Leveraging electronic health records to examine differential clinical outcomes in people with Alzheimer's Disease

Shruthi Venkatesh, BS[1] *, Linshanshan Wang, MS[2,3] *, Michele Morris, BS[4], Mohammed Moro, MS[2], Ratnam Srivastava, MS[2], Yunqing Han, MS[5], Riddhi Patira, MD[1], Sarah Berman, MD, PhD[1], Oscar Lopez, MD[1], Shyam Visweswaran, MD, PhD[4], Tianrun Cai, MD[5], Tianxi Cai, ScD[2,3] #, Zongqi Xia, MD, PhD[1] #

1. Department of Neurology, University of Pittsburgh, Pittsburgh, PA
2. Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA
3. Department of Biomedical Informatics, Harvard Medical School, Boston, MA
4. Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
5. Department of Rheumatology, Brigham and Women's Hospital, Boston, MA

*** Share co-first authorship**

 # Share co-senior authorship

Word Count: 2959

Corresponding Authors:

Zongqi Xia

3501 Fifth Ave, BST3-10.044, Pittsburgh, PA 15260

412-383-5377; zxia1@post.harvard.edu

Tianxi Cai

10 Shattuck Street, Room 434, Boston, MA 02115

617-432-4923; tcai@hsph.harvard.edu

ABSTRACT

BACKGROUND. Alzheimer's disease (AD) carries a high societal burden inequitably distributed across demographic groups.

OBJECTIVE. To examine demographic differences and drivers of AD decline using real-world electronic health record (EHR) data with accurate AD population identification.

METHODS. Leveraging EHR data (1994-2022) from two large independent healthcare systems, we applied a novel unsupervised phenotyping algorithm to predict AD diagnosis and validated using gold-standard chart-reviewed and registry-derived diagnosis labels. Among patients with ≥24 months of EHR data not living in nursing homes pre-AD diagnosis, we performed healthcare system-specific competing risk survival analyses to estimate the time to two readily ascertainable AD decline outcomes (*i.e.,* nursing home admission, death), stratified by demographic groups and accounting for baseline covariates (*e.g.,* age, gender, race, ethnicity, healthcare utilization, and comorbidities). We then performed a covariate-adjusted fixed-effects meta-analysis using inverse variance weighting to estimate the time to AD decline stratified by demographic groups.

RESULTS. The algorithm achieved robust prediction in identifying AD patients across both healthcare systems (AUROC score range: 0.835-0.923) and demographic groups. Of the 29,262 AD patients in both healthcare systems (61% women, 90% non-Hispanic White, 79.52±9.39 years of age at AD diagnosis), 49% entered nursing homes and 48% died during follow-up. In covariate-adjusted fixed-effects meta-analysis, women had a higher risk of nursing home admission (HR[95% CI]=1.061[1.024-1.100], p=.001) and lower risk of death (HR[95% CI]=0.856[0.811-0.904], p<.0001) than men. Non-Hispanic White individuals had similar nursing home risk (HR[95% CI]=1.006[0.952-1.063],

p=0.831) but higher death risk (HR[95% CI]=1.376[1.245-1.521], *p*<.0001) than racial and ethnic minorities. Older age at AD diagnosis and greater pre-existing comorbidity burden increased both nursing home admission and death risk.

CONCLUSION. Findings from two large EHR cohorts add to the real-world evidence of demographic differences in clinical AD decline and its drivers, which could potentially inform individual clinical management and future public health policies.

INTRODUCTION

Late-onset Alzheimer's disease (AD) is the leading cause of dementia and neurological disability affecting ~7.2 million people in the United States.[1–3] People with AD experience variable rates of decline with differences among demographic groups. The likelihood of nursing home admission, indicating decline in capacity for independent living, is higher among women and racial and ethnic minorities with AD.[4–7] In contrast, post-diagnosis survival is shorter for men and non-Hispanic White (NHW) individuals.[5,8–10]

Prior epidemiological or claims-based studies implicated sociodemographic factors and comorbidities as potential drivers of differential AD decline.[11–16] Comprehensive assessment of these modifiable factors and quantifying the extent to which they drive AD decline using large-scale *real-world* clinical data can augment prior research and enhance generalizability of the findings. Specifically, electronic health record (EHR) data containing the longitudinal clinical profiles of demographically broad patient populations could complement traditional data sources (*e.g.,* epidemiological studies that are costly and time-consuming to conduct, claims data that lack clinical granularity) given the growing adoption and standardization of EHR. Notably, EHR data have advantages over claims data (*e.g.,* Medicare) to examine differential AD decline. First, unstructured narrative data (*e.g.,* clinical narratives that contextualize diagnosis, treatment, and prognosis) adds significant value beyond structured codified data (*e.g.,* diagnoses and prescriptions). Second, EHR data can be more readily linked to local registry data (*e.g.,* Alzheimer's Disease Research Center) that provide the critical gold-standard labels for developing algorithms applicable at the point of care and for scaling to examine non-registry patients

in the EHR. Third, EHR data can undergo the crucial local quality control (*e.g.,* chart review), which would typically be infeasible for claims data.

Despite the advantages of EHR, accurate identification of target populations (*e.g.,* AD) from the EHR has limited the effective use of EHR data for examining drivers of differential AD decline (**S-Introduction-1**). While rule-based approaches (*e.g.,* identifying cases using diagnosis codes, diagnostic tests, prescriptions or the combination thereof) have achieved reasonable performance in identifying AD populations in claims data, gold-standard ground truth is rarely available with rare exceptions.[17–21] On the other hand, heterogeneity in diagnosis coding and clinical documentation leads to suboptimal performance of rule-based approaches in accurately identifying AD populations in the EHR.[22–25] Limitations of existing approaches to identify AD populations from the EHR include sparse gold-standard labels to validate AD diagnosis status (particularly via external validation) and the lack of validation across demographic groups despite known differences in clinical presentation.[25–36]

Here, we re-examined the demographic differences and drivers of AD decline using real-world EHR data with accurate AD population identification. We leveraged the EHR data of two independent healthcare systems, one of which was linked to an AD registry, applied a knowledge graph-guided unsupervised phenotyping algorithm to predict AD diagnosis across demographic groups, and validated algorithm performance using chart-reviewed and registry-derived gold-standard labels. Using these large real-world AD-EHR cohorts, we then examined differences in two *readily ascertainable* clinical outcomes of AD decline

(*i.e.,* nursing home admission, mortality) between demographic groups (*e.g.,* race and ethnicity, gender) and explored potential drivers of differential AD decline.

METHODS

Ethics Approval

The Institutional Review Board of the University of Pittsburgh (STUDY21020153) and Mass General Brigham (MGB, 2023P001450) approved the study protocol. The study using de-identified data was deemed exempt.

Harmonizing EHR data and creating AD-EHR data-marts

We obtained inpatient and outpatient EHR data from two healthcare systems, UPMC (Pittsburgh, PA; codified data: 2004-2022; narrative data: 2011-2022) and MGB (Boston, MA; codified and narrative data: 1994-2022). Codified data contained demographics, diagnoses (*e.g.*, International Classification of Disease [ICD] code), procedures (*e.g.*, Current Procedural Terminology [CPT] code), healthcare utilization metrics (*e.g.*, total number of ICD codes and encounter notes), medication prescriptions (*e.g.*, RxNorm codes), and laboratory tests (*e.g.*, Logical Observation Identifiers Names and Codes [LOINC] codes).[37,38] We organized diagnoses by mapping ICD-9 and ICD-10 codes to PheCodes,[39,40] consolidated procedures using the Clinical Classifications Software for Services and Procedures (Agency for Healthcare Research and Quality), and mapped electronic prescriptions to RxNorm ingredient level. We deployed an established natural language processing pipeline (Narrative Information Linear Extraction [NILE]) to derive concept unique identifiers (CUIs) from clinical narratives (*e.g.,* physician office notes, discharge summaries) according to unified medical language system (UMLS).[41,42] Aiming

high sensitivity to start, we built the *initial* **AD-EHR data-marts**, comprising UPMC (2011-2022) and MGB (1994-2022) patients with ≥1 ICD code for AD or related dementia (*e.g.,* ICD-9=290.x, 294.2x, 331.0; ICD-10=F03.9x, G30.x; **S-Table 1, S-Method-1**). We imputed missing race and ethnicity using age and gender.[43,44] We performed sensitivity analyses by excluding patients from the initial AD-EHR data-marts with missing race and ethnicity information (**S-Method-2, S-Figure 1**).

Inclusion Criteria

We applied a novel knowledge graph guided unsupervised phenotyping algorithm, Knowledge driven Online Multimodal Automated Phenotyping (**KOMAP**), to assign AD diagnosis (*i.e.,* probable or possible AD vs *not* AD) for all patients in the initial AD-EHR data marts (**Figure 1A, S-Table 2**).[45] We validated KOMAP-predicted AD diagnosis status using gold-standard chart-reviewed labels (UPMC, n=200; MGB, n=100, **S-Method-2**) and registry labels (University of Pittsburgh Alzheimer's Disease Research Registry [ADRC], n=1916 with UPMC EHR linkage) across demographic groups. We classified individuals from AD-EHR data-marts with KOMAP-predicted AD diagnosis at 90% specificity as having AD (**Figure 2A**, **S-Method-2**). We held the specificity at 90% to minimize false positives while keeping reasonably large sample sizes. The first AD PheCode (*i.e.,* PheCode 290.11) occurrence was the index date or *baseline*. For downstream analyses, we excluded patients with <24 months of EHR data or admitted to nursing homes (*i.e.,* with ≥1 diagnosis code ***or*** CUI for nursing home admission) prior to the baseline.

Covariates

We accounted for covariates, including demographics (*e.g.*, age at AD diagnosis, gender, race, ethnicity) and *baseline* clinical profiles (*e.g.,* healthcare utilization, pre-existing comorbidity burden) in the 24 months preceding the index date. Age at AD diagnosis was operationally defined as the age at the first AD PheCode. Self-reported gender included men and women. Race included American Indian or Alaska Native, Asian, Black, Native Hawaiian or Pacific Islander, White, and others. Ethnicity included Hispanic or Latino and non-Hispanic. We dichotomized race and ethnicity as non-Hispanic White (*i.e.,* NHW) versus other minorities, which included individuals of Hispanic and non-European descent. Baseline healthcare utilization was the annualized total number of diagnosis codes and clinical encounters.[46] Baseline comorbidity burden was the Elixhauser comorbidity index (ECI).[47] Using the R *comorbidity* package, we mapped diagnosis codes to the 29 pre-existing health conditions comprising the ECI.[48] We considered each comorbidity as an individual covariate in the main analysis, and performed a sensitivity analysis using the ECI score (**S-Method-2**).[49] We included the most common and consistently available covariates in the EHR (*e.g.,* demographics, comorbidities, healthcare utilization) relevant to AD to balance clinical relevance and data completeness. While incorporating other factors (*e.g.,* social determinants of health, blood or neuroimaging biomarkers) may reduce residual confounding, they are currently not readily ascertainable from the EHR for most patients.

Clinical outcomes of AD decline

We assessed two clinically relevant indicators of AD decline *readily ascertainable* from EHR: (1) time to nursing home admission and (2) time to death. Nursing home admission is routinely documented in clinical practice.[6,7] We defined nursing home admission as having ≥1 code *or* CUI for admission to or clinical encounters in any type of residential institution (*e.g.,* skilled nursing facility, long-term care facility, end-of-life care facility, **S-Table 3**), validated with gold-standard chart-reviewed labels (**S-Method-3**). Time to nursing home admission was from the index date to the first occurrence of any nursing home code or mention. Death status in EHR was linked to the social security death index. Time to death was from the index date to death. For patients without reaching either endpoint, we used all available EHR data.

Statistical analysis

We used two-tailed t-tests or two-proportion Z-tests to compare differences in demographics and baseline clinical profiles (**S-Table 4**) and prevalence of AD clinical outcomes among demographic groups (**S-Method-4**). We estimated the time-to-outcome using healthcare system-specific covariate-adjusted Cox proportional hazards (PH) models that accounted for the competing risks of nursing home admission and death, stratified by demographic groups (**Figure 2B**). Using patient-level data from both sites, we performed a fixed-effects meta-analysis using inverse variance weighting to estimate the time-to-outcome stratified by demographic groups, adjusting for pooled covariates (**Figure 2C)**. (Random-effects models would be unstable with two sites).

Data and code availability

Codes for KOMAP and project analysis are available on Github (**S-Method-5**).[45,50]

RESULTS

AD diagnosis phenotyping algorithm performance

Using chart-reviewed labels (**S-Table 5**) to evaluate AD diagnosis phenotyping algorithm performance, KOMAP achieved a higher AUROC at MGB (n=100, 45% AD, AUROC=0.923) than UPMC (n=200, 49% AD, AUROC=0.854, **Figure 1B-E, S-Tables 6-9, S-Result-1**). While KOMAP achieved higher AUROC in women at both healthcare systems (UPMC: women=0.888, men=0.782; MGB: women=0.965, men=0.855), there was inconsistent pattern between two systems for race and ethnicity groups (UPMC: NHW=0.856, other=0.854; MGB: NHW=0.925, other=1.000). Using registry-derived labels to evaluate algorithm performance at UPMC (n=1916, 47% AD, AUROC=0.835, **S-Table 10**), KOMAP achieved higher AUROC in women than men and higher in racial and ethnic minorities than NHW individuals (women=0.846, men=0.824; NHW=0.821, other=0.922). In sensitivity analysis (without imputation of missing race and ethnicity), the algorithm achieved similar performances (**S-Table 11**).

AD cohort characteristics

The combined AD cohort comprised 29,262 patients (61% women, 90% NHW, mean[SD]=79.52[9.39] years at AD diagnosis, **Table 1**), including 13,408 UPMC patients (64% women, 92% NHW, 80.56[9.04] years at AD diagnosis) and 15,854 MGB patients (59% women, 89% NHW, 78.63[9.58] years at AD diagnosis). AD cohorts in the two

healthcare systems differed in characteristics. Compared to MGB patients, UPMC patients had older age at AD diagnosis, greater proportion of women and NHW individuals, lower baseline healthcare utilization, greater pre-existing comorbidity burden, greater proportion of nursing home admission and death during study follow-up, greater proportion of AD-related medication prescriptions (particularly before AD diagnosis), and shorter follow-up duration (**Table 1, S-Table 12, S-Result-2**).

In the combined AD cohort, demographic groups differed in patient profiles (**Table 1**). At AD diagnosis, women were older than men, while NHW individuals were older than racial and ethnic minorities (mean[SD] years: women=80.03[9.33], men=78.71[9.41], $p<.0001$; NHW=79.67[9.32], other=78.11[9.86], $p<.0001$). Baseline healthcare utilization was higher in men than women and higher in racial and ethnic minorities than NHW individuals (mean[SD] healthcare utilization counts: women=192.48[222.49], men=218.90[264.36], $p<.0001$; NHW=198.12[230.91], other=245.77[308.67], $p<.0001$). The pre-existing comorbidity burden was overall high, with men having higher burden than women and racial and ethnic minorities having greater burden than NHW individuals, respectively (**Table 1, S-Figure 2**). While most patients were prescribed AD-related medications, women and racial and ethnic minorities were prescribed at lower rates than men and NHW individuals, respectively (n[%] AD-related prescriptions: overall=21,451[73%]; women=12,952[72%], men=8,499[75%], $p<.0001$; NHW=19,555[74%], other=1,896[67%], $p<.0001$, **Table 1**). During the study follow-up period, 14,283 (49%) patients were admitted to nursing homes and 15,148 (52%) died. While a higher proportion of women than men were admitted to nursing homes, a higher proportion of

men and NHW individuals died during follow-up than women and racial and ethnic minorities (n[%] admitted to nursing homes: women=9,052[51%], men=5,231[46%], *p*<.0001; NHW=12,878[49%], other=1,405[49%], *p*=.05; n[%] death: women=8,963[50%], men=6,185[54%], *p*<.0001; NHW=13,871[53%], other=1,277[45%], *p*<.0001). In sensitivity analysis (without imputation of missing race and ethnicity), patient profiles were comparable to the main AD cohort (**S-Table 13, S-Result-3**).

Differential risk of nursing home admission and death

Using fixed-effects meta-analysis of the healthcare system-specific covariate-adjusted Cox-PH models, we compared the risk of two pragmatic clinical outcomes among demographic groups. First, women had a higher risk of nursing home admission than men (HR [95% CI]=1.061 [1.024-1.100], *p*=.001, **Figure 3, S-Table 14, S-Result-4**), while there was no significant differences between race and ethnicity groups (HR [95% CI]=1.006 [0.952-1.063], *p*=0.831). Second, NHW individuals had a higher risk of death than racial and ethnic minorities (HR [95% CI]=1.376 [1.245-1.521], *p*<.0001, **S-Table 15**), while women had a lower risk than men (HR [95% CI]=0.856 [0.811-0.904], *p*<.0001).

Key drivers of the increased nursing home admission risk (**Figure 4A**) included older age at AD diagnosis, being a woman, high baseline healthcare utilization, and high pre-existing comorbidity burden characterized by the presence of psychiatric disorders (*e.g.,* psychoses, depression), cardiometabolic and pulmonary disorders (*e.g.,* hypertension, diabetes, chronic pulmonary disease), or renal disorders (*e.g.,* fluid and electrolyte disorders). Key drivers of the increased death risk (**Figure 4B**) included older age at AD

diagnosis, being male, being NHW, low healthcare utilization, and high pre-existing comorbidity burden characterized by the presence of neoplastic disorders (*e.g.,* metastatic cancer, solid tumor without metastasis), cardiometabolic and vascular disorders (*e.g.,* diabetes, congestive heart failure, peripheral vascular disease), or psychiatric disorders (*e.g.,* psychoses). Sensitivity analyses using a single score for pre-existing comorbidity burden and excluding patients with missing race and ethnicity information yielded similar results (**S-Figures 3-6**, **S-Table 14**, **S-Table 15**).

DISCUSSION

We re-examined the previously reported demographic differences and drivers of AD decline using real-world clinical data from two large healthcare systems with accurate AD population identification. We identified AD populations using a novel unsupervised phenotyping algorithm (KOMAP), which achieved *overall* robust predictive performance with modest differences across healthcare systems and demographic groups (**S-Discussion-1**). Demographic groups in the AD-EHR cohorts exhibited differential risks of clinical decline. Women had a higher risk of nursing home admission, while NHW individuals and men had a higher risk of death. Older age at AD diagnosis and greater pre-existing comorbidity burden increased both nursing home admission and death risk.

In these real-world AD-EHR cohorts, the risk of AD decline differed across demographic groups, largely consistent with prior epidemiological or claims-based studies.[4–10] Women had a higher risk of nursing home admission than men, possibly due to longer life expectancy.[4,5,35,51] In contrast, men had a higher risk of death than women, potentially attributable to the greater pre-existing comorbidity burden in men, which would be consistent with their higher baseline healthcare utilization. Further, men with AD might have a greater AD symptom burden (not measured here), leading to more rapid decline.[52] While prior studies (using claims data) found racial and ethnic minorities having a higher risk of nursing home admission than NHW individuals, our study did not confirm this finding.[6,7] Interestingly, NHW patients had a higher risk of death than racial and ethnic minorities, mirroring the broader pattern in “mortality crossover” among older adults.[8–10,16,53] As a possible explanation, racial and ethnic minorities more likely remained

undiagnosed, underestimating their AD-related mortality.[54] This study also cannot address whether biological factors drive differences in clinical outcomes across demographic groups.[55]

We found older age at AD diagnosis and greater pre-existing comorbidity burden (particularly psychiatric and cardiometabolic disorders) as key drivers of higher risks of *both* nursing home admission and death.[11–16,56] Given that healthcare utilization and comorbidities generally increase with age, older age of diagnosis may reflect diagnostic delay (rather than later disease onset), possibly explaining its role in worse AD outcomes. Greater pre-existing comorbidity burden compounds AD management challenges and increases comprehensive and specialized care needs, which are inaccessible to many patients. Interestingly, high baseline healthcare utilization (likely reflecting greater baseline functional impairment) increased nursing home admission risk, whereas low healthcare utilization (possibly reflecting inadequate symptom and comorbidity management) increased death risk. Taken together, clinical investigations using real-world AD cohorts would benefit from using multiple interrelated endpoints and careful interpretations.

The study has several strengths. First, we overcame the limitations of existing approaches to identifying AD populations (*e.g.,* sparse gold-standard labels, lack of validation across demographic groups, lack of external validation) by validating the algorithm with gold-standard labels in two independent healthcare systems and across demographic groups (**S-Discussion-2**).[25–34] The overall robust performance in identifying

people with AD supports our AD cohorts as a valuable clinical research resource. Second, we demonstrated the feasibility of using these two large AD-EHR cohorts with long-term follow-up to examine the time to two pragmatic but clinically relevant endpoints, *i.e.,* nursing home admission and death. These real-world AD cohorts derived data from both academic and community practices within two geographically distinct large catchment areas. Moreover, these real-world AD cohorts are more representative of clinical AD populations than traditional epidemiological studies. Third, although >40% of people with AD eventually reside in nursing homes, nursing home admission has been underutilized in EHR-based AD studies.[26,57–59] We applied a rule-based definition (≥1 code or CUI related to nursing home) to identify nursing home admission from the EHR, which was validated using gold-standard chart-reviewed labels. While other key clinical outcomes such as cognition and function fluctuate over time, nursing home admission and death are concrete milestones readily ascertainable from the EHR. Finally, we comprehensively assessed key demographic factors and pre-existing comorbidities and quantified the extent to which drive AD decline. Specifically, we assessed common comorbidities using ECI, a well-validated measure of comorbidity burden.

Our study also has limitations. First, we grouped multiple racial and ethnic groups (*e.g.,* American Indian, Asian, Black, Hispanic) as one minority group due to the modest sample size of each. While statistically necessary, this approach might obscure meaningful group differences and limit the generalizability to these demographic groups. Second, in this first series of studies using these AD-EHR cohorts, we did not assess cognitive or functional decline. Given the sparse EHR documentation of cognition (*e.g.,* Mini Mental

Status Exam, Montreal Cognitive Assessment) and function (*e.g.,* Functional Activities Questionnaire) status,[60] we are developing novel methods to impute these longitudinal measures for future studies. Third, these AD cohorts may include probable and possible AD and various clinical presentations (*e.g.,* amnestic, nonamnestic). Additional methods will be necessary to subtype the heterogeneous AD population. Through linkage with ADRC (at UPMC), we will compare the algorithm-predicted AD diagnosis with biomarker-informed AD diagnosis in the future. Finally, biases inherent in EHR data such as potential data incompleteness (*e.g.,* leakage due to the fragmented healthcare landscape) and selection bias (*e.g.,* stemming from disparities in diagnosis rates, healthcare access, and survival) might influence results of AD diagnosis phenotyping and estimation of AD decline.[61]

CONCLUSION

We created two large real-world AD cohorts with long follow-up in two independent healthcare systems. This cohort study underscored the demographic differences in the risk of nursing home admission and death among the AD population and identified potential drivers. Future research using such real-world cohorts will examine the contribution of additional factors (*e.g.,* social determinants of health as their coding are gradually introduced into routine clinical care) beyond demographics, healthcare utilization and comorbidities, deploy formal causal inference approaches, and incorporate imputed cognitive, functional and frailty status to complement nursing home admission and death as outcomes. This research can inform targeted interventions to modify the risk of AD decline.

ACKNOWLEDGEMENTS

We would like to thank the patients whose data contributed to the research findings.

This study was supported by the National Institutes of Health under award numbers R01 NS098023 and R01 NS124882 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

TABLES

Table 1. Cohort characteristics.

	UPMC + MGB	UPMC	MGB		UPMC + MGB by Race and Ethnicity			UPMC + MGB by Gender		
	Overall	**Overall**	**Overall**	**P-value**	**NHW [g]**	**Other [h]**	**P-value**	**Men**	**Women**	**P-value**
Number of Patients, N (%)	29262 (100%)	13408 (100%)	15854 (100%)		26418 (100%)	2844 (100%)		11,384 (100%)	17,878 (100%)	
Age at AD diagnosis, mean (SD)	79.52 (9.39)	80.56 (9.04)	78.63 (9.58)	**<.0001**	79.67 (9.32)	78.11 (9.86)	**<.0001**	78.71 (9.41)	80.03 (9.33)	**<.0001**
Gender, N (%)				**<.0001**			**<.0001**			**<.0001**
Women	17,878 (61%)	8,549 (64%)	9,329 (59%)		15,995 (61%)	1,883 (66%)		0 (0%)	17,878 (100%)	
Men	11384 (39%)	4859 (36%)	6525 (41%)		10423 (39%)	961 (34%)		11,384 (100%)	0 (0%)	
Race and ethnicity [a], N (%)										
Non-Hispanic White	26,418 (90%)	12,299 (92%)	14,119 (89%)	**<.0001**	26418 (100%)	0 (0%)	**<.0001**	10,423 (92%)	15,995 (89%)	**<.0001**
American Indian	<20 (<0.1%)	<10 (<0.1%)	<20 (<0.1%)		0 (0%)	<20 (<0.1%)		<10 (<0.1%)	<20 (<0.1%)	
Asian	331 (1.1%)	62 (0.5%)	269 (1.7%)		0 (0%)	331 (1.1%)		142 (1.2%)	189 (1.1%)	
Black	1,688 (5.8%)	998 (7.4%)	690 (4.4%)		0 (0%)	1,688 (5.8%)		523 (4.6%)	1,165 (6.5%)	
Hawaiian or Pacific Islander	<10 (<0.1%)	<10 (<0.1%)	<10 (<0.1%)		0 (0%)	<10 (<0.1%)		<10 (<0.1%)	<10 (<0.1%)	
Hispanic or Latino	525 (1.8%)	41 (0.3%)	484 (3.1%)		0 (0%)	525 (1.8%)		177 (1.6%)	348 (1.9%)	
Other	276 (0.9%)	0 (0%)	276 (1.7%)		0 (0%)	276 (0.9%)		109 (1.0%)	167 (0.9%)	
Baseline healthcare utilization (number of events) [b], mean (SD)	202.75 (239.99)	187.69 (193.14)	215.50 (272.77)	**0.009**	198.12 (230.91)	245.77 (308.67)	**<.0001**	218.90 (264.36)	192.48 (222.49)	**<.0001**
Baseline Elixhauser comorbidity index [b], N (%)				**<.0001**			**<.0001**			**<.0001**
<0	2,517 (8.6%)	1,303 (9.7%)	1,214 (7.7%)		2,271 (8.6%)	246 (8.6%)		677 (5.9%)	1,840 (10%)	
0	6,357 (22%)	2,285 (17%)	4,072 (26%)		5,800 (22%)	557 (20%)		2,167 (19%)	4,190 (23%)	
1-5	5,736 (20%)	2,739 (20%)	2,997 (19%)		5,211 (20%)	525 (18%)		2,155 (19%)	3,581 (20%)	
6-13	8,673 (30%)	3,958 (30%)	4,715 (30%)		7,832 (30%)	841 (30%)		3,557 (31%)	5,116 (29%)	
14+	5,979 (20%)	3,123 (23%)	2,856 (18%)		5,304 (20%)	675 (24%)		2,828 (25%)	3,151 (18%)	
Baseline comorbidities (top 2 prevalent), N (%)										

Hypertension, uncontrolled	18,282 (62%)	9,623 (72%)	8,659 (55%)	**<.0001**	16,181 (61%)	2,101 (74%)	**<.0001**	7,003 (62%)	11,279 (63%)	**0.007**
Cardiac arrhythmia	9,398 (32%)	4,707 (35%)	4,691 (30%)	**<.0001**	8,501 (32%)	897 (32%)	0.5	4,370 (38%)	5,028 (28%)	**<.0001**
Survival, N (%)				**<.0001**			**<.0001**			**<.0001**
Alive	14,114 (48%)	5,336 (40%)	8,778 (55%)		12,547 (47%)	1,567 (55%)		5,199 (46%)	8,915 (50%)	
Death	15,148 (52%)	8,072 (60%)	7,076 (45%)		13,871 (53%)	1,277 (45%)		6,185 (54%)	8,963 (50%)	
Time to death, median months (1st and 3rd quantiles)	33 (15,59)	32 (16, 53)	34 (15, 62)	**0.002**	21 (7,41)	23 (8,46)	**<.0001**	30 (13,53)	36 (17,63)	**<.0001**
Nursing Home Admission, N (%)				**<.0001**			0.5			**<.0001**
Admitted	14,283 (49%)	8,841 (66%)	5,442 (34%)		12,878 (49%)	1,405 (49%)		5,231 (46%)	9,052 (51%)	
Not admitted	14,979 (51%)	4,567 (34%)	10,412 (66%)		13,540 (51%)	1,439 (51%)		6,153 (54%)	8,826 (49%)	
Time to nursing home admission, median months (1st and 3rd quantiles)	21 (7, 42)	21 (7,21)	21 (7,44)	**0.004**	33 (15, 58)	43 (18, 76)	**<.0001**	20 (7,41)	21 (7,42)	**<0.001**
AD-related medication [c], N (%)										
Not prescribed	7,811 (27%)	2,572 (19%)	5,239 (33%)	**<.0001 [d]**	6,863 (26%)	948 (33%)	**<.0001 [d]**	2,885 (25%)	4,926 (28%)	**<.0001 [d]**
Prescribed	21,451 (73%)	10,836 (81%)	10,615 (67%)		19,555 (74%)	1,896 (67%)		8,499 (75%)	12,952 (72%)	
Before AD diagnosis	10,789 (50%)	5,966 (55%)	4,823 (45%)	**<.0001 [e]**	9,937 (51%)	852 (45%)	**<.0001 [e]**	4,395 (52%)	6,394 (49%)	**<.0001 [e]**
On or after AD diagnosis	10,662 (50%)	4,870 (45%)	5,792 (55%)		9,618 (49%)	1,044 (55%)		4,104 (48%)	6,558 (51%)	
Follow-up duration [f], median months (1st and 3rd quantiles)	80 (45, 123)	78 (48, 109)	81 (45,129)	**<.0001**	80 (46, 123)	80 (45, 131)	**0.01**	78 (45, 119)	81 (46, 126)	**<.0001**

a. We imputed missing race and ethnicity using multiple imputation based on age and gender.
b. We calculated the baseline healthcare utilization and comorbidity burden using the Elixhauser index (with van Walraven weighting) in the 24 months prior to AD diagnosis.
c. The complete list of AD-related medications is available in S-Table 3.
d. We used a two-sample t-test to compare the difference between the number of patients prescribed and not prescribed AD-related medication.
e. We used a two-sample t-test to compare the difference between the number of patients prescribed AD-related medication before vs after AD diagnosis.
f. We estimated the median follow-up duration using the reverse Kaplan-Meier survival analysis.

g. NHW, non-Hispanic White individuals.
h. Other, racial and ethnic minorities.

FIGURES

Figure 1. AD diagnosis phenotyping algorithm. (A) Schematic overview. (B-C) Phenotyping algorithm performance at UPMC. (D-E) Phenotyping algorithm performance at MGB.

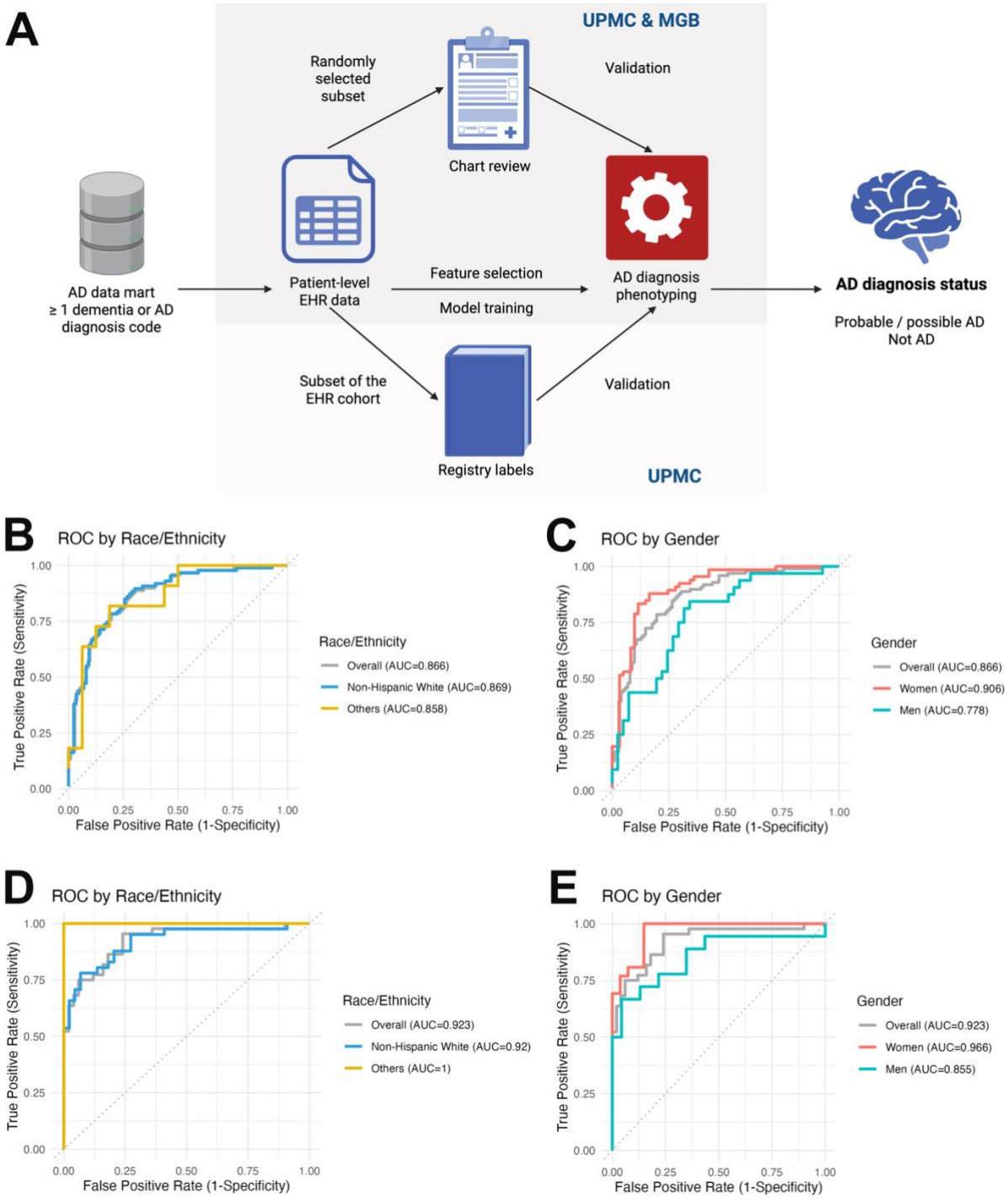


Figure 2. Study design. (A) Inclusion criteria. (B) Healthcare system-specific covariate-adjusted Cox proportional hazard models. (C) Fixed-effects meta-analysis of healthcare system-specific covariate-adjusted Cox proportional hazard models.

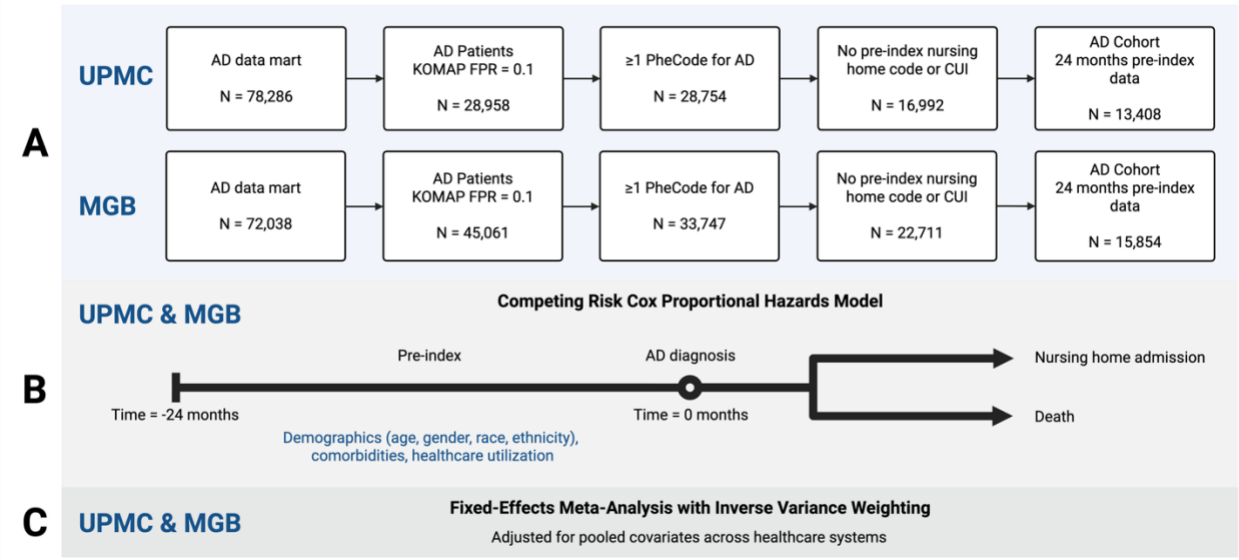


Figure 3. Fixed-effects meta-analysis of healthcare system-specific covariate-adjusted Cox proportional hazard models to estimate time to nursing home admission and death by demographic groups. The fixed-effects meta-analysis model was adjusted for pooled covariates across both healthcare systems from inverse-probability weighting. Covariates included demographics (*e.g.*, age at AD diagnosis, gender, race, ethnicity) and baseline clinical profiles (*e.g.,* healthcare utilization, pre-existing comorbidity burden) in the 24 months preceding the index date. (A-B) Nursing Home Admission. (C-D) Death. Reference group for race and ethnicity: non-Hispanic White. Reference group for gender: women.

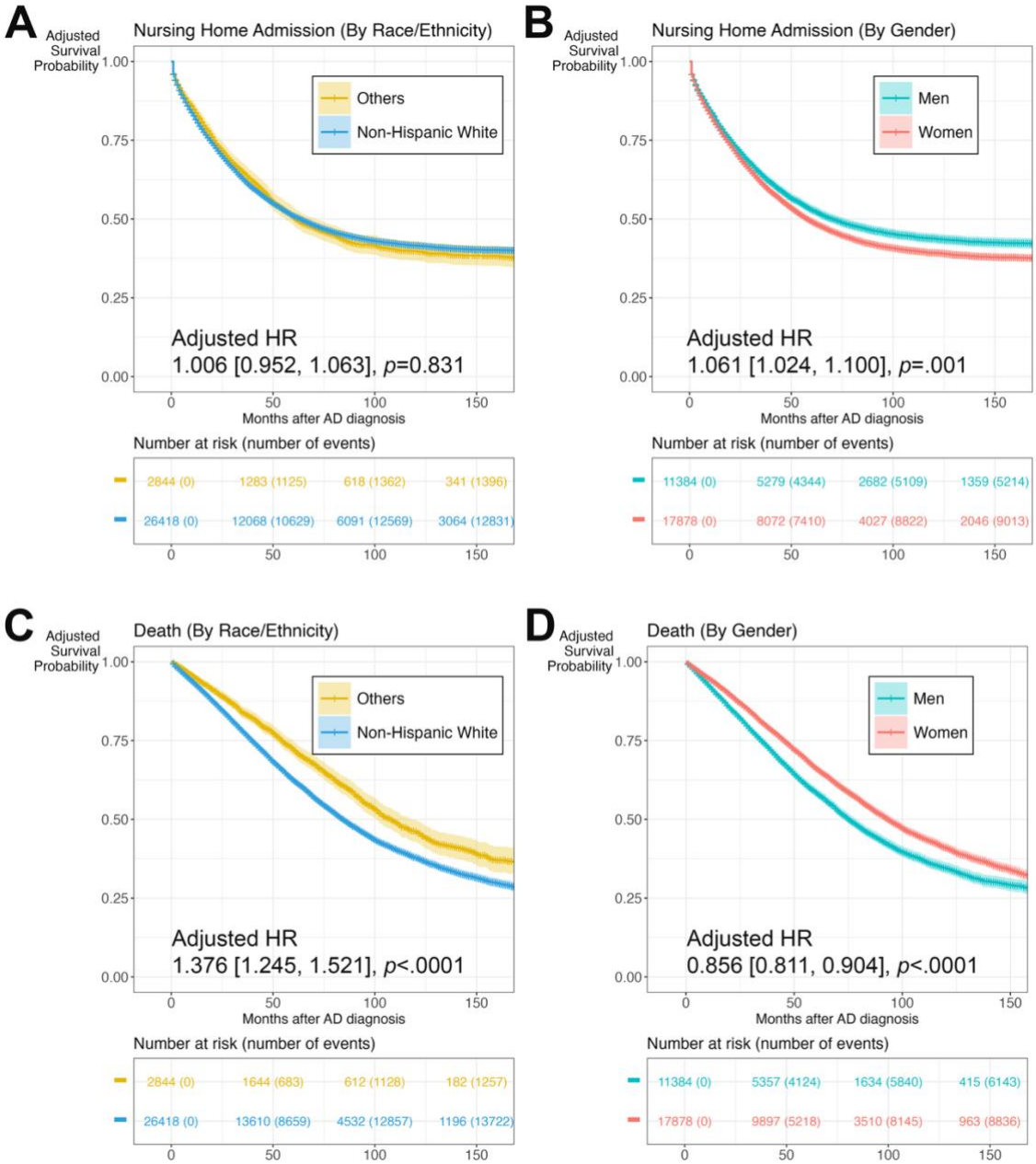
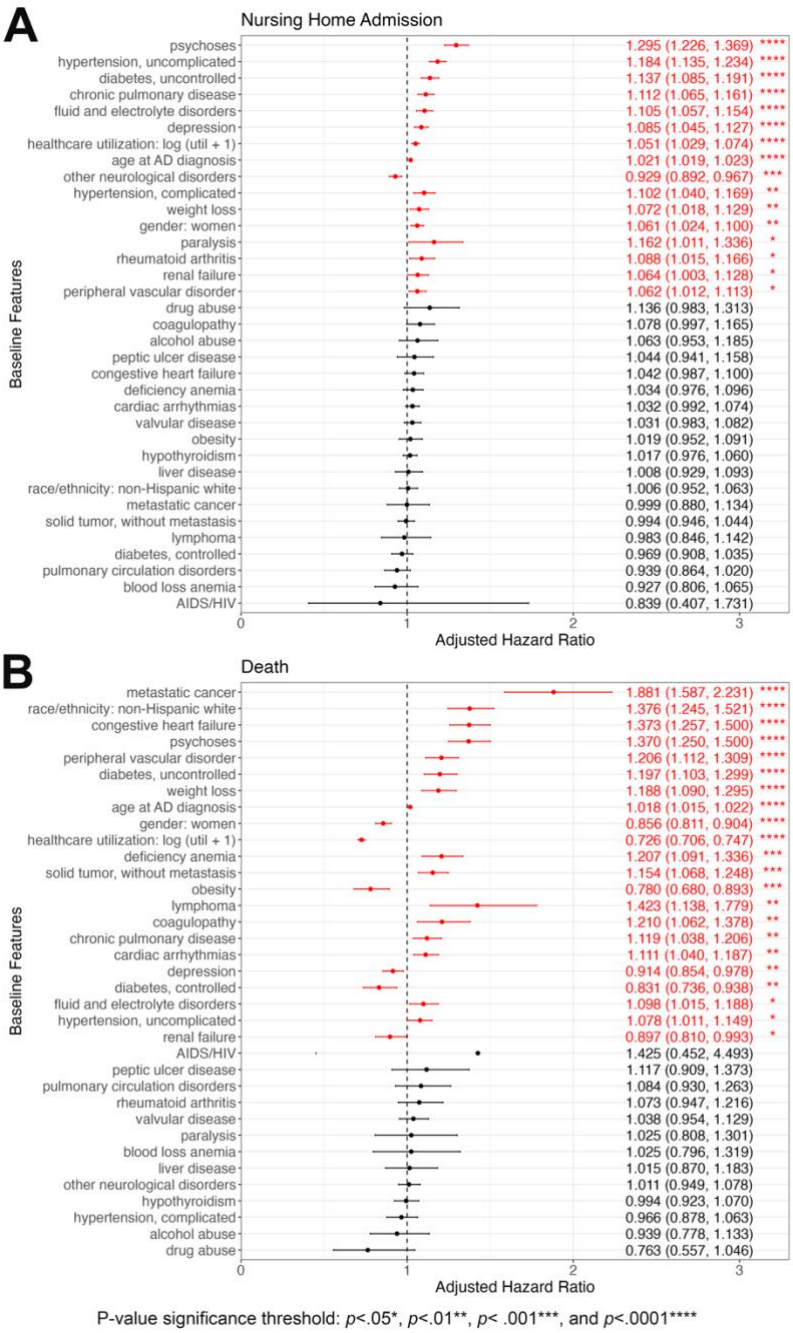


Figure 4. Adjusted hazard ratios of variables in the fixed-effects meta-analysis of healthcare system-specific covariate-adjusted Cox proportional hazard models of nursing home admission and death. The fixed-effects meta-analysis model was adjusted for pooled covariates across both healthcare systems from inverse-probability weighting. Covariates included demographics (*e.g.*, age at AD diagnosis, gender, race, ethnicity) and baseline clinical profiles (*e.g.,* healthcare utilization, pre-existing comorbidity burden) in the 24 months preceding the index date. Variables are ordered first by p-value significance threshold (p<.05*, p<.01**, p<.001***, p<.0001****) and then effect size. (A) Nursing Home Admission. (B) Death.



REFERENCES

1. Scheltens P, Strooper BD, Kivipelto M, et al. Alzheimer's disease. *Lancet*. 2021;397(EMBO J 36 2017):1577-1590. doi:10.1016/s0140-6736(20)32205-4

2. Jack CR, Andrews JS, Beach TG, et al. Revised criteria for diagnosis and staging of Alzheimer's disease: Alzheimer's Association Workgroup. *Alzheimer's Dement*. 2024;20(8):5143-5169. doi:10.1002/alz.13859

3. 2025 Alzheimer's disease facts and figures. *Alzheimer's Dement*. 2025;21(4):e70235. doi:10.1002/alz.70235

4. Gracner T, Stone PW, Agarwal M, Sorbero M, Mitchell SL, Dick AW. Advanced cognitive impairment among older nursing home residents. *BMC Geriatr*. 2021;21(1):382. doi:10.1186/s12877-021-02336-1

5. Lusk JB, Ford CB, Soneji S, et al. Sex Differences in Mortality and Health Care Utilization After Dementia Diagnosis. *JAMA Neurol*. 2025;82(10). doi:10.1001/jamaneurol.2025.2236

6. Rivera-Hernandez M, Kumar A, Epstein-Lubow G, Thomas KS. Disparities in Nursing Home Use and Quality Among African American, Hispanic, and White Medicare Residents With Alzheimer's Disease and Related Dementias. *J Aging Heal*. 2019;31(7):1259-1277. doi:10.1177/0898264318767778

7. Temkin-Greener H, Yan D, Wang S, Cai S. Racial disparity in end-of-life hospitalizations among nursing home residents with dementia. *J Am Geriatr Soc*. 2021;69(7):1877-1886. doi:10.1111/jgs.17117

8. Mayeda ER, Glymour MM, Quesenberry CP, Johnson JK, Pérez-Stable EJ, Whitmer RA. Survival after dementia diagnosis in five racial/ethnic groups. *Alzheimer's Dement*. 2017;13(7):761-769. doi:10.1016/j.jalz.2016.12.008

9. Chen Y, Crimmins E, Ferido P, Zissimopoulos JM. Racial/Ethnic Disparities in Length of Life after Dementia Diagnosis: an 18-Year Follow-up Study of Medicare Beneficiaries. *Lancet Reg Heal - Am*. 2022;8:100179. doi:10.1016/j.lana.2021.100179

10. Khan AH, Ijaz E, Ubaid B, et al. Analysis of Alzheimer's Disease-Related Mortality Rates Among the Elderly Populations Across the United States: An Analysis of Demographic and Regional Disparities from 1999 to 2020. *Curr Alzheimer Res*. 2024;21(6):384-394. doi:10.2174/0115672050338833240924113200

11. Yen TY, Beriwal N, Kaur P, et al. Medical Comorbidities and Association With Mortality Risk in Alzheimer's Disease: Population-Based Study of 132,405 Geriatric Inpatients. *Cureus J Medical Sci*. 2020;12(5):e8203. doi:10.7759/cureus.8203

12. Taudorf L, Nørgaard A, Brodaty H, Laursen TM, Waldemar G. Dementia increases mortality beyond effects of comorbid conditions: A national registry-based cohort study. *Eur J Neurol*. 2021;28(7):2174-2184. doi:10.1111/ene.14875

13. Rajamaki B, Hartikainen S, Tolppanen AM. The effect of comorbidities on survival in persons with Alzheimer's disease: a matched cohort study. *Bmc Geriatr*. 2021;21(1):173. doi:10.1186/s12877-021-02130-z

14. Armstrong MJ, Song S, Kurasz AM, Li Z. Predictors of Mortality in Individuals with Dementia in the National Alzheimer's Coordinating Center. *J Alzheimer's Dis*. 2022;86(4):1935-1946. doi:10.3233/jad-215587

15. Lanctôt KL, Hahn-Pedersen JH, Eichinger CS, et al. Burden of Illness in People with Alzheimer's Disease: A Systematic Review of Epidemiology, Comorbidities and Mortality. *J Prev Alzheimer's Dis*. 2024;11(1):97-107. doi:10.14283/jpad.2023.61

16. Zheng X, Wang S, Huang J, Li C, Shang H. Predictors for survival in patients with Alzheimer's disease: a large comprehensive meta-analysis. *Transl Psychiatry*. 2024;14(1):184. doi:10.1038/s41398-024-02897-w

17. Grodstein F, Chang CH, Capuano AW, et al. Identification of Dementia in Recent Medicare Claims Data, Compared With Rigorous Clinical Assessments. *J Gerontol: Ser A*. 2021;77(6):1272-1278. doi:10.1093/gerona/glab377

18. Bhattacharyya J, Barnes LL, Chen Y, et al. Evaluating linked ICD-10 Medicare claims data as a method of dementia case ascertainment in research settings. *Alzheimer's Dement*. 2025;21(5):e70200. doi:10.1002/alz.70200

19. McCarthy EP, Chang CH, Tilton N, Kabeto MU, Langa KM, Bynum JPW. Validation of Claims Algorithms to Identify Alzheimer's Disease and Related Dementias. *J Gerontol: Ser A*. 2021;77(6):1261-1271. doi:10.1093/gerona/glab373

20. Haye S, Thunell J, Joyce G, et al. Estimates of diagnosed dementia prevalence and incidence among diverse beneficiaries in traditional Medicare and Medicare Advantage. *Alzheimer's Dement: Diagn, Assess Dis Monit*. 2023;15(3):e12472. doi:10.1002/dad2.12472

21. Gianattasio KZ, Wachsmuth J, Murphy R, et al. Case Definition for Diagnosed Alzheimer Disease and Related Dementias in Medicare. *JAMA Netw Open*. 2024;7(9):e2427610. doi:10.1001/jamanetworkopen.2024.27610

22. Phung TKT, Andersen BB, Høgh P, Kessing LV, Mortensen PB, Waldemar G. Validity of Dementia Diagnoses in the Danish Hospital Registers. *Dement Geriatr Cogn Disord*. 2007;24(3):220-228. doi:10.1159/000107084

23. Salem LC, Andersen BB, Nielsen TR, et al. Overdiagnosis of Dementia in Young Patients – A Nationwide Register-Based Study. *Dement Geriatr Cogn Disord*. 2013;34(5-6):292-299. doi:10.1159/000345485

24. Rizzuto D, Feldman AL, Karlsson IK, Aslan AKD, Gatz M, Pedersen NL. Detection of Dementia Cases in Two Swedish Health Registers: A Validation Study. *J Alzheimer's Dis*. 2018;61(4):1301-1310. doi:10.3233/jad-170572

25. Tjandra D, Migrino RQ, Giordani B, Wiens J. Cohort discovery and risk stratification for Alzheimer's disease: an electronic health record-based approach. *Alzheimer's Dement: Transl Res Clin Interv*. 2020;6(1):e12035. doi:10.1002/trc2.12035

26. Han E, Kharrazi H, Shi L. Identifying Predictors of Nursing Home Admission by Using Electronic Health Records and Administrative Data: Scoping Review. *JMIR Aging*. 2023;6:e42437. doi:10.2196/42437

27. Walling AM, Pevnick J, Bennett AV, Vydiswaran VGV, Ritchie CS. Dementia and electronic health record phenotypes: a scoping review of available phenotypes and opportunities for future research. *J Am Méd Inform Assoc*. 2023;30(7):1333-1348. doi:10.1093/jamia/ocad086

28. Oh IY, Schindler SE, Ghoshal N, Lai AM, Payne PRO, Gupta A. Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing. *JAMIA Open*. 2023;6(1):ooad014. doi:10.1093/jamiaopen/ooad014

29. Estiri H, Azhir A, Blacker DL, Ritchie CS, Patel CJ, Murphy SN. Temporal characterization of Alzheimer's Disease with sequences of clinical records. *eBioMedicine*. 2023;92:104629. doi:10.1016/j.ebiom.2023.104629

30. Chen Z, Yang Y, Zhang D, et al. Predicting the Risk of Alzheimer's Disease and Related Dementia in Patients with Mild Cognitive Impairment Using a Semi-Competing Risk Approach. *Informatics*. 2023;10(2):46. doi:10.3390/informatics10020046

31. Li Q, Yang X, Xu J, et al. Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. *Alzheimer's Dementia*. Published online 2023. doi:10.1002/alz.12967

32. He X, Wei R, Huang Y, et al. Develop and validate a computable phenotype for the identification of Alzheimer's disease patients using electronic health record data. *Alzheimer's Dement: Diagn, Assess Dis Monit*. 2024;16(3):e12613. doi:10.1002/dad2.12613

33. Zhu W, Tang H, Zhang H, et al. Predicting Risk of Alzheimer's Diseases and Related Dementias with AI Foundation Model on Electronic Health Records. *medRxiv*. Published online 2024:2024.04.26.24306180. doi:10.1101/2024.04.26.24306180

34. Tang AS, Rankin KP, Cerono G, et al. Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. *Nat Aging*. 2024;4(3):379-395. doi:10.1038/s43587-024-00573-8

35. Mielke MM, Aggarwal NT, Vila-Castelar C, et al. Consideration of sex and gender in Alzheimer's disease and related disorders from a global perspective. *Alzheimer's Dement*. 2022;18(12):2707-2724. doi:10.1002/alz.12662

36. Shaaban CE, Rosso AL. Current Topics in Behavioral Neurosciences, Curr.Topics Behav.Neurosci., Racial, Ethnic, and Geographic Diversity in Population Neuroscience. *Curr Top Behav Neurosci*. Published online 2024:1-19. doi:10.1007/7854_2024_475

37. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Méd Inform Assoc*. 2011;18(4):441-448. doi:10.1136/amiajnl-2011-000116

38. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clin Chem*. 2003;49(4):624-633. doi:10.1373/49.4.624

39. Liao KP, Sun J, Cai TA, et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *J Am Med Inform Assn*. 2019;26(11):1255-1262. doi:10.1093/jamia/ocz066

40. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc*. 2019;14(12):3426-3444. doi:10.1038/s41596-019-0227-6

41. Yu S, Cai T, Cai T. NILE: Fast Natural Language Processing for Electronic Health Records. *Arxiv*. Published online 2013. doi:10.48550/arxiv.1311.6063

42. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(suppl_1):D267-D270. doi:10.1093/nar/gkh061

43. Buuren S van, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45(3). doi:10.18637/jss.v045.i03

44. Zhang G, Rose CE, Zhang Y, et al. Multiple Imputation of Missing Race and Ethnicity in CDC COVID-19 Case-Level Surveillance Data. *Int J Stat Méd Res*. 2022;11:1-11. doi:10.6000/1929-6029.2022.11.01

45. Xiong X, Sweet SM, Liu M, et al. Knowledge-Driven Online Multimodal Automated Phenotyping System. *medRxiv*. Published online 2023:2023.09.29.23296239. doi:10.1101/2023.09.29.23296239

46. Hou J, Kim N, Cai T, et al. Comparison of Dimethyl Fumarate vs Fingolimod and Rituximab vs Natalizumab for Treatment of Multiple Sclerosis. *Jama Netw Open*. 2021;4(11):e2134627. doi:10.1001/jamanetworkopen.2021.34627

47. Elixhauser* A, Steiner† C, Harris‡ DR, Coffey§ RM. Comorbidity Measures for Use with Administrative Data. *Méd Care*. 1998;36(1):8-27. doi:10.1097/00005650-199801000-00004

48. Gasparini A. comorbidity: An R package for computing comorbidity scores. *Journal of Open Source Software*. 2018;3(23):648. doi:10.21105/joss.00648

49. Walraven C van, Austin PC, Jennings A, Quan H, Forster AJ. A Modification of the Elixhauser Comorbidity Measures Into a Point System for Hospital Death Using Administrative Data. *Méd Care*. 2009;47(6):626-633. doi:10.1097/mlr.0b013e31819432e5

50. Leveraging electronic health records to examine differential clinical outcomes in people with Alzheimer's Disease. https://github.com/xialab2016/adphenotyping

51. Mjørud M, Selbæk G, Bjertness E, et al. Time from dementia diagnosis to nursing-home admission and death among persons with dementia: A multistate survival analysis. *PLoS ONE*. 2020;15(12):e0243513. doi:10.1371/journal.pone.0243513

52. Ferretti MT, Iulita MF, Cavedo E, et al. Sex differences in Alzheimer disease — the gateway to precision medicine. *Nat Rev Neurol*. 2018;14(8):457-469. doi:10.1038/s41582-018-0032-9

53. Manton KG, Poss SS, Wing S. The Black/White Mortality Crossover: Investigation from the Perspective of the Components of Aging. *Gerontol*. 1979;19(3):291-300. doi:10.1093/geront/19.3.291

54. Hinton L, Tran D, Peak K, Meyer OL, Quiñones AR. Mapping racial and ethnic healthcare disparities for persons living with dementia: A scoping review. *Alzheimer's Dement*. 2024;20(4):3000-3020. doi:10.1002/alz.13612

55. Research C on the U of R and E in B, Policy B on HS, Practice B on PH and PH, Services B on HC, Division H and M, Medicine NA of S Engineering, and. Rethinking Race and Ethnicity in Biomedical Research. Published online 2024. doi:10.17226/27913

56. Sharma S, Liu J, Abramowitz AC, et al. Leveraging multi-site electronic health data for characterization of subtypes: a pilot study of dementia in the N3C Clinical Tenant. *JAMIA Open*. 2024;7(3):ooae076. doi:10.1093/jamiaopen/ooae076

57. Knapp M, Chua KC, Broadbent M, et al. Predictors of care home and hospital admissions and their costs for older people with Alzheimer's disease: findings from a large London case register. *BMJ Open*. 2016;6(11):e013591. doi:10.1136/bmjopen-2016-013591

58. Joling KJ, Janssen O, Francke AL, et al. Time from diagnosis to institutionalization and death in people with dementia. *Alzheimer's Dement*. 2020;16(4):662-671. doi:10.1002/alz.12063

59. Kosar CM, Mor V, Werner RM, Rahman M. Risk of Discharge to Lower-Quality Nursing Homes Among Hospitalized Older Adults With Alzheimer Disease and Related Dementias. *JAMA Netw Open*. 2023;6(2):e2255134. doi:10.1001/jamanetworkopen.2022.55134

60. Maserejian N, Krzywy H, Eaton S, Galvin JE. Cognitive measures lacking in EHR prior to dementia or Alzheimer's disease diagnosis. *Alzheimer's Dement*. 2021;17(7):1231-1243. doi:10.1002/alz.12280

61. Al-Sahab B, Leviton A, Loddenkemper T, Paneth N, Zhang B. Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview. *J Healthc Inform Res*. 2024;8(1):121-139. doi:10.1007/s41666-023-00153-2