

Evaluating Patient Representation Strategies for Early Prediction of Alzheimer's Disease and Related Dementias from Longitudinal EHRs

Shahla Farzana

sfarzana@uchicago.edu

Institute for Population and Precision Health

Ash Arian

Pritzker School of Medicine

Moïse Desvarieux³

Columbia University

Tatjana Rundek⁴ Rundek

University of Miami

Habibul Ahsan

Institute for Population and Precision Health

Article

Keywords: Alzheimer's, Patient representation, Temporal modeling, EHRs

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10988118/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Evaluating Patient Representation Strategies for Early Prediction of Alzheimer’s Disease and Related Dementias from Longitudinal EHRs

Shahla Farzana^{1*}, Ash Arian^{1,2}, Moïse Desvarieux³,
Tatjana Rundek⁴, Habibul Ahsan^{1,5}

^{1*}Institute for Population and Precision Health, University of Chicago,
Chicago, IL, USA.

²Pritzker School of Medicine, University of Chicago, Chicago, IL, USA.

³Department of Epidemiology, Mailman School of Public Health,
Columbia University, New York, USA.

⁴Department of Neurology, Miller School of Medicine, University of
Miami, Miami, Florida, USA.

⁵Department of Family Medicine, Biological Sciences Division,
University of Chicago Medicine, Chicago, IL, USA.

*Corresponding author(s). E-mail(s): sfarzana@uchicago.edu;
Contributing authors: ashfaqarian@uchicago.edu;
mdesvarieux@columbia.edu; TRundek@med.miami.edu;
hahsan@bsd.uchicago.edu;

Abstract

The neurodegenerative processes underlying Alzheimer’s disease and related dementias (ADRD) begin years before clinical diagnosis, motivating scalable approaches for earlier identification. Electronic health records (EHRs) provide longitudinal clinical information for early ADRD prediction but are sparse, temporal, and hierarchical, and alternative EHR representations encode these characteristics differently. We systematically evaluated how EHR representation and temporal distance from diagnosis influence ADRD prediction and whether representation-specific performance generalizes across health systems. Using All of Us EHRs, we compared interpretable count-based representations with pretrained clinical foundation-model representations at clinically meaningful lead times across multiple ADRD case definitions. Count-based models

achieved the strongest overall discrimination and calibration. Performance generally declined with increasing prediction horizons, while differences between representations narrowed; at 36 months, pretrained representations achieved comparable AUROC (0.719 versus 0.738) with higher sensitivity and F1 at a fixed operating threshold. Temporal feature attribution showed greater model reliance on healthcare-utilization indicators at longer horizons and increasing importance of cognitive phenotypes, including amnesia and mild cognitive impairment, nearer diagnosis. Zero-shot evaluation on independent UChicago Medicine EHRs showed substantial performance degradation across representations, demonstrating challenges in cross-institutional transportability. These findings characterize representation-dependent tradeoffs in early ADRD prediction across temporal horizons and healthcare systems.

Keywords: Alzheimer’s, Patient representation, Temporal modeling, EHRs

1 Introduction

Alzheimer’s disease and related dementias (ADRD) affect more than six million people 65 years and older in the United States and remain among the leading causes of disability and mortality in older adults [1]. ADRD develops over years and often within complex trajectories of multimorbidity. Because neuropathological changes precede clinical diagnosis, identifying individuals at elevated risk earlier may support management of modifiable risk factors, care planning, clinical-trial recruitment, and timely initiation of emerging therapies. Yet early recognition in routine care remains difficult because cognitive manifestations can be subtle, heterogeneous, and under-recognized.

Electronic health records (EHRs) offer a scalable source of longitudinal information for clinical prediction without requiring additional imaging or biomarker acquisition. However, EHR histories are variable-length, high-dimensional, sparse, temporally irregular, and hierarchically structured, making patient representation a central modeling challenge. Conventional approaches transform longitudinal histories into fixed-length engineered features, such as concept counts and temporal bins, preserving clinical semantics but compressing temporal context. In contrast, representation-learning approaches derive latent patient embeddings from longitudinal event sequences to capture richer temporal and hierarchical structure [2]. Greater representational complexity, however, does not consistently translate into superior clinical prediction: strong count-based baselines remain competitive for several tasks, while pretrained representations provide advantages in particular outcomes and data regimes [3]. Pretrained EHR representations have also shown greater robustness than count-based approaches under temporal distribution shift for several clinical outcomes, although improvements vary across tasks and metrics [4]. Collectively, these findings suggest that the utility of an EHR representation depends on the prediction task, temporal context, and deployment setting.

EHR-based ADRD studies have similarly employed diverse representation and modeling strategies. Data-driven high-dimensional EHR features have outperformed

knowledge-driven risk models at prespecified intervals before diagnosis [5], while subsequent work has incorporated temporal models, pretrained EHR representations, knowledge-network-based interpretation, and large-language-model-assisted prediction [6–8]. More recently, dynamic frameworks have modeled irregular longitudinal observations to update ADRD risk as clinical information accumulates, enabling time-resolved prediction and interpretation [9]. However, prior studies have largely compared learning algorithms within engineered representations, evaluated individual learned-representation paradigms, or interpreted predictors at selected time points. How alternative patient representations perform relative to one another as prediction moves farther from documented ADRD onset therefore remains insufficiently characterized. This question is particularly relevant because increasing lead time changes both the availability of proximal disease-specific signals and the composition of observable clinical history. Moreover, whether representation-specific performance persists under cross-institutional distribution shift remains unclear, despite evidence of representation-dependent robustness to temporal shift in other clinical prediction tasks.

We therefore systematically evaluated interpretable count-based and pretrained clinical foundation-model representations for ADRD onset prediction at 6-, 12-, 24-, and 36-month lead times across multiple ADRD case definitions. We asked whether the relative predictive utility of these representations changes with temporal distance from documented diagnosis and whether temporally aligned self-reported lifestyle, personal and family health-history factors provide incremental information beyond routinely collected longitudinal EHR data. We further characterized how the clinically interpretable information supporting count-based predictions varies across prediction horizons and evaluated whether representation-specific performance generalizes to an independent health system under institution-specific training and zero-shot cross-institutional transfer.

2 Results

2.1 Study Cohort Characteristics

Because our CD1 (i.e., have at least one encounter with a valid ADRD diagnosis code or at least one encounter of anti-dementia medication) has the most relaxed rules, Table 1 shows the descriptive statistics of the case and control groups identified using CD1 for All of US (AoU) and UChicago Medicine (UCM) (External validation) datasets. Descriptive statistics of the secondary-confirmed cohort (have at least one encounter with a valid ADRD diagnosis code followed by anti-dementia medication prescription) CD2 are detailed in the Supplementary Table S3. The demographic composition between AoU and UCM dataset differs substantially. UCM includes an older population (median index age approximately 79 years vs. 71 years in AoU), a substantially higher proportion of Black patients (62% vs. 13%), fewer White patients (33% vs. 65%), and fewer Hispanic participants. There is also striking difference in clinical history density between two datasets. Compared with AoU, UCM patients have substantially shorter longitudinal follow-up (median history 8.2 vs. 13.8 years for cases; 7.7 vs. 10.7 years for controls) and markedly fewer recorded encounters (62.4 vs.

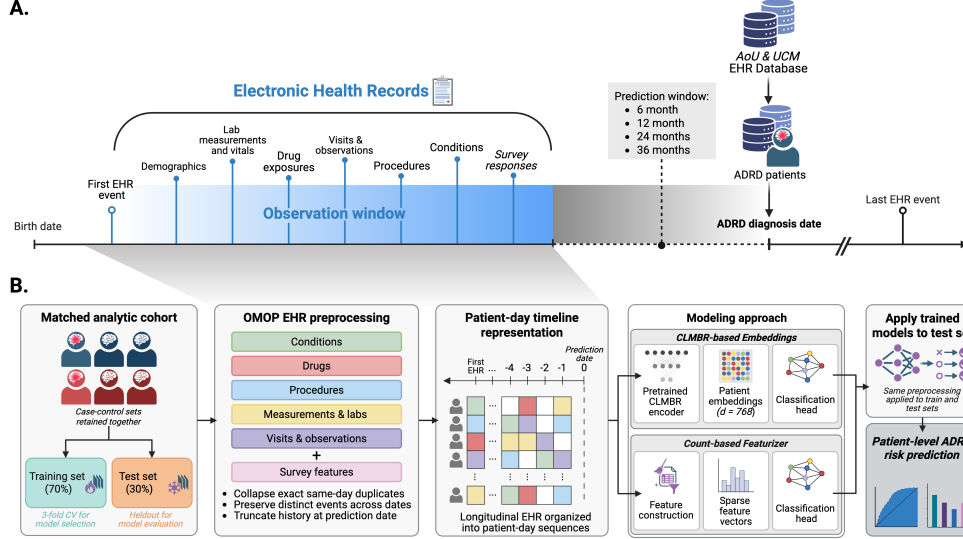


Fig. 1: Study design and modeling framework for EHR-based ADRD onset prediction. (A) Temporal design of the prediction task, defining the longitudinal observation period and prediction date relative to ADRD onset; only information recorded before the prediction date was available to the models. (B) End-to-end modeling workflow in which matched case-control sets were transformed from longitudinal OMOP-format EHR data into patient-day representations and partitioned at the matched-set level to prevent data leakage. Models were trained using either pre-trained Clinical Language Model-Based Representations (CLMBR) with a supervised classification head or engineered count-based features with classical machine-learning classifiers. Model and hyperparameter selection were performed exclusively within the training partition, after which the learned mapping between patient representations and ADRD outcome was applied without retraining to the held-out test set to generate patient-level risk estimates.

325.3 encounters for cases; 55.4 vs. 180.2 for controls). Similar reductions are observed across ambulatory, inpatient, and emergency encounters. These differences indicate that the external validation dataset represents a distinct patient population rather than a simple replication of AoU, providing a rigorous assessment of generalizability.

2.2 ADRD Risk Prediction Performance

In this subsection, we present the overall results of the experimental settings for ADRD risk prediction.

Discriminative performance across prediction horizons. Figure 2 shows the comparative performance of each representation strategy on held-out test set selected by their performance on 3-fold cross validation. Across the four prediction horizons, the count-based model consistently achieved the best overall predictive performance

Table 1: Demographic characteristics, healthcare utilization, and EHR history for the CD1 cohort. Number of encounters are reported as *mean (SD)*.

		AoU		UCM	
Variable	Sub-categories	Case (N=3,496)	Control (N=6,984)	Case (N=3,629)	Control (N=7,260)
Demographics					
Age	at index date, median	71.1	71.1	78.9	78.9
Gender	Female	53.3%	53.3%	63.5%	63.5%
	Male	45.1%	45.2%	36.5%	36.5%
	Unknown	1.5%	1.5%	-	-
Race	White	64.8%	64.8%	33.0%	33.0%
	Unknown	16.8%	16.8%	3.3%	3.2%
	Black or African American	13.1%	13.1%	62.1%	62.2%
	Others	4.3%	4.2%	0.3%	0.2%
Ethnicity	Asian	1.1%	1.1%	1.3%	1.3%
	NonHispanic	81.9%	82.0%	95.0%	95.2%
	Hispanic	14.3%	14.2%	3.3%	2.3%
	Unknown	3.8%	3.8%	1.6%	2.6%
Medical history length					
History years	median	13.8 ^a	10.7 ^a	8.2 ^a	7.7 ^a
Number of years, range	[4,6)	6.4% ^b	11.3% ^b	19.5% ^b	22.7% ^b
	[6,+)	93.6% ^b	88.7% ^b	80.5% ^b	77.3% ^b
Number of encounters	All types of encounters	325.3 ^a (452.1)	180.2 ^a (282.8)	62.4 ^a (75.3)	55.4 ^a (65.8)
	Ambulatory encounters	219.1 ^a (350.8)	121.1 ^a (229.1)	55.4 (70.3)	50.9 (62.3)
	Inpatient encounters	12.2 ^a (52.6)	4.8 ^a (20.1)	3.5 ^a (7.7)	2.3 ^a (5.6)
	Emergency encounters	7.8 ^a (21.5)	3.8 ^a (12.0)	3.6 ^a (8.6)	2.2 ^a (4.9)

^a Significant difference between case and control using Mann-Whitney U test.

^b Significant difference between case and control using Pearson chi-square test.

for the CD1 case definition. At the 6-month horizon, it substantially outperformed all clinical foundation models, achieving an AUROC of 0.827 and an AUPRC of 0.722, compared with AUROCs of 0.752–0.784 and AUPRCs of 0.608–0.662 for the embedding-based models. This advantage persisted at the 12-, 24-, and 36-month horizons, although the performance gap gradually narrowed as the prediction horizon increased. By 36 months, the count-based model retained the highest AUROC (0.738 vs. 0.682–0.719) and AUPRC (0.589 vs. 0.517–0.560), but the differences relative to the strongest embedding-based models were notably smaller than at the 6-month horizon. In the secondary-confirmed cohort (CD2), which contained the fewest case and control participants among the three case definitions, all representations exhibited lower discriminative performance than in CD1 across prediction horizons. Unlike the other case

definitions, discrimination did not decline monotonically with increasing prediction horizon. The highest AUROC and AUPRC were observed at the 12-month horizon for the count-based model (AUROC 0.802, AUPRC 0.690), followed by a gradual decline at the 24- and 36-month horizons. Among embedding-based models, GPT-2048 and LLaMA-512 generally provided the strongest discrimination, with AUROC for GPT-2048 and LLaMA-512 ranging from 0.767 and 0.759 at 6-month horizon to 0.659 and 0.647 at 36-month horizon respectively. Similar to CD1, the performance gap between count-based and embedding-based approaches narrowed modestly at longer prediction horizons in CD2. Institution-specific training on UChicago EHRs showed broadly similar horizon-dependent declines in discrimination, with count-based models generally achieving the strongest AUROC and AUPRC, although relative performance varied by case definition and prediction horizon (Supplementary Figure S2). Zero-shot transfer from AoU to UCM resulted in lower discrimination than institution-specific training and AoU evaluation overall; pretrained representations were generally more competitive than count-based representations in zero-shot AUROC, although the relative advantage varied by architecture, case definition, and prediction horizon, while AUPRC differences were less consistent (Supplementary Figure S3). This performance decline is expected given the substantial differences between two datasets, including older patients, a markedly different racial composition, shorter longitudinal clinical histories, and differences in healthcare utilization patterns in UCM dataset compared with AoU (Table 1, Supplementary Table S3). Overall, these findings suggest that manually engineered count-based features capture short-term disease signals particularly effectively, while the relative discriminative performance of pretrained clinical foundation models improves for longer-term risk prediction. Across all representation strategies, incorporating survey-derived features produced only marginal improvements over EHR-only models. Across case definitions and prediction horizons, AUROC gains did not exceed 1.4% (for CD1), with similarly modest changes in AUPRC and F1 score. The limited benefit likely reflects the substantial missingness of pre-index survey data, although smoking-related variables contributed the greatest incremental value because they were the most frequently available survey features.

Threshold-dependent classification performance. In contrast to AUROC and AUPRC, the F1-score showed a different trend across all case definitions. Although the count-based model achieved the highest F1-score at the 6-month horizon (0.617) for CD1, embedding-based models became increasingly competitive at longer prediction horizons. At 24 and 36 months, several foundation models, particularly GPT-2048, and Llama-2048, achieved F1-scores which exceeded those of the count-based model. F1-score exhibited the similar pattern across CD2. For example, while the count-based model achieved the highest F1 score at the 6-month horizon (0.597), its performance declined substantially to 0.429 at 36 months for CD2. In contrast, several embedding-based models maintained relatively stable F1 scores across prediction horizons, with GPT-2048 and LLaMA-512 achieving 0.539 at 36 months, exceeding the count-based model despite lower discriminative performance. F1 scores declined overall under zero-shot transfer relative to institution-specific training and AoU evaluation, but the magnitude and ordering of this decline varied across representations, pretrained architectures, case definitions, and prediction horizons

(Supplementary Figures S2 and S3). These findings suggest that, despite lower overall discriminative performance, foundation-model embeddings provide a more favorable balance between precision and recall under a fixed decision threshold for longer-term prediction. To further evaluate the robustness of model performance across decision thresholds, threshold sensitivity analyses are presented in the Supplementary Figure S4.

Model calibration. Calibration trends were similar across all three case definitions. Models were generally well calibrated at shorter prediction horizons but exhibited progressively poorer calibration as the prediction horizon increased from 6 to 36 months. Reliability diagram for representative models for CD1 are provided in Supplementary Figure S5. Correspondingly, Brier scores increased with prediction horizon for all representation strategies. For example, for CD1 at 6 months the brier score for the count-based model vs. the best-performing embedding-based model was 0.169 and 0.189, whereas at 36 months horizon the score increased to 0.198 and 0.225 respectively, indicating greater uncertainty and reduced calibration accuracy for long-term ADRD risk prediction. Similarly, for CD2, The count-based model consistently demonstrated the lowest Brier scores, increasing from 0.172 at 6 months to 0.196 at 36 months, whereas embedding-based models exhibited higher Brier scores throughout (approximately 0.211–0.240). This trend indicates increasing predictive uncertainty for longer-term ADRD risk prediction, while also confirming that count-based models remained better calibrated than embedding-based approaches across all prediction horizons.

Effect of foundation model context length on ADRD prediction performance. Figure 2 compares GPT-512, GPT-2048, Llama-512, and Llama-2048, which differ only in their pretrained maximum context lengths. All models were constructed from the same pre-cutoff clinical histories. For patients whose tokenized histories exceeded the model’s context length, only the most recent events up to the model’s maximum input length (512 or 2048 tokens) were retained. Consequently, this comparison evaluates the combined effect of pretrained architecture and available historical context. Across case definitions and prediction horizons, increasing the maximum context length resulted in only modest and architecture-dependent differences. In CD1, Llama-2048 generally achieved slightly higher discriminative performance than Llama-512 across prediction horizons. In CD2, performance differences between the two models were smaller and less consistent; however, Llama-2048 tended to outperform Llama-512 at the longer prediction horizons (24 and 36 months), whereas little difference was observed at the shorter horizons. Similarly, GPT-2048 exhibited only marginal and horizon-dependent differences relative to GPT-512, without a consistent performance advantage across either cohort. These patterns were consistent between cross-validation and held-out test evaluations. Overall, increasing the maximum context length of pretrained models did not consistently improve ADRD risk prediction.

2.3 Model Reliance on Clinically Interpretable EHR Features

Figure 3 summarizes the relative contribution of clinically interpretable features (top 10) to the XGBoost models across case definitions and prediction horizons, while

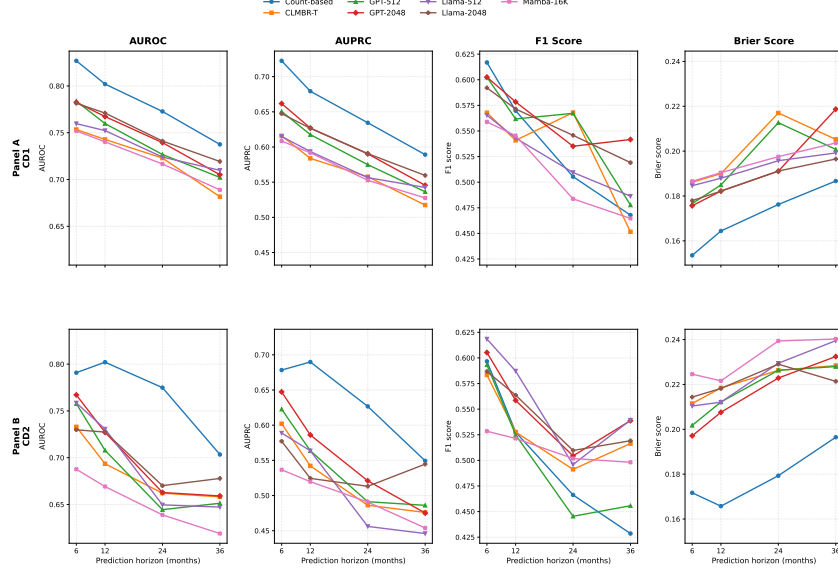


Fig. 2: Held-out test set performance of the best-performing feature variant within each representation strategy across all case definitions and prediction horizons for AoU dataset.

Supplementary Figure S6 presents the complete set of the top 20 predictors. Across cohorts, memory-related diagnoses (e.g., amnesia and minimal cognitive impairment) became increasingly important closer to the prediction date. Across all case definitions, memory-related diagnoses, including amnesia and minimal cognitive impairment, became progressively more influential as the prediction date approached, consistent with increasing manifestation of cognitive symptoms before ADRD diagnosis. In contrast, broader indicators of healthcare utilization, including drug utilization, visit frequency, observation frequency, and diagnosis frequency, contributed relatively more at longer prediction horizons, suggesting that patterns of healthcare engagement capture early risk before disease-specific manifestations emerge. Psychiatric conditions, particularly depressive disorder and mental disorder, showed persistent but modest contributions across horizons, whereas laboratory measurements (e.g., vitamin B12), socioeconomic status (household income), and lifestyle-related variables (smoking history) exhibited smaller and more cohort-specific effects. Additional clinically relevant predictors, including, Parkinson’s disease, dermatological condition appeared among the top 20 predictors for specific case definitions or prediction horizons. Overall, these findings suggest that the models assign greater importance to nonspecific indicators of health complexity and healthcare utilization at longer prediction horizons, whereas disease-specific cognitive phenotypes increasingly dominate prediction closer to ADRD onset.

Relative model reliance on clinically interpretable features across ADRD prediction horizons

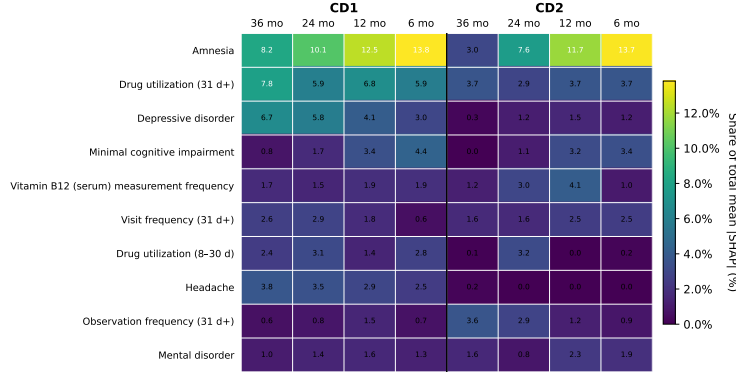


Fig. 3: Cross-horizon SHAP heatmap across case definitions for AoU dataset. Cell values represent each feature’s percentage contribution to the total mean absolute SHAP attribution within the corresponding model. Features were selected based on their rankings across case definitions and prediction horizons; however, heatmap values were obtained from the complete feature attribution set for each model, including instances in which a displayed feature did not individually rank among that model’s top 10 predictors. 31 d+ represents all events recorded more than 30 days before the prediction date, whereas 1–7 d and 8–30 d capture more proximal, non-overlapping time windows. A comprehensive heatmap including the top 20 features is provided in Supplementary Figure S6.

What do utilization features represent clinically? Because healthcare utilization features consistently ranked among the most important predictors in the global SHAP analysis (Figure 3), we performed a detailed clinical decomposition of these features in the CD2 (secondary-confirmed) cohort at the 36-month prediction horizon. For the visit-frequency SHAP feature, patient-prevalence analysis indicated that positive model attribution was primarily associated with hospital-, inpatient-, and emergency-related encounters (Supplementary Figure S7 (a)), suggesting that the model captured patterns of intensive healthcare utilization. For the drug-utilization SHAP feature, an initial patient-prevalence analysis showed that nearly all therapeutic classes were less prevalent among patients with high positive SHAP attribution. Examination of the SHAP dependence plot revealed that positive SHAP values were primarily associated with lower overall drug utilization, whereas extensive medication use contributed negatively to the model prediction. Thus, differences in therapeutic-class prevalence largely reflected differences in the overall volume of medication exposure, rather than differences in medication composition. To account for this, we normalized each patient’s medication profile by their total number of drug events and compared the mean within-patient proportion of each therapeutic class between SHAP groups. After normalization, patients with high positive SHAP attribution exhibited a smaller average share of gastrointestinal medications, antidepressants, vitamins/supplements, and antidiabetic medications than patients with low-SHAP

patients, whereas differences in the remaining therapeutic classes were not statistically significant after false-discovery-rate correction (Supplementary Figure S7 (b)). These findings suggest that the predictive contribution of the drug-utilization feature was driven primarily by overall medication utilization, with only modest differences in medication-class composition after accounting for total drug exposure.

2.4 Performance Across Clinically Relevant Subcohorts

Performance across clinically relevant subgroups was evaluated for the primary CD1 cohort at 6-month prediction horizon (Figure 4), with analyses for the longest prediction horizon (36-month) provided in Supplementary Figure S8. Across all demographic and clinical history subgroups, predictive performance declined as the prediction horizon increased from 6 to 36 months. The count-based model consistently achieved the highest AUROC and AUPRC, followed by the GPT-2048 and Llama-2048 models. Performance was relatively similar between males and females, whereas substantial heterogeneity was observed across age groups, with individuals aged ≥ 85 years exhibiting the lowest discrimination and greatest variability. Racial subgroup differences were modest, with White participants generally demonstrating slightly higher performance than Black and Other participants. Stratification by EHR history revealed a horizon-dependent pattern: at the 6-month horizon, individuals with 4–6 years of history achieved the highest performance, whereas at the 36-month horizon, patients with ≥ 6 years of history consistently showed superior discrimination, suggesting that longer longitudinal records become increasingly important for earlier prediction of ADRD onset.

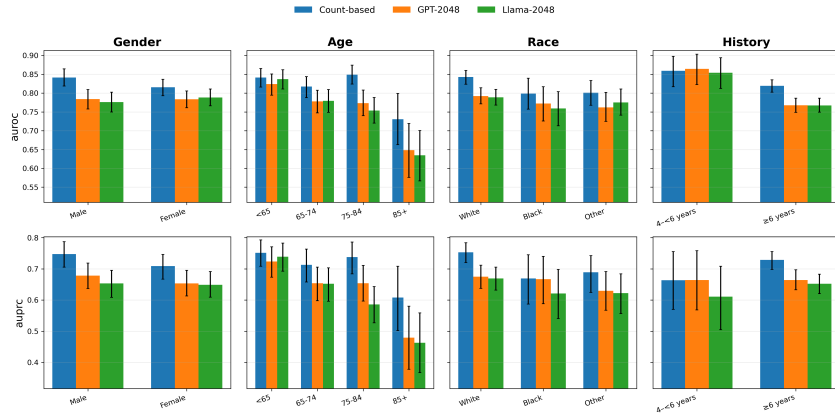


Fig. 4: Performance across clinically relevant patient subgroups for AoU dataset. AUROC (top) and AUPRC (bottom) of the representative Count-based, GPT-2048, and LLaMA-2048 models on the held-out test set for the CD1 cohort at the 6-month prediction horizon, stratified by sex, age, race, and pre-prediction clinical history length. Error bars denote 95% bootstrap percentile intervals estimated within each subgroup.

3 Discussion

Early identification of individuals at risk for ADRD remains a major challenge for population-scale clinical decision support. Leveraging longitudinal EHR data from two independent health systems, we systematically evaluated how patient representation influences ADRD prediction across multiple prediction horizons and case definitions. Count-based representations generally achieved the strongest discrimination and calibration, whereas pretrained foundation-model representations remained competitive, with differences between representations narrowing at longer prediction horizons. Increasing context length from 512 to 2048 tokens yielded only modest and architecture-dependent improvements, indicating that longer input sequences alone did not consistently improve prediction using frozen pretrained representations. Survey-derived lifestyle and personal and family health-history factors provided only modest incremental predictive value beyond routinely collected EHR data, partly reflecting limited availability before prediction cutoffs. Among interpretable count-based models, the clinical information supporting prediction also varied with temporal proximity to documented ADRD onset: broad healthcare-utilization indicators were more prominent at longer horizons, whereas cognitive phenotypes, particularly amnesia and mild cognitive impairment, became increasingly important nearer diagnosis. These patterns reflect model reliance rather than causal effects. Finally, cross-institutional evaluation revealed substantial degradation under zero-shot transfer from AoU to UCM compared with institution-specific training. Although pretrained representations showed advantages over count-based representations in some transfer settings, no representation demonstrated consistently superior transportability across architectures, case definitions, prediction horizons, and performance measures. Prior work under temporal distribution shift found that pretrained CLMBR representations generally achieved stronger discrimination than count-based representations in both in-distribution and out-of-distribution populations, although relative performance degradation was task-dependent and calibration results were heterogeneous [4]. Our findings extend this question to cross-institutional shift, where representation-specific advantages were less consistent, suggesting that performance under temporal distribution shift may not predict transportability across healthcare systems with differences in patient populations, clinical practice, coding, and documentation.

Together, these findings indicate that the utility of EHR patient representations for early ADRD prediction is context-dependent rather than determined by representational complexity alone, varying with prediction horizon, available information, case definition, and healthcare setting. The observed shift from broad healthcare-utilization signals at longer horizons toward cognitive phenotypes nearer documented diagnosis further motivates studying ADRD within the broader trajectories of multimorbidity rather than as an isolated clinical outcome. Multi-institutional initiatives such as the Multimorbidity Three-City Alzheimer’s Disease EHR (M3AD) Study provide a natural framework for examining how the type, sequence, and evolution of chronic conditions and care patterns relate to ADRD onset and progression [10]. Such analyses could help distinguish clinically meaningful multimorbidity trajectories from nonspecific utilization signals and inform earlier screening models. Future work should also evaluate diverse representation strategies across health systems and investigate

complementary modalities, including clinical notes, genetic risk factors, biomarkers, neuroimaging, and wearable data. Ultimately, prospective multi-institutional validation will be necessary to establish the clinical utility, robustness, and equitable implementation of ADRD prediction models across diverse healthcare systems.

4 Methods

4.1 Ethical Approval Declarations

This study was approved by the Columbia University Medical Center Central Institutional Review Board (CIRB; CIRB24-2050), which served as the IRB of record for the University of Chicago under an IRB reliance arrangement. A waiver of the informed consent process and documentation of consent was granted. All methods were performed in accordance with the relevant guidelines and regulations.

4.2 Data Source, ADRD Phenotype and Case Definitions.

We conducted a retrospective matched case-control study using two longitudinal electronic health record (EHR) databases: the National Institutes of Health All of Us (AoU) (633k patients as of 06/1/3036) Research Program Controlled Tier Version 8, and UCM EHR (1.6M patients as of 07/01/2026). AoU is a national research program that integrates participant demographics, surveys, physical measurements, EHRs, biospecimens, and genomic data giving access to the registered and controlled tier researchers through a cloud-based workbench [11]. For the present study, only longitudinal EHR, demographic, and self-reported survey data from AoU were used. UCM EHR provides only longitudinal EHR, demographic data. EHR data were harmonized to the Observational Medical Outcomes Partnership Common Data Model (OMOP) in both databases.

ADRD cases were identified using curated clinical concept sets representing Alzheimer’s disease, vascular dementia, Lewy body dementia, frontotemporal dementia, mixed dementia, and related dementia phenotypes, as well as exposure to medications commonly used for dementia treatment. Qualifying anti-dementia medications included donepezil, galantamine, memantine, and rivastigmine. Diagnosis concepts were identified using standard SNOMED concepts [12–16] and their mapped ICD-9-CM and ICD-10-CM source codes [17], while medication exposures were identified using corresponding RxNorm concepts [18]. Complete diagnosis and medication concept sets are provided in Supplementary Table S1.

Mild cognitive impairment (MCI) was not included in the ADRD outcome because it represents a clinically heterogeneous state with variable trajectories: some individuals progress to dementia, whereas others remain stable or revert to normal cognition. Restricting the outcome to ADRD diagnoses and ADRD-directed medication exposure provided a more specific target for predicting clinically recognized dementia onset.

We evaluated two rule-based ADRD case definitions to examine the effect of phenotype ascertainment on model performance. Case definition 1 (CD1; diagnosis or medication) was the broadest definition and included participants with either a qualifying ADRD diagnosis or exposure to a qualifying anti-dementia medication. For CD1,

the ADRD index date was defined as the earliest qualifying diagnosis or medication exposure. The index criteria distribution for CD1 cohort is presented in Supplementary Table S2. Case definition 2 (CD2; secondary-confirmed) included participants whose first qualifying ADRD event was a diagnosis represented by an eligible ICD- or SNOMED-mapped concept followed by anti-dementia drug exposure. The index date was the first qualifying ADRD diagnosis (Supplementary Figure S1).

The two definitions were evaluated as alternative case-ascertainment strategies. CD2 represents a narrower phenotype derived from the broader CD1 cohort and was not intended to form an exhaustive partition of CD1. Consequently, some participants included in CD1, such as those identified initially through medication exposure or whose diagnosis–medication sequence did not satisfy the CD2 criteria, were not included in CD2. The exact temporal logic for assigning participants to each case definition followed a prespecified cohort-generation algorithm. Although CD2 was expected to improve phenotypic specificity, it substantially reduced the number of eligible cases and therefore statistical power.

Using the broad ADRD phenotype (CD1), the initial case pool comprised 4,074 participants with at least one qualifying ADRD diagnosis or anti-dementia medication exposure. The initial control pool comprised 398,850 participants with no qualifying ADRD diagnosis and no qualifying anti-dementia medication exposure throughout their available EHR history. Initial cohort identification and case definitions are summarized in Figure 5.

4.3 Case-Control Cohort Construction and Matching

The initial case pool was required to have at least two clinical events more than five years before the case index date. Similarly, the initial control cohort was filtered to have at least two clinical events before the last recorded EHR event date. This criterion made the analytic cohorts more representative of individuals with sustained healthcare-system engagement than the broader AoU population. This resulted in the eligible pool of ADRD cases ($n=3,496$) and control ($n=388,850$) for final matching.

Each case was matched up to two controls without replacement. Candidate controls were restricted to participants with the same recorded sex and race as the corresponding case. For each candidate control, a pseudo-index date was calculated such that the control’s age at pseudo-index equaled the matched case’s age at ADRD index. To reduce temporal bias in retrospective case-control prediction studies by aligning the temporal context and longitudinal observation history between cases and controls, a control was considered eligible for matching only if: (1) the pseudo-index date fell within the control’s observed EHR interval; (2) the control had at least two clinical events recorded more than five years before the pseudo-index date, ensuring sufficient pre-index longitudinal history; and (3) the absolute difference between the case index date and the control pseudo-index date was no greater than two calendar years. These criteria were designed to align the prediction horizon and longitudinal observation history between matched cases and controls, consistent with recommendations for mitigating temporal bias in case-control prediction studies and with epidemiologic principles of risk-set sampling. [6, 19, 20]. Eligible cohort identification, matching process, and final analytic cohort sizes across case definitions for AoU database are

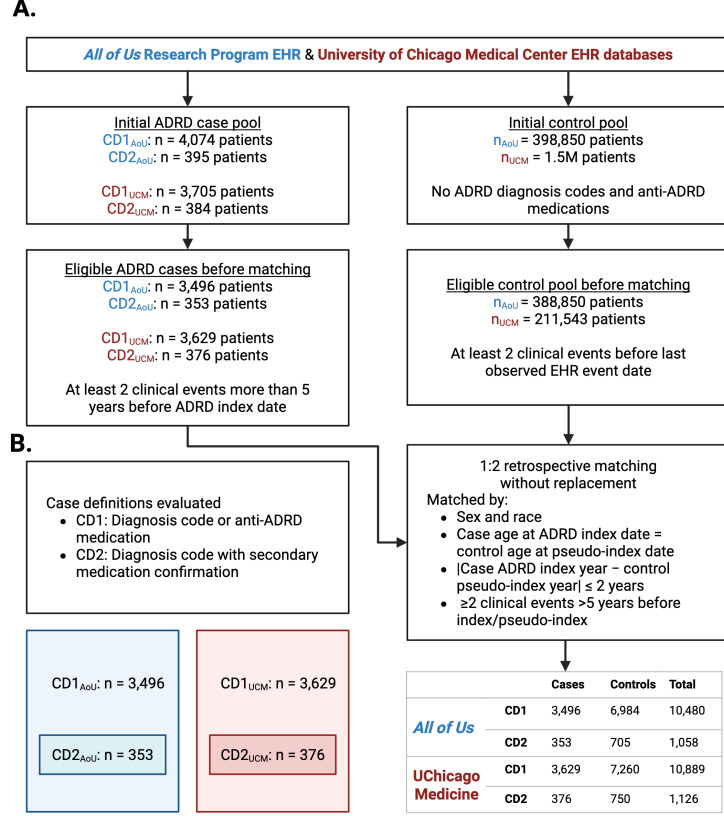


Fig. 5: Construction of matched ADRD case-control cohorts across alternative phenotype definitions. (A) Participants were screened from the AoU Controlled Tier V8 EHR population (external dataset: UChicago Medicine EHR) and evaluated for sufficient longitudinal history before cohort-specific retrospective matching. Matching was conducted on the eligible case-control population, with control pseudo-index dates assigned to align age and calendar time while preserving the required pre-index observation period. (B) The two rule-based phenotypes represent distinct levels and sources of ADRD ascertainment: CD1 provides the broadest definition, while CD2 is the secondary-confirmed cohort. The resulting samples were used as separate analytic cohorts in subsequent prediction experiments.

summarized in Figure 5. For UCM EHR dataset, we applied the same inclusion and exclusion criteria used for AoU database and identified 3,629 ADRD cases and 9,260 matched controls.

4.4 Observation Period and Prediction Horizons

For cases, the index date was defined according to the first qualifying event specified by the corresponding ADRD case definition. For controls, the pseudo-index date

assigned during matching served as the corresponding reference date. We evaluated ADRD onset prediction at 6-, 12-, and 24 and 36-month horizons before the case index or control pseudo-index date. For each prediction horizon, the prediction cutoff was calculated by subtracting the specified horizon from the index or pseudo-index date.

The observation period extended from the participant’s first eligible recorded clinical event through the prediction cutoff. Only information recorded on or before the prediction cutoff was eligible for model input. Clinical events and other time-varying information recorded between the prediction cutoff and the index or pseudo-index date were withheld. This temporal censoring was intended to prevent models from learning directly from diagnostic evaluations, treatment initiation, or other healthcare activity occurring during the period immediately preceding clinically recognized ADRD onset.

The requirement for at least two clinical events more than five years before the index or pseudo-index date was used to establish sufficient longitudinal EHR history and was distinct from the prediction horizon. Only information available by the prediction cutoff was used to generate model inputs. Accordingly, any time-varying demographic, survey, or clinical variables were required to have been recorded on or before the relevant cutoff. Variables recorded after the cutoff were excluded from that prediction-horizon analysis. The temporal study design is illustrated in Figure 1 A.

4.5 Data Extraction, Preprocessing, and Feature Representation

4.5.1 Clinical EHR Extraction and Preprocessing

Longitudinal clinical data were extracted from the OMOP condition_occurrence, drug_exposure, procedure_occurrence, measurement, and visit_occurrence domains. Domain-specific records were transformed into a common event representation containing the participant identifier, event date or datetime, standardized clinical concept, clinical domain, and, when available, recorded measurement values, units, and reference-range information. Events were ordered chronologically within each participant’s record and censored at the prediction cutoff for each prediction-horizon analysis.

Exact same-day duplicate records were collapsed only when all available identifying attributes were identical, including the standardized clinical concept, clinical domain, recorded value, unit, and reference-range information. Repeated occurrences on different dates were retained. Distinct same-day records were also retained, including events from different domains and measurements with different values, units, or reference ranges, because these represented separate clinical information rather than redundant database entries.

Clinical preprocessing included checks for unparseable or implausible dates, events occurring after the prediction cutoff, duplicated participants, invalid standardized concepts, and missing or nonfinite numeric values. Patient timelines were examined to confirm chronological ordering and the presence of usable clinical information before the prediction cutoff. Participants without any usable model-compatible history after preprocessing were identified before model fitting.

4.5.2 Survey-data Extraction and Processing

Unlike longitudinal EHR data, which capture repeated clinical events over time, survey data in the AoU dataset consist of cross-sectional self-reported questionnaires completed by a subset of participants. UCM dataset does not include any survey data. To prevent information leakage, only survey responses recorded on or before each participant’s prediction cutoff date were eligible for inclusion, and survey-derived features were generated independently for each prediction horizon and train–test split.

Survey variables were selected from two domains previously associated with ADRD risk identified from the existing literature [21–25]: lifestyle behaviors and personal and family health history. Rather than directly using individual questionnaire items, clinically related responses were curated and aggregated into composite variables to improve robustness and reduce redundancy. Lifestyle features captured smoking and alcohol exposure, including binary indicators of lifetime and current smoking or alcohol use, as well as quantitative measures such as cigarettes smoked per day, years of smoking, and daily alcohol consumption. Smoking-related binary variables were constructed by integrating responses from multiple tobacco-use questionnaires (e.g., cigarettes, cigars, hookah, smokeless tobacco, and electronic cigarettes). Personal and family health history features included self-reported mental health and substance use disorders, sleep apnea, social phobia, and family history of dementia or other neurodegenerative disorders. Composite indicators were similarly constructed for clinically related concepts, such as active social phobia, by combining responses regarding physician diagnosis, treatment, and medication use.

Survey responses indicating skipped questions, "prefer not to answer," "don’t know," or logically inapplicable responses were treated as missing. Because survey completion varied substantially across participants, explicit missingness indicator variables were created for each survey-derived feature, allowing the models to distinguish between the absence of a reported characteristic and the absence of survey information. When multiple responses to the same questionnaire item were available, the most recent response obtained before the participant’s prediction cutoff was retained.

Continuous survey variables were standardized together with other continuous model inputs during classifier training. In contrast to the longitudinal EHR features, which were represented as cumulative counts of clinical events occurring before the prediction time, survey variables were modeled as static participant-level attributes representing the most recent available pre-prediction self-reported status.

4.5.3 Count-based Clinical Representation

The count-based clinical representation summarizes each participant’s longitudinal EHR history using interpretable features derived from the OMOP Common Data Model (CDM). Following established EHR count-featurization approaches [4, 26], patient representations were constructed independently for each prediction horizon using only information available before the participant’s prediction cutoff.

Clinical information was extracted from five OMOP CDM domains mentioned in 4.5.1. The resulting representation consisted of three groups of features: 1) **Clinical concept indicators**, representing the presence of individual OMOP concepts

before the prediction time; 2) **Healthcare utilization features**, including the total number of recorded clinical events, the number of unique clinical concepts, and domain-specific event counts (e.g., diagnosis, medication, procedure, laboratory, and observation counts); 3) **Demographic features**, including sex, race, ethnicity, and age categorized into 5-year intervals.

To preserve temporal information while maintaining a fixed-length representation, clinical events were aggregated within four time windows defined relative to the prediction time: 0–24 hours, 1–7 days, 8–30 days, and >30 days before prediction. Features from each temporal window were concatenated to form the final longitudinal patient representation.

Continuous laboratory measurements were converted into binary abnormality indicators based on whether the recorded value exceeded or fell below the reference range documented in the EHR, rather than using raw numeric values.

To prevent information leakage, the feature vocabulary (i.e., the set of retained OMOP concepts and feature ordering) was derived exclusively from the training data. The learned vocabulary and column ordering were subsequently applied unchanged to the validation and held-out test sets. Consequently, concepts not observed during training were ignored during feature construction in the evaluation datasets.

4.5.4 CLMBR-compatible Timelines and Representation Extraction

For the clinical foundation-model analyses, each participant’s longitudinal EHR history prior to the prediction cutoff was transformed into a chronological sequence of clinical events compatible with the Foundation Medical Record (FEMR) framework. Eligible events from the OMOP CDM were mapped to the standardized clinical vocabulary used by each pretrained CLMBR model [3, 27], while preserving their temporal ordering. Events occurring after the prediction cutoff were excluded to prevent information leakage. The complete pre-cutoff timeline was provided as input to the CLMBR tokenizer and encoder without manual truncation to a predefined context length. Sequence padding, truncation, and tokenization were performed internally by the pretrained CLMBR framework according to the model-specific context-length configuration. Consequently, each model automatically retained the appropriate portion of the longitudinal history consistent with its architecture.

Unlike the count-based representation, which summarizes clinical history using manually engineered features, the clinical foundation model represents the patient’s complete longitudinal record as a sequence of timestamped clinical tokens. Each participant’s sequence was tokenized using the fixed vocabulary distributed with the corresponding pretrained CLMBR model, without modifying or extending the tokenizer vocabulary during this study. Clinical concepts not recognized by the pretrained tokenizer were discarded prior to representation extraction.

A patient-level representation was obtained by applying the frozen pretrained CLMBR encoder to the pre-cutoff clinical sequence. For each participant and prediction horizon, the encoder produced a fixed-length 768-dimensional embedding summarizing the entire longitudinal clinical history available up to the prediction cutoff. These embeddings were used directly as input features for downstream classifiers,

while the CLMBR encoder remained frozen throughout model development. Each pre-trained model retained its native maximum context length. Sequences shorter than the maximum context length were left-padded, whereas longer sequences were truncated to retain the most recent eligible clinical events.

Representation-level quality control verified that exactly one embedding was generated for every participant and prediction horizon, with the expected embedding dimensionality, no missing or non-finite values, and complete consistency among participant identifiers, case-control labels, matched-set identifiers, and train-validation-test assignments.

4.6 Data Splitting and Leakage Prevention

Model training and evaluation were conducted separately for each ADRD case definition. Each matched case-control cohort was divided into a 70% training set and a 30% held-out test set. Splitting was performed at the matched-set level rather than at the individual-participant level. Each case and all of its matched controls shared a common matched-set identifier and were assigned together to either the training or test partition. This prevented a case from appearing in one partition while demographically and temporally matched controls appeared in another.

Within each case definition, the train-test split was designed to preserve the distribution of the outcome and key demographic characteristics. Stratified sampling was performed using sex, race, 5-year index-age categories, and the ADRD ascertainment criterion (first qualifying diagnosis concept or first qualifying anti-dementia medication concept), with sparse strata collapsed when necessary. The resulting train and test assignments were fixed and reused across all feature representations and downstream classifiers, ensuring fair model comparisons on identical participants.

Within the 70% training set, group-aware three-fold cross-validation was used for hyperparameter tuning, and model selection. Matched sets remained intact within each cross-validation fold. No matched set was permitted to contribute participants to both the training and validation portions of a fold.

All learned preprocessing and feature-construction procedures were restricted to the training data. These included clinical vocabulary construction, rare-concept filtering, regularization tuning, and class-weight selection. During cross-validation, procedures that could influence model selection were fitted using the corresponding inner-training folds and applied to the associated validation fold. The held-out test set was not used to construct features, select hyperparameters, choose decision thresholds, determine training duration, or select model configurations.

After the final configuration was selected (cross-validation result in Supplementary Table S4), its preprocessing procedures and model parameters were refitted using the complete training set. The resulting model was then evaluated once on the held-out test set. Split quality-control analyses verified that no participant identifier or matched-set identifier overlapped between the training and test sets or between the training and validation portions of each cross-validation fold. Each partition was also checked to confirm the presence of both cases and controls and preservation of the approximate case-control matching ratio.

4.7 Model Development and Evaluation

We evaluated two primary approaches to represent longitudinal EHR information: sparse count-based clinical representations and dense representations generated using pretrained CLMBR clinical foundation models. We evaluated pretrained CLMBR models with four backbone architectures, including Transformer-base, LLaMA-style, GPT-style, and Mamba sequence models. For the LLaMA-style and GPT-style CLMBR architectures, we extracted embeddings from pretrained models with varied context length (512 and 2048) to assess whether retaining a larger number of preceding clinical events improved prediction performance. Sparse clinical-concept count vectors and CLMBR derived embeddings were used to train logistic regression (with L2 [28] and elasticnet [29] regularization) [30], XGBoost [31], and LightGBM [32] models. Logistic-regression hyperparameters included regularization strength, class weight, L1 ratio (for elasticnet). Hyperparameters for XGBoost and LightGBM included max_depth (XGBoost), num_leaves (LightGBM), learning_rate. Hyperparameters were selected using group-aware three-fold cross-validation within the development set.

Survey-derived variables were concatenated with the longitudinal EHR count-based features and the foundation model embeddings using early fusion strategy to form a unified feature representations for downstream classification. Rather than learning latent survey representations using a dedicated neural encoder, survey variables were incorporated directly as engineered features. Given the relatively small number of survey variables, substantial missingness, and the study’s emphasis on interpretable multimodal integration, early feature fusion provides a simpler and more robust modeling strategy while reducing the risk of overfitting. The count-based model and CLMBR models were modeled with and without different variants of survey features (e.g. smoking, alcohol, survey (all survey features combined that are mentioned in 4.5.2)). These comparisons were used to assess whether participant-reported information improved prediction beyond longitudinal structured EHR information alone.

4.7.1 Experimental Setup and Model Selection

Count-based logistic regression and XGBoost were used as classical baselines for each case definition and prediction horizon. CLMBR experiments assessed whether pretrained sequential EHR representations improved performance relative to sparse clinical-concept counts. Additional analyses compared backbone architectures and context lengths. Model selection was performed exclusively within the training set using 3-fold cross-validation. For each variant of the model (with or without survey data and the varying context length of the CLMBR models), the configuration was selected that maximizes the AUROC stability score ¹ across folds. The selected model was then retrained on the full training set and evaluated once on the independent held-out test set.

¹of mean and SD of the AUROC value

4.7.2 Performance Evaluation

Discrimination was evaluated using the area under the receiver operating characteristic curve and area under the precision-recall curve. Probabilistic prediction accuracy was evaluated using the Brier score. Threshold-based performance measures included sensitivity, specificity, precision, F1 score, and balanced accuracy. AUROC, AUPRC, and Brier score were calculated from continuous predicted probabilities and did not depend on a classification threshold.

During development, empirically selected decision thresholds generally ranged from 0.30 to 0.50. To standardize comparisons across model variants, case definitions, and prediction horizons, a fixed probability threshold of 0.50 was used for primary threshold-based evaluation on the held-out test set.

4.8 Top Feature Interpretation

To improve the interpretability of the count-based models, we quantified feature importance using SHAP (SHapley Additive exPlanations) [33] values computed from the final XGBoost models. Although LightGBM achieved the highest predictive performance among the count-based classifiers, followed closely by XGBoost, XGBoost was selected for interpretation because it provides robust and widely adopted TreeSHAP explanations with well-established support for feature-level attribution in tree-based models. For each case definition and prediction horizon, TreeSHAP values were computed from the final retrained XGBoost model on the held-out test set. Feature importance was quantified as the mean absolute SHAP value across all patients, reflecting the average contribution of each feature to model predictions irrespective of direction. Features were ranked independently within each case definition and prediction horizon according to their mean absolute SHAP values, and the top 20 features from each model were retained for downstream comparison. To facilitate visualization across case definitions and prediction horizons, the union of all selected top-ranked features was assembled into a combined heatmap. Heatmap values represent the mean absolute SHAP value for each feature within each model. Features that did not rank among the top 20 for a given case definition and prediction horizon were assigned a value of 0.0 in the heatmap, indicating that they were not selected among the highest-contributing features in that model rather than having zero predictive contribution. To better interpret the healthcare utilization features identified by the global SHAP analysis, we performed secondary analyses on the held-out test of a selected cohort for the highest-performing count-based XGBoost models. For the visit frequency feature, patients were stratified into high- and low-SHAP groups using the median SHAP value of the feature. Within each group, we quantified the prevalence of healthcare encounters across OMOP visit domains, including inpatient, outpatient, emergency department, home health, and other visit types. Visit-type prevalence was calculated as the proportion of patients with at least one encounter of each type before the prediction cutoff, allowing us to determine which encounter settings primarily contributed to elevated visit frequency SHAP attribution. To further characterize the global drug utilization feature, we mapped individual RxNorm ingredient concepts to higher-level

therapeutic drug classes using the Anatomical Therapeutic Chemical (ATC) classification hierarchy. Major therapeutic classes were selected based on overall frequency within the study population. Patients were similarly stratified into high- and low-SHAP groups according to the median SHAP value for the drug utilization feature. For each patient, the contribution of each therapeutic class was quantified as the proportion of all pre-cutoff medication exposures belonging to that class, thereby normalizing for differences in overall medication burden. Therapeutic-class distributions were then compared between the two SHAP groups to identify medication categories associated with positive or negative model attribution while accounting for variation in total drug utilization.

4.8.1 Subgroup Evaluation

For the selected model, performance was evaluated on the held-out test set within pre-specified subgroups defined by sex, race, age at prediction, and clinical history length. The same trained model was applied to each subgroup without subgroup-specific retraining. Age was calculated at the prediction cutoff and categorized as younger than 65 years, 65–74 years, 75–84 years, and 85 years or older. Pre-prediction clinical history length was categorized as 4–6 years and ≥ 6 years of observable EHR history. To illustrate the variability of subgroup-specific performance estimates, we computed bootstrap percentile intervals by resampling participants within each subgroup with replacement. The original matched case-control structure was not preserved during resampling; therefore, these intervals should be interpreted as descriptive measures of sampling variability rather than formal inferential confidence intervals for comparing subgroup performance.

4.9 Computing Platform

Data analysis was conducted on a Unix system equipped with 8 CPU having 4 cores (Intel® Xeon® CPU @ 2.30GHz) and 50 GB of memory. Data curation, visualization, and modeling were carried out using Python 3.10.19 with packages scikit-learn 1.6.1, seaborn 0.13.2, matplotlib 3.10.1, PyTorch 2.0.1+cu118, transformers 5.16.1.

5 Code availability

The code for the modeling approach can be found in https://github.com/treena908/ADRD_onset_prediction

6 Data Availability

The data used in this study contain sensitive patient information and therefore are not publicly available. Access to data is restricted to protect patient privacy and confidentiality. Researchers interested in accessing the data for academic purposes can contact the corresponding author for more information on the terms and conditions of access to the data.

7 Limitations

This study has several limitations. First, our primary case definition (CD1) used a broad ADRD phenotype that included individuals identified through ADRD diagnosis code or anti-dementia medication exposure. Because these medications may occasionally be prescribed for conditions other than ADRD, some outcome misclassification is possible. To address this concern, we additionally evaluated a more specific secondary-confirmed case definition (CD2), in which ADRD diagnosis preceded anti-dementia medication use, and observed consistent overall trends across both case definitions. Second, our SHAP analyses were designed to characterize global temporal changes in model reliance across prediction horizons and case definitions rather than sex-specific predictor differences. Consequently, we did not perform sex-stratified feature attribution analyses, although such analyses may provide additional insights into potential biological or clinical heterogeneity. Third, SHAP values quantify feature contributions to model predictions rather than biological mechanisms or causal pathways; therefore, the reported feature importance should be interpreted as model reliance within longitudinal EHR data rather than evidence of disease etiology. Finally, interpretability analyses were limited to the count-based models. We did not investigate the internal representations learned by the pretrained clinical foundation models, and understanding the clinical information encoded within these embedding spaces remains an important direction for future research.

References

- [1] 2026 alzheimer’s disease facts and figures. Alzheimer’s & Dementia **22**(4), 71345 (2026) <https://doi.org/10.1002/alz.71345> <https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/alz.71345>
- [2] Steinberg, E., Jung, K., Fries, J.A., Corbin, C.K., Pfohl, S.R., Shah, N.H.: Language models are an effective representation learning technique for electronic health record data. J. of Biomedical Informatics **113**(C) (2021) <https://doi.org/10.1016/j.jbi.2020.103637>
- [3] Wornow, M., Thapa, R., Steinberg, E., Fries, J.A., Shah, N.H.: Ehrshot: an ehr benchmark for few-shot evaluation of foundation models. NIPS ’23. Curran Associates Inc., Red Hook, NY, USA (2023)
- [4] Guo, L.L., Steinberg, E., Fleming, S., Posada, J., Lemmon, J., Pfohl, S., Shah, N., Fries, J., Sung, L.: Ehr foundation models improve robustness in the presence of temporal distribution shift (2022) <https://doi.org/10.1101/2022.04.15.22273900>
- [5] Li, Q., Yang, X., Xu, J., Guo, Y., He, X., Hu, H., Lyu, T., Marra, D., Miller, A., Smith, G., DeKosky, S., Boyce, R., Schliep, K., Shenkman, E., Maraganore, D., Wu, Y., Bian, J.: Early prediction of alzheimer’s disease and related dementias using real-world electronic health records. Alzheimer’s & dementia : the journal of the Alzheimer’s Association **19** (2023) <https://doi.org/10.1002/alz.12967>

- [6] Tang, A.S., Rankin, K.P., Ceronio, G., Miramontes, S., Mills, H., Roger, J., Zeng, B., Nelson, C., Soman, K., Woldemariam, S., Li, Y., Lee, A., Bove, R., Glymour, M., Aghaeepour, N., Oskotsky, T.T., Miller, Z., Allen, I.E., Sanders, S.J., Baranzini, S., Sirota, M.: Leveraging electronic health records and knowledge networks for alzheimer’s disease prediction and sex-specific biological insights. *Nature Aging* **4**(3), 379–395 (2024) <https://doi.org/10.1038/s43587-024-00573-8>
- [7] Zhu, W., Tang, H., Zhang, H., Rajamohan, H., Huang, S.-L., Ma, X., Chaudhari, A., Madaan, D., Almahmoud, E., Chopra, S., Dodson, J., Brody, A., Masurkar, A., Razavian, N.: Predicting risk of alzheimer’s diseases and related dementias with ai foundation model on electronic health records. *medRxiv : the preprint server for health sciences* (2024) <https://doi.org/10.1101/2024.04.26.24306180>
- [8] Wang, J., Ahn, S., Dalal, T., Zhang, X., Pan, W., Zhang, Q., Chen, B., Dodge, H., Wang, F., Zhou, J.: Augmented risk prediction for the onset of alzheimer’s disease from electronic health records with large language models (2024) <https://doi.org/10.48550/arXiv.2405.16413>
- [9] Ruan, X., Lu, S., Fu, S., Ahn, J., Chen, F., Li, R., Wen, A., Wang, L., Onyema, E., Tang, V., Liu, H.: A dynamic risk prediction framework for alzheimer’s disease and related dementias with interpretability. *npj Digital Medicine* **9**, 446 (2026) <https://doi.org/10.1038/s41746-026-02732-0>
- [10] Desvarieux, M., Rundek, T., Ahsan, H., Narvaez, J., Diaz, F., Malinsky, D., Ruiz, L.M., Topaz, M., Falconer, T., Natarajan, K., Noble, J., Entwisle, B., Puram, D., Anand, T., Chen, H.Y., Jiang, X., Gu, Y., Cohen, A., Terry, M.B., Pierce, B., Andrews, H., Rogalsky, E., Farzana, S., Gulotta, G., Beard, J., Landron, D., Volchenboum, S.L., Ravaud, P., Johnson, J., Susser, E., Rundle, A., Wei, Y., Tsinoremas, N., Loewenstein, D., Fried, L., Aiello, A., Mayeux, R., Hripacsak, G.: Accelerating real-world prediction and research in alzheimer’s: The m3ad study. *Alzheimer’s & Dementia* **22**(3), 71174 (2026) <https://doi.org/10.1002/alz.71174> <https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/alz.71174>
- [11] Mayo, K.R., Basford, M.A., Carroll, R.J., Dillon, M., Fullen, H., Leung, J., Master, H., Rura, S., Sulieman, L., Kennedy, N., *et al.*: The all of us data and research center: creating a secure, scalable, and sustainable ecosystem for biomedical research. *Annual review of biomedical data science* **6**(1), 443–464 (2023)
- [12] Lee, D., Cornet, R., Lau, F., de Keizer, N.: A survey of snomed ct implementations. *Journal of Biomedical Informatics* **46**(1), 87–96 (2013) <https://doi.org/10.1016/j.jbi.2012.09.006>
- [13] SNOMED International: SNOMED Clinical Terms (SNOMED CT). Accessed: 2026-08-18 (2026). https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html
- [14] Lee, D., Keizer, N., Lau, F., Cornet, R.: Literature review of snomed

- ct use. *Journal of the American Medical Informatics Association* **21**(e1), 11–19 (2014) <https://doi.org/10.1136/amiajnl-2013-001636> <https://academic.oup.com/jamia/article-pdf/21/e1/e11/17375846/21-e1-e11.pdf>
- [15] Duarte, J., Castro, S., Santos, M., Abelha, A., Machado, J.: Improving quality of electronic health records with snomed. *Procedia Technology* **16**, 1342–1350 (2014) <https://doi.org/10.1016/j.protcy.2014.10.151> . CENTERIS 2014 - Conference on ENTERprise Information Systems / ProjMAN 2014 - International Conference on Project MANagement / HCIST 2014 - International Conference on Health and Social Care Information Systems and Technologies
 - [16] Cornet, R., Keizer, N.: Forty years of snomed: a literature review. *BMC medical informatics and decision making* **8**(Suppl 1), 2 (2008)
 - [17] Abeysinghe, R., Black, A., Kaduk, D., Li, Y., Reich, C., Davydov, A., Yao, L., Cui, L.: Towards quality improvement of vaccine concept mappings in the omop vocabulary with a semi-automated method. *Journal of Biomedical Informatics* **134**, 104162 (2022) <https://doi.org/10.1016/j.jbi.2022.104162>
 - [18] Zeng, K., Kilbourne, J., Powell, T., Moore, R.: Normalized names for clinical drugs: Rxnorm at 6 years. *Journal of the American Medical Informatics Association* **18**, 441–448 (2011) <https://doi.org/10.1136/amiajnl-2011-000116>
 - [19] Yuan, W., Beaulieu-Jones, B., Yu, K.-H., Lipnick, S., Palmer, N., Loscalzo, J., Cai, T., Kohane, I.: Temporal bias in case-control design: preventing reliable predictions of the future. *Nature Communications* **12** (2021) <https://doi.org/10.1038/s41467-021-21390-2>
 - [20] Iwagami, M., Shinozaki, T.: Introduction to matching in case-control and cohort studies. *Annals of Clinical Epidemiology* **4**, 33–40 (2022) <https://doi.org/10.37737/ace.22005>
 - [21] Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C.G., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S.G., Dias, A., Fox, N.C., Gitlin, L.N., Howard, R.J., Kales, H.C., Kivimäki, M., Larson, E.B., Ogunniyi, A., Orgeta, V., Ritchie, K., Rockwood, K., Sampson, E.L., Samus, Q.M., Schneider, L.S., Selbæk, G., Teri, L., Mukadam, N.: Dementia prevention, intervention, and care: 2020 report of the lancet commission. *Lancet (London, England)* **396**, 413–446 (2020)
 - [22] Jones, A., Ali, M., Kenny, M., Mayhew, A., Mokashi, V., He, H., Lin, S., Yavari, E., Paik, K., Subramanian, D., Dydynsky, R., Aryal, K., Correia, R., Dash, D., Manis, D., O’Connell, M., Liu-Ambrose, T., Taler, V., McMillan, J., Griffith, L.: Potentially modifiable risk factors for dementia and mild cognitive impairment: An umbrella review and meta-analysis. *Dementia and Geriatric Cognitive Disorders* **53**, 91–106 (2024) <https://doi.org/10.1159/000536643>

- [23] Shi, L., Chen, S., Ma, M.-Y., Bao, Y., Han, Y., Wang, Y.-M., Shi, J., Vitiello, M., Lu, L.: Sleep disturbances increase the risk of dementia: A systematic review and meta-analysis. *Sleep medicine reviews* **40**, 4–16 (2017)
- [24] Wang, S., Zheng, X., Huang, J., Liu, J., Li, C., Shang, H.: Sleep characteristics and risk of alzheimer’s disease: a systematic review and meta-analysis of longitudinal studies. *Journal of Neurology* **271**, 1–12 (2024) <https://doi.org/10.1007/s00415-024-12380-7>
- [25] Yates, L., Ziser, S., Spector, A., Orrell, M.: Cognitive leisure activities and future risk of cognitive impairment and dementia: systematic review and meta-analysis. *International Psychogeriatrics* **28**, 1791–1806 (2016) <https://doi.org/10.1017/S1041610216001137>
- [26] Reps, J., Schuemie, M., Suchard, M., Ryan, P., Rijnbeek, P.: Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association* **25**, 969–975 (2018) <https://doi.org/10.1093/jamia/ocy032>
- [27] Wornow, M., Bedi, S., Hernandez, M.A.F., Steinberg, E., Fries, J.A., Ré, C., Koyejo, S., Shah, N.H.: Context clues: Evaluating long context models for clinical prediction tasks on ehcs (2024) [2412.16178](https://arxiv.org/abs/2412.16178)
- [28] Cortes, C., Mohri, M., Rostamizadeh, A.: L2 regularization for learning kernels. *UAI ’09*, pp. 109–116. AUAI Press, Arlington, Virginia, USA (2009)
- [29] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**(2), 301–320 (2005) <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
- [30] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression: Hosmer/applied logistic regression. (2005). <https://api.semanticscholar.org/CorpusID:119351629>
- [31] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., Cortes, D.: Xgboost: Extreme Gradient Boosting. (2026). R package version 3.4.1.1. <https://github.com/dmlc/xgboost>
- [32] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 30, pp. 3146–3154. Curran Associates, Inc., ??? (2017). <http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>

- [33] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017)

8 Acknowledgements

This research is funded by grant R56AG082167 from the National Institute on Aging, of the National Institutes of Health (NIH), OT2 OD036445 from All of Us Research Program, NIH, P30 ES027792 from National Institute of Environmental Health Sciences.

9 Author contributions

S.F.: Study design, NLP-related data curation, model design, model interpretability assessment, data analysis, manuscript writing. A.A.: Study design, NLP-related data curation, model interpretability assessment, manuscript writing, T.R., M.D.: Expert opinion from a clinician’s perspective, model interpretability assessment, manuscript revision, H.A.: Study design, Expert opinion on cohort selection, manuscript revision.

10 Competing Interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarymaterial.pdf](#)