ARTICLE IN PRESS

Communications Medicine

https://doi.org/10.1038/s43856-026-01443-7

Article in Press

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

Received: 27 May 2025

Accepted: 5 February 2026

Cite this article as: Venkatesh, S., Wang, L., Morris, M. *et al.* Leveraging electronic health records to examine differential clinical outcomes in people with Alzheimer's disease. *Commun Med* (2026). https://doi.org/10.1038/s43856-026-01443-7

Shruthi Venkatesh, Linshanshan Wang, Michele Morris, Mohammed Moro, Ratnam Srivastava, Yunqing Han, Riddhi Patira, Sarah Berman, Oscar L. Lopez, Shyam Visweswaran, Tianrun Cai, Tianxi Cai & Zongqi Xia

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

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

Shruthi Venkatesh, BS[1] *, Linshanshan Wang, PhD[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, USA
2. Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA
3. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
4. Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
5. Department of Rheumatology, Brigham and Women's Hospital, Boston, MA, USA

*** Share co-first authorship**
 # Share co-senior authorship

Word Count: 5832

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. Using real-world electronic health record (EHR) data with accurate population identification, we examine demographic differences and potentially modifiable drivers of AD decline.

METHODS. Leveraging EHR data (1994-2022) from two large independent healthcare systems, we applied an 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 estimated the time-to-decline (nursing home admission, death) in healthcare system-specific covariate-adjusted competing risk survival analyses stratified by demographic groups. We then performed covariate-adjusted fixed-effects meta-analyses using inverse variance weighting.

RESULTS. The algorithm demonstrates robust performance in identifying AD populations across healthcare systems and demographic groups (AUROC score range: 0.835-0.923). 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% transition to nursing homes and 48% die during follow-up. In covariate-adjusted fixed-effects meta-analysis, women have higher nursing home admission risk (HR[95% CI]=1.061[1.024-1.100], *p*=1.203×10^{-3}) but lower death risk (HR[95% CI]=0.856[0.811-0.904], *p*=2.434×10^{-8}) than men. Non-Hispanic White individuals have similar nursing home risk (HR[95% CI]=1.006[0.952-1.063], *p*=8.306×10^{-1}) but higher death risk (HR[95% CI]=1.376[1.245-1.521], *p*=$4.084 \mathrm{X} 10^{-10}$) than racial and ethnic minorities. Older age at AD diagnosis and greater comorbidity burden increase both nursing home admission and death risk.

CONCLUSIONS. We provide real-world evidence of drivers of demographic differences in AD decline that could inform individual clinical management and public health policies.

PLAIN LANGUAGE SUMMARY

People with Alzheimer's disease have memory loss and behavior changes. They experience varying rates of decline, with some facing higher risks of entering nursing homes or dying. In this study, we examined which individuals were at greater risk of decline by analyzing medical records. We identified 29,262 people with Alzheimer's disease using an accurate patient identification tool. We found that women were more likely to enter nursing homes, while men and non-Hispanic White individuals were more likely to die. Patients diagnosed at an older age and those with additional health conditions faced increased risks of both poor outcomes. These insights from real-world clinical data may help clinicians tailor individualized care and promote more equitable Alzheimer's disease management.

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] Systematic 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) add significant value beyond structured codified data (*e.g.,* diagnoses and prescriptions), enriching the assessment of conditions beyond the primary disease of interest (*e.g.,* comorbidities among people with AD). 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. 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] Recent efforts using structured codified data and/or unstructured clinical information (*i.e.,* clinical narrative) largely relied on the EHR data of a single healthcare system and lacked external validation.[26–34] 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, which may exacerbate algorithm bias.[25–30,32–37] Taken together, this leaves an unmet need for more accurate identification of AD populations from the EHR with robust validation of algorithm performance.

Here, we pursue two complementary goals: (1) to accurately identify AD populations from real-world EHR data, and (2) to re-examine demographic differences and systematically assess potentially modifiable drivers of AD decline in these populations. We leverage the EHR data of two independent healthcare systems (that include both academic and community practices), one of which is linked to an AD registry, apply a knowledge graph-guided unsupervised phenotyping algorithm to predict AD diagnosis across demographic groups, and validate algorithm performance using chart-reviewed and registry-derived gold-standard labels to address the methodological challenge of AD population identification. Using these large real-world AD-EHR cohorts, we then examine differences in two *readily ascertainable* clinical outcomes of AD decline (*i.e.,* nursing home admission, mortality) between demographic groups (*e.g.,* gender, race and ethnicity) and examine potentially modifiable drivers of differential AD decline to demonstrate clinical utility. The unsupervised phenotyping algorithm achieves robust performance in identifying AD populations across both healthcare systems. In these real-world AD cohorts, women have higher nursing home admission risk but lower death risk than men. Non-Hispanic White individuals demonstrate similar nursing home admission risk, but higher death risk compared to racial and ethnic minorities. Older age at AD diagnosis and higher comorbidity burden increase both nursing home admission and death risk.

METHODS

Ethics Approval

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

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). As narrative data were collected at UPMC only after January 1, 2011, we included UPMC patients with codified and narrative data from 2011 to 2022. We included MGB patients with codified and narrative data from 1994 to 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).[38,39] We organized diagnoses by mapping ICD-9 and ICD-10 codes to PheCodes,[40,41] 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).[42,43] 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**). This inclusion criteria have high sensitivity for identifying true AD patients as the negative predictive value of dementia diagnosis codes is very high.[40] We imputed missing race and ethnicity using age and gender.[44,45] We performed sensitivity analyses by excluding patients from the initial AD-EHR data-marts with missing race and ethnicity information (**S-Figure 1**).

Identifying AD populations using a phenotyping algorithm

To address the challenge of accurate AD population identification, we applied a 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, Supplementary Data 1**).[46,47] KOMAP leverages an online narrative and codified feature search engine (ONCE) powered by multi-source knowledge-graph representation learning to generate a list of informative codified and narrative features relevant to the target clinical concept, *i.e.,* AD (**Supplementary Data 1**). Using these ONCE-selected features, we trained KOMAP independently at UPMC and MGB to account for population heterogeneity. We combined probable and possible AD into a single category for prediction of AD diagnosis status as they represented the heterogeneous presentations of AD in the real world. We classified individuals from the initial AD-EHR data-marts with KOMAP-predicted AD diagnosis at 90% specificity as having AD (**Figure 2A**). We held the specificity at 90% to minimize false positives while keeping reasonably large sample sizes.

Validating algorithm-predicted AD diagnosis status

Although KOMAP itself is unsupervised, we validated KOMAP-predicted AD diagnosis status for a subset of the patients in the initial AD-EHR data-marts (comprising patients ≥1 dementia or AD PheCode) using gold-standard chart-reviewed labels (UPMC, n=200; MGB, n=100) and registry labels (University of Pittsburgh Alzheimer's Disease Research Registry [ADRC], n=1916 with UPMC EHR linkage) across demographic groups.

Validating algorithm-predicted AD diagnosis status with chart-reviewed labels

After excluding patients in the University of Pittsburgh ADRC registry, we randomly sampled 200 patients from the initial UPMC AD-EHR data-mart for chart review. We also randomly sampled 100 patients from the initial MGB AD-EHR data-mart for chart review. We sampled fewer patients from MGB due to better clinical documentation. After training by an experienced cognitive and behavioral neurologist (RP), a clinical domain expert (SV) reviewed the EHR of these patients to determine AD diagnosis status according to the 2011 diagnostic criteria.[48] Patients whose clinical documentation met the core clinical criteria for probable or possible AD were classified as having a diagnosis of AD. Patients who did not meet these criteria were classified as not having AD. We did not assess any measures of chart review reliability or validity since one clinical domain expert reviewed the medical records. These chart-reviewed labels were used as the gold-standard ground truth for validating KOMAP-predicted AD diagnosis status.

Validating algorithm-predicted AD diagnosis status with registry-derived labels

The University of Pittsburgh ADRC has maintained a longitudinal prospective registry of AD patients for over 25 years. For the 1,916 patients in the UPMC AD data-mart who were enrolled in the ADRC registry, we used an honest broker system to link their EHR data with ADRC registry elements, including clinician-determined AD diagnosis status. We validated KOMAP-predicted AD diagnosis status with clinician-determined AD diagnosis status from the ADRC registry.

Comparing algorithm performance to benchmark phenotyping methods

We compared the performance of KOMAP to common benchmark phenotyping methods constructed from surrogate features predictive of AD diagnosis. These surrogate features included the number of PheCodes for “Alzheimer’s disease” (PheCode.290.11), “dementia” (PheCode.290.1), “delirium, dementia, and amnestic and other cognitive disorders” (PheCode.290), and the number of mentions of the main CUI (*i.e.,* narrative feature) for “Alzheimer’s disease” (C0002395). Further, we trained KOMAP separately by race and ethnicity and by gender groups (“KOMAP separate”). We compared the performance of KOMAP and KOMAP separate to assess if phenotyping needed to be performed independently based on demographic factors.

Phenotyping algorithm performance

We obtained performance measures of all phenotyping methods stratified by race and ethnicity and by gender. We obtained the area under the receiver operating characteristic (AUROC), area under the precision-recall curve (AUPRC), true positive rate (TPR), positive predictive value (PPV), and negative predictive value (NPV). We report TPR, PPV and NPV by thresholding the algorithm output at a value that obtains a specificity of 90%. Based on previously published protocol for phenotyping studies, we defined the AD cohort at 90% specificity as a pragmatic decision to strike a balance between sensitivity and PPV while minimizing false positives.[41] We obtained the AUPRC in addition to the AUROC for all phenotyping algorithms as AUPRC is robust to sample size imbalance.[49] To assess the tradeoff between specificity and PPV, we also provided a comparison of the model performance at different specificity thresholds.

Inclusion Criteria

For downstream analyses (**Figure 2A**), we included AD patients identified using KOMAP with ≥1 AD PheCode (*i.e.,* PheCode 290.11). 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 (first AD PheCode occurrence was the index date or 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.[50] Baseline comorbidity burden was characterized using the Elixhauser comorbidity index (ECI), which comprises comorbidities that are well-established predictors and known risk factors for functional decline and death.[51] Using the R *comorbidity* package, we mapped diagnosis codes to all pre-existing health conditions comprising the ECI.[52] We considered each comorbidity as an individual covariate in the main analysis to account for baseline patient heterogeneity in comorbidity burden, and performed a sensitivity analysis using the van Walraven weighted ECI score (a summed and weighted composite measure of individual comorbidity burden) to assess robustness to modeling comorbidities as a composite measure.[53] 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, **Supplementary Data 2**), validated with gold-standard chart-reviewed labels. A clinical domain expert (SV) reviewed the EHR of 100 randomly sampled patients at UPMC and MGB meeting the study inclusion criteria to determine nursing home admission status. These chart-reviewed labels were used as the gold-standard ground truth for validating nursing home admission. We also performed a sensitivity analysis using ≥1 code to define nursing home admission. 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.

Statistics and Reproducibility

Statistical methods

We used two-tailed t-tests to compare differences in demographics, baseline healthcare utilization, pre-existing comorbidity burden, prevalence of AD-related medication prescriptions (**S-Table 2**), and prevalence of outcomes of AD decline among demographic groups. We calculated the proportion of patients with each comorbidity in the ECI by demographic group and used two-proportion Z-tests to compare the difference in proportion of patients with each comorbidity among demographic groups.

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 then performed a *fixed*-effects meta-analysis of the competing risk Cox proportional hazard models across both healthcare systems (**Figure 2C**), as *random*-effects models would be highly unstable given the number of sites (*i.e.,* 2 healthcare systems). We used inverse variance weighting to combine the data from both sites. We adjusted for the pooled covariates across sites in the fixed-effects meta-analysis. We performed stratified analyses according to demographic groups by estimating the time-to-outcome for each patient while holding other covariates constant, except for the demographic variable of stratification (*i.e.,* gender, or race and ethnicity). This allowed for estimating the time-to-outcome for each demographic group, independent of the effect of additional covariates. We constructed covariate-adjusted survival curves stratified according to demographic groups. We obtained the adjusted hazard ratios for each covariate and visualized the fixed-effects meta-analysis results using forest plots.

Sensitivity analyses

We conducted two sensitivity analyses to confirm the robustness of our findings by: (1) including a weighted comorbidity index score instead of individual comorbidities, and (2) excluding patients with missing race and ethnicity information. In the first sensitivity analysis, we included a validated composite measure indicative of pre-existing comorbidity burden (Elixhauser comorbidity index with van Walraven weighting) instead of individual comorbidities.[51,53] In the second sensitivity analysis, we excluded patients with missing race and ethnicity information from the initial UPMC and MGB AD-EHR data-marts (*i.e., without* imputation of missing race and ethnicity) (**S-Figure 1**). Similar to the main AD cohorts, we created sensitivity analysis AD cohorts by including individuals predicted as having AD by KOMAP at 90% specificity. We excluded patients with <24 months of EHR data or admitted to nursing homes prior to the baseline.

Data availability

Anonymous summary-level registry data and EHR data will be made available upon reasonable request to the corresponding author. The rationale for not sharing patient-level data is that patient-level clinical data (either de-identified information or limited protected health information containing dates of clinical events or even if anonymous due to concern for re-identification) are universally subject to the rules and regulation of each healthcare system, which may only be affiliated with but are not the same as the primary academic institutions of the study investigators. Sharing of de-identified EHR data with qualified external researchers by each of the study performance site may be permissible only after the approval of the respective Institutional Review Boards (IRBs), regulatory oversight agents of the healthcare systems (that own the clinical data) as well as the appropriate Data Usage Agreements (DUA) between institutions.

Code availability

All statistical analyses were conducted using R (version 4.4.1). Codes for KOMAP and project analysis are publicly available on Github.[46,54] Data harmonization procedures, covariate definitions, and validation of outcomes are described in the Methods to enable replication in other healthcare systems with EHR data.

RESULTS

AD diagnosis phenotyping algorithm performance

Validation with chart-reviewed labels

The demographic profiles of the randomly selected patients for chart review at UPMC (64% women, 87% NHW, mean[SD]=81.55[9.00] years at AD diagnosis) and MGB (55% women, 90% NHW, mean[SD]=79.18[10.43] years at AD diagnosis) were similar to the larger UPMC (64% women, 92% NHW, mean[SD]=80.56[9.04] years at AD diagnosis) and MGB (59% women, 89% NHW, mean[SD]=78.63[9.58] years at AD diagnosis) AD-EHR cohorts (**S-Table 3**). Using chart-reviewed labels 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-Table 4, S-Table 5, S-Table 6, S-Table 7**). At 90% specificity, the sensitivity and PPV were reasonably balanced at both UPMC and MGB, similar to prior phenotyping studies (**S-Table 4**).[41] (Neither AUROC nor AUPRC changed when altering the specificity thresholds.) 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). AUPRC, which accounts for imbalance in the sample size, was higher among women than men at both UPMC (AUPRC: women=0.866, men=0.709) and MGB (AUPRC: women=0.926, men=0.868).

There was an inconsistent pattern in AUPRC between two sites for race and ethnicity groups (UPMC: NHW=0.843, other=0.732; MGB: NHW=0.903, other=1.000).

Validation with registry-derived labels

Using registry-derived labels to evaluate algorithm performance at UPMC (n=1916, 47% AD, AUROC=0.835, **S-Table 8**), 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).

Comparison to benchmark phenotyping methods

KOMAP performance evaluated with both gold-standard chart-reviewed and registry-derived labels outperformed all the benchmark methods constructed from surrogate features of AD diagnosis in terms of AUROC and AUPRC, demonstrating the advantage of incorporating information from knowledge graph-selected features using KOMAP (**S-Table 6, S-Table 7, S-Table 8**). Training KOMAP separately for each demographic group (KOMAP separate) resulted in similar performance. KOMAP performance in the sensitivity analysis AD cohorts (without imputation of missing race and ethnicity) evaluated using chart-reviewed labels was consistent with the main AD cohorts (**S-Table 9**).

Validation of nursing home admission status definition

We used the definition of ≥1 code *or* CUI for nursing home admission in the main analysis since this definition outperformed ≥1 code for nursing home admission at both UPMC (≥1 code: sensitivity=0.493, specificity=0.968, PPV=0.971, NPV=0.462; ≥1 code *or* CUI: sensitivity=0.913, specificity=0.903, PPV=0.955, NPV=0.824, **S-Table 10**) and MGB (≥1 code: sensitivity=0.709, specificity=1, PPV=1, NPV=0.738; ≥1 code or CUI: sensitivity=0.945, specificity=0.911, PPV=0.929, NPV=0.932). The definition of ≥1 code *or* CUI greatly increased the sensitivity and NPV over the definition of ≥1 code, while slightly decreasing the specificity and PPV.

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, **Supplementary Data 3**), 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 (**Supplementary Data 3**).

In the combined AD cohort, demographic groups differed in patient profiles (**Supplementary Data 3**). 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=7.819X10^{-35}; NHW=79.67[9.32], other=78.11[9.86], p=8.869X10^{-20}). 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=2.411X10^{-}

^{14}; NHW=198.12[230.91], other=245.77[308.67], p=1.539X10^{-21}). 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 (**Supplementary Data 3, 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=2.987X10^{-5}; NHW=19,555[74%], other=1,896[67%], p=1.609X10^{-16}, **Supplementary Data 3**). 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=5.740X10^{-15}; NHW=12,878[49%], other=1,405[49%], p=.515; n[%] death: women=8,963[50%], men=6,185[54%], p=2.632X10^{-12}; NHW=13,871[53%], other=1,277[45%], p=1.443X10^{-14}). In sensitivity analysis (without imputation of missing race and ethnicity), patient profiles were comparable to the main AD cohort (**S-Table 11**).

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=1.203X10^{-3}, **Figure 3, S-Table 12**), while there was no significant differences between race and ethnicity groups (HR [95% CI]=1.006 [0.952-

1.063], p=8.306X10^{-1}). Second, NHW individuals had a higher risk of death than racial and ethnic minorities (HR [95% CI]=1.376 [1.245-1.521], p=4.084X10^{-10}, **S-Table 13**), while women had a lower risk than men (HR [95% CI]=0.856 [0.811-0.904], p=2.434X10^{-8}).

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-Figure 3, S-Figure 4, S-Figure 5, S-Figure 6**, **S-Table 12**, **S-Table 13**).

Sensitivity analyses

AD cohort characteristics in the sensitivity analyses

In the first sensitivity analysis using the van Walraven weighted Elixhauser score (instead of individual comorbidities) as a covariate, we included all patients from the main combined AD cohort (*with* imputation of missing race and ethnicity, **Supplementary Data 3**). In the second sensitivity analysis, the combined AD cohort (*without* imputation of missing race and ethnicity) comprised 24,075 patients (61% women, 90% NHW, mean[SD]=79.55[9.45] years at AD diagnosis, **S-Table 11**), including 12,938 UPMC patients (64% women, 92% NHW, 80.58[9.01] years at AD diagnosis) and 11,137 MGB patients (59% women, 88% NHW, 78.35[9.81] years at AD diagnosis). Similar to the main AD cohorts, when 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 (**S-Figure 2**), greater proportion of nursing home admission and death during study follow-up, and higher proportion of AD-related medication prescriptions (particularly before AD diagnosis).

Differential risk of nursing home admission and death

For sensitivity analyses, we again used fixed-effects meta-analysis of covariate-adjusted Cox proportional hazard models to compare the risk of two readily ascertainable clinical outcomes among demographic groups.

In the first sensitivity analysis when using the van Walraven weighted Elixhauser score as covariate, the results were consistent results with the main findings. Women had a higher risk of nursing home admission than men (HR [95% CI]=1.084 [1.047-1.122], *p*=4.299X10^{-6}, **Figure 3, S-Table 12, S-Figure 3**, **S-Figure 4**), but there was no significant difference between racial and ethnic groups (HR [95% CI]=0.962 [0.912-1.016], *p*=1.626X10^{-1}). Women had a lower risk of death than men and NHW individuals had a higher risk of death than racial and ethnic minorities (women HR [95% CI]=0.860 [0.816-0.907], *p*=2.239X10^{-8}; NHW HR [95% CI]=1.360 [1.233-1.501], *p*=8.247X10^{-10}). Consistent with the main findings, higher Elixhauser scores (indicative of greater cumulative pre-existing comorbidity burden) were associated with increased risk of both nursing home admission and death (nursing home admission HR [95% CI]=1.008 [1.006-1.010], *p*=3.125X10^{-12}; death HR [95% CI]=1.026 [1.023-1.030], *p*=1.284X10^{-44}, **Figure 4**, **S-Figure 4**).

In the second sensitivity analysis, the combined AD cohort (*without* imputation of missing race and ethnicity) showed consistent results as the main combined AD cohort (*with* imputation of missing race and ethnicity). Women had a higher risk of nursing home admission than men (HR [95% CI]=1.065 [1.025-1.107], *p*=1.261X10^{-3}, **Figure 3, S-Table 12**, **S-Figure 5**, **S-Figure 6**), but there was no significant difference between racial and ethnic groups (HR [95% CI]=1.009 [0.952-1.069], *p*=7.674X10^{-1}). Women had a lower risk of death than men and NHW individuals had a higher risk of death than racial and ethnic minorities (women HR [95% CI]=0.823 [0.769-0.882], *p*=2.588X10^{-8}; NHW HR [95% CI]=1.458 [1.283-1.657], *p*=6.871X10^{-9}, **Figure 3**, **S-Table 13**, **S-Figure 5**, **S-Figure 6**). Consistent with the main combined AD cohort (*with* imputation of missing race and ethnicity), the key drivers of nursing home admission and death risk in the combined sensitivity analysis AD cohort (*without* imputation of missing race and ethnicity) included older age at AD diagnosis and greater pre-existing comorbidity burden (**S-Figure 6**).

DISCUSSION

We pursued two complementary goals: we accurately identified AD-EHR populations from two different healthcare systems and re-examined previously reported demographic differences and systematically assessed potentially modifiable drivers of AD decline in these well-defined populations. We identified AD populations using an unsupervised phenotyping algorithm (KOMAP), which achieved *overall* robust predictive performance with modest differences across healthcare systems and demographic groups. Demographic groups in the AD-EHR cohorts exhibited differential risks of clinical decline. Women had a higher risk of nursing home admission, while men and NHW individuals 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.

KOMAP demonstrated significantly better performance than rule-based methods, outperforming all the benchmark methods based on surrogate features of AD diagnosis (*i.e.*, main AD PheCode, main dementia PheCode, main AD CUI). Specifically, the performance of the rule-based approaches in this study using surrogate features was comparable to previously published studies of AD cohort identification.[25,26,32] KOMAP, while performing well overall, showed modest differences in healthcare system- and demographic group-specific performances, all of which still achieved reasonable AUROCs and AURPCs. To avoid co-training, we trained KOMAP independently at UPMC and MGB. KOMAP performed better at MGB compared to UPMC, potentially due to differences in clinical practices, documentation standards, population characteristics, and EHR data quality between the two healthcare systems. At both healthcare systems, KOMAP performed better among women than men, possibly attributable to women having more available EHR data for algorithm than men for several related reasons: (1) women had higher incidence and prevalence of AD;[1–3,36,55] (2) the real-world AD cohorts had older age at AD diagnosis than traditional epidemiological studies;[56,57] (3) women survived longer than men, including after AD diagnosis.[4,5] Men with AD might also have greater overlap with other related dementias.[58] There was no consistent pattern of performance differences by race and ethnicity. KOMAP performed equally well among racial and ethnic groups at UPMC, while KOMAP in racial and ethnic minorities outperformed NHW individuals at MGB. The different patterns of healthcare utilization between UPMC and MGB populations might be contributory. KOMAP deployment for AD diagnosis in future healthcare systems will require healthcare system-specific validation. Despite modest algorithm performance differences, the final combined AD cohort will be valuable for generating real-world evidence, particularly across demographic groups.

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.[4,5,36,59] 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.[58] 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] NHW patients had a higher risk of death than racial and ethnic minorities, mirroring the broader pattern in "mortality crossover" among older adults.[8–10,16,60] As a possible explanation, racial and ethnic minorities more likely remained undiagnosed, underestimating their AD-related mortality.[61] This study also cannot address whether biological factors drive differences in clinical outcomes across demographic groups.[62]

We found older age at AD diagnosis and greater pre-existing comorbidity burden increased the risk of *both* nursing home admission and death.[11–16,63] 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. Among comorbidities, psychiatric and cardiometabolic disorders were key drivers of AD decline. Psychosis was the strongest predictor of nursing home admission while also increasing the risk of death. These findings are consistent with prior epidemiological and claims-based studies,[64,65] but we captured psychosis through routine clinical care documentation (rather than neuropsychiatric assessments). Psychotic symptoms (*e.g.,* delusions, hallucinations) create safety concerns and care needs that often exceed the capacity of family caregivers, increasing the likelihood of nursing home admission. While chronic cardiovascular risk factors (*e.g.,* hypertension, diabetes) increased the risk of both nursing home admission and death, acute cardiovascular disease (*e.g.,* congestive heart failure, cardiac arrhythmia) increased risk of mortality.[11–16] Finally, 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 KOMAP performance with gold-standard labels in two independent healthcare systems and across demographic groups.[25–30,32–37] KOMAP had the following innovations that enhanced its application when compared to prior methods:[25–30,32–37] (1) it leveraged pre-trained, publicly available knowledge graphs of interconnected EHR concepts to rapidly select informative features, a strategy superior to reliance on literature review or expert curation of features while effectively reducing feature dimensionality; and (2) it did not require gold-standard labels for training, though we used gold-standard labels to provide a highly reliable estimate of algorithm performance, unlike prior studies that relied on clinical assessment, chart review, or surrogate EHR indicators for validation. The overall robust performance of KOMAP 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 of 29,262 patients 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. This study represents one of the largest EHR-based investigations of AD clinical outcomes using accurately identified AD population. Importantly, these real-world AD cohorts are more representative of real-world 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.[35,66–68] 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 simultaneously assessed multiple comorbidities using the ECI, a well-validated predictor of functional decline and death. We quantified the extent to which the potentially modifiable factors included in the ECI drove differences in AD decline.

Our study also has limitations. First, we examined differences in AD decline by broader categories of race and ethnicity (NHW, other racial and ethnic minorities), but could not examine differences across individual racial and ethnic minority groups (*e.g.,* American Indian, Asian, Black, Hispanic) due to the modest sample size of each. While statistically necessary, this approach might obscure meaningful 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,[69] we are developing 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.[70]

CONCLUSIONS

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.

AUTHOR CONTRIBUTIONS

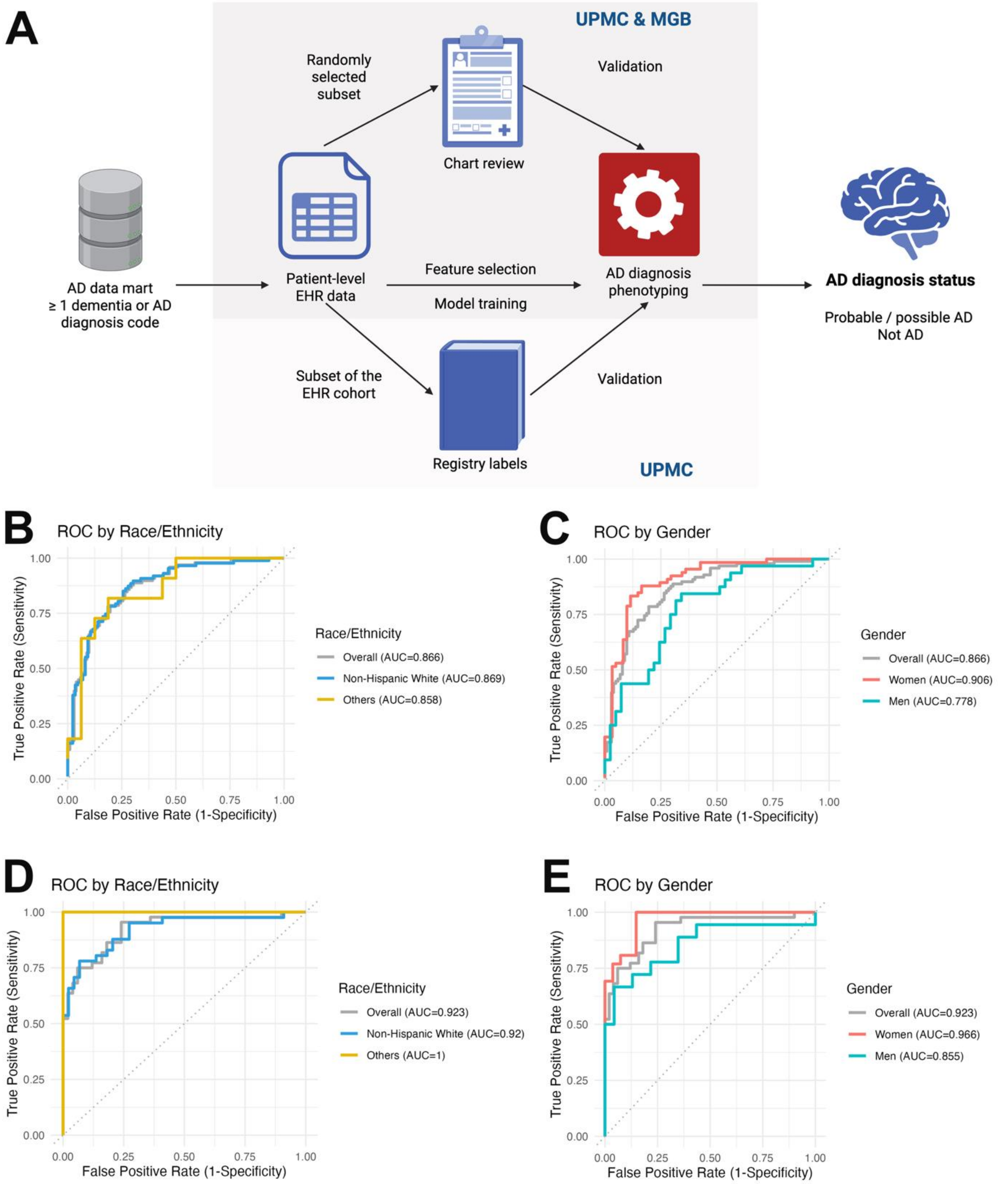
Shruthi Venkatesh, Linshanshan Wang, Michele Morris, Mohammed Moro, Ratnam Srivastava, Yunqing Han, Riddhi Patira, Sarah Berman, Oscar Lopez, Shyam Visweswaran, Tianrun Cai, Tianxi Cai, and Zongqi Xia contributed to the design and conceptualization of the study. Shruthi Venkatesh and Linshanshan Wang contributed equally as co-first authors to data analysis and manuscript writing. Michele Morris, Mohammed Moro, Ratnam Srivastava, Yunqing Han, Riddhi Patira, Sarah Berman, Oscar Lopez, Shyam Visweswaran, Tianrun Cai, Tianxi Cai, and Zongqi Xia contributed to data acquisition and manuscript writing. Tianxi Cai and Zongqi Xia contributed equally as co-senior authors and jointly supervised this work. All authors have reviewed and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

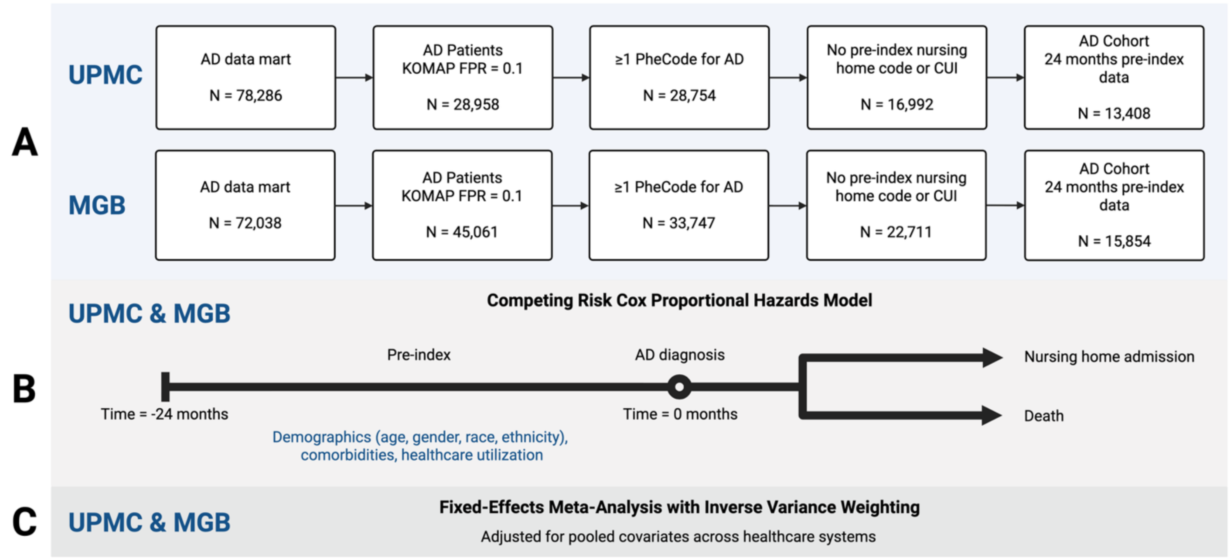
FIGURES

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



a. This panel was created in BioRender.[71]

Figure 2. Study design.[a] (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.



a. This figure was created in BioRender.[72]

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. Solid lines represent pooled survival probability estimates (average across healthcare systems) from the fixed-effects meta-analysis. Shaded regions represent 95% confidence intervals. 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. We used two-tailed t-tests; p-values are not adjusted for multiple comparisons since all covariates were evaluated simultaneously in a single multivariable model. (A-B) Nursing Home Admission (n=29,262 patients; 14,283 events). (C-D) Death (n=29,262 patients; 15,148 events). Reference group for race and ethnicity: non-Hispanic White. Reference group for gender: women.

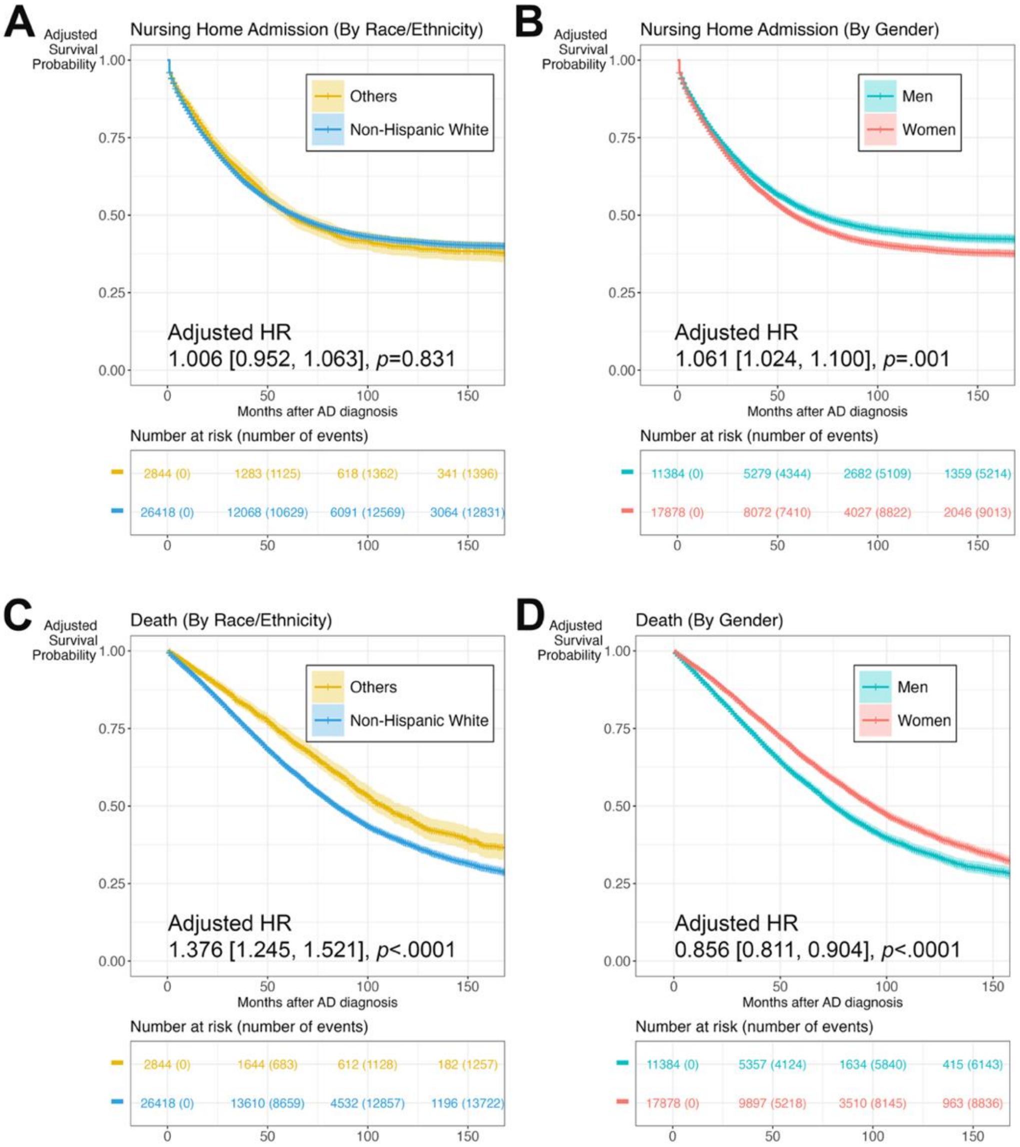
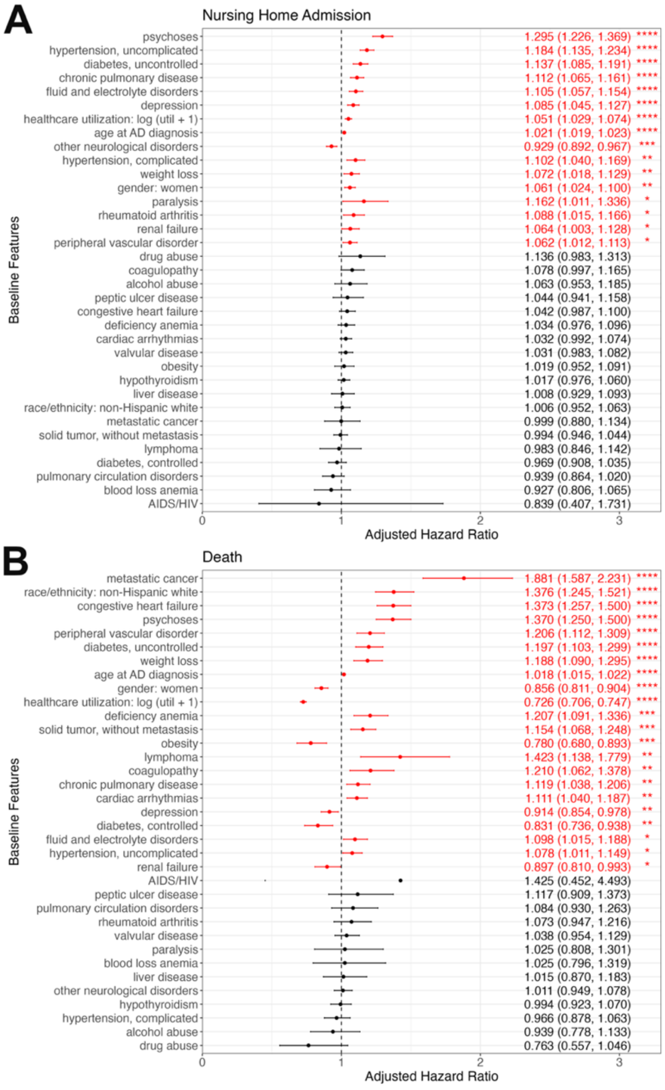


Figure 4. Adjusted hazard ratios of variables in the fixed-effects meta-analysis of healthcare system-specific covariate-adjusted Cox proportional hazard models of nursing home admission and death. Points represent adjusted hazard ratio estimates (measure of centre). Horizontal lines represent 95% confidence intervals of these estimates. The vertical dashed line at hazard ratio = 1.0 indicates no effect. 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. We used two-tailed t-tests; p-values are not adjusted for multiple comparisons since all covariates were evaluated simultaneously in a single multivariable model. (A) Nursing Home Admission (n=29,262 patients; 14,283 events). (B) Death (n=29,262 patients; 15,148 events).



P-value significance threshold: *p*<.05*, *p*<.01**, *p*< .001***, and *p*<.0001****

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. 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

27. 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

28. 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

29. 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

30. 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

31. Zhang Q, Coury R, Tang W. Prediction of conversion from mild cognitive impairment to Alzheimer's disease and simultaneous feature selection and grouping using Medicaid claim data. *Alzheimer's Res Ther*. 2024;16(1):54. doi:10.1186/s13195-024-01421-y

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. 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

36. 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

37. 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

38. 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

39. 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

40. 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

41. 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

42. 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

43. 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

44. 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

45. 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

46. 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

47. Wang L, Venkatesh S, Morris M, et al. Stratification of Alzheimer's Disease Patients Using Knowledge-Guided Unsupervised Latent Factor Clustering with Electronic Health Record Data. *medRxiv*. Published online 2024:2024.12.23.24319588. doi:10.1101/2024.12.23.24319588

48. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement*. 2011;7(3):263-269. doi:10.1016/j.jalz.2011.03.005

49. Pinker E. Reporting accuracy of rare event classifiers. *npj Digit Med*. 2018;1(1):56. doi:10.1038/s41746-018-0062-0

50. 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

51. 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

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

53. 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

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

55. Yan BW, Arias E, Geller AC, Miller DR, Kochanek KD, Koh HK. Widening Gender Gap in Life Expectancy in the US, 2010-2021. *JAMA Intern Med*. 2024;184(1):108-110. doi:10.1001/jamainternmed.2023.6041

56. Rentería MA, Mobley TM, Evangelista ND, et al. Representativeness of samples enrolled in Alzheimer's disease research centers. *Alzheimer's Dement: Diagn, Assess Dis Monit*. 2023;15(2):e12450. doi:10.1002/dad2.12450

57. Gianattasio KZ, Bennett EE, Wei J, et al. Generalizability of findings from a clinical sample to a community-based sample: A comparison of ADNI and ARIC. *Alzheimer's Dement*. 2021;17(8):1265-1276. doi:10.1002/alz.12293

58. 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

59. 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

60. 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

61. 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

62. 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

63. 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

64. Miller EA, Schneider LS, Rosenheck RA. Predictors of nursing home admission among Alzheimer's disease patients with psychosis and/or agitation. *Int Psychogeriatr*. 2011;23(1):44-53. doi:10.1017/s1041610210000244

65. Forns J, Danysh HE, McQuay LJ, et al. Clinical outcomes and treatment patterns of older adults with dementia-related psychosis by dementia type in the United States. *BMC Geriatr*. 2022;22(1):784. doi:10.1186/s12877-022-03489-3

66. 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

67. 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

68. 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

69. 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

70. 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

71. Venkatesh S. *Figure 1A. Created in BioRender.*; 2025. https://BioRender.com/xwiof9h

72. Venkatesh S. *Figure 2. Created in BioRender.*; 2025. https://BioRender.com/58sbkky

Editorial Summary:
Venkatesh, Wang et al. identify Alzheimer's disease populations from electronic health records and examine demographic differences and modifiable drivers of nursing home admission and death risk. Women have higher nursing home admission risk while non-Hispanic white individuals and men have higher death risk, with comorbidities as key drivers.

Peer review information: *Communications Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work.