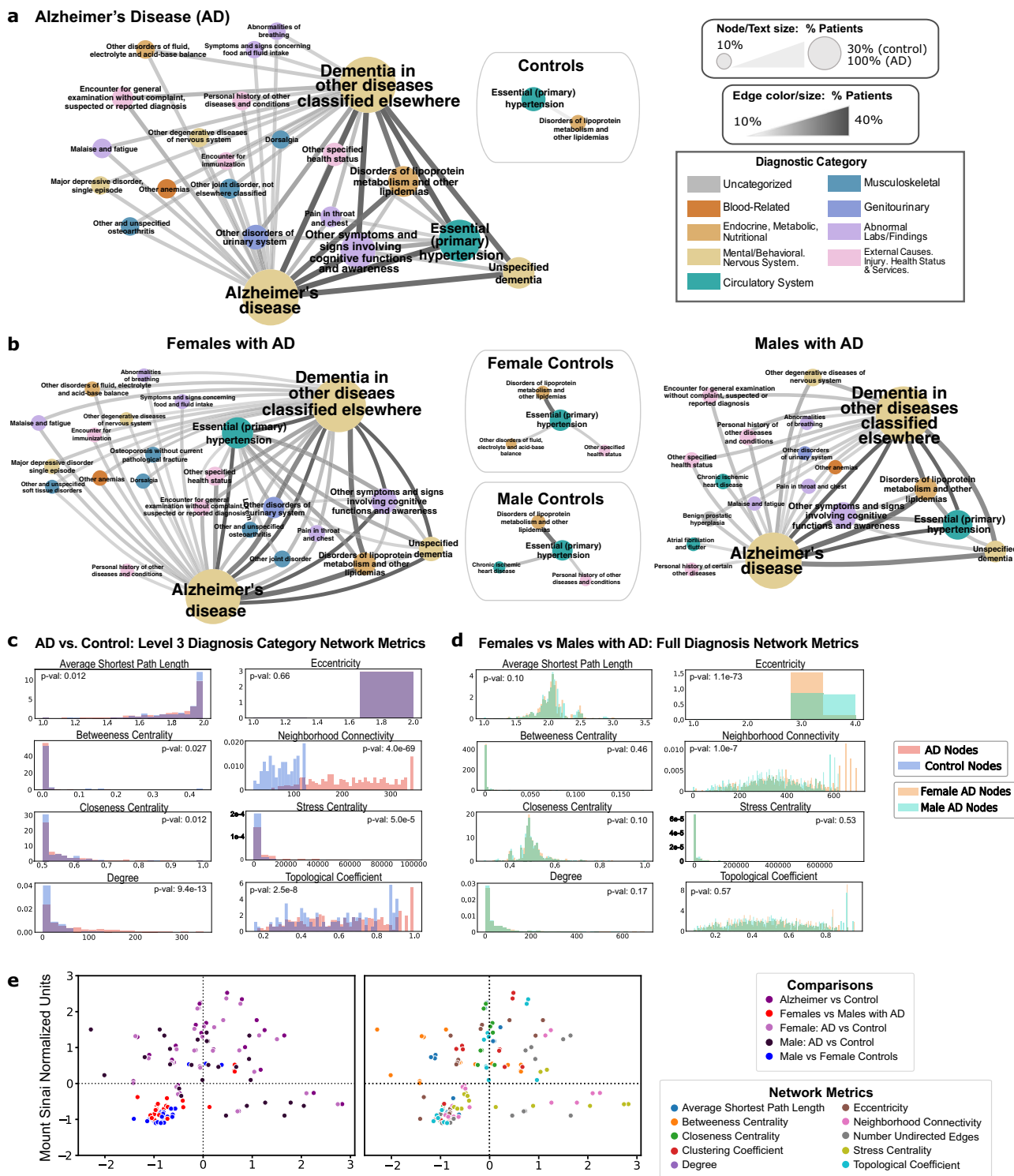


Fig. 3 Comorbidity networks show greater co-diagnosis in patients with AD vs. controls, and in females with AD vs males with AD. **a, b** Network diagrams: For each network, the node size, text size, edge size, and edge color represent the number of patients sharing a diagnosis or diagnosis pair. Node colors are based on ICD-10-CM category. A threshold of 10% sharing was applied. **a** Network for Level 3 diagnosis categories in patients with AD vs. controls. Nodes and edges represent >10% of diagnosis or diagnosis pairs shared in each cohort, respectively. **b** Female and male network of Level 3 diagnosis categories for patients with AD and controls. Each node and edge represent a diagnosis or diagnosis pairs shared by >10% of males or females in the AD or control group. **c** Comparison of Level 3 diagnosis category network metrics between patients with AD and controls. Statistical tests are performed with a two-sided Mann-Whitney *U*-test. Significant metrics with *p*-value < 0.01: degree (9.4e-13), neighborhood connectivity (4.0e-69), stress centrality (5.0e-5), and topological coefficient (2.5e-8). **d** Comparison of network metrics between male and female Alzheimer's disease full diagnostic name networks. Statistical tests are performed with a two-sided Mann-Whitney *U*-test. Significant metrics with *p*-value < 0.01: eccentricity (1.1e-73) and neighborhood connectivity (1.0e-7). **e** Correlation of network metrics compared with validation EMR network metrics, normalized by the metric. Colors represent comparison type (left) or the specific network metric (right), Spearman's $\rho = 0.55$, *p*-value < 1e-4.

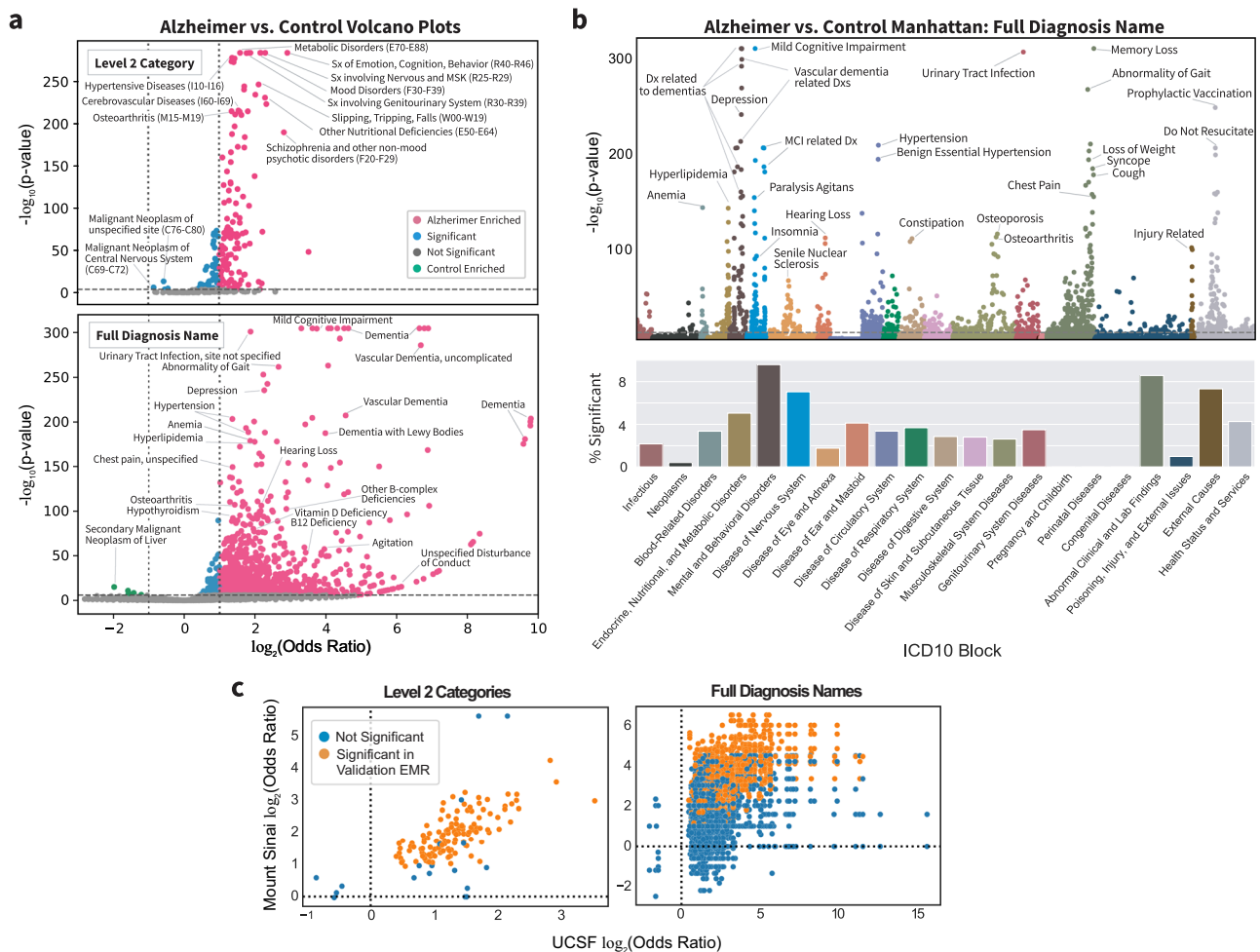


Fig. 4 Comorbidity enrichment analysis identifies enriched diagnosis in AD vs. control cohorts. **a** Volcano plot for Level 2 categories (top) and full diagnosis names (bottom) compared between AD and control cohorts using two-sided Fisher's exact or Chi-squared test. p -value cutoff is Bonferroni-corrected (p -value $< 2e-8$ and $1e-6$) with \log_2 odds ratio cutoff of 1 for AD-enriched (pink) or \log_2 odds ratio cutoff of -1 for control-enriched (green) and remaining significant diagnoses in blue. Some of the top significant diagnoses are labeled. **b** Top, a Manhattan plot with full diagnosis names colored by ICD-10-CM categories with significance determined by two-sided Fisher's exact or Chi-squared test with Bonferroni-corrected p -value threshold of 0.05. Some of the top diagnoses in each category are labeled. Bottom, the percentage of diagnosis in each ICD-10-CM category is significant. **c** Diagnosis AD vs. control odds ratio correlation plots between UCSF and Mount Sinai for Level 2 diagnosis categories and full diagnosis names that are significant at UCSF (two-sided Fisher's exact or Chi-squared test, Bonferroni-corrected p -value threshold of 0.05). Each dot represents a category or diagnosis, and dots in orange are significant at Mount Sinai with (two-sided Fisher's exact or Chi-squared test with Bonferroni-corrected p -value threshold of 0.05 based on the number of significant UCSF diagnoses).

For both males and females, there are 136, 338, and 714 shared significant diagnostic categories or diagnoses for Level 2, Level 3, and full diagnosis names, respectively. In a sex-stratified analysis, there were 29, 164, and 699 female-only significant hits and 5, 18, and 91 male-only significant hits for Level 2, Level 3, and full diagnosis names (two-sided Fisher's exact or Chi-squared test, Bonferroni-corrected p -value < 0.05 , Fig. 5a, Supplementary Data 1). Compared to males among Level 2 diagnostic categories, females have a greater percent of significant diagnoses in blood-related disorders (e.g., nutritional anemia, coagulation defects) and congenital disorders and also have greater enrichment of pervasive and specific developmental disorders, musculoskeletal disorders (e.g., chondropathies, other osteopathies), injuries (e.g., injuries to the hip and thigh, injuries to the ankle and foot), infections with a predominantly sexual mode of transmission, and metabolic disorders (Supplementary Data 1). When comparing Level 2 categories in the validation cohort, among females, 153 out of 165 mapped with 60 (30.22%) significant, and among

males, 133 out of 141 mapped with 64 (48.12%) significant (two-sided Fisher's exact or Chi-squared test, Bonferroni-corrected p -value < 0.05 based on the number of significant UCSF diagnoses). In general, Level 2 category sex-specific odds ratios are correlated between institutions (Females: Spearman's $\rho = 0.77$, p -value $< 1e-5$; males: Spearman's $\rho = 0.83$, p -value $< 1e-5$). In the validation cohort, females have similar enrichment of blood-related disorders (e.g., nutritional anemia) and injuries (e.g., injuries to the hip and thigh), while males have enrichment of behavioral/emotional disorders.

Within full diagnosis names, unique significant diagnoses of female patients with AD include asthma, atrial fibrillation, arthritis, fractures, and accidents while unique significant diagnoses of male patients with AD include parkinsonism, sleep apnea, hypersomnia, neuropathy, irritability, and imbalance (two-sided Fisher's Exact or Chi-squared test, Bonferroni-corrected p -value < 0.05 , Fig. 5a, b, Supplementary Data 1). Among full diagnosis names significant in both males and

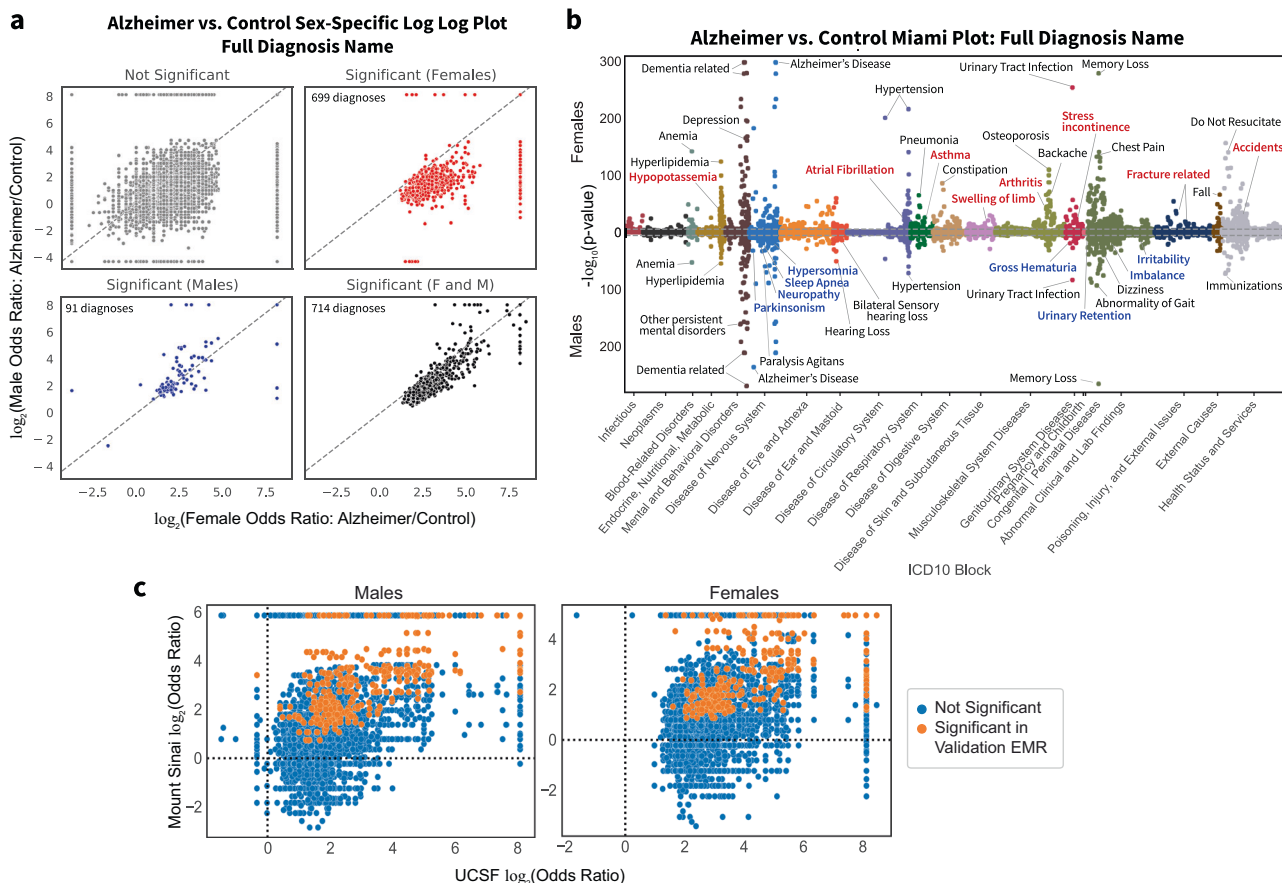


Fig. 5 Comorbidity enrichment analysis identifies sex-specific enriched diagnoses in AD vs. control cohorts. **a** Full diagnosis names compared between patients with AD and controls within each sex. The log₂ of the odds ratio for each sex is plotted on the axis, and points are colored by significance (Bonferroni-corrected, *p*-value cutoff > 3e−6). **b** Miami plot of the diagnosis names grouped by sex and ICD-10-CM categories. Select top diagnoses are labeled, with diagnosis names colored by significance as female-only (red), male-only (blue), or significant in both sexes (black). **c** Correlation plots of AD vs. control odds ratios between UCSF and Mount Sinai for diagnoses that are significant at UCSF for each sex group (two-sided Fisher’s exact or Chi-squared test, Bonferroni-corrected *p*-value threshold of 0.05). Each dot represents a diagnosis, and dots in orange are significant at Mount Sinai (two-sided Fisher’s exact or Chi-squared test with Bonferroni-corrected *p*-value threshold of 0.05 based on the number of significant UCSF diagnoses for each sex group).

females, female patients with AD have a greater association in depression, hypertension, hyperlipidemia, urinary tract infections, upper respiratory infections, anemia, osteoporosis, and pneumonia, while male patients with AD have greater effect size with behavioral phenotypes, hearing loss, and agitation (Supplementary Data 1). Among the full diagnosis names in the validation cohort, for females, 1149 out of 1383 significant diagnoses mapped, of which 240 (20.89%) were significant, and for males, 702 out of 805 significant diagnoses mapped, of which 216 (30.77%) were significant. In general, sex-specific diagnosis odds ratios were correlated for both females (Spearman’s $\rho = 0.77$, *p*-value < 1e−4) and males (Spearman’s $\rho = 0.83$, *p*-value < 1e−4, Fig. 5c). In the validation cohort, similarly, female patients with AD have a greater association in depression, hypertension, and osteoporosis while male patients with AD have a greater association in hearing loss and agitation (Supplementary Data 1).

Few comorbidities change with sensitivity analysis on encounters. For our sensitivity analysis that included only patients with ≥10 encounters (i.e., recorded outpatient, inpatient, or emergency room visits to UCSF Health) in EMR and with visits spanning >1 year, there were 6612 patients with AD (2382 males, 4223 females) and 13,224 control patients (4674 males,

8539 females) identified by PS-matching on the number and timespan of encounters in addition to demographic characteristics and death status. A summary of the demographic characteristics of these cohorts is shown in Supplementary Table 1. We identified 100, 222, and 561 significant Level 2, Level 3, and full diagnosis names respectively (two-sided Fisher’s exact or Chi-squared test, Bonferroni-corrected *p*-value threshold of 0.05), and an increase in the odds ratio for chromosomal abnormalities and cerebrovascular disorders in patients with AD (Supplementary Data 2). With sex-stratified enrichment analysis, encounter controlling increased enrichment of cerebrovascular disease in females, and increased significant enrichment of behavioral disorders, vision problems, and vascular dementia in males (Supplementary Data 2). An interactive visualization of Figs. 3 and 4 is made available in an Rshiny app vizad.org.

Medication association analysis identifies dexamethasone as enriched in controls. In addition to comorbidities, we performed medication enrichment analysis in order to phenotype patients and investigate medication prescriptions enriched in patients with AD and controls. Medications found enriched (two-sided Fisher’s exact or Chi-squared test, Bonferroni-corrected *p*-value < 0.05, OR > 2 or < 0.5) in patients with AD include current treatments

like donepezil and memantine, but also vitamin B12, antidepressants (escitalopram, citalopram, sertraline, mirtazapine, trazodone), antipsychotics (quetiapine, risperidone, olanzapine), carbidopa/levodopa, vitamin D3, and melatonin. Medications found enriched in control patients include dexamethasone, ondansetron, and alteplase. Significant medications in controls with lesser effect size (two-sided Fisher's exact or Chi-squared test, Bonferroni-corrected p -value < 0.05 , $0.5 < OR < 1$) include midazolam, propofol, opioids (oxycodone, fentanyl citrate), and furosemide (Fig. 6a). From the validation cohort, 116 out of 121 medications mapped, of which 66 (56.90%) were significant (two-sided Fisher's Exact or Chi-Squared test, Bonferroni-corrected p -value < 0.05 based upon significant medications at UCSF). In general, odds ratios of medications are significantly correlated (Spearman's $\rho = 0.85$, p -value $< 1e-4$, Fig. 6c). Dexamethasone is significant among controls in both institutions, and multiple medications including vitamin B12, antidepressants, and antipsychotics are significant in patients with AD among both institutions.

In a sex-stratified analysis, medications enriched in males with AD include Tdap vaccine, melatonin, and carbidopa/levodopa while methylprednisolone and phenylephrine are enriched in control males. Female patients with AD have enrichments in diazepam, antipsychotics (risperidone, aripiprazole), buspirone, antidepressants (sertraline, mirtazapine, trazodone, bupropion), vitamin D2, and levothyroxine while control females are enriched in norepinephrine bitartrate and fentanyl citrate (Fig. 6b). In the validation EMR, 18 of 23 (78.25%) significant medications found at UCSF are significant in females at Mount Sinai, and 13 of 16 (81.25%) in males (two-sided Fisher's exact or Chi-squared test, Bonferroni-corrected p -value < 0.05 based upon significant medications at UCSF within a group). Overall, there is significant correlation of sex-specific medication odds ratios in females (Spearman's $\rho = 0.7$, p -value = 0.001) and males (Spearman's $\rho = 0.62$, p -value = 0.001, Fig. 6c). Among both institutions, carbidopa/levodopa is significant in males with AD only.

Comparing labs between sex-specific AD and control groups identifies clusters of lab value differences. We also performed an unbiased analysis of laboratory test result differences between patients with AD and controls to phenotype patient groups. Among significantly different median lab values in both UCSF and Mount Sinai, patients with AD have higher levels of hematocrit, serum calcium, RBC count, serum albumin, and cholesterol and lower levels of glucose, activated partial thromboplastin time (aPTT), alanine transaminase (ALT), and aspartate transaminase (AST) compared to controls (two-sided Mann-Whitney U -test, Bonferroni-corrected p -value threshold of 0.05, Fig. 6d, Supplementary Fig. 4A).

Average significant median lab values across sex-stratified groups (females with AD, males with AD, control females, control males) and across institutions were clustered into 7 significant clusters (Family-wise Error Rate (FWER) corrected p -value 0.05 cutoff, Fig. 6d). Clusters 1, 4, and 7 show discordant results between UCSF and Mount Sinai. Cluster 2 represents groups of significant median lab values lowest in control males, and highest either in all patients with AD (e.g., albumin, sodium, and carbon dioxide) or highest in females with AD (e.g., HDL cholesterol, lymphocytes, calcium). Cluster 3 represents significant labs with greater median values in females and in controls (e.g., Free T4, sedimentation rate). Cluster 5 represents labs with lower significant median values in patients with AD than controls for either the whole group (e.g., B-Type Natriuretic Peptide, AST) or in a sex-specific way where significant median lab values for

males are greater than for females (e.g., aPTT, ALT, ferritin). Cluster 6 shows labs greater in AD compared to controls in a sex-specific way where overall males have greater significant median lab values than females (e.g., hemoglobin, RBC count). Across the board, the normalized lab values are correlated between the institutions (Female control: Spearman's $\rho = 0.45$, p -value < 0.001 ; male control: 0.46, p -value < 0.001 ; female AD: 0.59, p -value $< 1e-5$; Male AD: 0.64, p -value $< 1e-5$; Supplementary Fig. 4B).

Discussion

In this work, we demonstrate the capability of utilizing data from EMRs in order to perform deep phenotyping of a complex and heterogeneous disease, Alzheimer's Disease (AD), and derive insights into associations with AD in a combined and sex-stratified analysis.

First, we performed low-dimensional topographical embedding of patients using diagnoses as features in order to visualize patients spatially. We see that AD status is significantly correlated with the first two UMAP components at both institutions, suggesting that phenotypic representation of patients using diagnosis data can demonstrate separation of patients with AD and control patients. The UMAP representation demonstrates a progressive spectrum between control patients and patients with AD, as well as representing variance and heterogeneity at individual patient resolution. Furthermore, with the UMAP representation, we can visualize topographically the distribution of age, sex, and other variables among patients.

We then generated comorbidity networks between patients with AD and control patients which provide a phenotypic representation of disease interactions among patient groups and a difference in connectivity between diseases in patients with AD and control patients. AD networks contain a greater number of edges and network metrics that point to higher rates of comorbid conditions among patients with AD at both institutions, particularly with stronger links of hypertension (HTN)—lipidemias and HTN—urinary disorders. Indeed, other studies have found multimorbidities (such as neuropsychiatric and cardiovascular patterns) to increase the risk for dementia³⁹, and to contribute to AD pathological heterogeneity^{40,41} displaying the larger complexity and heterogeneous nature of AD.

With enrichment analysis, we applied an integrative, unbiased, big data approach to EMR and identified previously known associations and possible novel connections with AD. Some diagnoses found enriched in patients with AD compared to control patients from our analysis at both institutions that have been previously identified as linked with AD include midlife hypertension^{16,42}, diabetes mellitus^{18,43}, anemia^{44,45}, vascular pathology^{17,46}, osteoporosis^{47,48}, and urinary tract infections (UTI)⁴⁹. Enrichment of hypertension and vascular risk factors supports many current hypotheses of potential vascular pathologies and inflammatory factors that may lead to AD^{17,50-52} or "unmask" the symptoms of AD by decreasing cognitive reserve by causing vascular brain disease. Enrichment of diabetes and dyslipidemia supports existing literature that found links with diabetes mellitus and dyslipidemia⁵³, with proposed hypotheses involving energy metabolism⁵⁴⁻⁵⁶, inflammation⁵⁷⁻⁵⁹, or the integrity of the blood-brain barrier⁶⁰⁻⁶². Enrichment of degenerative diseases of age, such as osteoporosis, osteoarthritis, urinary issues, and sensory issues may align with theories of AD as being a disease linked with frailty⁶³⁻⁶⁵. This analysis, therefore, provides an unbiased integrative way to identify multifactorial associations with AD. Our enrichment analysis also identified neoplasms as enriched in controls at UCSF, especially cancer of the brain and liver. While this is an associative finding, this supports ideas that cancer and AD co-occur less

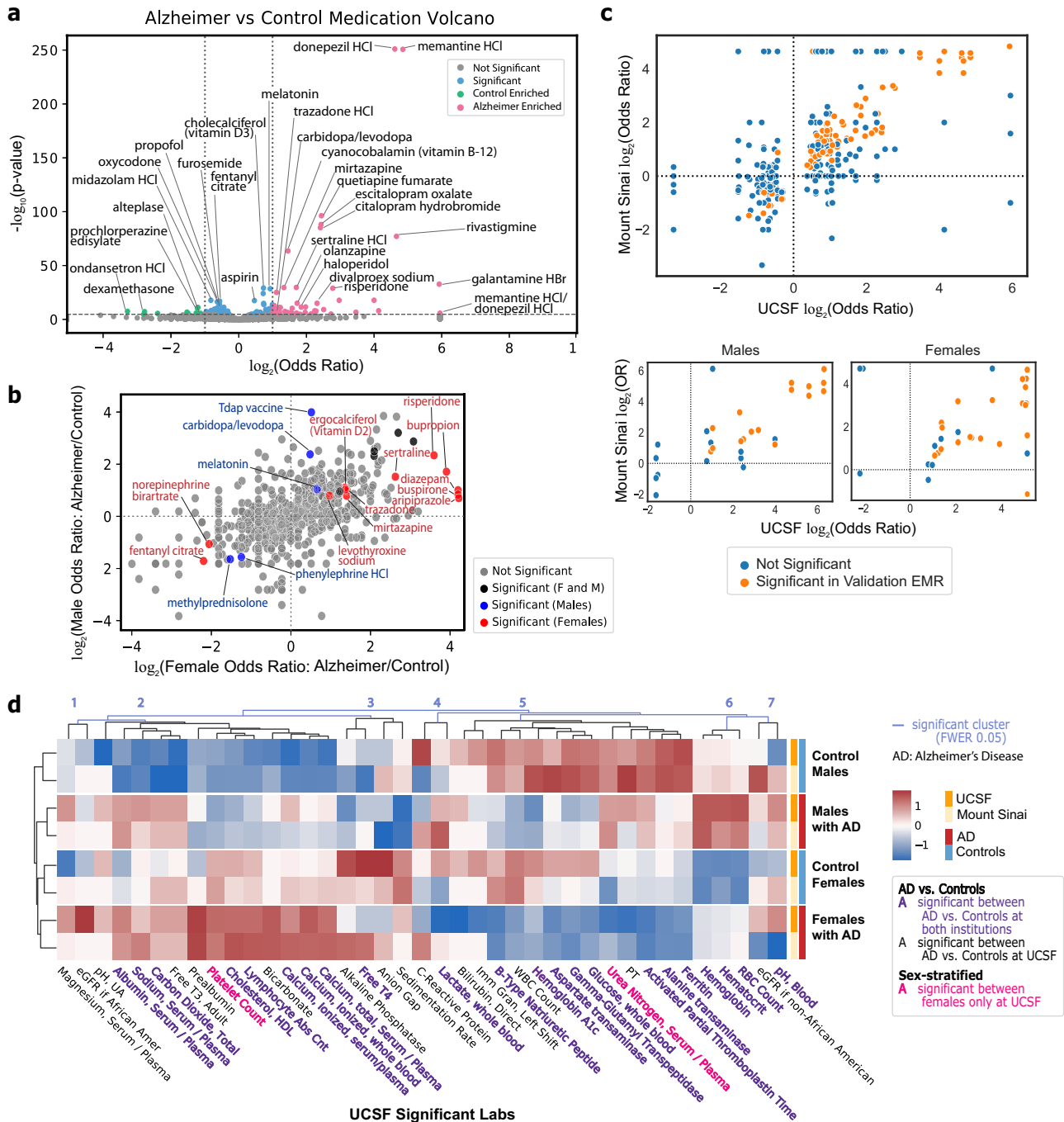


Fig. 6 Medication and lab analysis shows medication enrichments and median lab value differences between AD and control cohorts. a Volcano plot for generic medication names compared between AD and control cohorts using two-sided Fisher’s exact or Chi-squared test. p -value cutoff is Bonferroni-corrected (p -value $< 2e-5$) with odds ratio cutoff at 2 for AD-enriched (pink) or 1/2 for control-enriched (green). Remaining significant diagnoses are in blue. **b** Log-log plot of generic medication names compared between AD and control cohorts within each sex. The log of the odds ratio for each sex is plotted on the axis, with points colored by significance based upon two-sided Fisher’s exact or Chi-squared test with Bonferroni-corrected threshold of 0.05 if female-only (red), male-only (blue), or both (black). **c** AD vs control (top) and sex-specific (bottom) odds ratio correlation plots between UCSF and Mount Sinai for medications significant at UCSF (two-sided Fisher’s exact or Chi-squared test with Bonferroni-corrected p -value threshold of 0.05). Each dot represents a medication, and the dots in orange are significant at Mount Sinai (two-sided Fisher’s exact or Chi-squared test with Bonferroni-corrected p -value threshold of 0.05 based on the number of significant UCSF diagnoses in each group). **d** Heatmap of lab values filtered on significance at UCSF in AD vs control comparison across sex-specific groups at UCSF and Mount Sinai. Labs are clustered with light blue lines representing significant cluster breaks (family-wise error rate (FWER)-corrected p -value 0.05). Text color represents significant labs at both institutions (purple), significant among females only at UCSF (red), or significant between AD vs controls at UCSF only (black). Heatmap colors represent z-score of the average median value across the 4 groups at each institution.

frequently than the general population^{66,67}. Some theories propose that AD and cancer have similar mechanisms and molecular pathways, but are dysregulated in different directions^{68,69}.

Next, we generated sex-specific comorbidity networks to provide insight into sex differences in the complexity of the disease. In both EMRs, female AD networks contain more nodes with network metrics suggesting greater connectivity than female controls or male AD networks. This may support the association with greater combined diagnoses and multimorbidity in female patients with AD compared to males⁷⁰. These associations would be consistent with theories of greater risk of dementia in females as a result of multiple diseases or the theory of greater cognitive and pathological resilience to AD in females due to the burden of more comorbidities. Furthermore, sex-stratified networks show secondary interactions between comorbidities and AD, such as links of HTN-UTI and HTN-chest pain among female AD populations, but not in male patients with AD. These findings give higher-order comorbidity interactions associated with AD that have not been examined previously.

When performing enrichment analysis, we identify sex-specific enrichments that may be linked to AD that have not been previously explored in depth. Male patients with AD show enrichment of neurological and sensory disorders (sleep disorder, parkinsonism, and irritability), and among diagnoses significant in both sexes, males with AD have a stronger effect size with behavioral diagnoses, agitation, and hearing loss. These disorders are also mostly shown to be significant and associated with greater effect size compared to females in our validation cohort. Prior studies have found hearing loss to increase risk of dementia diagnosis^{71,72} or cognitive decline^{73,74} in men. The enrichment of behavioral and neurological disorders found in male patients with AD may indicate lessened resilience or higher occurrence of co-pathology. Furthermore, this analysis found the psychiatric phenotype associated with AD to be related to behavioral phenotypes in males compared to females, which is consistent with prior studies^{75,76}.

Female patients with AD have enrichment of unique significant diagnoses in musculoskeletal categories (arthritis, fractures), atrial fibrillation, and accidents, and among diagnoses significant in both sexes, females with AD show stronger effect size with depression, hypertension, urinary tract infections, and osteoporosis. Some of these disorders are similarly significant and associated with greater effect sizes compared to males in our validation cohort. The diagnoses of hypertension and atrial fibrillation would be in line with the hypothesis of potential cardiovascular risk factors and pathology that may affect females more. Indeed, there is evidence supporting cardiovascular fitness to be protective or vascular risk factors to be harmful towards cognitive decline and dementia in women^{42,77–79}. Furthermore, these diagnoses suggest a phenotype for females with AD along with other degenerative diseases of aging and frailty. In particular, the increase in musculoskeletal and bone disorders in females with AD, as well as high calcium and vitamin D deficiency, may point to a potential bone metabolism pathology or aberrant calcium metabolism in females with AD. From a psychiatric standpoint, the female AD phenotype is more associated with depression compared to males as supported by studies that found depression associated with greater hippocampal volume loss in women⁸⁰, and is more likely to be a manifestation of mild cognitive impairment or AD in females^{81,82}.

We performed sensitivity analysis by taking the number of encounters for each group into account. In general, we see a decrease in statistical significance in our enrichment analysis consistently across all diagnoses. This is likely due to decreased power from a lower sample size, and a bias toward the selection of patients with more severe disease due to encounter thresholding.

Overall, enriched diagnoses are relatively similar, with an increase in cerebrovascular disorders observed in AD, and particularly females with AD. Neuroimaging studies have identified differences in AD phenotypes and brain networks depending on the presence of cerebrovascular disease^{83,84}, which may support cerebrovascular events as an associated phenotype for a different or severe phenotype of AD.

Medication enrichments show expected associations with AD, as the top medication hits are current therapies used to modify symptoms of AD (e.g., memantine, donepezil), or are associated with diagnoses found in comorbidity analysis (e.g., anti-depressants for depression).

These medications are also identified as AD-enriched in our validation cohort, although many of these medications are expected as they are associated with conditions of aging. Medications enriched in controls provide a more interesting story, as they not only suggest an ‘opposite AD’ phenotype, but control-enriched hits may provide a way to hypothesize potential targets for further exploration of protective drug effects or drug repurposing. From our medication analysis, we see control enrichments of opioids, sedatives, dexamethasone, and furosemide, with dexamethasone, also found significant in our validation cohort. The negative association with opioids is inconsistent with prior studies that found associations between prescription opioid use and AD risk⁸⁵, although control enrichment of opioids could possibly be due in part to decreased ability to communicate pain and decreased opioid prescriptions after AD⁸⁶. Nevertheless, studies have implicated the role of opioid system dysregulation in tau hyperphosphorylation and AD⁸⁷. Dexamethasone is a corticosteroid that has been suggested to help reduce inflammation in AD^{88,89}, although the data on efficacy is still uncertain and may depend upon the need for combination therapy⁹⁰ or control of other factors that complicate the relationship between hormonal levels and the brain^{91,92}. Furosemide is a diuretic drug used to treat hypertension and may confer a protective effect through the control of comorbid conditions that contribute to cardiovascular risk factors. Furosemide also reduces the production of CSF by inhibiting carbonic anhydrase, which may impact CSF dynamics and help decrease the risk of AD⁹³. Prior studies have shown possible protective effects from diuretic drugs and AD^{94–97}, and one study identified furosemide as a potential probe molecule for reducing neuroinflammation⁹⁸.

Characterizing patients by lab values provides another way to phenotype patient groups. Through our analysis, greater calcium levels were identified, especially in females with AD. A small observational study found calcium supplementation to increase the risk of dementia in women with cerebrovascular disease⁹⁹. Calcium dysregulation and homeostasis have been implicated in AD neuronal signaling pathology, and identified as a target for drug development^{99,100}. Control-enriched labs may also be related to gastrointestinal cancers or liver/pancreatic dysfunction, as we observe increased AST, ALT, and glucose levels in controls and particularly among males. This result is not consistent with a study observing greater glucose levels to increase dementia risk¹⁰¹, although one study did find low ALT¹⁰² to be associated with AD, and some publications implicate altered glucose metabolism^{103,104} and liver dysfunction in AD pathology^{102,105,106}. Furthermore, since our control cohort has been matched on age and death status, control patients may encompass a population with a terminal disease. Lab clusters also demonstrate phenotypes specific to a sex group. A lower clotting time (aPTT, PT) and greater platelet count, prealbumin, lymphocytes, and cholesterol levels in females with AD may provide a multivariate way to identify potential AD phenotype in females. Prior studies have shown high thrombin^{107,108}, abnormalities of hemostasis^{109,110}, and abnormal platelet activation^{111–113} in patients with AD that may contribute

to a pro-thrombotic state in AD¹¹⁴, leading to microinfarcts and cerebrovascular dysfunction^{115,116}, although sex-specific associations have not been studied previously. Furthermore, control sex phenotype may demonstrate protective labs or biomarkers that decrease the risk of AD. We see lower free T3 in control males, and greater free T4 in control females. Indeed, studies on AD populations have shown high TSH and low free T4 to be associated with the disease^{117–119}, although sex-specific associations have not been explored in depth.

Some limitations do exist in our study. First, AD is an insidious and heterogeneous disorder, and is frequently misdiagnosed even in specialized dementia centers. Clinically, Alzheimer's dementia is suspected when disease biomarker status is unknown, whereas Alzheimer's disease is diagnosed when biomarker status is confirmed. Our current study did not rely on biomarker-positive cases of Alzheimer's disease, and we did not exclude patients with other pathologies that can also impact brain health through different pathways, such as Parkinson's disease. Nevertheless, Alzheimer's disease often co-occurs with other dementias^{120,121}. Second, EMRs, while a rich data source, is a very sparse data set with a lot of missing data, such as sociological factors (e.g., income, education, etc). Nevertheless, the number of patients represented in the EMR is exceptionally large and provides robust opportunities for deriving meaningful insights or hypotheses. This limitation also applies to our validation EMR. Additionally, some associations may be different across the two systems due to differences in the underlying patient populations or standards of care. Therefore, it is possible that the UCSF EMR does not capture an association that may be more prevalent in a different population in New York, and vice versa. How other covariates including socioeconomic factors modify specific AD associations is a question that can be followed up in future work. Third, our definition of controls comes with limitations, as it is difficult to identify "healthy" controls in the EMR. The institutions represented in our data include both primary and tertiary care, which includes patients that seek hospital care for a variety of reasons. As such, there may be bias in the underlying patient population who chooses to seek medical care at a metropolitan medical center. Regardless, the power in utilizing EMR allows us to generate hypotheses with a large number of patients and versatility in choice of controls compared to many current AD studies. Lastly, our analysis only identifies associations with AD and does not take temporal factors into consideration, therefore causal relationships cannot be concluded. This will be the main focus of future work, as the temporal association can categorize an association as a risk/protective factor (if early in age), a diagnostic clue (if during AD diagnosis), or as a manifestation of AD progression or severity (if after AD diagnosis). Nevertheless, given AD is an insidious disorder, there can be brain perturbations a decade or more before a diagnosis is determined and documented in clinical records. While we made the assumption of independence in our statistical methods to identify significant associations, this method can be further extended to alternative statistical models that take covariates into account. Our current work allows the unbiased identification of associations and phenotyping, which can then be used to generate hypotheses for guiding follow-up studies.

Overall, our analyses leveraged an extensive clinical data set to (1) phenotype and represent AD and (2) perform enrichment analysis to identify known or suggested novel associations with AD, as well as elicit sex-specific differences. We were therefore able to apply an integrative, unbiased, big data approach to identify associations with AD and provide phenotypic representations of an otherwise complex disease. With this approach,

we can generate many new hypotheses to better motivate future work to understand AD complexity and develop diagnostic strategies and therapeutic interventions. Future work will include temporal analysis in order to identify longitudinal relationships and predictive modeling for AD risk, diagnosis, or progression. More extensive analysis of medication and lab values, especially among opposite phenotypes in controls, may lead to better strategies for the prevention or treatment of AD. Besides elucidating sex differences, the next steps for phenotyping can include investigating race/ethnicity differences or differences based upon other covariates to better characterize Alzheimer's Disease heterogeneity. Furthermore, the incorporation of molecular or genetic data with clinical data can help better elucidate potential mechanisms underlying identified associations.

Methods

All analysis of UCSF and Mount Sinai EMR data was performed under the approval of respective Institutional Review Boards. All clinical data were de-identified and written informed consent was waived by the institutions.

In this study, we performed deep phenotyping and association analysis of patients with AD and controls. First, AD and control cohorts were identified from the UCSF EMR and topographically visualized via a low-dimensional projection of comorbidities. Comorbidity networks were created, and association and enrichment analyses were performed on all diagnoses, medications, and lab values. These analyses were further performed in a sex-stratified manner to identify sex-specific associations, and validation was performed on the Mount Sinai EMR. An overview of the workflow is shown in Fig. 1.

Patient cohort identification. Patient cohorts were identified from over five million patients in the UCSF EMR database, which includes clinical data from 1982 to 2020. Due to the de-identification process, dates are shifted by at most a year (with relative dates preserved) and all birth dates before 1930 (=estimated age 90) are shifted to be no earlier than 1930. Patients with AD were identified by inclusion criteria of estimated age >64 years, and ICD-10-CM codes G30.1, G30.8, or G30.9, where estimated age is determined from the birth date. Male and female groups were identified by the most recent sex assignment in the EMR. To identify a control group, we used propensity score (PS) matching method (matchit R package¹¹⁵) by a logistic regression model to match controls to patients with AD. The control group was selected from patients >64 years old without AD diagnosis, matched on sex, estimated age, race, and death status at a 1:2 AD:control ratio using a nearest neighbors method. The validation cohort was identified similarly in the Mount Sinai EMR database, which includes clinical data from 2003 to 2020. The demographic properties of the UCSF and Mount Sinai cohorts are shown in Table 1.

Dimensionality reduction patient visualization. All identified patients were represented using one-hot encoding of diagnoses, excluding encoding of diagnoses with Alzheimer's in the name (list in Supplementary Table 2 and Fig. 2). Patients were then visualized in a lower dimension using Uniform Manifold Approximation and Projection¹²² (UMAP) with the umap-learn package from Python. Correlations between variables and UMAP coordinates were analyzed using Mann-Whitney *U*-test for categorical variables, and Pearson's correlation coefficient for continuous variables.

AD vs. control enrichment analysis of comorbidities. To evaluate comorbidities, all diagnoses recorded from patient cohorts were identified with the earliest entry of every diagnosis. Comparisons were made at different ICD-10-CM hierarchical levels, specifically Level 2 categories (e.g., G30-G32: Other degenerative diseases of the nervous system), Level 3 categories (e.g., G30: Alzheimer's Disease), or full diagnosis names (e.g., G30.9 Alzheimer's disease, unspecified). Level 2, Level 3, and full diagnosis names are also grouped by ICD-10-CM blocks (e.g., G00-G99: Diseases of the Nervous System). More information on ICD-10-CM codes can be found here: <https://www.cms.gov/Medicare/Coding/ICD10/ICD-10Resources>.

Diagnosis networks were created based upon a diagnosis category or diagnosis shared by >1% patients in a group (node) or pair of diagnosis categories or diagnoses shared by >1% of patients in a group (edge). Network metrics were computed using Cytoscape app Network Analyzer¹²³. Metrics were then compared between AD and control networks using Mann-Whitney *U*-test, with and without singleton nodes removed. Nodes and edges were thresholded by 5% of patients in a group for visualization purposes.

Enrichment analysis of diagnosis was compared between AD and control cohorts. For each diagnosis, the proportions of patients in each group were compared using Fisher's exact (if <5 patients in a category) or Chi-squared test. Significant diagnoses were determined by a Bonferroni-corrected threshold of *p*-value < 0.05, and directionality determined with odds ratio (OR). With

inspiration from genetic and molecular approaches, the results were visualized using Manhattan plots by categorizing diagnoses in ICD-10-CM blocks.

Sex-stratified AD vs. control enrichment analysis of comorbidities. Diagnostic networks were created for each sex, with diagnosis categories or diagnoses shared by >1% of patients in a group (node), and diagnosis category/diagnosis pair shared by >1% of patients in a group (edge). Network metrics were then computed using Cytoscape Network Analyzer app, and compared between sex-stratified patients with AD and controls, and between males and females for both AD and control cohorts separately with a Mann–Whitney *U*-test. Nodes and edges were thresholded by 5% of patients in a group for visualization.

Sex-specific enrichment analysis of diagnoses between AD and control cohorts were compared with a subset of equal numbers of patients with AD and controls within each sex. For each diagnosis, the proportions of patients in each group were compared using the Fisher's exact (if <5 patients in a category) or Chi-squared test. Significance was determined by applying a threshold of 0.05 for Bonferroni-corrected *p*-values. Log-log plots were generated from odds ratios between females and males with AD and controls, and Miami plots were created by categorizing diagnoses in ICD-10-CM blocks.

Sensitivity analysis taking encounters into account. Sensitivity analysis of diagnosis enrichment analysis was performed with a subgroup of patients with AD and a second control cohort to account for variability in the number of visits for each patient. AD cohorts were subgrouped by identifying patients with over 10 encounters in the EMR and records spanning over a year. The encounter-filtered control cohort was identified by additionally matching the number of encounters and years between the first and last record in the EMR. Diagnosis enrichment analysis was carried out as described above for general comorbidities and sex-specific analysis.

AD vs. control enrichment analysis of medications. All medications ordered for patients with AD and controls were extracted and grouped based upon the generic medication name, with route and dosage information removed. The proportions of patients with AD and controls prescribed each medication were compared using Fisher's exact (if <5 patients in a category) or Chi-squared tests. Significantly enriched medications were identified by a Bonferroni-corrected threshold of *p*-value 0.05, and directionality was determined with an odds ratio. Sex-specific medication comparisons were also performed within a subset of equal numbers of patients with AD and controls for each sex and plotted with cutoffs based upon a Bonferroni-corrected *p*-value threshold of 0.05 and odds ratios threshold of <0.5 or >2.

AD vs. control comparisons of lab values. For laboratory values, median values for all numerical lab test results for each patient were identified. Lab tests missing data among 95% or more patients were removed. Lab value distributions were compared using Mann–Whitney *U*-test across three comparisons (AD vs. controls, females with AD vs. female controls, and males with AD vs. male controls) in order to identify significantly different lab values.

For clustering analysis, significant lab tests above a threshold of 0.05 for Bonferroni-corrected *p*-value were isolated, and mean values were then identified for each group (females with AD, males with AD, control females, control males) and normalized across groups as a Z-score. Clustering was then performed using the *sigclust2* R package¹²⁴ to determine the significance of each cluster break using permutations (Euclidean distance metric and average linkage).

Validation in external EMR. AD and PS-matched control patients were identified in the Mount Sinai EMR in the same fashion as described in [Patient Identification] in the UCSF EMR. All aforementioned analyses with dimensionality reduction, comorbidity networks, diagnosis/medication enrichments, sex-specific enrichments, and lab value comparisons were performed in the Mount Sinai data set as they have been in the UCSF EMR data set.

For network comparisons, network metrics were standard normalized across the 12 networks (6 at UCSF, 6 at Mount Sinai) by the metric and Spearman-rank correlation coefficient and significance determined. For diagnosis comparison, Level 2, Level 3, and full diagnosis names were mapped and compared by the sub-chapter, three-digit codes, and full code of the ICD-10-CM hierarchy, respectively. Significant diagnosis in the validation cohort was determined by a Bonferroni-corrected threshold of 0.05 based upon the number of mapped UCSF-significant diagnoses. Correlations between odds ratios were determined by a Spearman-rank correlation coefficient and significance. Medications were mapped based upon the generic name, and correlations between odds ratios determined with the Spearman-rank correlation coefficient.

For comparison of labs, the normalized lab values for each institution were combined, and clustering was performed using Euclidean distance and average linkage to identify groups of labs with similar trends between AD/sex/institution stratified patient groups. The R package *sigclust2* was used to determine significant clusters of labs.

Data visualization using RShiny. An interactive visualization of comorbidity enrichments and networks between AD and control groups and with sex stratification was implemented in an Rshiny¹²⁵ app: vizad.org.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The UCSF EHR database is available to individuals affiliated with UCSF who can contact the UCSF's Clinical and Translational Science Institute (CTSI) (ctsi@ucsf.edu) or the UCSF's Information Commons team for more information (Info.Commons@ucsf.edu). The Mount Sinai EHR database is available to individuals affiliated with Mount Sinai who can contact the Mount Sinai Intellectual Partners (MSIP) for more information (MSIPInfo@mssm.edu). If the reader is not affiliated with the aforementioned institutions, they can set up an official collaboration with an investigator affiliated with the target institution(s) by contacting the PIs Marina Sirota (marina.sirota@ucsf.edu) and Benjamin Glicksberg (benjamin.glicksberg@mssm.edu). Requests should be processed within a couple of weeks. Summary data is available in supplementary files and, for UCSF, can be explored at <https://vizad.org>.

Code availability

The code is available at https://github.com/al1563/adehr_phenotype.

Received: 15 April 2021; Accepted: 18 January 2022;

Published online: 03 February 2022

References

1. Alzheimer's Association. 2020 Alzheimer's disease facts and figures. *Alzheimers Dement.* **16**, 391–460 (2020).
2. Ferreira, D., Wahlund, L.-O. & Westman, E. The heterogeneity within Alzheimer's disease. *Aging* **10**, 3058–3060 (2018).
3. Neu, S. C. et al. Apolipoprotein E genotype and sex risk factors for Alzheimer disease: a meta-analysis. *JAMA Neurol.* **74**, 1178 (2017).
4. Cognitive Function and Ageing Studies (CFAS) Collaboration et al. A two decade dementia incidence comparison from the Cognitive Function and Ageing Studies I and II. *Nat. Commun.* **7**, 11398 (2016).
5. Dubal, D. B. in *Handbook of Clinical Neurology* Vol. 175 (eds Lanzenberger, R., Kranz, G. S. & Savic, I.) Ch.16, 261–273 (Elsevier, 2020).
6. Davis, E. J. et al. A second X chromosome contributes to resilience in a mouse model of Alzheimer's disease. *Sci. Transl. Med.* **12**, eaaz5677 (2020).
7. Ossenkoppe, R. et al. Assessment of demographic, genetic, and imaging variables associated with brain resilience and cognitive resilience to pathological tau in patients with Alzheimer disease. *JAMA Neurol.* **77**, 632 (2020).
8. Digma, L. A. et al. Women can bear a bigger burden: ante- and post-mortem evidence for reserve in the face of tau. *Brain Commun.* **2**, fcaa025 (2020).
9. Nebel, R. A. et al. Understanding the impact of sex and gender in Alzheimer's disease: a call to action. *Alzheimers Dement. J. Alzheimers Assoc.* **14**, 1171–1183 (2018).
10. Gilsanz, P. et al. Female sex, early-onset hypertension, and risk of dementia. *Neurology* **89**, 1886–1893 (2017).
11. Fan, C. C. et al. Sex-dependent autosomal effects on clinical progression of Alzheimer's disease. *Brain* **143**, 2272–2280 (2020).
12. Arnold, M. et al. Sex and APOE ε4 genotype modify the Alzheimer's disease serum metabolome. *Nat. Commun.* **11**, 1–12 (2020).
13. Zhao, N. et al. Alzheimer's risk factors age, APOE genotype, and sex drive distinct molecular pathways. *Neuron* **106**, 727–742.e6 (2020).
14. Paranjpe, M. D. et al. Sex-specific cross tissue meta-analysis identifies immune dysregulation in women with Alzheimer's disease. *Front. Aging Neurosci.* **13**, 735611 (2021).
15. Belonwu, S. A. et al. Sex-stratified single-cell rna-seq analysis identifies sex-specific and cell type-specific transcriptional responses in Alzheimer's disease across two brain regions. *Mol. Neurobiol.* <https://doi.org/10.1007/s12035-021-02591-8> (2021).
16. Ou Ya-Nan et al. Blood pressure and risks of cognitive impairment and dementia. *Hypertension* **76**, 217–225 (2020).
17. Nucera, A. & Hachinski, V. Cerebrovascular and Alzheimer disease: fellow travelers or partners in crime? *J. Neurochem.* **144**, 513–516 (2018).
18. Santiago, J. A., Bottero, V. & Potashkin, J. A. Transcriptomic and network analysis highlight the association of diabetes at different stages of Alzheimer's disease. *Front. Neurosci.* **13**, 1273 (2019).
19. Pugazhenth, S., Qin, L. & Reddy, P. H. Common neurodegenerative pathways in obesity, diabetes, and Alzheimer's disease. *Biochim. Biophys. Acta* **1863**, 1037–1045 (2017).

20. Duthie, A., Chew, D. & Soiza, R. L. Non-psychiatric comorbidity associated with Alzheimer's disease. *QJM Mon.* **104**, 913–920 (2011).
21. Santiago, J. A. & Potashkin, J. A. The impact of disease comorbidities in Alzheimer's disease. *Front. Aging Neurosci.* **13**, 631770 (2021).
22. Liao, J.-Y., Lee, C. T.-C., Lin, T.-Y. & Liu, C.-M. Exploring prior diseases associated with incident late-onset Alzheimer's disease dementia. *PLoS ONE* **15**, e0228172 (2020).
23. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
24. Nazarian, A., Yashin, A. I. & Kulminski, A. M. Genome-wide analysis of genetic predisposition to Alzheimer's disease and related sex disparities. *Alzheimers Res. Ther.* **11**, 5 (2019).
25. Chen, W.-T. et al. Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182**, 976–991.e19 (2020).
26. Qorri, B., Tsay, M., Agrawal, A., Au, R. & Gracie, J. Using machine intelligence to uncover Alzheimer's disease progression heterogeneity. *Explor. Med.* **1**, 100126 (2020).
27. Davis, E. J. et al. Sex-specific association of the X chromosome with cognitive change and tau pathology in aging and Alzheimer disease. *JAMA Neurol.* <https://doi.org/10.1001/jamaneurol.2021.2806> (2021).
28. Alzheimer's Disease Neuroimaging Initiative et al. Multimodal phenotyping of Alzheimer's disease with longitudinal magnetic resonance imaging and cognitive function data. *Sci. Rep.* **10**, 5527 (2020).
29. Vardy, E. R. L. C. et al. Cognitive phenotypes in Alzheimer's disease and genetic variants in ACE and IDE. *Neurobiol. Aging* **33**, 1486.e1 (2012).
30. Jaakkimainen, R. L. et al. Identification of physician-diagnosed Alzheimer's disease and related dementias in population-based administrative data: a validation study using family physicians' electronic medical records. *J. Alzheimers Dis.* **54**, 337–349 (2016).
31. The Office of the National Coordinator for Health Information Technology (ONC) & Office of Secretary, United States Department of Health and Human Services. *2016 Report to Congress on Health IT Progress: Examining the HITECH Era and the Future of Health IT* (2016).
32. Glicksberg, B. S. et al. Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks. *Bioinformatics* **32**, i101–i110 (2016).
33. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
34. Abraham, A. et al. Dense phenotyping from electronic health records enables machine-learning-based prediction of preterm birth. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.07.15.20154864> (2020).
35. Norgeot, B. et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Netw. Open* **2**, e190606 (2019).
36. Zhang, R., Simon, G. & Yu, F. Advancing Alzheimer's research: a review of big data promises. *Int. J. Med. Inf.* **106**, 48–56 (2017).
37. Delude, C. M. Deep phenotyping: the details of disease. *Nature* **527**, S14–S15 (2015).
38. Weng, C., Shah, N. H. & Hripcsak, G. Deep phenotyping: Embracing complexity and temporality—towards scalability, portability, and interoperability. *J. Biomed. Inform.* **105**, 103433 (2020).
39. Grande, G. et al. Multimorbidity burden and dementia risk in older adults: the role of inflammation and genetics. *Alzheimers Dement.* <https://doi.org/10.1002/alz.12237> (2021).
40. Vassilaki, M. et al. Multimorbidity and neuroimaging biomarkers among cognitively normal persons. *Neurology* **86**, 2077–2084 (2016).
41. Jellinger, K. A. & Attems, J. Challenges of multimorbidity of the aging brain: a critical update. *J. Neural Transm.* **122**, 505–521 (2015).
42. Hörder, H. et al. Midlife cardiovascular fitness and dementia: a 44-year longitudinal population study in women. *Neurology* **90**, e1298–e1305 (2018).
43. Carlsson, C. M. Type 2 diabetes mellitus, dyslipidemia, and Alzheimer's disease. *J. Alzheimers Dis.* **20**, 711–722 (2010).
44. Jeong, S.-M. et al. Anemia is associated with incidence of dementia: a national health screening study in Korea involving 37,900 persons. *Alzheimers Res. Ther.* **9**, 94 (2017).
45. Hong, C. H. et al. Anemia and risk of dementia in older adults: findings from the Health ABC study. *Neurology* **81**, 528–533 (2013).
46. Goldstein, F. C. et al. Effects of hypertension and hypercholesterolemia on cognitive functioning in patients with Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* **22**, 336–342 (2008).
47. Chen, Y.-H. & Lo, R. Y. Alzheimer's disease and osteoporosis. *Ci Ji Yi Xue Za Zhi Tzu-Chi Med. J.* **29**, 138–142 (2017).
48. Lv, X.-L. et al. Association between osteoporosis, bone mineral density levels and Alzheimer's disease: a systematic review and meta-analysis. *Int. J. Gerontol.* **12**, 76–83 (2018).
49. Chiang, C.-H. et al. Lower urinary tract symptoms are associated with increased risk of dementia among the elderly: a nationwide study. *BioMed Res. Int.* **2015**, 187819 (2015).
50. de la Torre, J. C. Alzheimer disease as a vascular disorder: nosological evidence. *Stroke* **33**, 1152–1162 (2002).
51. Attems, J. & Jellinger, K. A. The overlap between vascular disease and Alzheimer's disease - lessons from pathology. *BMC Med.* **12**, 206 (2014).
52. Rius-Pérez, S., Tormos, A. M., Pérez, S. & Taléns-Visconti, R. Vascular pathology: cause or effect in Alzheimer disease? *Neurol. Barc. Spain* **33**, 112–120 (2018).
53. Reitz, C. Dyslipidemia and the risk of Alzheimer's disease. *Curr. Atheroscler. Rep.* **15**, 307 (2013).
54. de la Monte, S. M. & Wands, J. R. Alzheimer's disease is type 3 diabetes-evidence reviewed. *J. Diabetes Sci. Technol.* **2**, 1101–1113 (2008).
55. Kandimalla, R., Thirumala, V. & Reddy, P. H. Is Alzheimer's disease a Type 3 diabetes? A critical appraisal. *Biochim. Biophys. Acta* **1863**, 1078–1089 (2017).
56. Sun, Y. et al. Metabolism: a novel shared link between diabetes mellitus and Alzheimer's disease. *J. Diabetes Res.* **2020**, 1–12 (2020).
57. Mushtaq, G., Khan, J. A., Kumosani, T. A. & Kamal, M. A. Alzheimer's disease and type 2 diabetes via chronic inflammatory mechanisms. *Saudi J. Biol. Sci.* **22**, 4–13 (2015).
58. Deleidi, M., Jäggle, M. & Rubino, G. Immune aging, dysmetabolism, and inflammation in neurological diseases. *Front. Neurosci.* **9**, 172 (2015).
59. Lue, L.-F., Andrade, C., Sabbagh, M. & Walker, D. Is there inflammatory synergy in type II diabetes mellitus and Alzheimer's disease? *Int. J. Alzheimers Dis.* **2012**, 1–9 (2012).
60. Bowman, G. L., Kaye, J. A. & Quinn, J. F. Dyslipidemia and blood-brain barrier integrity in Alzheimer's disease. *Curr. Gerontol. Geriatr. Res.* **2012**, 1–5 (2012).
61. Goldwaser, E. L., Acharya, N. K., Sarkar, A., Godsey, G. & Nagele, R. G. Breakdown of the cerebrovasculature and blood-brain barrier: a mechanistic link between diabetes mellitus and Alzheimer's disease. *J. Alzheimers Dis.* **54**, 445–456 (2016).
62. Nelson, A. R., Sweeney, M. D., Sagare, A. P. & Zlokovic, B. V. Neurovascular dysfunction and neurodegeneration in dementia and Alzheimer's disease. *Biochim. Biophys. Acta* **1862**, 887–900 (2016).
63. Borda, M. G. et al. Frailty in older adults with mild dementia: dementia with Lewy bodies and Alzheimer's disease. *Dement. Geriatr. Cogn. Disord. Extra* **9**, 176–183 (2019).
64. Buchman, A. S., Schneider, J. A., Leurgans, S. & Bennett, D. A. Physical frailty in older persons is associated with Alzheimer disease pathology. *Neurology* **71**, 499–504 (2008).
65. Wallace, L. M. K. et al. Investigation of frailty as a moderator of the relationship between neuropathology and dementia in Alzheimer's disease: a cross-sectional analysis of data from the Rush Memory and Aging Project. *Lancet Neurol.* **18**, 177–184 (2019).
66. Lanni, C., Masi, M., Racchi, M. & Govoni, S. Cancer and Alzheimer's disease inverse relationship: an age-associated diverging derailment of shared pathways. *Mol. Psychiatry* **26**, 280–295 (2021).
67. Okereke, O. I. & Meadows, M.-E. More evidence of an inverse association between cancer and Alzheimer disease. *JAMA Netw. Open* **2**, e196167 (2019).
68. Behrens, M. I., Lendon, C. & Roe, C. M. A common biological mechanism in cancer and Alzheimer's disease? *Curr. Alzheimer Res.* **6**, 196–204 (2009).
69. Majd, S., Power, J. & Majd, Z. Alzheimer's disease and cancer: when two monsters cannot be together. *Front. Neurosci.* **13**, 155 (2019).
70. Goldstein, J. M., Langer, A. & Lesser, J. A. Sex differences in disorders of the brain and heart—a global crisis of multimorbidity and novel opportunity. *JAMA Psychiatry* **78**, 7 (2021).
71. Osler, M. et al. Hearing loss, cognitive ability, and dementia in men age 19–78 years. *Eur. J. Epidemiol.* **34**, 125–130 (2019).
72. Ford, A. H. et al. Hearing loss and the risk of dementia in later life. *Maturitas* **112**, 1–11 (2018).
73. Curhan, S. G., Willett, W. C., Grodstein, F. & Curhan, G. C. Longitudinal study of hearing loss and subjective cognitive function decline in men. *Alzheimers Dement.* **15**, 525–533 (2019).
74. Huang, B. et al. Gender differences in the association between hearing loss and cognitive function. *Am. J. Alzheimers Dis. Dement.* **35**, 153331751987116 (2020).
75. Kitamura, T., Kitamura, M., Hino, S., Tanaka, N. & Kurata, K. Gender differences in clinical manifestations and outcomes among hospitalized patients with behavioral and psychological symptoms of dementia. *J. Clin. Psychiatry* **73**, 1548–1554 (2012).
76. Resnick, B. et al. Gender differences in presentation and management of behavioral and psychological symptoms associated with dementia among nursing home residents with moderate to severe dementia. *J. Women Aging* **1–18**, <https://doi.org/10.1080/08952841.2020.1735925> (2020).

77. Dufouil, C., Seshadri, S. & Chêne, G. Cardiovascular risk profile in women and dementia. *J. Alzheimers Dis.* **42**, S353–S363 (2014).
78. Pajak, A., Kawalec, E. & Szczudlik, A. [Cognitive impairment and cardiovascular disease risk factors. Project CASCADE Kraków. I. Project to test exposure to risk factors for cardiovascular disease in the studied sample]. *Przegl. Lek.* **55**, 676–682 (1998).
79. Haring, B. et al. Cardiovascular disease and cognitive decline in postmenopausal women: results from the Women's Health Initiative Memory Study. *J. Am. Heart Assoc.* **2**, e000369 (2013).
80. Elbejjani, M. et al. Depression, depressive symptoms, and rate of hippocampal atrophy in a longitudinal cohort of older men and women. *Psychol. Med.* **45**, 1931–1944 (2015).
81. Lee, J., Lee, K. J. & Kim, H. Gender differences in behavioral and psychological symptoms of patients with Alzheimer's disease. *Asian J. Psychiatry* **26**, 124–128 (2017).
82. Goveas, J. S., Espeland, M. A., Woods, N. F., Wassertheil-Smoller, S. & Kotchen, J. M. Depressive symptoms and incidence of mild cognitive impairment and probable dementia in elderly women: The Women's Health Initiative Memory Study: depression and incident MCI and dementia. *J. Am. Geriatr. Soc.* **59**, 57–66 (2011).
83. Chong, J. S. X. et al. Influence of cerebrovascular disease on brain networks in prodromal and clinical Alzheimer's disease. *Brain* **140**, 3012–3022 (2017).
84. Vipin, A. et al. Cerebrovascular disease influences functional and structural network connectivity in patients with amnesic mild cognitive impairment and Alzheimer's disease. *Alzheimers Res. Ther.* **10**, 82 (2018).
85. Dublin, S. et al. Prescription opioids and risk of dementia or cognitive decline: a prospective cohort study. *J. Am. Geriatr. Soc.* **63**, 1519–1526 (2015).
86. Hamina, A. et al. Differences in analgesic use in community-dwelling persons with and without Alzheimer's disease. *Eur. J. Pain.* **21**, 658–667 (2017).
87. Cai, Z. & Ratka, A. Opioid system and Alzheimer's disease. *NeuroMol. Med.* **14**, 91–111 (2012).
88. Alisky, J. Intrathecal corticosteroids might slow Alzheimer's disease progression. *Neuropsychiatr. Dis. Treat.* **4**, 831–833 (2008).
89. Beeri, M. S. et al. Corticosteroids, but not NSAIDs, are associated with less Alzheimer neuropathology. *Neurobiol. Aging* **33**, 1258–1264 (2012).
90. Hui, Z. et al. The combination of acyclovir and dexamethasone protects against Alzheimer's disease-related cognitive impairments in mice. *Psychopharmacology* **237**, 1851–1860 (2020).
91. Murialdo, G. et al. Dexamethasone effects on cortisol secretion in Alzheimer's disease: Some clinical and hormonal features in suppressor and nonsuppressor patients. *J. Endocrinol. Invest.* **23**, 178–186 (2000).
92. Belanoff, J. K., Gross, K., Yager, A. & Schatzberg, A. F. Corticosteroids and cognition. *J. Psychiatr. Res.* **35**, 127–145 (2001).
93. Sahar, A. & Tsipstein, E. Effects of mannitol and furosemide on the rate of formation of cerebrospinal fluid. *Exp. Neurol.* **60**, 584–591 (1978).
94. Chuang, Y.-F. et al. Use of diuretics is associated with reduced risk of Alzheimer's disease: the Cache County Study. *Neurobiol. Aging* **35**, 2429–2435 (2014).
95. Tully, P. J., Hanon, O., Cosh, S. & Tzourio, C. Diuretic antihypertensive drugs and incident dementia risk: a systematic review, meta-analysis and meta-regression of prospective studies. *J. Hypertens.* **34**, 1027–1035 (2016).
96. Wang, J. et al. Unintended effects of cardiovascular drugs on the pathogenesis of Alzheimer's disease. *PLoS ONE* **8**, e65232 (2013).
97. Taubes, A. et al. Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nat. Aging* **1**, 932–947 (2021).
98. Wang, Z., Vilekar, P., Huang, J. & Weaver, D. F. Furosemide as a probe molecule for the treatment of neuroinflammation in Alzheimer's disease. *ACS Chem. Neurosci.* **11**, 4152–4168 (2020).
99. Wang, Y., Shi, Y. & Wei, H. Calcium dysregulation in Alzheimer's disease: a target for new drug development. *J. Alzheimers Dis. Park.* **7**, 374 (2017).
100. Tong, B. C.-K., Wu, A. J., Li, M. & Cheung, K.-H. Calcium signaling in Alzheimer's disease & therapies. *Biochim. Biophys. Acta* **1865**, 1745–1760 (2018).
101. Crane, P. K. et al. Glucose levels and risk of dementia. *N. Engl. J. Med.* **369**, 540–548 (2013).
102. Nho, K. et al. Association of altered liver enzymes with alzheimer disease diagnosis, cognition, neuroimaging measures, and cerebrospinal fluid biomarkers. *JAMA Netw. Open* **2**, e197978 (2019).
103. An, Y. et al. Evidence for brain glucose dysregulation in Alzheimer's disease. *Alzheimers Dement.* **14**, 318–329 (2018).
104. Iadecola, C. Sugar and Alzheimer's disease: a bittersweet truth. *Nat. Neurosci.* **18**, 477–478 (2015).
105. Bassendine, M. F., Taylor-Robinson, S. D., Fertleman, M., Khan, M. & Neely, D. Is Alzheimer's disease a liver disease of the brain? *J. Alzheimers Dis.* **75**, 1–14 (2020).
106. Wang, J., Gu, B. J., Masters, C. L. & Wang, Y.-J. A systemic view of Alzheimer disease — insights from amyloid- β metabolism beyond the brain. *Nat. Rev. Neurol.* **13**, 612–623 (2017).
107. Akiyama, H., Ikeda, K., Kondo, H. & McGeer, P. L. Thrombin accumulation in brains of patients with Alzheimer's disease. *Neurosci. Lett.* **146**, 152–154 (1992).
108. Iannucci, J., Renahan, W. & Grammas, P. Thrombin, a mediator of coagulation, inflammation, and neurotoxicity at the neurovascular interface: implications for Alzheimer's disease. *Front. Neurosci.* **14**, 762 (2020).
109. Mari, D. et al. Hemostasis abnormalities in patients with vascular dementia and Alzheimer's disease. *Thromb. Haemost.* **75**, 216–218 (1996).
110. Gupta, A. et al. Coagulation and inflammatory markers in Alzheimer's and vascular dementia: Alzheimer's and vascular dementia. *Int. J. Clin. Pract.* **59**, 52–57 (2004).
111. Stellos, K. et al. Predictive value of platelet activation for the rate of cognitive decline in Alzheimer's disease patients. *J. Cereb. Blood Flow. Metab.* **30**, 1817–1820 (2010).
112. Sevush, S. et al. Platelet activation in Alzheimer disease. *Arch. Neurol.* **55**, 530 (1998).
113. Gowert, N. S. et al. Blood platelets in the progression of Alzheimer's disease. *PLoS ONE* **9**, e90523 (2014).
114. Strickland, S. Impact of the coagulation system on the pathogenesis of Alzheimer's disease. *Blood* **130**, SCI-3 (2017).
115. Merlini, M. et al. Fibrinogen induces microglia-mediated spine elimination and cognitive impairment in an Alzheimer's disease model. *Neuron* **101**, 1099–1108.e6 (2019).
116. Klohs, J. An integrated view on vascular dysfunction in Alzheimer's disease. *Neurodegener. Dis.* **19**, 109–127 (2019).
117. For the KBASE Research Group. et al. Associations of thyroid hormone serum levels with in-vivo Alzheimer's disease pathologies. *Alzheimers Res. Ther.* **9**, 64 (2017).
118. Choi, B. W., Kang, S. & Kim, H. W. Relationship between serum TSH level and Alzheimer disease pathology: Human neuropathology/clinico-pathologic correlations. *Alzheimers Dement.* **16**, e041210 (2020).
119. Choi, B. W. et al. Relationship between thyroid hormone levels and the pathology of Alzheimer's disease in euthyroid subjects. *Thyroid* **30**, 1547–1555 (2020).
120. Brenowitz, W. D. et al. Mixed neuropathologies and estimated rates of clinical progression in a large autopsy sample. *Alzheimers Dement.* **13**, 654–662 (2017).
121. Jørgensen, I. F., Aguayo-Orozco, A., Lademann, M. & Brunak, S. Age-stratified longitudinal study of Alzheimer's and vascular dementia patients. *Alzheimers Dement.* **16**, 908–917 (2020).
122. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at <https://arxiv.org/abs/1802.03426> (2020).
123. Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282–284 (2008).
124. Kimes, P. K., Liu, Y., Neil Hayes, D. & Marron, J. S. Statistical significance for hierarchical clustering. *Biometrics* **73**, 811–821 (2017).
125. Rstudio, Inc. *Shiny: Easy Web Applications in R* (2014).

Acknowledgements

Primary support through Grant # NIA R01AG060393, R01AG057683 (A.T., T.O., C.W.S., M.S.). Additional support was provided by NIA RF1AG068325 (D.B.D) and Medical Scientist Training Program T32GM007618 (A.T.). B.G. and M.B. are supported by grant 1 RF1 AG059319-01. We'd like to acknowledge Zachary Cutts, Stella Belonwu, and other members of the Sirota Lab for their suggestions and help.

Author contributions

A.T. and M.S. designed the question, experiments, and analytic plan. B.Z. and Z.H. helped with data acquisition, cleaning, and interpretation. A.T., C.W.S. and B.O. helped with creation of Rshiny app. A.T., W.M., T.O., C.W.S. and D.D. interpreted results. J.H., M.B. and B.G. helped acquire and analyze validation data. S.W. and I.A. aided in statistical methods. A.T. wrote the manuscript with editing from all the authors. All the authors edited and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-28273-0>.

Correspondence and requests for materials should be addressed to Alice S. Tang or Marina Sirota.

Peer review information *Nature Communications* thanks Mohammad Ali Moni, Chun Chieh Fan and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022