Stratification of Alzheimer's Disease Patients Using Knowledge-Guided Unsupervised Latent Factor Clustering with Electronic Health Record Data

Linshanshan Wang, MS[1,2] *, Shruthi Venkatesh, BS[3] *, Michele Morris, BS[4], Mengyan Li, PhD[5], Ratnam Srivastava, MS[2], Shyam Visweswaran, MD, PhD[4], Oscar Lopez, MD[3], Zongqi Xia, MD, PhD[3] #, Tianxi Cai, ScD[1,2] #

1. Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA
2. Department of Biomedical Informatics, Harvard Medical School, Boston, MA
3. Department of Neurology, University of Pittsburgh, Pittsburgh, PA
4. Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
5. Department of Mathematical Sciences, Bentley University, Waltham, MA

*** Share co-first authorship**

 # Share co-senior authorship

Corresponding Authors:

Zongqi Xia zxia1@post.harvard.edu

Tianxi Cai tcai.hsph@gmail.com

SHORT ABSTRACT

Prognostication for people with Alzheimer's disease (AD) at the point of care could improve clinical management. Applying a novel unsupervised latent factor clustering approach guided by knowledge graph embeddings of relevant clinical features from electronic health records, we stratified 16,411 AD patients into two groups at diagnosis and prognosticated their risk of AD-related outcomes (*i.e.,* nursing home admission, mortality), adjusting for baseline confounders. To reflect real-world evolution in clinical trajectories, we updated patient stratification for 12,606 AD patients remaining at risk 1-year post-diagnosis and repeated prognostication. At both timepoints, one group had a higher nursing home admission risk and exhibited characteristics suggesting greater symptom burden, but the mortality risk remained comparable between groups. This study supports that patient stratification can enable outcome prognosis for AD patients. While baseline prognostication can guide early treatment and tailored management, dynamic prognostication may inform more timely interventions to improve long-term outcomes.

ABSTRACT

Background: People with Alzheimer's disease (AD) exhibit varying clinical trajectories. There is a need to predict future AD-related outcomes such as morbidity and mortality using clinical profile at the point of care.

Objective: To stratify AD patients based on baseline clinical profiles (up to two years prior to AD diagnosis) and update the model after AD diagnosis to prognosticate future AD-related outcomes.

Methods: Using the electronic health record (EHR) data of a large healthcare system (2011-2022), we first identified patients with ≥1 diagnosis code for AD or related dementia and applied a validated unsupervised phenotyping algorithm to assign AD diagnosis status. Next, we applied an unsupervised latent factor clustering approach, guided by knowledge graph embeddings of relevant EHR features up to the baseline, to cluster patients into two groups at AD diagnosis. We then prognosticated the risk of two readily ascertainable and clinically relevant AD-related outcomes (*i.e.,* nursing home admission indicating greater need for assistance and mortality), adjusting for baseline confounders (*e.g.,* age, gender, race, ethnicity, healthcare utilization, and comorbidities). For patients remaining at risk one year post-diagnosis, we updated their group membership and repeated the prognostication.

Results: We stratified 16,411 algorithm-identified AD patients into two groups based on their baseline clinical profiles (41% Group 1, 59% Group 2). Patients in Group 1 were marginally older at AD diagnosis (age Mean [SD]: 81.4 [9.3] vs 81.0 [8.7], p=.007), exhibited greater comorbidity burden (Elixhauser comorbidity index Mean [SD]: 11.3 [10.3] vs 7.5 [8.6], p<.0001), and more frequently received AD-related medications (47.7% vs 40.9%, p<.0001) than those in Group 2. Compared to Group 1, Group 2 had a lower risk of nursing home admission (HR [95% CI]=0.804 [0.765, 0.844], p<.001), while the two groups had similar mortality risk (HR [95%

CI]=1.008 [0.963, 1.056], *p*=.733). One year after AD diagnosis, 12,606 patients remained at risk (45.7% Group 1, 54.3% Group 2). Consistent with baseline findings, Group 2 had a lower risk of nursing home admission than (HR [95% CI]=0.815 [0.766, 0.868], *p*<.001) and similar mortality risk as (HR [95% CI]=0.977 [0.922, 1.035], *p*=0.430) Group 1 in the updated model.

Conclusions: It is feasible to stratify patients based on readily available clinical profiles before AD diagnosis and crucially to update the model one year after diagnosis to effectively prognosticate future AD-related outcomes.

INTRODUCTION

Alzheimer's disease (AD) is the leading cause of dementia and neurological disability in the aging population, causing high societal burdens [1–3]. As promising treatment options emerge and given their potential tradeoff (*e.g.,* adverse events, cost) [4,5], there is a heightened interest in patient stratification and prognostication at the point of care. Multifactorial contributors of the varying patient trajectories pose a major challenge to effective clinical prediction [6–11].

The increasing availability of large-scale clinical datasets and advances in machine learning (ML) algorithms enable data-driven investigations of heterogeneity in AD clinical trajectories. Unsupervised ML techniques such as clustering can stratify AD patient populations into more homogeneous groups based on latent clinical features relevant to AD. Iterative patient stratifications using baseline and subsequently evolving clinical features can inform contributors of differential risks in clinical trajectories, and pragmatically improve clinical prediction and ultimately treatment guidance [12].

Prior AD clustering studies used genetics [13], biomarkers [14–16], neuropathology neuroimaging [17–22], and neuropsychological data [16,23,24]. However, these studies were limited in follow-up duration, scalability, generalizability and practicality. Electronic health records (EHRs) contain large-scale, real-world, longitudinal data for clinical investigations of AD. Several studies have leveraged EHR data to cluster AD patients for prediction [25–29]. Two studies clustered patients based on symptomatic presentation and comorbid conditions preceding AD onset for predicting time to AD onset [25,26]. Two other studies clustered patients based on pre-selected lists of features for predicting AD outcomes such as mortality [27,28], but pre-selected clinical features lack scalability and generalizability. Patient clustering can be enhanced by automated selection of clinically informative features based on data-derived prior knowledge that incorporates

complementary features embedded in the EHR such as comorbid conditions, medication prescriptions, test results, and clinical narratives.

Here, we aim to stratify AD patients at and one year after diagnosis to prognosticate readily ascertainable AD-related outcomes of nursing home admission and mortality, using baseline clinical features. Leveraging previously trained knowledge graph embeddings of EHR concepts, we automated the selection of AD-relevant clinical features to guide a latent class model of patient clustering [30]. Patient cluster memberships based on clinical profile before AD diagnosis and updated cluster membership one year post-diagnosis predicted future AD outcomes.

METHODS

Ethics approval

The Institutional Review Board of the University of Pittsburgh (STUDY21020153) approved the study protocol.

AD data mart

We obtained inpatient and outpatient EHR data for years 2011 to 2022 from the UPMC healthcare system. Codified EHR data contain demographics (*e.g.*, age, gender, race, ethnicity), diagnosis and procedure codes (*e.g.*, International Classification of Disease [ICD] code, 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 test results (*e.g.*, Logical Observation Identifiers Names and Codes [LOINC] codes). To systematically organize diagnoses, we mapped all ICD-9 and ICD-10 codes to PheCodes [31,32], which are manually curated to capture clinically relevant concepts for research. We consolidated procedure codes (CPT) using the Clinical Classifications Software (CCS) for Services and

Procedures (CCS-Services and Procedures), a tool developed by the Agency for Healthcare Research and Quality (AHRQ) that categorizes medical procedures into clinically meaningful groups.

Narrative EHR data comprise free-form text information from clinical encounters (*e.g.,* physician office visit notes, discharge summaries) that were converted to structured data by natural language processing (NLP). Using codified and narrative EHR data, we built an AD data mart comprising patients with at least one 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**). To obtain standardized clinical concepts from narrative EHR data, we deployed a previously validated NLP pipeline (Narrative Information Linear Extraction [NILE]) to generate concept unique identifiers (CUIs) according to unified medical language system (UMLS) [35,36].

AD cohort

We applied a phenotyping algorithm, Knowledge driven Online Multimodal Automated Phenotyping (KOMAP) [30], to assign AD diagnosis status (probable or possible AD vs not AD) for all patients in the AD data mart (**Figure 1A**). We combined probable and possible AD into a single category for prediction of AD diagnosis status as they represent heterogenous presentations of AD. KOMAP is a two-step unsupervised algorithm that enables retrieval of diagnosis labels from EHR data. First, 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 AD. Second, we trained the KOMAP algorithm based on ONCE-selected features. We obtained the AD cohort by including patients assigned by KOMAP as having probable or possible AD at 90% specificity. We included patients with at least one AD PheCode and 24 months of EHR data preceding the index date (date of the first AD PheCode, which was operationally defined as the

date of AD diagnosis). We excluded patients who had at least one code for nursing home admission preceding AD diagnosis. We imputed missing demographics (*e.g.,* race and ethnicity) by performing multiple imputation based on age and gender [37,38].

Clustering patients at AD diagnosis

Feature selection and knowledge graph embeddings

To cluster patients at AD diagnosis, we selected all AD-related codified and narrative features using ONCE and aggregated counts of EHR features for the 24 months preceding the index date [30]. We obtained embeddings of selected concepts, denoted by E, through multi-source knowledge graph representation learning. This knowledge graph leverages the pre-trained ONCE embeddings of large scale EHR concepts learned from summary level EHR data at two independent healthcare systems (1,25,00,000 Veterans Health Administration patients and 60,000 Mass General Brigham Biobank patients) and the textual descriptions of the concepts embedded using CODER pre-trained language model [39]. Importantly, this multi-source representation of EHR concepts contains rich information on the concept-concept relationships, which are generalizable across diseases and health systems.

Identifying clusters and model specification

To identify the underlying patient clusters, we adopted the multivariate Poisson-LogNormal Mixture Model (PLMM) to jointly model the EHR feature counts. We set the number of clusters to two before fitting the model to represent a low-risk and a high-risk group, dividing patients into two groups (Group 1 and Group 2) at AD diagnosis. We specified the model as:

$$\boldsymbol{X}_i \mid \boldsymbol{Z}_i \sim Poisson(e^{\boldsymbol{Z}_i}), \quad X_{ij} \mid Z_{ij} \ indep. \quad ,$$
$$\boldsymbol{Z}_i \mid Y_i, \boldsymbol{U}_i = \boldsymbol{V}\left\{[(1, \boldsymbol{U}_i^{\top}, Y_i)\boldsymbol{B}]^{\top} + \boldsymbol{W}_i\right\} \quad ,$$
$$Y_i \mid \boldsymbol{U}_i \sim Ber(\pi_i), \quad \pi_i = \mathrm{expit}(a + \boldsymbol{b}^{\top}\boldsymbol{U}_i)$$
$$\boldsymbol{W}_i \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{\Lambda})$$

In this model, the observed variable X_{ij} represents the counts of *j* th feature (*e.g.,* PheCode, NLP mentions) for the *i* th patient. U_i denotes the healthcare utilization (calculated as the total number of clinical encounters). $Y_i = \{0, 1\}$ is a latent variable denoting the two underlying patient clusters, where each cluster is assumed to have differential distributions of feature counts and downstream AD-related outcomes. $Z_i \in R^p$ represents the latent patient-level embedding for *i* th patient, with each cluster having a different mean. A graphical representation of the model is shown in **S-Figure 1**.

Unlike previous ecological and genetics studies that use PLMM, we introduce an additional assumption: the observed EHR counts are influenced by a small number of latent factors. These latent factors capture underlying patterns of similarity (hierarchical relationships, *e.g.,* "PheCode:290", "PheCode:290.1" and "PheCode:290.11") and relatedness (*e.g.,* "associated with", "may treat", "co-occur with" etc.) among EHR features, reflecting shared relationships or common processes in the data. Further, we incorporate knowledge graph embeddings, which provide rich contextual information about the relationships between concepts, to guide the identification of these factors. The latent factors serve as a bridge, summarizing complex, high-dimensional data into a simpler structure that still retains key information. Model parameters B_{2j} can be interpreted as the effect of the *j* th factor on cluster membership, after adjustment for healthcare utilization. The free model parameters Λ can then be viewed as a

fine-tuning parameter to allow for re-weighting of each factor based on the observed data. See supplementary methods for details on model training.

Clustering patients one year after AD diagnosis

To illustrate the capacity of the approach to account for potential changes in patient profile that could influence trajectory, we updated the clustering model one year post-diagnosis for the patients remaining at risk (**Figure 1C**). The patients that remained at risk included those who were: (1) not admitted to nursing home, (2) not censored, (3) alive one year post-diagnosis, and (4) assigned to Group 2 at baseline. We reclustered the patients into Groups 1 and 2 leveraging 24 months of pre-index data and 12 months of post-index data. The updated Group 1 comprised patients that were reassigned from Group 2 to Group 1 one year post-diagnosis and patients in Group 1 at AD diagnosis. The updated Group 2 comprised patients that continued to be assigned to Group 2 one year post-diagnosis.

AD-related outcomes

We assessed two AD-related outcomes as endpoints readily ascertainable from EHR data: time-to-nursing home admission and time-to-death. Admission to a nursing home, which indicates loss of functional independence, is routinely documented in clinical practice and EHR [6,7]. We defined nursing home admission as having at least one ICD code for admission to any residential institution (*e.g.,* assisted living facility, long term care facility, end-of-life care, see **S-Table 2** for list of nursing home codes). Death status in the UPMC EHR data is linked to the social security death index. For patients that died during the follow-up period, we calculated the time-to-death from the index date. For patients that did not progress to either endpoint, we utilized the data available till the last record of clinical visit.

Validation analyses

Characterizing the AD cohort

We report baseline demographics, clinical characteristics, pre-existing comorbidity burden, and prescription of AD-related medications (*e.g.,* donepezil, memantine, see **S-Table 3** for full list of medications) during the follow-up period stratified by model-predicted clusters. We assessed the pre-existing comorbidity burden using the Elixhauser Comorbidity Index (ECI). Using the comorbidity package, we mapped the diagnosis codes for each patient to 29 pre-existing health conditions comprising the ECI [40]. Each comorbidity in the ECI was used as an individual feature in downstream analyses. We used two-tailed t-tests to evaluate differences in baseline characteristics between clusters. To examine the differences in pre-existing comorbidity burden by predicted clusters, we calculated the proportion of patients with each comorbidity and used two-proportion Z-tests to evaluate the group difference.

Prognostication of AD-related outcomes at time of diagnosis and 1-year post diagnosis

We examined the association between cluster membership at diagnosis and AD-related outcomes in competing risk Cox proportional hazard models. We assessed the time-to-nursing home admission alive using a competing risk model to control for death. Further, we compared the 2-year, 5-year, and 10-year rate of nursing home admission and death between clusters at diagnosis. We performed the same analyses after updating cluster membership one year post-diagnosis.

Feature importance

We examined the salient EHR features that distinguish patients in each baseline and updated cluster. We first obtained the counts of each AD-related feature (**S-Table 6, S-Table 7**). Next, for each feature, we regressed the log-transformed (x -> log(x+1)) feature count against

log-transformed healthcare utilization, and calculated the *adjusted intensity* as the mean of the residuals in each cluster. We obtained a feature importance score for each feature as the difference in adjusted intensity between the two clusters. This score reflects the difference in count frequencies between the clusters while accounting for the variation in healthcare utilization across patients.

Software availability

The code to train and apply the PLMM is implemented in R and C++ and is publicly available at https://github.com/wanglss-2024/ADstratification, and the multi-source knowledge graph embeddings is at https://shiny.parse-health.org/ONCE/ [30].

RESULTS

Prediction of outcomes at AD diagnosis

Demographic characteristics

The overall AD cohort comprised 16,411 patients with mean age 81.2±9.0 years at diagnosis (**Table 1**). Using the baseline clinical profile, the model clustered the AD cohort into two groups at diagnosis (Group 1: 6,734, 41.0%; Group 2: 9,677, 59.0%). Group 1 was marginally older at diagnosis (age at AD diagnosis Mean [SD]: Group 1: 81.4 [9.3], Group 2: 81.0 [8.7], p=.007). While the proportion of men and women was comparable between groups (Group 1: 4,314, 64.1%; Group 2: 6,245, 64.5%; p=.546), the proportion of non-Hispanic White individuals was higher in Group 2 (Group 1: 6,138, 91.1%; Group 2: 8,967, 92.7%; p=.014). The overall comorbidity burden was higher in Group 1 (ECI, Mean [SD]: Group 1: 11.3 [10.3]; Group 2: 7.5 [8.6]; p<.0001), with hypertension, cardiac arrhythmias, depression, and fluid and electrolyte disorders being the most prevalent comorbidities (**Table 1, Figure 2**).

In the overall AD cohort, 12,702 patients (77.4%) were prescribed AD-related medications (**Table 1**): 7,166 (43.7%) preceding diagnosis and 5,536 (33.7%) after diagnosis. A greater proportion of patients in Group 1 were prescribed AD-related medications than Group 2 (Group 1: 3,212, 47.7%; Group 2: 3,954, 40.9%; $p<.0001$), suggesting higher symptom severity in Group 1. Group 1 also had a shorter follow-up duration (median [IQR]: Group 1: 62 [35, 85], Group 2: 86 [59, 125], $p<.0001$), suggesting that individuals in Group 1 progressed to the endpoints (nursing home admission or death) faster than individuals in Group 2.

Time to AD-related outcomes

We compared the rates of nursing home admission and mortality at the 2-year, 5-year, and 10-year milestone. While Group 1 had a higher 2-year and 5-year nursing home admission rate, the groups did not differ in the 10-year nursing home admission rate (**Table 2**). Group 1 also had a higher 2-year and 10-year mortality rate than Group 2, but there was no difference in the 5-year mortality rate between the groups.

In survival analysis after adjusting for baseline covariates (age, gender, race, ethnicity, healthcare utilization, and comorbidities), Group 2 had a significantly lower risk of nursing home admission (HR [95% CI]=0.804 [0.765, 0.844], $p<.001$, **Figure 3**) than Group 1. However, the risk of mortality was comparable among the groups (HR [95% CI]=1.008 [0.963, 1.056], $p=0.733$) after covariate adjustment.

Feature importance

We identified features with the largest absolute feature importance score, *i.e.,* largest differences of adjusted intensity between the groups. Features with high adjusted intensity are characteristic of a given group. Group 1 had high adjusted intensity of dementia-related diagnoses (*e.g.,* dementias, delirium dementia and amnestic and other cognitive disorders), prescriptions of AD-related medication (*e.g.,* donepezil, memantine, and rivastigmine), vascular comorbidities (*e.g.,* cerebrovascular disease, cerebrovascular disorders, cerebral ischemia,

cerebral artery occlusion, cerebrovascular accident, late effects of cerebrovascular disease), neurological disorders, and epilepsy (**Figure 4A**). Group 2 had high adjusted intensity of signs and symptoms of dementia (*e.g.,* confusion, mental recall), general symptoms of decline that may indicate an increased need for care (*e.g.,* asthenia, weight decreased, malaise and fatigue, clinical decision making), factors pertaining to living situation (*e.g.,* at home, household composition, living alone), vitamin B12 / cyanocobalamin prescriptions, certain comorbidities (*e.g.,* vertigo), and medications used to treat these comorbidities (*e.g.,* amlodipine for treating hypertension, sertraline for treating depression).

We further examined the correlation between the residualized features using hierarchical clustering (**Figure 4B**). Related concepts (*e.g.,* vascular comorbidities, factors pertaining to living situation) were often correlated with each other. The features associated with each group were strongly correlated with other features in the same group.

Prediction of outomes one year after AD diagnosis

Demographic characteristics

Of the 16,411 patients in the AD cohort, 12,606 patients remained at risk for the two endpoints one year post-diagnosis (*i.e.,* alive one year post-diagnosis, not admitted to nursing home, not censored, **S-Table 4**). After updating the clustering model one year post-diagnosis, 5,759 (45.7%) patients were assigned to Group 1 and 6,847 (54.3%) to Group 2. The patients remaining at risk and included in the updated model were younger at diagnosis (age at AD diagnosis mean [SD]: Group 1: baseline: 81.4 [9.3], updated: 80.8 [9.4]; Group 2: baseline: 81.0 [8.7], updated: 80.1 [8.8]). The proportion of women remained comparable between the updated groups (updated Group 1: 3,748, 65.1%; updated Group 2: 4,413, 64.5%, p=.473), and the proportion of non-Hispanic White individuals was higher in the updated Group 2 (updated Group 1: 5,240, 91.0%; updated Group 2: 6,367, 93.0%; p<.001). The overall comorbidity burden was lower in the updated clusters in comparison to baseline (ECI mean [SD] Group 1: baseline: 11.3

[10.3], updated: 10.4 [10.0]; Group 1: baseline: 7.5 [8.6], updated: 6.4 [8.0]), but the updated Group 1 still had a higher comorbidity burden than the updated Group 2 (**S-Figure 2**).

AD-related medications were overall more frequently prescribed for patients in the updated clusters when compared to the baseline clusters (baseline: 12,702, 77.4%; updated: 10,193, 80.9%): 5,703 (45.2%) patients receiving AD-related medications before diagnosis and 4,490 (35.6%) after diagnosis (**S-Table 4**). Consistent with baseline findings, a higher percentage of the updated Group 1 was prescribed AD-related medication preceding AD diagnosis than the updated Group 2 (updated Group 1: 2,759, 47.9%; updated Group 2: 2,944, 43.0%; $p<.0001$). The updated Group 1 also had a shorter follow-up duration (follow up duration median [IQR]: updated Group 1: 63 [35, 86], updated Group 2: 88 [61, 126], $p<.0001$) than the updated Group 2.

Time to AD-related outcomes

For the 12,606 patients remaining at risk one year post-diagnosis, we compared the 2-year, 5-year, and 10-year rate of nursing home admission and mortality. Without updating the cluster membership assigned at AD diagnosis, Group 1 exhibited higher rates of nursing home admission and mortality at 2-year, 5-year, and 10-year intervals than Group 2 (**S-Table 5**). After updating the cluster membership one year post-diagnosis, the updated Group 1 had a higher 2-year and 5-year rate of nursing home admission, and higher 2-year, 5-year, and 10-year mortality rate than the updated Group 2.

In survival analysis, the updated Group 2 had a lower risk of nursing home admission (HR [95% CI]=0.815 [0.766, 0.868], $p<.001$, **S-Figure 3**) than the updated Group 1, after adjusting for covariates, consistent with the baseline model. The mortality risk was again comparable between updated groups (HR [95% CI]=0.977 [0.922, 1.035], $p=.430$) after covariate adjustment.

Feature importance

The updated Group 1 continued to exhibit a high adjusted intensity of dementia-related diagnoses and AD-related medication prescriptions (**S-Figure 4A**). Interestingly, the updated Group 1 also showed a high adjusted intensity of features associated with Parkinson disease (*e.g.,* Parkinson disease, Parkinsonian disorders, carbidopa/levodopa), which was not observed in either of the baseline groups. Further, medications used to treat benign prostatic hyperplasia (*e.g.,* tamsulosin, finasteride) and antipsychotics (*e.g.,* olanzapine) also had high adjusted intensity in the updated Group 1. In contrast, other neurological disorders and epilepsy, which were characteristic of baseline Group 1, were not prominent in the updated Group 1. Moreover, vascular comorbidities (*e.g.,* cerebrovascular disease, cerebrovascular accident), which were typical of baseline Group 1, had higher adjusted intensity in the updated Group 2. The updated Group 2 continued to be characterized by signs and symptoms of dementia, general symptoms of decline, factors pertaining to living situation, vitamin B12 / cyanocobalamin prescriptions, concomitant comorbidities, and medications used to treat these comorbidities among individuals with AD. When examining the correlation between the residualized features using hierarchical clustering, features associated with the updated groups had stronger correlations with other features in their respective groups (**S-Figure 4B**).

DISCUSSION

This study demonstrates the feasibility of stratifying AD patients based on clinical profile available at the point of care using an unsupervised latent factor clustering model guided by knowledge-graph embeddings of EHR concepts. We stratified AD patients into two groups at AD diagnosis and crucially repeated the process one year after diagnosis for prognosticating AD-related outcomes, illustrating the potential of dynamically updating the model with more data accumulation beyond the baseline to iteratively inform prognostication. Our study has several

key clinical findings. First, the patient stratification predicted one group having greater risk of future nursing home admission than the other group at AD diagnosis and again one year post-diagnosis. Second, the predicted mortality risk was comparable between the two groups stratified at AD diagnosis and one year post-diagnosis. Third, Group 1 (with higher risk for nursing home admission) was characterized by a high frequency of dementia-related diagnoses and AD-related medication prescriptions, while Group 2 was characterized by a high frequency of signs and symptoms of dementia, generalized symptoms of decline, and decreased capacity for independent living.

Novel application of latent representation clustering for AD patient stratification

Our approach of using latent representation clustering overcomes limitations in existing approaches of using EHR data to stratify AD patients to predict AD-related outcomes. Xu et al. [25] performed hierarchical clustering to derive AD subphenotypes for predicting the risk of AD diagnosis (but not outcomes after diagnosis) in patients identified using SNOMED-CT codes alone. Landi et al. [29] applied deep learning to derive representations of AD patients and then hierarchical clustering to stratify AD patients into three subgroups, though the clinical utility of predicting AD-related outcomes around AD diagnosis (critical for clinical decision and prognostication discussion) was not shown. To fill this gap, He et al. [27] identified clinically meaningful clusters predictive of mortality, by applying spectral clustering to EHR data preceding AD diagnosis, but the method relies on 40 manually selected AD-related features. Automated selection of informative features from a broad range of procedures, medications, laboratory findings, and information embedded in the clinical narratives may potentially improve the patient stratification model. Our study streamlined the process of patient stratification while improving scalability and generalizability by starting with all codified and narrative features available in the EHR and performing automated feature selection based on data-derived prior knowledge. Prior knowledge on the related features through their representations trained using

separate large-scale EHR data and language models effectively harnessed existing clinical knowledge. The resulting AD patient stratification (into patient groups) based on posterior probability enabled downstream AD-related outcome prognostication.

Prediction models of AD-related outcomes that can be efficiently updated after AD diagnosis are critical for eventual clinical implementation in the real-world setting. First, diagnostic delay may influence patient stratification at AD diagnosis. Iterative updates of the prediction model after diagnosis could improve prediction of future outcomes. Second, it is unclear whether patients experience the same trajectory of decline (or variable changes in trajectory) throughout the long disease course. To systematically investigate factors that accelerate or slow the trajectory of decline, it is crucial to periodically update the prediction of AD-related outcomes. Third, with emerging disease modifying therapies such as anti-amyloid agents (*e.g.,* lecanemab, donanemab), updating the AD-related outcome prediction that incorporates AD treatment status will critically assess treatment response and clinically meaningful benefit in real-world settings [4,5,24].

Prognostication of the risk of nursing home admission and mortality

In survival analysis, Group 1 had a higher risk of nursing home admission (suggesting earlier need for assistance) than Group 2, when clustering patient groups at AD diagnosis and one year post-diagnosis, while the risk of mortality was comparable between the two groups when clustering at either point. Group 1 had a higher overall comorbidity burden than Group 2, with cardiometabolic diseases and other neurological disorders being the most prevalent comorbidities. Our findings are consistent with prior studies that demonstrated the association between comorbidity burden and AD disease progression [10,41,42]. Our study sets a framework for future investigations using EHR data to address whether early effective management of common modifiable comorbidities would improve long-term cognitive and functional outcomes and reduce the risk of nursing home admission and mortality in AD patients.

Group 1 characteristics

Dementia-related diagnoses (*e.g.,* delirium dementia and amnestic and other cognitive disorders) and prescriptions for AD-related medications to improve cognition (*e.g.,* acetylcholinesterase inhibitors) [43,44] had high baseline feature frequencies in Group 1, when clustering at AD diagnosis and one year later. Further, AD-related medications were more frequently prescribed before diagnosis in Group 1 than Group 2. These findings suggest that patients in Group 1 likely experienced high symptom burden preceding AD diagnosis and possibly diagnostic delay (*i.e.,* diagnosed at later stages). Diagnostic delay is common among AD patients given the variable and insidious nature of the disease, with higher risk in certain demographic groups (*e.g.,* non-White individuals, lower socioeconomic status) [45,46].

Vascular comorbidities (*e.g.,* cerebrovascular disease), other neurological disorders, and epilepsy, which are known to accelerate AD decline, had high baseline frequencies in Group 1 [47–53] when clustering at AD diagnosis, but not one year later. Cerebral infarcts and vascular pathologies are independent risk factors for cognitive decline and can accelerate AD-related cognitive decline. Epilepsy often occurs in early stages of AD, and can worsen AD pathology and cognitive decline. These comorbidities, particularly when co-occurring at early AD stages, may contribute to the higher risk of nursing home admission in Group 1.

Parkinson disease and certain non-AD medications (*e.g.,* finasteride, olanzapine) had high frequency in updated Group 1, when clustering one year post-diagnosis (but not at AD diagnosis). Concomitant Parkinson disease and AD pathology can accelerate the rate of decline [54]. While evidence supporting medications for benign prostatic hyperplasia (*e.g.,* finasteride) as disease-modifying in AD is limited [55], it warrants future investigation. Antipsychotics (*e.g.,* olanzapine) are used for treating psychosis in AD [56] as psychosis is common among AD patients and is associated with worse prognosis [57]. While olanzapine was highlighted in our study, aripiprazole and risperidone are also used for AD psychosis [58,59].

Group 2 characteristics

Group 2 had a high baseline frequency of signs and symptoms of dementia (*e.g.,* confusion) when clustering at AD diagnosis and one year later. Individuals in Group 2 with signs and symptoms of dementia might be identified earlier in disease and were less likely to experience diagnostic delay than individuals in Group 1. Likewise, individuals in Group 2 had high baseline frequency of generalized symptoms of decline (*e.g.,* asthenia), suggesting decreased capacity to complete daily activities and possibly also explaining the high frequency of factors pertaining to living situations (*e.g.,* mentions of household composition) in Group 2.

Vitamin B12 and cyanocobalamin were frequently prescribed to individuals in Group 2. Randomized controlled trials and meta-analyses support vitamin B12 supplementation in improving cognitive performance in older adults with and without cognitive impairment [60]. Although vestibular impairment may occur more frequently in AD patients than healthy older adults [61], the clinical significance of the high frequency of vertigo in Group 2 needs further studies. Amlodipine is a first-line antihypertensive that can reduce mortality rate but not the rate of cognitive decline in AD patients [62]. Sertraline is a commonly used antidepressant for AD patients [63], while untreated depression could worsen cognition. The high frequencies of amlodipine and sertraline in Group 2 suggest treatments for common comorbidities that could otherwise worsen AD outcomes in these patients.

Strengths and limitations

This study has several strengths. First, our latent clustering approach leveraged the rich complexity of the available EHR data, particularly the knowledge-driven feature selection of a variety of AD-related codified and narrative features, in model training. To account for similarities and relatedness among the selected features, we utilized the knowledge graph embeddings of EHR concepts to guide the projection of the observed feature counts onto a lower dimensional

space, while leaving flexibility for re-weighting of each factor based on observed data. This step properly addresses the high correlations among selected features when assigning cluster membership. Second, our method is unsupervised, specifically not requiring otherwise time-consuming chart reviews. As a proof of principle, we demonstrated the feasibility to efficiently update the AD-related outcome prediction one year post-diagnosis as more clinical data become available. Third, this study uses two well-validated, generalizable and relevant clinical endpoints for AD. Nursing home admission reflects a loss of functional independence that leads to an increased need for care. Although over 40% of AD patients reside in nursing homes, nursing home admission has been underutilized as an outcome, particularly in EHR-based AD studies [64]. While other important AD-related outcomes such as cognitive and functional measures fluctuate over time and are sparsely documented in EHR, nursing home admission and death are both concrete milestones readily ascertainable from the EHR. Finally, by prognosticating AD-related outcomes at diagnosis and then subsequently updating the prediction with more clinical data, we illustrate a pragmatic real-world clinical application. While baseline prognostication can guide early treatment initiation and tailored care management, dynamically updating the prediction model can inform more timely interventions that improve long-term outcomes.

Our study has some limitations. First, we clustered AD patients at AD diagnosis based on 24 months of pre-index data to strike a balance for adequate sample size and follow-up duration given the dataset contained ~11 years of EHR data (2011-2022). Limited longitudinal pre-index data may incompletely inform baseline profile as signs and symptoms of AD precede onset by years and may inadequately capture diagnostic delay [65,66]. Even if we had longer pre-index EHR data, there may still be a limitation of potential data incompleteness given the fragmented healthcare landscape in the US. The potential for more accurate patient stratification using longitudinal EHR data may increase with wider adoption of EHR systems and health information exchange. Second, this study did not evaluate cognitive and functional decline as endpoints. We

will address this limitation in future studies when we apply novel phenotyping efforts that require minimal gold-standard labels to impute large-scale longitudinal cognitive and functional measures for AD patients. Finally, the findings from this single-institution study may not be representative of the broad AD population. We plan to replicate the study using multi-institutional EHR data to enable more accurate patient stratification.

Future directions

Patient stratification is important for clinical management in AD [2,67]. First, patient stratification enables systematic detection of diagnostic delay, which can inform timely diagnosis and prompt initiation of treatments that are more effective during early disease stages [4,5]. Second, patient stratification may uncover modifiable risk factors that can be targeted to alter AD course. Third, patient stratification and prognostication of AD-related outcomes can inform timely initiation of future care planning conversations with the patient and family while the patient still has decision-making capacity. Third, patient stratification enables the recruitment of targeted sub-populations for clinical trials and may improve trial design. Varying disease trajectory adds to the challenges of assessing treatment response for specific patient groups [68]. There is a heightened need to identify AD patient subpopulations that may derive the greatest therapeutic benefit from novel disease-modifying therapies, given their known adverse events and high financial costs [4,5,69].

CONCLUSION

In summary, we applied a novel knowledge-guided unsupervised latent factor clustering approach to stratify AD patients into two groups based on baseline clinical profile at AD diagnosis. To refine the stratification, we updated the prognostication models one year post-diagnosis, aiming for a more accurate representation of the true underlying clinical

trajectory. Group 1 had a higher risk of nursing home admission, though the risk of mortality was comparable between the two groups. Group 1 exhibited characteristics indicative of greater symptom burden, which may suggest diagnostic delay. In contrast, factors suggestive of gradual decline seemed more frequent in Group 2. Our study supports that patient stratification can enable outcome prognosis and possibly guide clinical management for AD patients. Future studies leveraging EHRs from multiple healthcare systems will further improve AD patient stratification.

ACKNOWLEDGEMENTS

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

L.W., S.Ven., M.M., M.L., R.S., S.Vis., O.L., Z.X., and T.C. contributed to the design and conceptualization of the study. L.W. and S.Ven. contributed equally as co-first authors to data analysis and manuscript writing. M.M., R.S., S.Vis., O.L., Z.X., and T.C. contributed to data acquisition and manuscript writing. Z.X. and T.C. contributed equally as co-senior authors and jointly supervised this work. All authors have reviewed and approved the final manuscript.

MAIN FIGURES AND TABLES

Table 1. Demographic and clinical profile of the overall study population and patient groups clustered at AD diagnosis.

	Overall	Group 1	Group 2	P-value
Number of Patients, N (%)	16411 (100%)	6734 (41.0%)	9677 (59.0%)	
Age at AD diagnosis, mean (SD), years	81.2 (9.0)	81.4 (9.3)	81.0 (8.7)	**.007**
Gender, N (%)				.546
Women	10559 (64.3%)	4314 (64.1%)	6245 (64.5%)	
Men	5852 (35.7%)	2420 (35.9%)	3432 (35.5%)	
Race/ethnicity [a], N (%)				**.014**
Non-Hispanic white	15105 (92.0%)	6138 (91.1%)	8967 (92.7%)	
Others	1306 (8.0%)	596 (8.9%)	710 (7.3%)	
American Indian	8 (0.0%)	3 (0.0%)	5 (0.1%)	
Asian	65 (0.4%)	32 (0.5%)	33 (0.3%)	
Black	1183 (7.2%)	535 (7.9%)	648 (6.7%)	
Hawaiian or Pacific Islander	3 (0.0%)	2 (0.0%)	1 (0.0%)	
Hispanic or Latino	47 (0.3%)	24 (0.4%)	23 (0.2%)	
Other	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Elixhauser Comorbidity Index [b], mean (SD)	9.0 (9.5)	11.3 (10.3)	7.5 (8.6)	**<.0001**
Baseline comorbidities (top 4 prevalent), N (%)				**<.0001**
Hypertension, uncomplicated	12052 (73.4%)	5305 (78.8%)	6747 (69.7%)	
Cardiac arrhythmias	6326 (38.5%)	3124 (46.4%)	3202 (33.1%)	
Depression	5405 (32.9%)	2772 (41.2%)	2633 (27.2%)	
Fluid and electrolyte disorders	4730 (28.8%)	2526 (37.5%)	2204 (22.8%)	
AD-related medication [c], N (%)				
Not prescribed	3709 (22.6%)	1794 (26.6%)	3742 (38.7%)	**<.0001**

Prescribed	12702 (77.4%)	4940 (73.4%)	5935 (61.3%)	
Before AD diagnosis	7166 (43.7%)	3212 (47.7%)	3954 (40.9%)	**<.0001**
On or after AD diagnosis	5536 (33.7%)	1794 (26.6%)	3742 (38.7%)	
Follow-up months, Median (IQR)	78 (48, 109)	62 (35, 85)	86 (59, 125)	**<.0001**

a. We imputed missing race and ethnicity based on age and gender.

b. We calculated the comorbidity burden using 24 months of pre-index data (*i.e.,* in the two years preceding AD diagnosis).

c. List of AD-related medications can be found in S-Table 3.

Table 2. Comparison of 2-, 5- and 10-year rate of nursing home admission and death for AD patient groups clustered at AD diagnosis (N = 16,411).

	Nursing Home Admission			**Death**		
	Group 1	**Group 2**	**P-value**	**Group 1**	**Group 2**	**P-value**
2 years	31.0% (29.9%, 32.1%)	23.5% (22.6%, 24.3%)	**<.0001**	23.2% (22.2%, 24.3%)	19.1% (18.3%, 19.9%)	**<.0001**
5 years	45.5% (44.3%, 46.7%)	40.2% (39.3%, 41.2%)	**<.0001**	44.1% (42.9%, 45.3%)	43.6% (42.6%, 44.6%)	.566
10 years	48.8% (47.6%, 50.0%)	47.5% (46.5%, 48.5%)	.0919	49.6% (48.4%, 50.8%)	61.6% (60.6%, 62.6%)	**<.0001**

Figure 1. Schematic overview. (A) Inclusion criteria. (B) Study workflow.

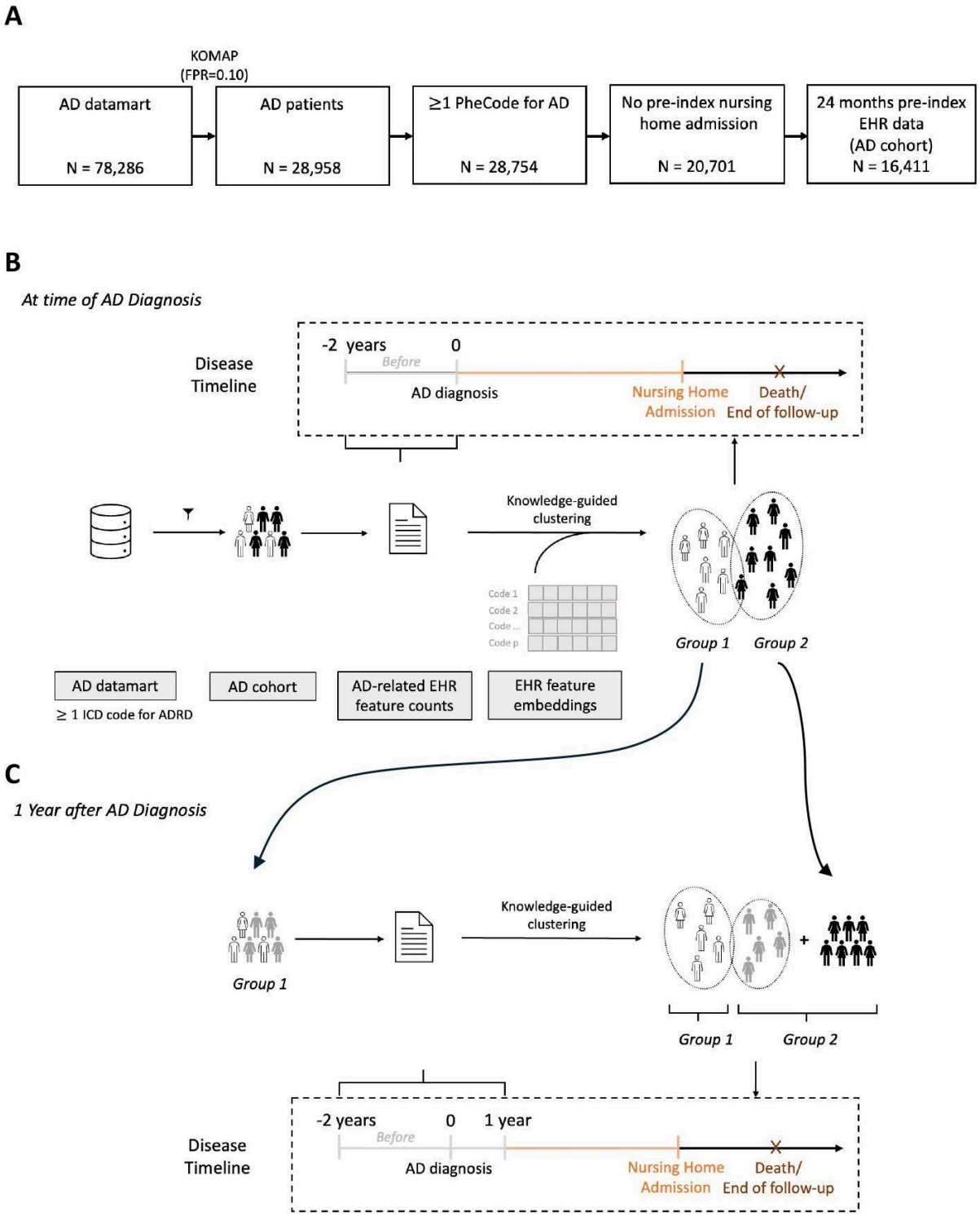


Figure 2. Baseline comorbidity burden of AD patient groups clustered at AD diagnosis.

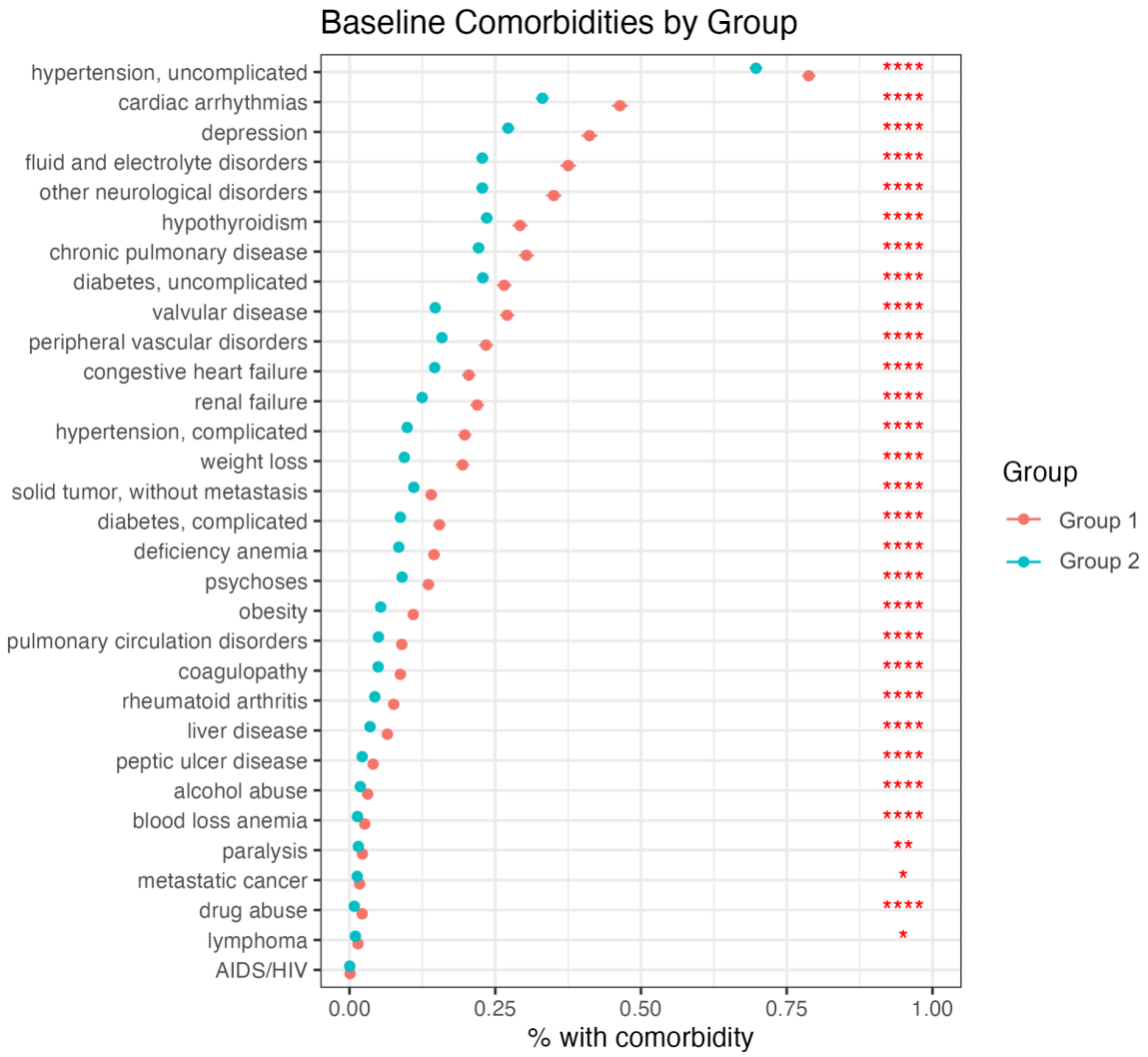


Figure 3. Adjusted survival curve of (A) time to nursing home and (B) time to death for AD patient groups clustered at AD diagnosis.

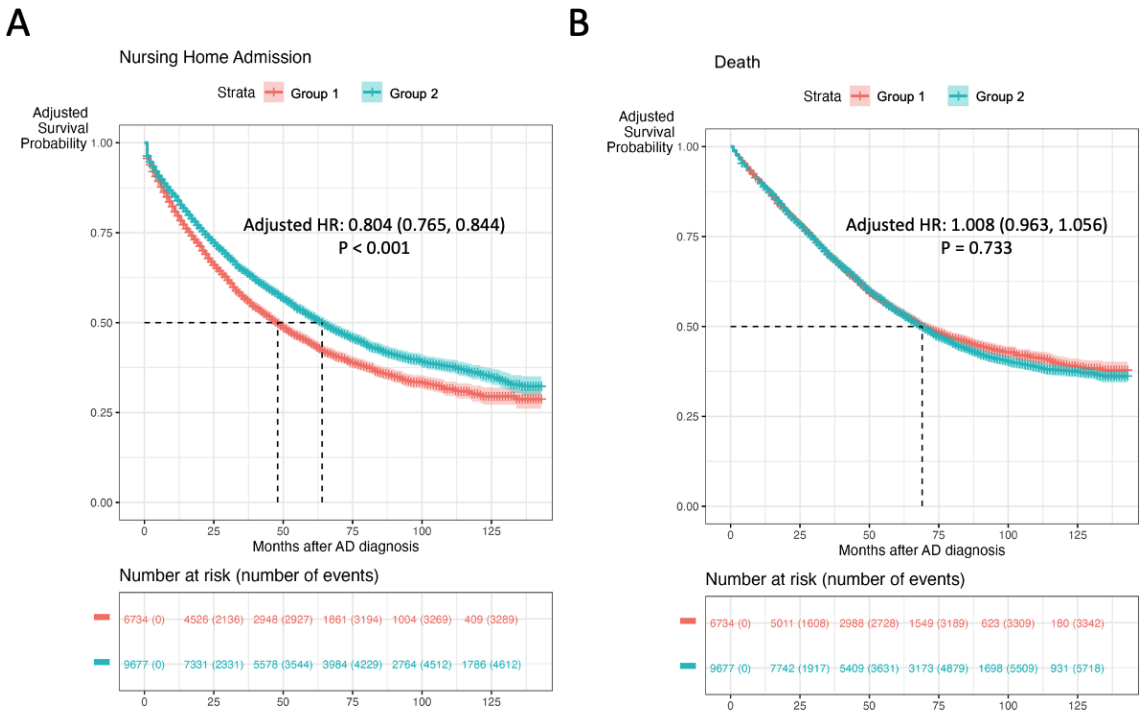
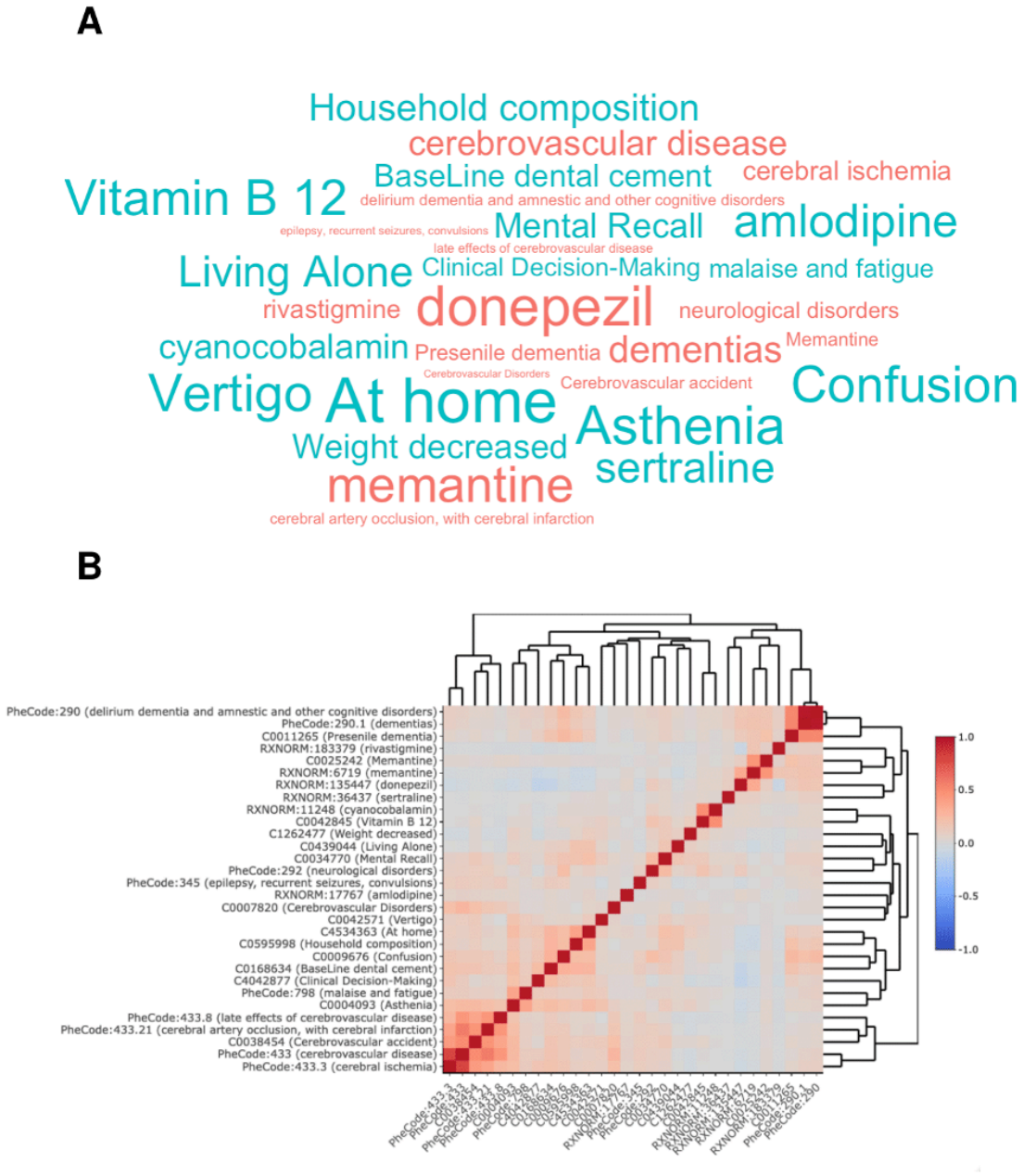


Figure 4. Top 15 features contributing to differences between AD patient groups clustered at AD diagnosis. Features in salmon are positively associated with Group 1 membership, while features in teal are negatively associated with Group 1 membership. (A) Word cloud of the important features. Text size represents the relative importance of the feature. (B) Correlation matrix of the features.



SUPPLEMENTARY MATERIAL

S-Table 1. AD data mart specifications.

Code	Description
ICD-9 290.0	Senile dementia, uncomplicated
ICD-9 290.1	Presenile dementia
ICD-9 290.10	Presenile dementia, uncomplicated
ICD-9 290.11	Presenile dementia with delirium
ICD-9 290.12	Presenile dementia with delusional features
ICD-9 290.13	Presenile dementia with depressive features
ICD-9 290.20	Senile dementia with delusional features
ICD-9 290.21	Senile dementia with depressive features
ICD-9 290.3	Senile dementia with delirium
ICD-9 294.20	Dementia, unspecified, without behavioral disturbance
ICD-9 294.21	Dementia, unspecified, with behavioral disturbance
ICD-9 331.0	Alzheimer’s disease
ICD-10 F03.90	Unspecified dementia without behavioral disturbance
ICD-10 F03.91	Unspecified dementia with behavioral disturbance
ICD-10 G30	Alzheimer's disease
ICD-10 G30.0	Alzheimer's disease with early onset
ICD-10 G30.1	Alzheimer's disease with late onset
ICD-10 G30.8	Other Alzheimer's disease
ICD-10 G30.9	Alzheimer's disease, unspecified

Note: Patients were included in the data mart if they had at least one of the AD related diagnosis codes. The data mart is the starting point to more accurately identify AD patients.

S-Table 2. List of nursing home admission codes.

Code	Description
Assisted living	
CPT 99321	Domiciliary or rest home visit for the evaluation and management of a new patient; the presenting problems are of low severity
CPT 99322	Domiciliary or rest home visit for the evaluation and management of a new patient; the presenting problems are of moderate severity
CPT 99323	Domiciliary or rest home visit for the evaluation and management of a new patient; the presenting problems are of high complexity
CPT 99331	Domiciliary or rest home visit for the evaluation and management of a new patient; the patient is stable, recovering or improving
CPT 99332	Domiciliary or rest home visit for the evaluation and management of a new patient; the patient is responding inadequately to therapy or has developed a minor complication
CPT 99333	Domiciliary or rest home visit for the evaluation and management of a new patient; the patient is unstable or has developed a significant complication or a significant new problem
ICD-9 V60.89	Other specified housing or economic circumstances
ICD-10 Z74	Problems related to care provider dependency
ICD-9 V63.2	Person awaiting admission to adequate facility elsewhere
ICD-10 Z75.1	Person awaiting admission to adequate facility elsewhere
Long-term care	
CPT 99301	Evaluation and management of a new or established patient involving an annual nursing facility assessment: 30 minutes at the bedside
CPT 99302	Evaluation and management of a new or established patient involving an annual nursing facility assessment of a complication or a new problem: 40 minutes at the bedside
CPT 99303	Evaluation and management of a new or established patient involving an annual nursing facility assessment at the time of initial admission to the facility: 50 minutes at the bedside
CPT 99304	Initial nursing facility care 25 min
CPT 99305	Initial nursing facility care 35 min
CPT 99306	Initial nursing facility care 45 min
CPT 99307	Subsequent nursing facility care, per day, for the evaluation and management of a patient: 10 minutes at the bedside
CPT 99308	Subsequent nursing facility care, per day, for the evaluation and management of a patient: 15 minutes at the bedside
CPT 99309	Subsequent nursing facility care, per day, for the evaluation and management of a patient: 25 minutes at the bedside
CPT 99310	Subsequent nursing facility care, per day, for the evaluation and management of a patient: 35 minutes at the bedside
CPT 99311	Subsequent nursing facility care, per day, for the evaluation and management of a new or established patient: 15 minutes at the bedside
CPT 99312	Subsequent nursing facility care, per day, for the evaluation and management of a new or established patient who is responding inadequately to therapy or has developed a minor complication: 25 minutes at the bedside

CPT 99313	Subsequent nursing facility care, per day, for the evaluation and management of a new or established patient who has developed a significant complication or a new problem: 35 minutes at the bedside
CPT 99315	Nursing facility discharge day management; 30 minutes or less
CPT 99316	Nursing facility discharge day management; more than 30 minutes
CPT 99318	Evaluation and management of a patient involving an annual nursing facility assessment
CPT 99379	Physician supervision of a nursing facility patient (patient not present) requiring complex and multidisciplinary care; 15–29 minutes
CPT 99380	Physician supervision of a nursing facility patient (patient not present) requiring complex and multidisciplinary care modalities; 30 minutes or more
CPT 99199	Unlisted special service, procedure, or report
ICD-9 V60.6	Person living in residential institution
ICD-10 Z59.3	Problems related to living in residential institution
ICD-9 E849.7	Accidents occurring in residential institution
ICD-10 Y92.12	Nursing home as the place of occurrence of the external cause
ICD-9 E849.0	Accidents occurring in hospital
ICD-10 Y92.23	Hospital as the place of occurrence of the external cause
ICD-9 V70.3	Other general medical examination for administrative purposes
ICD-10 Z02.2	Encounter for examination for admission to residential institution
HCPCS G0066	Physician supervision of a nursing facility patient (patient not present); 30 minutes or more per month
HCPCS G9685	Physician service or other qualified health care professional for the evaluation and management of a beneficiary's acute change in condition in a nursing facility
End-of-life care	
CPT 99377	Under Care Plan Oversight Services. Report this service when the provider supervises and coordinates the care provided to a hospice patient.
CPT 99378	Under Care Plan Oversight Services. Report this service when the provider supervises and coordinates the care provided to a hospice patient.
ICD-9 E876.8	Other specified misadventures during medical care
ICD-10 Y66	Non-administration of surgical and medical care
ICD-9 V65.49	Other specified counseling
ICD-9 V65.8	Other reasons for seeking consultation
ICD-10 Z76.89	Persons encountering health services in other specified circumstances
ICD-9 V66.7	Encounter for palliative care
ICD-10 Z51.5	Encounter for palliative care
HCPCS G0299	Direct skilled nursing services of a registered nurse (RN) in the home health or hospice setting, each 15 minutes
HCPCS G0300	Direct skilled nursing services of a licensed practical nurse (LPN) in the home health or hospice setting, each 15 minutes
HCPCS Q5001	Hospice or home health care provided in patient's home/residence
HCPCS Q5002	Hospice or home health care provided in assisted living facility
HCPCS Q5003	Hospice care provided in nursing long term care facility (LTC) or non-skilled nursing facility (NF)
HCPCS Q5004	Hospice care provided in skilled nursing facility (SNF)
HCPCS Q5005	Hospice care provided in inpatient hospital
HCPCS Q5006	Hospice care provided in inpatient hospice facility
HCPCS Q5007	Hospice care provided in long term care facility

HCPCS Q5008	Hospice care provided in inpatient psychiatric facility
HCPCS Q5009	Hospice or home health care provided in place not otherwise specified (NOS)
HCPCS Q5010	Hospice home care provided in a hospice facility
HCPCS T2042	Hospice routine home care per diem
HCPCS T2043	Hospice continuous home care per hour
HCPCS T2044	Hospice inpatient respite care per diem
HCPCS T2045	Hospice general inpatient care per diem
HCPCS T2046	Hospice long term care, room and board only per diem

S-Table 3. List of AD-related medications.

Brand Name	Generic Name	Treatment Class
Adlarity	Donepezil	Acetylcholinesterase Inhibitor
Aricept	Donepezil	Acetylcholinesterase Inhibitor
Exelon	Rivastigmine	Acetylcholinesterase Inhibitor
Razadyne ER	Galantamine	Acetylcholinesterase Inhibitor
Razadyne [DSC]	Galantamine	Acetylcholinesterase Inhibitor
Namenda	Memantine	N-Methyl-D-Aspartate (NMDA) Receptor Antagonist
Namenda Titration Pak	Memantine	N-Methyl-D-Aspartate (NMDA) Receptor Antagonist
Namenda XR	Memantine	N-Methyl-D-Aspartate (NMDA) Receptor Antagonist
Namenda XR Titration Pack [DSC]	Memantine	N-Methyl-D-Aspartate (NMDA) Receptor Antagonist
Aduhelm	Aducanumab	Anti-Amyloid Monoclonal Antibody
Leqembi	Lecanemamb-irmb	Anti-Amyloid Monoclonal Antibody
Namzaric	Donepezil and memantine	Acetylcholinesterase Inhibitor; N-Methyl-D-Aspartate (NMDA) Receptor Antagonist

S-Table 4. Baseline demographic and clinical profile of AD patient groups clustered one year after AD diagnosis.

	Overall	Group 1	Group 2	P-value
Number of Patients, N (%)	12606 (100%)	5759 (45.7%)	6847 (54.3%)	
Age at AD diagnosis, mean (SD)	80.4 (9.1)	80.8 (9.4)	80.1 (8.8)	**<.0001**
Gender, N (%)				.473
Women	8161 (64.7%)	3748 (65.1%)	4413 (64.5%)	
Men	4445 (35.3%)	2011 (34.9%)	2434 (35.5%)	
Race/ethnicity [a], N (%)				**<.001**
Non-Hispanic white	11607 (92.1%)	5240 (91.0%)	6367 (93.0%)	
Others				
American Indian or Alaska Native	8 (0.1%)	5 (0.1%)	3 (0.0%)	
Asian	59 (0.5%)	35 (0.6%)	24 (0.4%)	
Black	893 (7.1%)	460 (8.0%)	433 (6.3%)	
Hawaiian or Pacific Islander	2 (0.0%)	0 (0.0%)	2 (0.0%)	
Hispanic or Latino	37 (0.3%)	19 (0.3%)	37 (0.3%)	
Other	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Elixhauser Comorbidity Index [b], mean (SD)	8.2 (9.2)	10.4 (10.0)	6.4 (8.0)	**<.0001**
Baseline comorbidities (top 4 prevalent), N (%)				**<.0001**
Hypertension, uncomplicated	9061 (71.9%)	4499 (78.1%)	4562 (66.6%)	
Cardiac arrhythmias	4483 (35.6%)	2538 (44.1%)	1945 (28.4%)	
Depression	4115 (32.6%)	2258 (39.2%)	1857 (27.1%)	
Other neurological disorders	3360 (26.7%)	1878 (32.6%)	1482 (21.6%)	
AD-related medication [c], N (%)				
Not prescribed	2413 (19.1%)	1310 (22.7%)	1103 (16.1%)	**<.0001**
Prescribed	10193 (80.9%)	4449 (77.3%)	5744 (83.9%)	
Before AD diagnosis	5703 (45.2%)	2759 (47.9%)	2944 (43.0%)	**<.0001**

On or after AD diagnosis	4490 (35.6%)	1690 (29.3%)	2800 (40.9%)	
Follow-up months from baseline, Median (IQR)	80 (50, 111)	63 (35, 86)	88 (61, 126)	**<.0001**

a. We imputed missing race and ethnicity based on age and gender.

b. We calculated the comorbidity burden using 24 months of pre-index data (*i.e.,* during the two years prior to AD diagnosis).

c. List of AD-related medications can be found in S-Table 3.

S-Table 5. Comparison of 2-, 5- and 10-year rate of nursing home admission and death for AD patient groups clustered one year after AD diagnosis (N=12606).

Time of Clustering		Nursing Home Admission			Death		
		Group 1	Group 2	P-value	Group 1	Group 2	P-value
At AD diagnosis	**2 years**	14.9% (13.9%, 15.9%)	11.6% (10.9%, 12.3%)	**<0.0001**	10.4% (9.6%, 11.3%)	8.9% (8.3%, 9.6%)	**0.00465**
	5 years	34.6% (33.3%, 35.9%)	32.5% (31.4%, 33.5%)	**0.0144**	33.7% (32.4%, 35.0%)	35.7% (34.7%, 36.8%)	**0.0192**
	10 years	39.1% (37.8%, 40.5%)	41.4% (40.3%, 42.5%)	**0.0107**	40.3% (39.0%, 41.7%)	52.7% (51.6%, 53.8%)	**<0.0001**
1 year after AD diagnosis	**2 years**	14.6% (13.7%, 15.5%)	11.5% (10.7%, 12.2%)	**<0.0001**	11.0% (10.3%, 11.9%)	8.2% (7.6%, 8.9%)	**<0.0001**
	5 years	34.5% (33.3%, 35.8%)	32.3% (31.2%, 33.4%)	**0.0082**	35.1% (33.9%, 36.3%)	34.8% (33.6%, 35.9%)	**<0.0001**
	10 years	39.6% (38.3%, 40.9%)	41.3% (40.1%, 42.5%)	0.060	52.4% (51.2%, 53.6%)	42.5% (41.2%, 43.7%)	**<0.0001**

S-Table 6. Importance and prevalence of AD-related features included in the clustering model at AD diagnosis. Feature importance is calculated by the difference in mean adjusted feature intensity between the clusters. Features with positive feature importance values are associated with Group 1 whereas features with negative feature importance values are associated with Group 2.

Feature	Description	Feature Importance	Prevalence		
			Overall	Group 1	Group 2
RXNORM:135447	Donepezil	0.1453	42.57%	45.28%	40.06%
PheCode:433	Cerebrovascular disease	0.1098	36.31%	40.07%	32.82%
RXNORM:6719	Memantine	0.0979	20.09%	18.34%	21.71%
PheCode:290.1	Dementias	0.0518	100.00%	100.00%	100.00%
PheCode:290	Delirium dementia and amnestic and other cognitive disorders	0.0501	100.00%	100.00%	100.00%
PheCode:433.3	Cerebral ischemia	0.0446	27.87%	31.75%	24.27%
PheCode:292	Neurological disorders	0.0365	52.63%	62.22%	43.76%
PheCode:433.21	Cerebral artery occlusion, with cerebral infarction	0.0221	8.44%	9.34%	7.60%
C0011265	Presenile dementia	0.0158	66.44%	76.37%	57.25%
RXNORM:183379	Rivastigmine	0.0141	6.70%	6.68%	6.71%
C0038454	Cerebrovascular accident	0.0079	26.17%	35.39%	17.63%
PheCode:433.8	Late effects of cerebrovascular disease	0.0069	4.10%	4.57%	3.66%
PheCode:345	Epilepsy, recurrent seizures, convulsions	0.0060	6.78%	7.78%	5.86%
C0025242	Memantine	0.0058	8.69%	9.86%	7.60%
C0007820	Cerebrovascular Disorders	0.0048	3.87%	4.95%	2.87%
C0007787	Transient Ischemic Attack	0.0046	7.85%	10.48%	5.42%
C0002395	Alzheimer's Disease	0.0045	54.34%	61.39%	47.81%
PheCode:332	Parkinson's disease	0.0043	4.86%	5.30%	4.44%
RXNORM:114477	Levetiracetam	0.0033	3.01%	3.61%	2.46%
C0527316	Donepezil	0.0031	16.98%	21.19%	13.08%
PheCode:430	Intracranial hemorrhage	0.0028	2.57%	2.78%	2.38%
C0003537	Aphasia	0.0015	4.26%	5.55%	3.07%

C0011269	Dementia, Vascular	0.0013	3.65%	4.81%	2.57%
PheCode:818	Intracranial hemorrhage (injury)	0.0010	1.80%	1.84%	1.76%
PheCode:430.2	Intracerebral hemorrhage	0.0006	1.09%	1.31%	0.89%
C0242422	Parkinsonian Disorders	0.0004	2.68%	3.64%	1.80%
RXNORM:103990	Carbidopa/levodopa	0.0002	2.81%	3.53%	2.13%
C0085220	Cerebral Amyloid Angiopathy	0.0001	0.19%	0.21%	0.18%
C0751587	CADASIL Syndrome	0.0000	0.06%	0.06%	0.07%
C3662068	Static encephalopathy	0.0000	0.04%	0.06%	0.03%
C0016957	Galactosylceramidase	0.0000	0.00%	0.00%	0.00%
C0393483	Brainstem encephalitis	0.0000	0.00%	0.00%	0.00%
C4763868	Cocaine use disorder	0.0000	0.01%	0.03%	0.00%
PheCode:290.11	Alzheimer's disease	0.0000	99.94%	99.91%	99.96%
C0260942	Encounter due to screening for depression	0.0000	0.09%	0.15%	0.04%
RXNORM:1791685	Pimavanserin	0.0000	0.04%	0.06%	0.01%
PheCode:333.3	Tics and choreas	0.0000	0.06%	0.12%	0.00%
C0030567	Parkinson disease	-0.0001	5.59%	7.26%	4.04%
RXNORM:4637	Galantamine	-0.0001	0.69%	0.73%	0.65%
C0041296	Tuberculosis	-0.0003	0.91%	1.41%	0.44%
C3714756	Intellectual disability	-0.0009	0.90%	1.32%	0.51%
C0011127	Pressure ulcer	-0.0011	2.13%	3.40%	0.95%
C0123091	Quetiapine	-0.0018	2.05%	3.50%	0.70%
C4484264	Clark	-0.0018	2.03%	3.16%	0.99%
C0022983	Laminectomy	-0.0025	2.96%	4.60%	1.43%
RXNORM:61381	Olanzapine	-0.0034	2.81%	4.34%	1.39%
RXNORM:25025	Finasteride	-0.0042	3.52%	4.71%	2.42%
PheCode:292.4	Altered mental status	-0.0043	23.53%	29.91%	17.63%
C0518460	Bathing self care	-0.0052	6.33%	10.05%	2.87%
C0008845	Citalopram	-0.0053	4.78%	7.44%	2.31%
C0278061	Abnormal mental state	-0.0075	14.65%	20.52%	9.21%
RXNORM:43611	Latanoprost	-0.0090	3.14%	4.89%	1.53%
C0019360	Herpes zoster disease	-0.0094	8.64%	13.38%	4.25%
C0011168	Deglutition disorders	-0.0095	7.94%	12.67%	3.56%
RXNORM:35636	Risperidone	-0.0098	3.88%	5.90%	2.01%
PheCode:295	Schizophrenia and other psychotic disorders	-0.0104	9.47%	12.68%	6.50%

C0428977	Bradycardia	-0.0110	8.52%	13.60%	3.81%
RXNORM:221147	Polyethylene glycol 3350	-0.0116	10.02%	16.29%	4.21%
C0430533	Dental diagnostic procedure	-0.0132	13.22%	20.58%	6.39%
PheCode:350.2	Abnormality of gait	-0.0137	15.24%	21.01%	9.89%
RXNORM:321988	Escitalopram	-0.0209	6.20%	9.10%	3.51%
C0034991	Rehabilitation therapy	-0.0217	14.14%	23.60%	5.39%
RXNORM:51272	Quetiapine	-0.0225	5.87%	9.24%	2.75%
RXNORM:77492	Tamsulosin	-0.0226	8.76%	12.73%	5.09%
RXNORM:15996	Mirtazapine	-0.0237	5.50%	9.03%	2.24%
RXNORM:10737	Trazodone	-0.0259	5.75%	9.33%	2.45%
PheCode:591	Urinary tract infection	-0.0295	23.25%	30.56%	16.49%
C0011053	Deafness	-0.0306	14.89%	23.43%	6.97%
C0757844	TNFSF13 protein, human	-0.0319	20.75%	32.76%	9.63%
C0206275	Widowhood	-0.0325	11.70%	19.02%	4.92%
C3809991	Immunodeficiency, common variable	-0.0336	17.84%	27.58%	8.81%
C0042029	Urinary tract infection	-0.0364	17.59%	27.43%	8.47%
C1856053	Hydranencephaly with renal aplasia-dysplasia	-0.0375	21.25%	33.75%	9.67%
PheCode:798	Malaise and fatigue	-0.0447	29.07%	41.27%	17.78%
C4042877	Clinical decision-making	-0.0521	42.17%	57.71%	27.79%
C0034770	Mental recall	-0.0553	34.61%	51.40%	19.06%
RXNORM:36437	Sertraline	-0.0576	10.80%	16.05%	5.94%
RXNORM:11248	Cyanocobalamin	-0.0577	14.48%	22.82%	6.75%
C0042571	Vertigo	-0.0638	26.54%	41.00%	13.15%
C0439044	Living alone	-0.0642	20.13%	33.41%	7.82%
C1262477	Weight decreased	-0.0648	22.79%	36.52%	10.08%
C0168634	BaseLine dental cement	-0.0677	35.58%	52.17%	20.23%
RXNORM:17767	Amlodipine	-0.0761	16.79%	24.65%	9.52%
C0009676	Confusion	-0.0817	43.53%	60.50%	27.82%
C0004093	Asthenia	-0.0906	32.95%	51.96%	15.35%
C0042845	Vitamin B 12	-0.1030	28.20%	42.06%	15.37%
C5425799	All other	-0.1042	47.99%	67.57%	29.85%
C4534363	At home	-0.1405	46.55%	69.16%	25.62%
C0595998	Household composition	-0.1820	49.69%	73.23%	27.89%

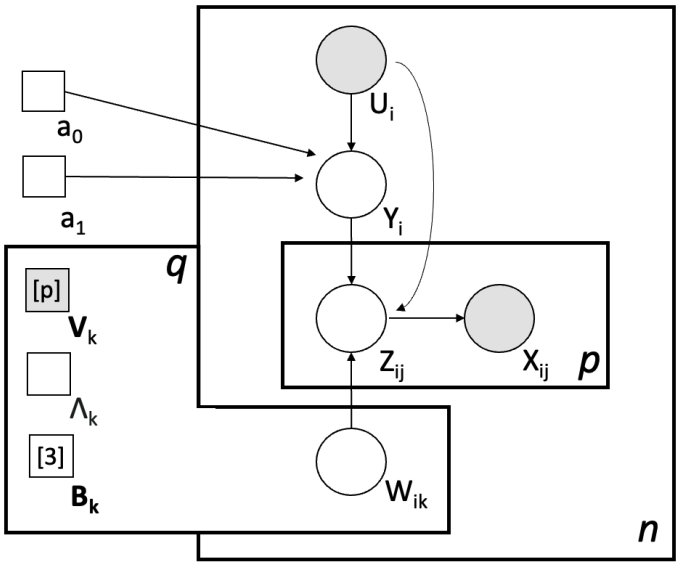
S-Table 7: Importance and prevalence of AD-related features included in the clustering model one year after AD diagnosis. Feature importance is calculated by the difference in mean adjusted feature intensity between the clusters. Features with positive feature importance values are associated with Group 1 whereas features with negative feature importance values are associated with Group 2.

Feature	Description	Feature Importance	Prevalence		
			Overall	Group 1	Group 2
RXNORM:6719	Memantine	0.2194	34.99%	39.05%	31.24%
RXNORM:135447	Donepezil	0.1926	59.24%	58.59%	59.84%
PheCode:290.1	Dementias	0.1591	98.65%	98.14%	99.11%
PheCode:290	Delirium dementia and amnestic and other cognitive disorders	0.1570	98.66%	98.16%	99.11%
PheCode:332	Parkinson's disease	0.0862	5.57%	8.60%	2.78%
RXNORM:183379	Rivastigmine	0.0801	11.24%	13.82%	8.86%
PheCode:290.11	Alzheimer's disease	0.0753	98.37%	97.62%	99.06%
C0030567	Parkinson disease	0.0745	7.60%	9.99%	5.40%
RXNORM:103990	Carbidopa/levodopa	0.0575	3.53%	5.99%	1.25%
C0242422	Parkinsonian disorders	0.0187	4.07%	5.54%	2.71%
RXNORM:77492	Tamsulosin	0.0145	12.66%	9.82%	15.28%
C0025242	Memantine	0.0121	17.76%	15.73%	19.64%
C0011265	Presenile dementia	0.0048	84.64%	77.68%	91.07%
RXNORM:61381	Olanzapine	0.0026	5.27%	3.89%	6.55%
RXNORM:25025	Finasteride	0.0022	4.93%	4.28%	5.52%
PheCode:292	Neurological disorders	0.0019	61.28%	50.70%	71.07%
RXNORM:4637	Galantamine	0.0010	1.37%	1.48%	1.27%
PheCode:295	Schizophrenia and other psychotic disorders	0.0005	11.83%	9.24%	14.22%
RXNORM:1791685	Pimavanserin	0.0001	0.09%	0.17%	0.02%
C0149843	Punch drunk syndrome	0.0000	0.01%	0.02%	0.00%

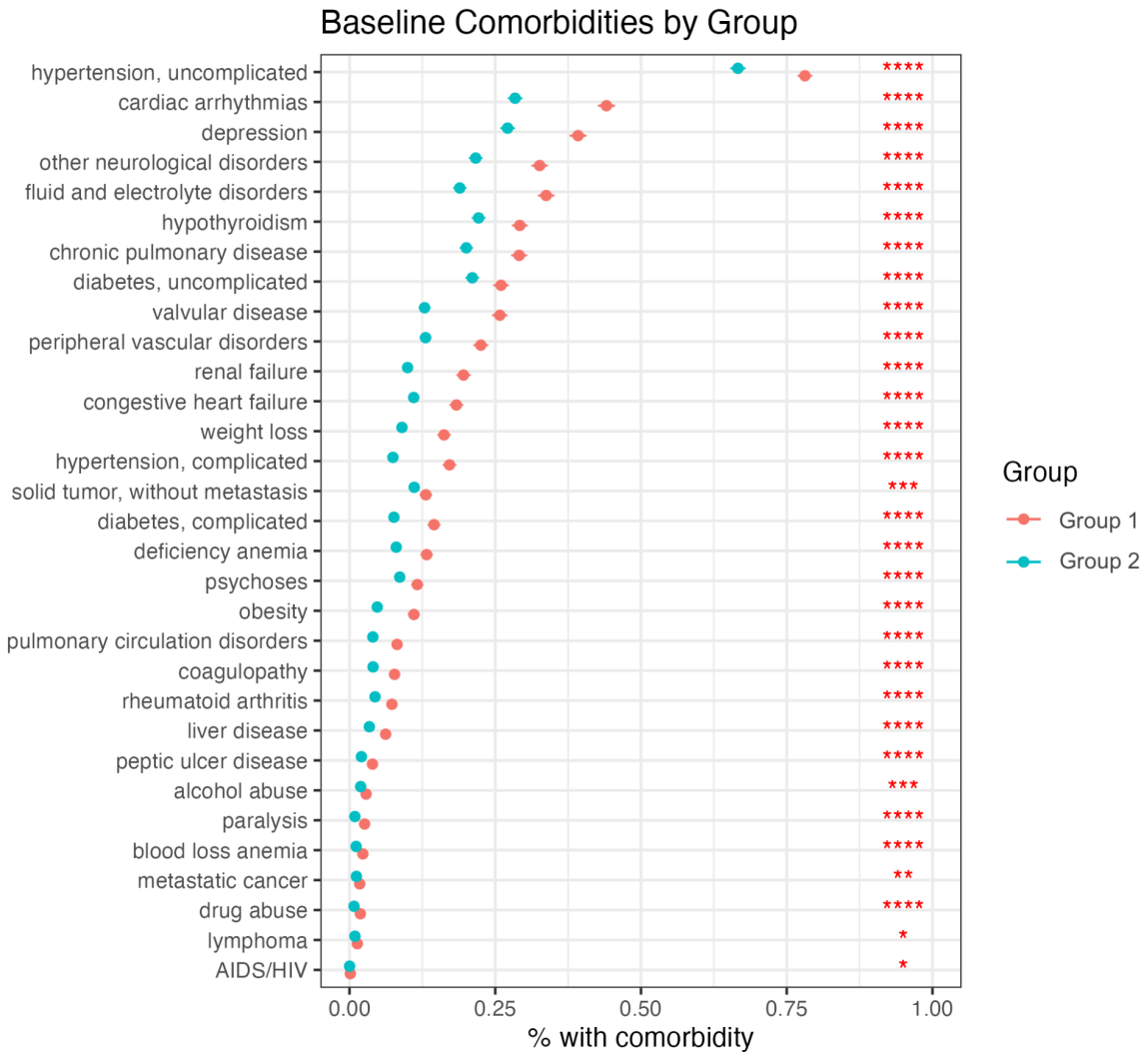
C0751587	CADASIL Syndrome	0.0000	0.12%	0.11%	0.12%
C0016957	Galactosylceramidase	0.0000	0.00%	0.00%	0.00%
C0393483	Brainstem encephalitis	0.0000	0.00%	0.00%	0.00%
C4763868	Cocaine use disorder	0.0000	0.02%	0.00%	0.03%
C0238111	Lennox-Gastaut syndrome	0.0000	0.01%	0.00%	0.02%
C3662068	Static encephalopathy	0.0000	0.03%	0.00%	0.05%
C0260942	Encounter due to screening for depression	0.0000	0.22%	0.17%	0.26%
C0085220	Cerebral amyloid angiopathy	0.0000	0.30%	0.19%	0.40%
PheCode:333.3	Tics and choreas	-0.0001	0.07%	0.02%	0.12%
C0041296	Tuberculosis	-0.0003	1.28%	0.73%	1.79%
C3714756	Intellectual disability	-0.0014	1.24%	0.71%	1.74%
C4484264	Clark levels	-0.0018	2.98%	1.95%	3.92%
C0011127	Pressure ulcer	-0.0018	3.58%	1.91%	5.12%
C0123091	Quetiapine	-0.0019	4.48%	2.55%	6.27%
PheCode:430.2	Intracerebral hemorrhage	-0.0020	1.52%	0.62%	2.34%
PheCode:818	Intracranial hemorrhage (injury)	-0.0037	2.32%	1.11%	3.44%
C0007820	Cerebrovascular disorders	-0.0037	5.64%	2.95%	8.13%
C0011269	Dementia, vascular	-0.0048	6.31%	3.77%	8.66%
C0022983	Laminectomy	-0.0052	4.05%	2.18%	5.78%
C0003537	Aphasia	-0.0071	6.36%	3.36%	9.13%
PheCode:430	Intracranial hemorrhage	-0.0073	3.51%	1.75%	5.14%
RXNORM:43611	Latanoprost	-0.0079	4.43%	2.87%	5.87%
RXNORM:35636	Risperidone	-0.0097	7.25%	5.95%	8.46%
PheCode:433.8	Late effects of cerebrovascular disease	-0.0099	5.50%	2.03%	8.72%
C0002395	Alzheimer's disease	-0.0100	83.32%	76.65%	89.49%
PheCode:345	Epilepsy, recurrent seizures, convulsions	-0.0101	9.00%	6.06%	11.72%
PheCode:350.2	Abnormality of gait	-0.0104	21.99%	15.37%	28.11%
PheCode:292.4	Altered mental status	-0.0105	28.96%	20.05%	37.19%
C0008845	Citalopram	-0.0109	7.43%	4.26%	10.37%
RXNORM:114477	Levetiracetam	-0.0111	4.90%	2.67%	6.96%
RXNORM:221147	Polyethylene glycol 3350	-0.0118	18.46%	11.08%	25.30%
C0518460	Bathing self care	-0.0126	10.92%	5.80%	15.66%
C0011168	Deglutition disorders	-0.0132	13.05%	6.91%	18.74%
RXNORM:321988	Escitalopram	-0.0141	9.68%	7.38%	11.81%

PheCode:591	Urinary tract infection	-0.0145	29.33%	21.10%	36.95%
C0278061	Abnormal mental state	-0.0153	21.01%	12.80%	28.60%
C0019360	Herpes zoster disease	-0.0158	12.10%	7.02%	16.79%
RXNORM:51272	Quetiapine	-0.0165	11.33%	8.71%	13.75%
C0007787	Transient ischemic attack	-0.0202	13.02%	7.32%	18.28%
C0430533	Dental diagnostic procedure	-0.0237	19.95%	10.76%	28.46%
C0428977	Bradycardia	-0.0242	14.23%	7.79%	20.18%
RXNORM:15996	Mirtazapine	-0.0245	9.25%	6.25%	12.03%
RXNORM:10737	Trazodone	-0.0252	9.45%	6.01%	12.64%
PheCode:433.21	Cerebral artery occlusion, with cerebral infarction	-0.0282	11.00%	5.26%	16.32%
C0527316	Donepezil	-0.0377	30.91%	22.86%	38.36%
C0011053	Deafness	-0.0404	20.57%	12.77%	27.78%
PheCode:798	Malaise and fatigue	-0.0406	37.10%	24.57%	48.69%
C0034991	Rehabilitation therapy	-0.0458	23.18%	11.02%	34.43%
RXNORM:36437	Sertraline	-0.0473	15.71%	11.51%	19.60%
C0042029	Urinary tract infection	-0.0528	25.16%	14.42%	35.09%
C0757844	TNFSF13 protein, human	-0.0624	31.67%	17.42%	44.85%
C0206275	Widowhood	-0.0625	16.80%	8.49%	24.48%
PheCode:433.3	Cerebral ischemia	-0.0633	35.31%	23.47%	46.28%
C3809991	Immunodeficiency common variable, 10	-0.0680	26.37%	14.29%	37.54%
C1856053	Hydranencephaly with renal aplasia-dysplasia	-0.0710	32.28%	17.33%	46.12%
RXNORM:11248	Cyanocobalamin	-0.0738	22.19%	14.21%	29.57%
C1262477	Weight decreased	-0.0764	31.96%	19.95%	43.06%
C4042877	Clinical decision-making	-0.0874	52.18%	34.48%	68.55%
C0042571	Vertigo	-0.0886	36.95%	22.51%	50.30%
PheCode:433	Cerebrovascular disease	-0.0983	44.20%	32.33%	55.18%
C0439044	Living alone	-0.1022	28.59%	14.77%	41.36%
C0034770	Mental recall	-0.1065	51.81%	34.41%	67.91%
RXNORM:17767	Amlodipine	-0.1265	23.35%	15.37%	30.73%
C0009676	Confusion	-0.1273	58.89%	43.27%	73.35%
C0168634	BaseLine dental cement	-0.1315	50.72%	31.88%	68.15%
C0038454	Cerebrovascular accident	-0.1366	35.72%	20.71%	49.61%
C0004093	Asthenia	-0.1435	44.88%	24.87%	63.38%
C5425799	All other	-0.1476	61.91%	42.80%	79.58%
C0042845	Vitamin B 12	-0.1610	38.65%	25.89%	50.46%

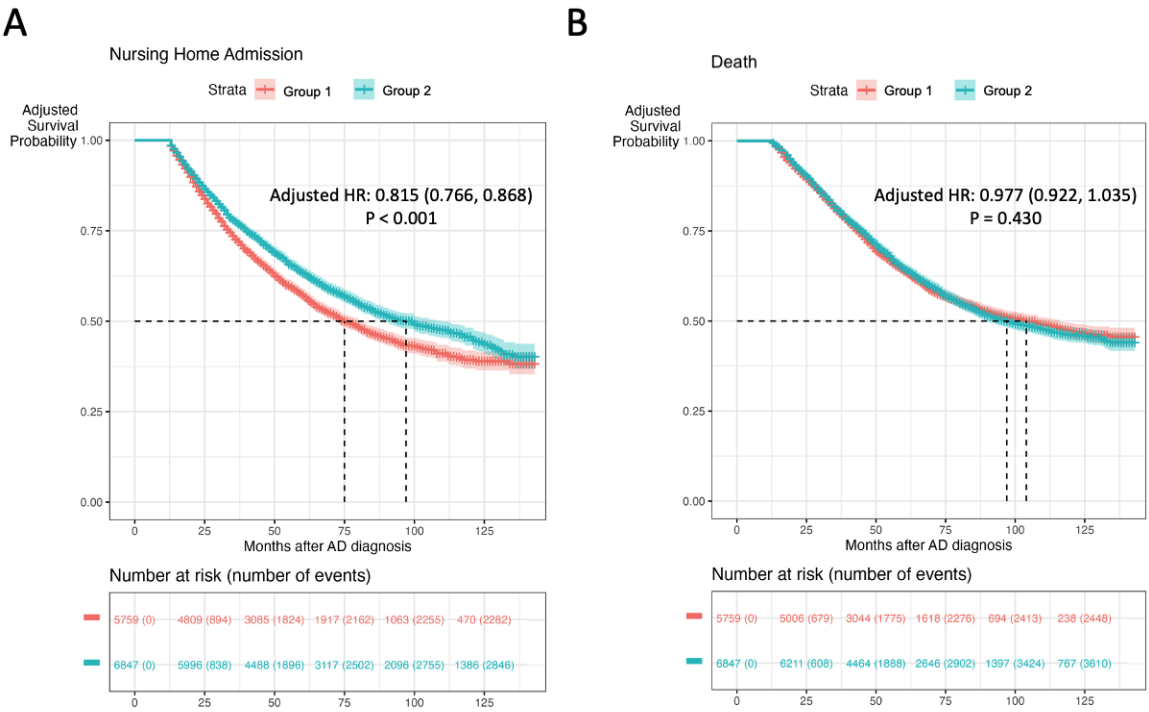
S-Figure 1. Graphical representation of our model. Squares indicate model parameters; circles indicate random variables; Filled-in shapes indicate observed variables. The indication [p] means a vector of size p.



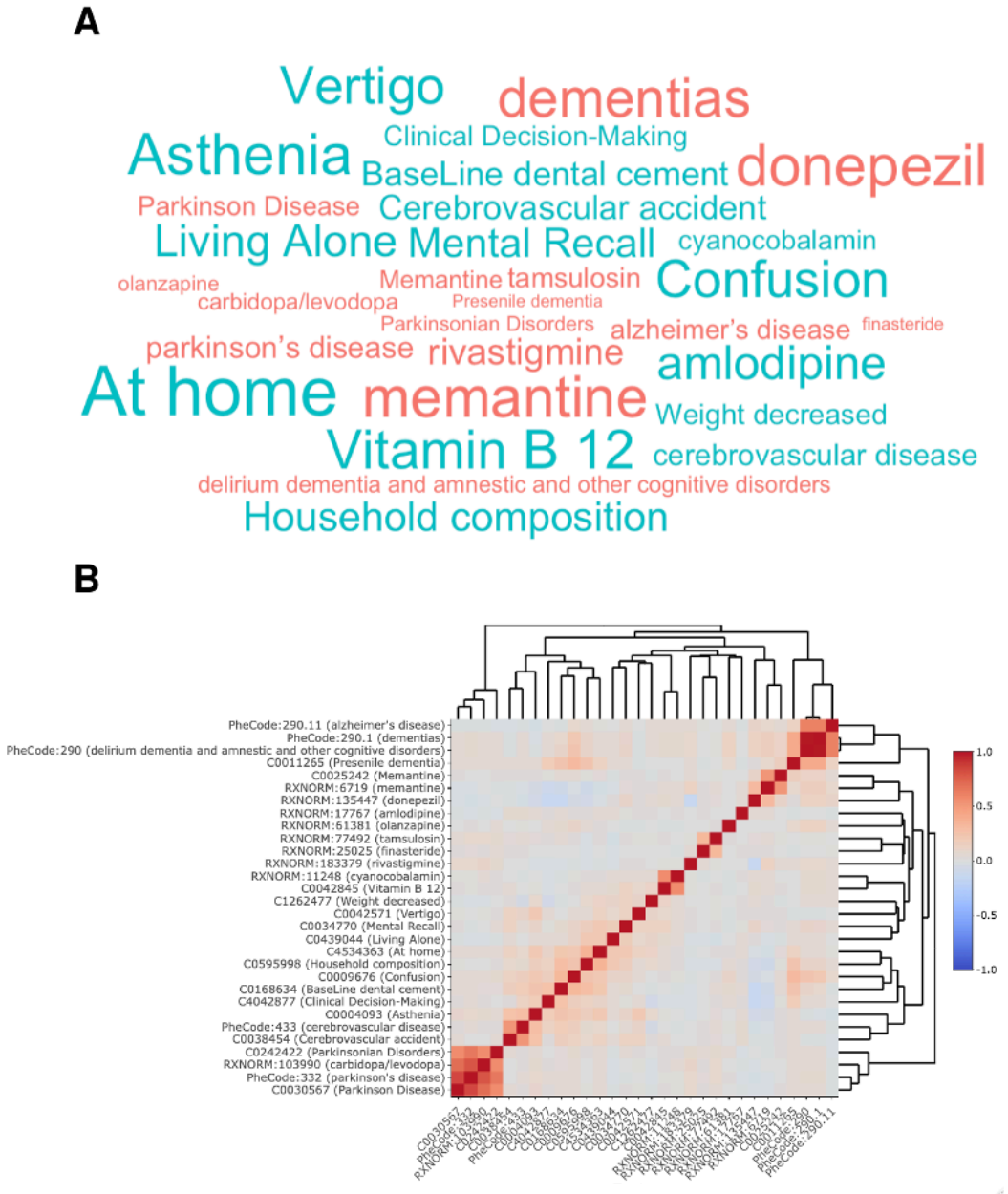
S-Figure 2. Baseline comorbidity burden of AD patient groups clustered one year after AD diagnosis.



S-Figure 3. Adjusted survival curve of (A) time to nursing home and (B) time to death for AD patient groups clustered one year after AD diagnosis.



S-Figure 4. Top 15 features contributing to differences between AD patient groups clustered one year after AD diagnosis. Features in salmon are positively associated with Group 1 membership, while features in teal are negatively associated with Group 1 membership. (A) Word cloud of the important features. Text size represents the relative importance of the feature. (B) Correlation matrix of the features.



SUPPLEMENTARY METHODS

To ensure the scalability of our model training, we implemented an algorithm based on variational Gaussian approximation and Expectation-Maximization [1]. At model convergence, the posterior probability of membership of the latent classes was calculated for each patient. The patient was assigned exclusively to the class with the highest posterior probability. The details are outlined below:

The complete-data log likelihood is

$$\log p(\boldsymbol{X},\boldsymbol{W},\boldsymbol{Y};\boldsymbol{B},\boldsymbol{\Lambda},a_0,a_1) = \sum_{i=1}^{N}\{\log p(\boldsymbol{X}_i \mid \boldsymbol{W}_i,U_i,Y_i;\boldsymbol{B},\boldsymbol{\Lambda},\pi_i) + \log p(\boldsymbol{W}_i) + \log p(Y_i;\pi_i)\}$$

Since $\boldsymbol{W}$ and $\boldsymbol{Y}$ are latent, the evaluations of the log-likelihood of the observed data

$$\ell_L(\boldsymbol{B},\boldsymbol{\Lambda},a_0,a_1) = \log p(\boldsymbol{X};\boldsymbol{B},\boldsymbol{\Lambda},a_0,a_1) = \log \int\int p(\boldsymbol{X},\boldsymbol{W},\boldsymbol{Y};\boldsymbol{B},\boldsymbol{\Lambda},a_0,a_1)d\boldsymbol{W}d\boldsymbol{Y}$$

is intractable, as well as its maximization with respect to $(\boldsymbol{B},\boldsymbol{\Lambda})$.

Variational approximation presents an alternative parameter estimation framework by using a computationally convenient approximating density in place of a more complex posterior density [2]. Using computationally convenient Gaussian densities, complex posterior distributions are approximated by minimizing the Kullback-Leibler (KL) divergence between the true and the approximating densities and therefore, reducing the computational overhead. Here instead of directly maximizing the log likelihood, we maximize the evidence lower bound (ELBO):

$$\begin{aligned}\mathscr{O}(\boldsymbol{B},\boldsymbol{\Lambda},q) &= \log p(\boldsymbol{X} \mid \boldsymbol{U};\boldsymbol{B},\boldsymbol{\Lambda}) - KL[q(\boldsymbol{W},\boldsymbol{Y} \mid \boldsymbol{X},\boldsymbol{U}), p(\boldsymbol{W},\boldsymbol{Y} \mid \boldsymbol{X},\boldsymbol{U};\boldsymbol{B},\boldsymbol{\Lambda}))] \\ &= \mathbb{E}_q[\log p(\boldsymbol{X},\boldsymbol{W},\boldsymbol{Y} \mid \boldsymbol{U};\boldsymbol{B},\boldsymbol{\Lambda})] - \mathbb{E}_q[\log q(\boldsymbol{W},\boldsymbol{Y} \mid \boldsymbol{X},\boldsymbol{U})]\end{aligned}$$

We used a mixture Gaussian approximation for

$$q(\boldsymbol{W}_i,Y_i=y \mid \boldsymbol{X}_i,U_i) = \underbrace{q(\boldsymbol{W}_i \mid Y_i=y,\boldsymbol{X}_i,U_i)}_{q_y(\boldsymbol{W}_i)}\underbrace{P(Y_i=y \mid \boldsymbol{X}_i,U_i)}_{\gamma_i}$$

with

$$q_y(\boldsymbol{W}_i \mid Y_i=y,\boldsymbol{X}_i,U_i) \sim \mathscr{N}(\boldsymbol{m}_i^{(y)},\boldsymbol{s}_i^{(y)}) \;\; for \;\; y=0,1.$$

Let $\boldsymbol{\Theta} = \{\boldsymbol{B}, \boldsymbol{\Lambda}, a_0, a_1\}$ and $\boldsymbol{\Omega} = (\boldsymbol{M}^{(0)}, \boldsymbol{S}^{(0)}, \boldsymbol{M}^{(1)}, \boldsymbol{S}^{(1)})$, our objective function is

$$
\begin{aligned}
\mathscr{O}(\boldsymbol{\Theta}, \boldsymbol{\Omega}; \boldsymbol{X}, \boldsymbol{U}) = \sum_{y=0}^{1} \mathbf{1}_n^\top \Big[\widehat{\boldsymbol{\Gamma}}_p^y (\mathbf{1} - \widehat{\boldsymbol{\Gamma}}_p)^{1-y} \odot \Big(\boldsymbol{X} \odot [(\boldsymbol{D}^{(y)} \boldsymbol{B} + \boldsymbol{M}^{(y)}) \boldsymbol{\Lambda}^{1/2} \boldsymbol{V}^\top] - \mathbb{E}_{q_y} \Big\{ \exp[(\boldsymbol{D}^{(y)} \boldsymbol{B} + \boldsymbol{W}) \boldsymbol{\Lambda}^{1/2} \boldsymbol{V}^\top] \Big\} \Big) \Big] \mathbf{1}_p \\
- \frac{1}{2} \mathbf{1}_n^\top \Big\{ \widehat{\boldsymbol{\Gamma}}_q^y (\mathbf{1} - \widehat{\boldsymbol{\Gamma}}_q)^{1-y} \odot (\boldsymbol{M}^{(y)} \odot \boldsymbol{M}^{(y)} + \boldsymbol{S}^{(y)} \odot \boldsymbol{S}^{(y)} - 2\log(\boldsymbol{S}^{(y)}) - \mathbf{1}_{n \times q}) \Big\} \mathbf{1}_q \\
+ \widehat{\boldsymbol{\Gamma}}^y (\mathbf{1} - \widehat{\boldsymbol{\Gamma}})^{1-y} \big[y \log\{expit(a_0 + a_1 \boldsymbol{U})\} + (1-y) \log\{1 - expit(a_0 + a_1 \boldsymbol{U})\} \big] \\
- \widehat{\boldsymbol{\Gamma}}^y (\mathbf{1} - \widehat{\boldsymbol{\Gamma}})^{1-y} \big[\log\{\widehat{\boldsymbol{\Gamma}}^y (1 - \widehat{\boldsymbol{\Gamma}}^y)\} \big]
\end{aligned}
$$

where $\boldsymbol{\Gamma} = (\gamma_1, \cdots, \gamma_N) \in \mathbb{R}^N$.

The algorithm alternates between two steps until convergence:

- **E-Step:** Let $\widehat{\mathscr{J}_i^{(y)}} = \mathscr{J}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}; \boldsymbol{X}_i, \boldsymbol{D}_i^{(y)}) = \mathbb{E}_{q_y}[\log p(\boldsymbol{X}_i, \boldsymbol{W}_i, Y_i = y \mid U_i)] - \mathbb{E}_{q_y}[\log q_y(\boldsymbol{W}_i \mid \boldsymbol{X}_i, Y_i = y, U_i)]$, we update
$$\widehat{\gamma}_i = \frac{\exp(\widehat{\mathscr{J}_i^{(1)}})}{\exp(\widehat{\mathscr{J}_i^{(1)}}) + \exp(\widehat{\mathscr{J}_i^{(0)}})}$$
- **M-Step:** Optimize for parameters $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) = \arg\max_{\boldsymbol{\Theta}, \boldsymbol{\Omega}} \mathscr{O}(\boldsymbol{\Theta}, \boldsymbol{\Omega}; \boldsymbol{X}, \boldsymbol{U})$, by alternating between the following two steps until convergence:
  - **Optimizing for $\boldsymbol{B}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Omega}$:** We first optimize for these parameters using the globally convergent method of moving asymptotes (MMA) algorithm for gradient-based local optimization implemented in the `NLOPT` optimization library in `C++` [3,4]. We use box constraints for the positive model parameters $\boldsymbol{\Lambda}$ and the variational parameters $\boldsymbol{S}$. The blockwise gradients are derived as follows:

$$
\begin{aligned}
\frac{\partial \mathscr{O}}{\partial \boldsymbol{M}^{(y)}} &= \big\{ \boldsymbol{\Gamma}^y (\mathbf{1} - \boldsymbol{\Gamma})^{1-y} \mathbf{1}_p^\top \odot (\boldsymbol{X} - \boldsymbol{A}^{(y)}) \big\} \boldsymbol{V} \boldsymbol{\Lambda}^{1/2} - \boldsymbol{M} \\
\frac{\partial \mathscr{O}}{\partial \boldsymbol{S}^{(y)}} &= \boldsymbol{S}^{\oslash} - \boldsymbol{S} - \big\{ \boldsymbol{\Gamma}^y (\mathbf{1} - \boldsymbol{\Gamma})^{1-y} \mathbf{1}_p^\top \odot \boldsymbol{A}^{(y)} \big\} (\boldsymbol{V} \boldsymbol{\Lambda}^{1/2} \odot \boldsymbol{V} \boldsymbol{\Lambda}^{1/2}) \odot \boldsymbol{S} \\
\frac{\partial \mathscr{O}}{\partial \boldsymbol{B}} &= \boldsymbol{\Lambda}^{1/2} \boldsymbol{V} \big\{ (\boldsymbol{X} - \boldsymbol{A}^{(0)})^\top (\mathbf{1} - \boldsymbol{\Gamma}) \mathbf{1}_d^\top \odot \boldsymbol{D}^{(0)} + (\boldsymbol{X} - \boldsymbol{A}^{(1)})^\top \boldsymbol{\Gamma} \mathbf{1}_d^\top \odot \boldsymbol{D}^{(1)} \big\} \\
\frac{\partial \mathscr{O}}{\partial \boldsymbol{\Lambda}^{1/2}} &= diag \Bigg\{ \boldsymbol{V}^\top \Bigg[(\boldsymbol{X} - \big\{ (\mathbf{1} - \boldsymbol{\Gamma}) \mathbf{1}_p^\top \odot \boldsymbol{A}^{(0)} + \boldsymbol{\Gamma} \mathbf{1}_p^\top \odot \boldsymbol{A}^{(1)} \big\})^\top (\boldsymbol{D}_L \boldsymbol{B} + \boldsymbol{M}_L) - \\
&\qquad [\big\{ (\mathbf{1} - \boldsymbol{\Gamma}) \mathbf{1}_p^\top \odot \boldsymbol{A}^{(0)} + \boldsymbol{\Gamma} \mathbf{1}_p^\top \odot \boldsymbol{A}^{(1)} \big\}^\top (\boldsymbol{S} \odot \boldsymbol{S})] \odot \boldsymbol{V} \boldsymbol{\Lambda}^{1/2} \Bigg] \Bigg\}
\end{aligned}
$$

– **Optimizing for a_0 and a_1:** We then update (a_0, a_1) by optimizing the objective function

$$(\widehat{a}_0, \widehat{a}_1) = \underset{a_0, a_1}{\arg\max} \sum_{i=1}^{N} \widehat{\gamma}_i \log\{expit(a_0 + a_1 U_i)\} + (1 - \widehat{\gamma}_i) \log\{1 - expit(a_0 + a_1 U_i)\}$$

We can do this by fitting weighted logistic regression with outcomes $(\mathbf{0}_N, \mathbf{1}_N)$, weights $(\mathbf{1} - \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\gamma}})$ and covariates $(\boldsymbol{U}, \boldsymbol{U})$.

REFERENCES FOR SUPPLEMENTARY MATERIALS

1. Moon TK. The expectation-maximization algorithm. *IEEE Signal processing magazine*. 1996 Nov;13(6):47-60.

2. Wainwright MJ, Jordan MI. Graphical Models, Exponential Families, and Variational Inference. *Found Trends Mach Learn*. 2007;1(1-2):1-305. doi:10.1561/220000000170.

3. Svanberg K. The method of moving asymptotes-a new method for structural optimization. *International journal for numerical methods in engineering*. 1987 Feb;24(2):359-73.

4. Steven G. Johnson, The NLopt nonlinear-optimization package, http://github.com/stevengj/nlopt.

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. 2024 Alzheimer's disease facts and figures. *Alzheimer's Dement*. Published online 2024. doi:10.1002/alz.13809

3. 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

4. Dyck CH van, Swanson CJ, Aisen P, et al. Lecanemab in Early Alzheimer's Disease. *N Engl J Med*. 2022;388(1):9-21. doi:10.1056/nejmoa2212948

5. Sims JR, Zimmer JA, Evans CD, et al. Donanemab in Early Symptomatic Alzheimer Disease. *JAMA*. 2023;330(6):512-527. doi:10.1001/jama.2023.13239

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. 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

11. Adkins‐Jackson PB, George KM, Besser LM, et al. The structural and social determinants of Alzheimer's disease related dementias. Alzheimer's Dement. 2023;19(7):3171-3185. doi:10.1002/alz.13027

12. Abdelnour C, Agosta F, Bozzali M, et al. Perspectives and challenges in patient stratification in Alzheimer's disease. *Alzheimer's Res Ther*. 2022;14(1):112. doi:10.1186/s13195-022-01055-y

13. Hernández-Lorenzo L, García-Gutiérrez F, Solbas-Casajús A, Corrochano S, Matías-Guiu JA, Ayala JL. Genetic-based patient stratification in Alzheimer's disease. *Sci Rep*. 2024;14(1):9970. doi:10.1038/s41598-024-60707-1

14. Sadiq MU, Kwak K, Dayan E, Initiative for the ADN. Model-based stratification of progression along the Alzheimer disease continuum highlights the centrality of biomarker synergies. Alzheimer's Res Ther. 2022;14(1):16. doi:10.1186/s13195-021-00941-1

15. Luckett PH, Chen C, Gordon BA, et al. Biomarker clustering in autosomal dominant Alzheimer's disease. Alzheimer's Dement. 2023;19(1):274-284. doi:10.1002/alz.12661

16. Tijms BM, Gobom J, Reus L, et al. Pathophysiological subtypes of Alzheimer's disease based on cerebrospinal fluid proteomics. *Brain*. 2020;143(12):3776-3792. doi:10.1093/brain/awaa325

17. Habes M, Grothe MJ, Tunc B, McMillan C, Wolk DA, Davatzikos C. Disentangling Heterogeneity in Alzheimer's Disease and Related Dementias Using Data-Driven Methods. *Biol Psychiatry*. 2020;88(1):70-82. doi:10.1016/j.biopsych.2020.01.016

18. Vogel JW, Young AL, Oxtoby NP, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med*. 2021;27(5):871-881. doi:10.1038/s41591-021-01309-6

19. Poulakis K, Pereira JB, Muehlboeck JS, et al. Multi-cohort and longitudinal Bayesian

clustering study of stage and subtype in Alzheimer's disease. *Nat Commun*. 2022;13(1):4566. doi:10.1038/s41467-022-32202-6

20. Petersen KK, Lipton RB, Grober E, Nallapu BT, Ezzati A. MRI-guided Clustering of Alzheimer's Disease patients: A post-hoc analysis of Phase 3 Trial of Solanezumab for Mild Dementia Due to Alzheimer's Disease. Alzheimer's Dement. 2022;18(S6). doi:10.1002/alz.063313

21. Alashwal H, Halaby ME, Crouse JJ, Abdalla A, Moustafa AA. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Front Comput Neurosci*. 2019;13:31. doi:10.3389/fncom.2019.00031

22. Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol*. 2011;10(9):785-796. doi:10.1016/s1474-4422(11)70156-9

23. Wang D, Ma X, Schulz PE, Jiang X, Kim Y. Knowledge-guided Deep Temporal Clustering for Alzheimer's Disease Subtypes in Completed Clinical Trials. *medRxiv*. Published online 2023:2023.10.13.23296985. doi:10.1101/2023.10.13.23296985

24. Birkenbihl C, Jong J de, Yalchyk I, Fröhlich H, Initiative the ADN. Deep learning-based patient stratification for prognostic enrichment of clinical dementia trials. medRxiv. Published online 2023:2023.11.25.23299015. doi:10.1101/2023.11.25.23299015

25. Xu J, Wang F, Xu Z, et al. Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn Heal Syst*. 2020;4(4):e10246. doi:10.1002/lrh2.10246

26. Tang AS, Oskotsky T, Havaldar S, et al. Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. Nat Commun. 2022;13(1):675. doi:10.1038/s41467-022-28273-0

27. He Z, Tian S, Erdengasileng A, Charness N, Bian J. Temporal Subtyping of Alzheimer's

Disease Using Medical Conditions Preceding Alzheimer's Disease Onset in Electronic Health Records. *Amia Annu Symposium Proc Amia Symposium*. 2022;2022:226-235.

28. Alexander N, Alexander DC, Barkhof F, Denaxas S. Identifying and evaluating clinical subtypes of Alzheimer's disease in care electronic health records using unsupervised machine learning. *Bmc Med Inform Decis*. 2021;21(1):343. doi:10.1186/s12911-021-01693-6

29. Landi I, Glicksberg BS, Lee HC, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *npj Digit Med*. 2020;3(1):96. doi:10.1038/s41746-020-0301-z

30. 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

31. 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

32. 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

33. 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

34. 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

35. 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

36. 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

37. 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

38. 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

39. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *J Biomed Inform*. 2022;126:103983. doi:10.1016/j.jbi.2021.103983

40. Gasparini A. comorbidity: An R package for computing comorbidity scores. *J Open Source Softw*. 2018;3(23):648. doi:10.21105/joss.00648

41. Haaksma ML, Vilela LR, Marengoni A, et al. Comorbidity and progression of late onset Alzheimer's disease: A systematic review. *PLoS ONE*. 2017;12(5):e0177044. doi:10.1371/journal.pone.0177044

42. Santiago JA, Potashkin JA. The Impact of Disease Comorbidities in Alzheimer's Disease. *Front Aging Neurosci*. 2021;13:631770. doi:10.3389/fnagi.2021.631770

43. McGleenon, Dynan, Passmore. Acetylcholinesterase inhibitors in Alzheimer's disease. *Br J Clin Pharmacol*. 1999;48(4):471-480. doi:10.1046/j.1365-2125.1999.00026.x

44. Reisberg B, Doody R, Stöffler A, et al. Memantine in Moderate-to-Severe Alzheimer's Disease. *New Engl J Medicine*. 2003;348(14):1333-1341. doi:10.1056/nejmoa013128

45. Barthold D, Joyce G, Ferido P, et al. Pharmaceutical Treatment for Alzheimer's Disease and Related Dementias: Utilization and Disparities. *J Alzheimer's Dis*. 2020;76(2):579-589. doi:10.3233/jad-200133

46. Holm E, Jacobsen KK, Lony TB de, et al. Frequency of missed or delayed diagnosis in dementia is associated with neighborhood socioeconomic status. *Alzheimer's Dement: Transl Res Clin Interv*. 2022;8(1):e12271. doi:10.1002/trc2.12271

47. Arvanitakis Z, Capuano AW, Leurgans SE, Bennett DA, Schneider JA. Relation of cerebral vessel disease to Alzheimer's disease dementia and cognitive function in elderly people: a cross-sectional study. Lancet Neurol. 2016;15(9):934-943. doi:10.1016/s1474-4422(16)30029-1

48. Kapasi A, DeCarli C, Schneider JA. Impact of multiple pathologies on the threshold for clinically overt dementia. *Acta Neuropathol*. 2017;134(2):171-186. doi:10.1007/s00401-017-1717-7

49. Cederwald BF von, Josefsson M, Wåhlin A, Nyberg L, Karalija N. Association of Cardiovascular Risk Trajectory With Cognitive Decline and Incident Dementia. *Neurology*. 2022;98(20):e2013-e2022. doi:10.1212/wnl.0000000000200255

50. Coomans EM, Westen D van, Binette AP, et al. Interactions between vascular burden and amyloid-β pathology on trajectories of tau accumulation. Brain. 2023;147(3):949-960. doi:10.1093/brain/awad317

51. Baker J, Libretto T, Henley W, Zeman A. A Longitudinal Study of Epileptic Seizures in Alzheimer's Disease. *Front Neurol*. 2019;10:1266. doi:10.3389/fneur.2019.01266

52. Vossel KA, Tartaglia MC, Nygaard HB, Zeman AZ, Miller BL. Epileptic activity in Alzheimer's disease: causes and clinical relevance. Lancet Neurol. 2017;16(4):311-322. doi:10.1016/s1474-4422(17)30044-3

53. Vossel K. Putting the Brakes on Accelerated Cognitive Decline in Alzheimer's Disease with Epileptic Activity. J Alzheimer's Dis. 2023;94(3):1075-1077. doi:10.3233/jad-230613

54. Aarsland D, Batzu L, Halliday GM, et al. Parkinson disease-associated cognitive impairment. *Nat Rev Dis Prim*. 2021;7(1):47. doi:10.1038/s41572-021-00280-3

55. Garcia-Argibay M, Hiyoshi A, Fall K, Montgomery S. Association of 5α-Reductase Inhibitors With Dementia, Depression, and Suicide. *JAMA Netw Open*. 2022;5(12):e2248135. doi:10.1001/jamanetworkopen.2022.48135

56. Street JS, Clark WS, Gannon KS, et al. Olanzapine Treatment of Psychotic and Behavioral Symptoms in Patients With Alzheimer Disease in Nursing Care Facilities: A Double-blind,

Randomized, Placebo-Controlled Trial. *Arch Gen Psychiatry*. 2000;57(10):968-976. doi:10.1001/archpsyc.57.10.968

57. Ismail Z, Creese B, Aarsland D, et al. Psychosis in Alzheimer disease — mechanisms, genetics and therapeutic opportunities. *Nat Rev Neurol*. 2022;18(3):131-144. doi:10.1038/s41582-021-00597-3

58. Tampi RR, Tampi DJ, Balachandran S, Srinivasan S. Antipsychotic use in dementia: a systematic review of benefits and risks from meta-analyses. *Ther Adv Chronic Dis*. 2016;7(5):229-245. doi:10.1177/2040622316658463

59. Yunusa I, Alsumali A, Garba AE, Regestein QR, Eguale T. Assessment of Reported Comparative Effectiveness and Safety of Atypical Antipsychotics in the Treatment of Behavioral and Psychological Symptoms of Dementia. JAMA Netw Open. 2019;2(3):e190828. doi:10.1001/jamanetworkopen.2019.0828

60. Chen H, Liu S, Ge B, et al. Effects of Folic Acid and Vitamin B12 Supplementation on Cognitive Impairment and Inflammation in Patients with Alzheimer's Disease: A Randomized, Single-Blinded, Placebo-Controlled Trial. *J Prev Alzheimer's Dis*. 2021;8(3):249-256. doi:10.14283/jpad.2021.22

61. Harun A, Oh ES, Bigelow RT, Studenski S, Agrawal Y. Vestibular Impairment in Dementia. *Otol Neurotol*. 2016;37(8):1137-1142. doi:10.1097/mao.0000000000001157

62. Peters R, Yasar S, Anderson CS, et al. Investigation of antihypertensive class, dementia, and cognitive decline: A meta-analysis. Neurology. 2020;94(3):e267-e281. doi:10.1212/wnl.0000000000008732

63. Zhang J, Zheng X, Zhao Z. A systematic review and meta-analysis on the efficacy outcomes of selective serotonin reuptake inhibitors in depression in Alzheimer's disease. *BMC Neurol*. 2023;23(1):210. doi:10.1186/s12883-023-03191-w

64. 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

65. Dubois B, Hampel H, Feldman HH, et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. Alzheimer's Dement. 2016;12(3):292-323. doi:10.1016/j.jalz.2016.02.002

66. Rafii MS, Aisen PS. Detection and treatment of Alzheimer's disease in its preclinical stage. *Nat Aging*. 2023;3(5):520-531. doi:10.1038/s43587-023-00410-4

67. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020;396(Dement Geriatr Cogn Dis 37 2014):413-446. doi:10.1016/s0140-6736(20)30367-6

68. Duara R, Barker W. Heterogeneity in Alzheimer's Disease Diagnosis and Progression Rates: Implications for Therapeutic Trials. *Neurotherapeutics*. 2022;19(1):8-25. doi:10.1007/s13311-022-01185-z

69. Reitz C. Toward precision medicine in Alzheimer's disease. *Ann Transl Med*. 2016;4(6):107-107. doi:10.21037/atm.2016.03.05