DualAlign: Generating Clinically Grounded Synthetic Data

Rumeng Li[1,2] **Xun Wang**[4] **Hong Yu**[3,2,1]
[1]University of Massachusetts Amherst
[2]VA Bedford Healthcare System
[3]University of Massachusetts Lowell
[4]Microsoft
rli@umass.edu wangxun@outlook.com Hong_yu@uml.edu

Abstract

Synthetic clinical data are essential for advancing AI in healthcare, given strict privacy constraints on electronic health records (EHRs), the scarcity of annotated data for rare or slowly progressing conditions, and demographic biases in observational cohorts. Large language models (LLMs) can generate fluent clinical text, but ensuring that such outputs are both clinically grounded and useful for downstream modeling remains challenging. We present DualAlign, a disease-agnostic framework for generating privacy-preserving, clinically faithful synthetic EHR narratives. DualAlign improves generation fidelity through two complementary alignment mechanisms: persona alignment, which conditions generation on patient demographics and risk factors, and symptom-trajectory alignment, which grounds narratives in empirically observed longitudinal symptom patterns. Using Alzheimer's disease (AD) as a case study, DualAlign produces context-aware, symptom-rich sentences that more closely reflect real-world clinical documentation. Augmenting limited gold-standard data with DualAlign substantially improves AD symptom classification, outperforming both gold-only training and unconstrained synthetic baselines. Overall, DualAlign provides a generalizable approach for generating high-utility synthetic clinical text in chronic and progressive diseases, reducing annotation burden while enabling scalable and privacy-conscious clinical NLP research.

1 Introduction

Synthetic data generation has become an important strategy for advancing machine learning in healthcare (Chen et al., 2021; Rujas et al., 2025; Giuffrè and Shung, 2023). Access to electronic health records (EHRs) is often constrained by privacy regulations, institutional policies, and compliance requirements (Liu and Panagiotakos, 2022; Rieke et al., 2020). By producing privacy-preserving datasets that approximate real patient data, synthetic text enables broader research collaboration, reproducibility, and model development in data-scarce settings. Recent work has increasingly explored large language model (LLM)–based approaches for synthetic clinical data generation, including privacy-preserving EHR text simulation, domain-focused surveys, and data augmentation for downstream modeling tasks (Chen et al., 2021; Rujas et al., 2025; Giuffrè and Shung, 2023; Smolyak et al., 2024; Barr et al., 2025). These studies demonstrate the feasibility of using LLMs to generate fluent clinical narratives. However, many existing approaches rely on narrowly scoped or task-specific text representations—such as tabular summaries, short note fragments, or isolated passages—that do not capture broader clinical context. As a result, generated outputs may lack contextual diversity, comprehensive symptom coverage, clinical coherence, and factual fidelity, and may diverge from real-world documentation patterns.

This limitation is particularly pronounced for chronic and progressive conditions. Alzheimer's disease (AD) exemplifies the challenge: symptoms span cognition, function, neuropsychiatric behavior, and physiology, often emerging years before diagnosis and evolving heterogeneously across care settings (van der Flier et al., 2023; Liss et al., 2021). Existing synthetic text for AD remains scarce and often lacks the diversity and temporal grounding needed for tasks such as early risk prediction and symptom identification (Li et al., 2023b; Loni et al., 2025).

To address this gap, we introduce **DualAlign**, a framework for disease-aware synthetic clinical text generation that combines two complementary alignment strategies: (1) *persona alignment*, which conditions generation on demographics and risk factors to produce diverse, demographically consistent patient profiles; and (2) *symptom-trajectory alignment*, which grounds narratives in empirically ob-

Findings of the Association for Computational Linguistics: ACL 2026, pages 28189–28208

served longitudinal symptom patterns. By anchoring LLM generation in real-world demographic priors and symptom dynamics, DualAlign produces context-aware, symptom-rich narratives that more closely reflect real clinical documentation (Park et al., 2023; Peralta Ramirez et al., 2025). Through both human assessment and automated evaluation, we show that DualAlign can generate large volumes of labeled, privacy-preserving clinical text with improved diversity and clinical grounding.

Contributions Our main contributions are:

- We propose DualAlign, a framework for LLM-based clinical text generation using persona-driven simulation and longitudinal symptom alignment, enabling the synthesis of realistic, diverse, and clinically grounded narratives.
- We apply DualAlign to AD and publicly release **DualAlign-AD**, a dataset of 233,014 privacy-preserving, LLM-annotated symptom mentions, available at Hugging Face. To our knowledge, this is the first publicly available dataset tailored for early AD research. Augmenting gold data with DualAlign improves classification performance compared to using gold data alone and unguided synthetic baselines.
- We analyze synthetic longitudinal EHR narratives, identify open challenges in temporal progression modeling, and outline directions for future work.

2 Related Work

Synthetic Data Generation in Healthcare Synthetic data has been widely studied as a means to enable machine learning in clinical settings where real patient data are restricted by privacy regulations and institutional barriers. Early efforts primarily focused on structured data synthesis using generative models such as GANs and VAEs to simulate demographics, diagnoses, and laboratory values while preserving statistical properties (Chen et al., 2021; Liu et al., 2025; Rujas et al., 2025; Hernandez et al., 2022; Gonzales et al., 2023). While effective for benchmarking and population-level analysis, these approaches do not extend naturally to unstructured clinical narratives, which require richer semantic representation and contextual coherence.

LLMs for Clinical Text Generation Recent large language models (LLMs), such as GPT-* models, have enabled the generation of synthetic clinical text including discharge summaries, radiology reports, and SOAP notes (Taloni et al., 2025; Peng et al., 2023; Mawaldi and Mladenov, 2024; Ganzinger et al., 2025; Williams et al., 2025). These models can produce fluent and stylistically appropriate text, but most existing work focuses on sentence- or short-passage generation without explicit longitudinal grounding. For example, CLIN-GEN (Xu et al., 2024) leverages structured clinical knowledge to guide sentence-level generation and improves performance on general clinical NLP tasks such as named entity recognition and relation extraction; In contrast, our work targets chronic disease modeling by generating multi-visit synthetic EHR narratives conditioned on patient personas and empirically grounded symptom patterns, enabling analysis of early-stage signals in AD (Kim et al., 2025; Shah, 2024).

Synthetic Data for Alzheimer's Disease Alzheimer's disease (AD) poses additional challenges for synthetic data generation due to its gradual, multifaceted progression across cognitive, functional, neuropsychiatric, and physiological domains. While prior work has developed models for AD risk stratification and phenotyping using real EHR data (Li et al., 2023a; Xu et al., 2020; Wang et al., 2021; Tjandra et al., 2020), synthetic text efforts for AD remain limited. Existing approaches largely rely on structured simulations or rule-based generation (Li et al., 2023b; Muniz-Terrera et al., 2021; Sajjad et al., 2021), and do not aim to model longitudinal, context-rich clinical narratives. These gaps motivate the need for synthetic methods that better capture demographic diversity, symptom heterogeneity, and temporal progression in complex chronic diseases.

3 Method

As a synthetic data generation framework, DualAlign converts real-world clinical patterns into demographically representative and symptom-rich narratives. As illustrated in Figure 1, the workflow integrates three components: (i) extraction of demographic statistics and symptom trajectories from real-world EHRs, (ii) LLM-based narrative generation guided by structured prompts, and (iii) automated symptom annotation. These components are instantiated in a six-stage pipeline from cohort construction to dataset release.

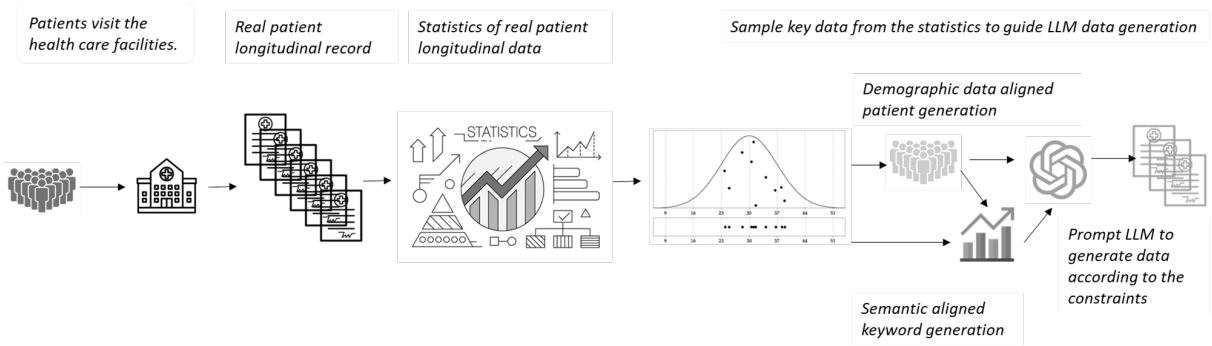

Figure 1: DualAlign narrative generation workflow. The figure depicts how real-world electronic health record (EHR) patterns are extracted and transformed to guide large language model (LLM)–based note generation. The process includes deriving patient-level statistics from real longitudinal records, sampling key demographic and symptom trends, and using these to constrain and guide the LLM in generating privacy-preserving, symptom-rich synthetic clinical narratives.

3.1 AD Patient Cohort Building

To guide synthetic data generation with real-world patterns, we analyzed EHRs from the VA Corporate Data Warehouse, one of the largest integrated healthcare databases in the U.S. The Veterans Health Administration (VHA) provides care across more than 1,200 facilities, and its EHR system includes demographics, diagnoses, clinical notes, and visit records.

We constructed an AD case cohort of 35,308 patients by identifying individuals with at least two separate AD diagnoses between October 1, 2015 (the start of ICD-10 coding), and September 30, 2022 (the end of our study period). To ensure expert-level confirmation, at least one diagnosis must have been recorded in a specialty clinic—such as neurology, geriatrics, GeriPACT, psychiatry, or psychology—by a provider specializing in neurology, psychiatry, neuropsychology, or geriatric medicine. The cohort had a mean (SD) age of 79.0 (8.4) years at diagnosis and was demographically skewed, reflecting the veteran population (2.0% women, 78.9% White, 12.9% Black or African American, 6.6% Hispanic/Latino).

We extracted longitudinal clinical notes spanning 2000–2022 across diverse care settings, including primary care, emergency visits, home-based primary care (HBPC), neurology, geriatrics, psychiatry, neuropsychology, memory clinics, cognitive care nursing, mental health, and compensation and pension examinations. These notes provided reference statistics for demographics, utilization patterns, and symptom documentation. The study was approved by the Institutional Review Board at the VHA Bedford Healthcare System.

3.2 Patient Profile Generation: Demographics and Risk Factors

Because the VA cohort is not representative of the broader AD population, we reweight the synthetic cohort's demographic distribution using national prevalence estimates from the 2024 Alzheimer's Disease Facts and Figures report (Better, 2023). Age, sex, and race/ethnicity are sampled from stratified probability distributions to promote population-level representativeness.

Aligning patient personas to population-level demographics while grounding generation in real-world clinical data necessarily blends distributions. This reflects a core motivation for synthetic data: correcting demographic imbalances inherent to any single health system while preserving realistic clinical context. Because clinical systems exhibit characteristic care-seeking behaviors and documentation styles, this design exposes models to cross-system heterogeneity, helping mitigate single-institution bias and supporting more robust, transferable clinical NLP models.

Beyond demographics, we incorporate a broad set of AD risk factors and social determinants of health (SDOH), covering medical comorbidities, neuropsychiatric conditions, lifestyle behaviors, psychosocial stressors, and structural barriers. Prevalence values are drawn from epidemiological studies (Livingston et al., 2024; Jones et al., 2025; Röhr et al., 2022; Stites et al., 2022), and the complete list with distributions is provided in Appendix AD Risk Factors. Risk factors are sampled categorically, producing heterogeneous patient personas that anchor LLM-driven narrative generation in clinically relevant variation.

3.3 Temporal Alignment of Clinical Notes: Visit Frequency and Type

For each synthetic patient, we simulate a ten-year trajectory of note generation prior to AD diagnosis. Note counts and types are aligned with empirical patterns observed in the VA cohort (see Appendix, Table 5), and mapped to AD stages (Appendix Temporal Context) since LLMs are insensitive to raw year indices. Unless otherwise specified, sampling follows empirical bootstrapping.

Visit frequencies are sampled from time-stratified distributions reflecting real-world utilization trends. Primary care dominates across all periods, while specialty visits (neurology, memory clinics, psychiatry, neuropsychology) increase in the final years, consistent with clinical progression.

Each visit is assigned a note type drawn from these time-dependent distributions (Appendix, Table 5), covering primary care, emergency, home-based primary care (HBPC), and specialty services. Notes are anchored to a year relative to diagnosis and tied to the patient persona and care context, ensuring trajectories remain clinically plausible.

3.4 Semantic Alignment of Clinical Notes: Constraint via Clinician-Curated Keywords

To enhance clinical validity, we guide narrative generation using a curated lexicon of AD-relevant signs and symptoms, aligning keyword sampling with real-world temporal and categorical trends observed in our created VA cohort.

Lexicon Construction. Building on prior expert-curated keyword efforts (Better, 2023; Livingston et al., 2024; Wang et al., 2021), we developed a lexicon of 122 Alzheimer's disease (AD)–relevant terms. Six domain experts—including clinicians and researchers in cognitive aging, neurology, and epidemiology—validated candidate terms across six major symptom domains: (1) cognition—speech/language (word-finding difficulty, aphasia, impaired comprehension), (2) cognition—memory (memory loss, forgetfulness, difficulty recognizing people or places), (3) cognition—learning/perception (impaired attention, visuospatial disorientation, executive dysfunction), (4) assistance needed (difficulties with ADLs/iADLs such as dressing, medication adherence, or managing finances), (5) physiological changes (gait instability, sleep disturbance, sensory deficits, swallowing issues), and (6) neuropsychiatric symptoms (depression, apathy, agitation, impaired insight, personality change). The full keyword list appears in Appendix Keywords for Each Category.

Keyword Pattern Alignment. To mimic realistic symptom emergence, we analyzed keyword distributions in the AD case cohort along two axes: (i) frequency per note across time windows and (ii) prevalence across categories. Stratified sampling tables derived from these trends (Appendix Tables 6, 7) guide generation.

Real-world clinical notes are typically lengthy but contain very sparse AD-related content: in the pre-diagnosis years, notes include only 2.7–4.2 symptom mentions on average, often confined to one or two sentences (Table 6). Replicating this level of sparsity in synthetic data would be inefficient and poorly suited for training extraction and classification models given the cost of generation. We therefore apply a $5\times$ density multiplier to increase symptom mention frequency while preserving category-level proportions, yielding denser and more informative text for downstream modeling.

For semantic alignment, we sequentially: (1) sample the number of mentions based on temporal trends, (2) assign categories using stratified distributions, and (3) select specific terms from the corresponding category lexicon. Sampled keywords are then embedded into structured LLM prompts.

3.5 Prompt Construction and Narrative Generation

Patient-level statistics—including demographics, visit frequency and type, and symptom category trends—are transformed into structured inputs for the LLM. The final prompt template (Figure 2) embeds clinical knowledge and empirically observed AD progression patterns, ensuring generated content is clinically plausible and aligned with real trajectories. An example generated note appears in Appendix Example Note.

These structured prompts are then provided to the LLM to generate SOAP-style clinical notes, with deliberate variation in phrasing, abbreviations, and documentation style to increase lexical diversity. The resulting synthetic cohort includes 100 patients with up to 10 years of simulated longitudinal EHR history; each patient contributes a median of 106 notes (interquartile range = 98–112) with an average length of 138 tokens (standard deviation = 42). Collectively, these narratives are symptom-rich, temporally coherent, and demographically cal-

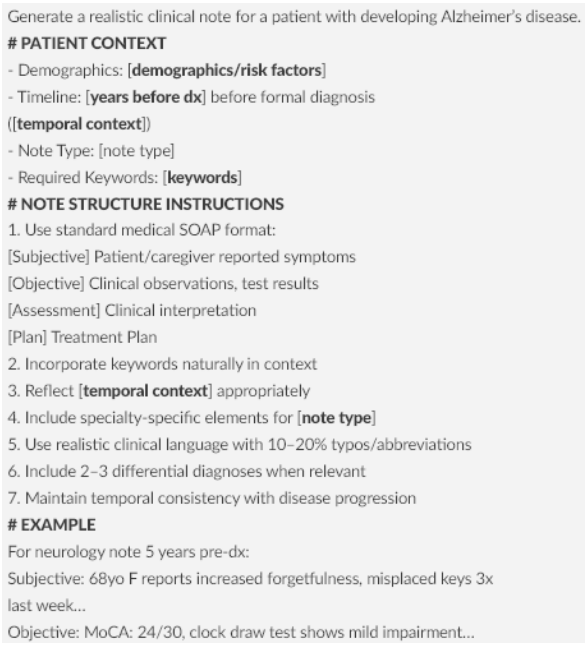

Figure 2: Prompt template used for note generation in DualAlign. The template specifies patient context (demographics, risk factors, timeline, note type, and required keywords) and structured instructions for generating SOAP-style clinical notes.

ibrated, providing a high-utility resource for downstream clinical NLP research.

3.6 Automated Annotation

To convert synthetic narratives into a structured resource, we employ an LLM-based annotator guided by clinician-curated labeling protocols. These protocols build on prior work (Li et al., 2023b), which defined nine categories including diagnostic tests and formal cognitive assessments. In this study, we follow Li et al. (Li et al., 2025), which refines the taxonomy to five clinically salient categories emphasizing observable signs and symptoms documented in routine care:

- **Cognitive impairment:** e.g., memory loss, confusion, word-finding difficulty
- **Concerns raised by others:** e.g., caregiver- or family-reported changes
- **Functional impairment:** e.g., decline in activities of daily living (ADLs) or instrumental ADLs (iADLs), supervision needs
- **Physiological changes:** e.g., motor or sensory decline, sleep disturbance, incontinence
- **Neuropsychiatric symptoms:** e.g., agitation, hallucinations, anxiety, apathy

This streamlined taxonomy targets subtle early indicators—particularly memory-related cognitive changes, caregiver-noted behavioral shifts, and emerging functional decline—that often precede formal diagnostic assessments yet appear in unstructured clinical notes. By excluding categories such as diagnostic tests and cognitive assessments, whose ordering typically reflects an existing clinician suspicion of AD, we focus on signals that are both more scalable and more representative of early-stage disease progression in routine documentation. Detailed annotation guidelines are provided in Appendix Annotation Guideline. The LLM-based annotator operates at the sentence level, tagging relevant mentions across the longitudinal narratives generated by DualAlign and yielding 233,014 labeled sentences.

Because DualAlign uses AD sign and symptom keywords to guide generation, there is a potential risk of introducing surface-level lexical cues. Our design substantially mitigates this risk. Prompt keywords serve only as high-level thematic anchors, which the LLM incorporates naturally into broader clinical descriptions, resulting in varied phrasing and contextually grounded narratives rather than template-like repetitions. Importantly, annotation decisions are based on clinical meaning and functional change—not keyword overlap—preventing keyword–label leakage (Appendix Annotation Guideline).

Compared with the Gold and Bronze baselines (Table 1), DualAlign provides substantially more labeled examples across all AD symptom categories, including underrepresented groups such as concerns raised by others and functional impairment, thereby greatly expanding coverage. Because these sentences are extracted from full synthetic notes rather than isolated fragments, they retain surrounding clinical context.

To quantify lexical diversity, we sampled 3,000 annotated sentences from DualAlign and 3,000 from the Bronze baseline, matching category distributions across the five symptom groups. We evaluated both sets using Distinct-4 (proportion of unique four-word sequences; higher indicates greater diversity) and Self-BLEU-4 (average n-gram overlap; lower indicates less redundancy). DualAlign achieved a Distinct-4 score of 0.9419 compared to 0.4995 for Bronze, and a Self-BLEU-4 of 0.2518 versus 0.8032, indicating substantially richer and less repetitive language. Together, these 233,014 lexically diverse, contextually grounded,

and category-rich labeled sentences form a high-coverage, privacy-preserving resource for downstream tasks such as AD symptom classification, disease trajectory modeling, and early risk prediction.

4 Experiments

4.1 Benchmark

To evaluate DualAlign-generated data, we use a gold-standard dataset of 11,571 human-annotated sentences drawn from 5,112 longitudinal notes of 76 real-world AD patients (Li et al., 2023b). Annotations were created under physician supervision following the same structured guidelines applied in our LLM-based annotation process (Cohen's κ = 0.868). The dataset was split into 80/10/10 training, validation, and test sets.

To measure the contribution of synthetic data—both alone and in combination with gold data—we designed classification experiments across five training configurations.

- **Gold Only:** Human-annotated baseline, serving as the reference without synthetic augmentation.
- **Gold + Bronze:** Synthetic augmentation baseline from (Li et al., 2023b), combining gold data with text generated by GPT-4 but lacked demographic or symptom constraints.
- **Gold + DualAlign (matched):** Size-matched subset of DualAlign to isolate the effect of improved realism and contextual grounding.
- **Gold + DualAlign (full):** Upper bound using the full DualAlign dataset to assess large-scale augmentation gains.
- **DualAlign (full) Only:** Tests if synthetic data alone provides sufficient signal for classification.

We exclude the Silver dataset from (Li et al., 2023b) since it involves LLM annotation of real-world EHRs (MIMIC-III) rather than fully synthetic generation, making it methodologically distinct from DualAlign. Table 1 summarizes sentence counts per symptom category across the gold, bronze, and DualAlign datasets.

Following prior work (Li et al., 2023b), we evaluated all configurations on two tasks using the same held-out test set for fair comparison.

- **Binary classification:** Check whether a sentence describes an AD-related sign or symptom. Negative examples were sampled from the same 76-patient corpus at a 5:1 negative-to-positive ratio.

Category	Gold	Bronze	DA-M	DA-F
Cog. impair.	6,240	2,704	3,100	82,359
Notice/others	785	1,710	865	22,951
Req. assist.	1,864	1,205	525	13,915
Physiol. chg.	1,340	1,769	2,480	65,943
Neuropsych. sym.	1,342	1,308	1,720	45,693
Total	11,571	8,696	8,690	**233,014**

Table 1: Sentence counts per category. DA-M = DualAlign (matched), DA-F = DualAlign (full).

- **Multi-class classification:** Assign each positive sentence to one of the five symptom categories defined in the Methods section.

4.2 Model and Training Setup

We fine-tune the Llama 3.1 8B model (Meta AI, 2024) with the following configuration: inputs tokenized and truncated/padded to 128 tokens; batch size 16 per GPU with gradient accumulation (effective batch size 64); AdamW optimizer with linear warmup ratio 0.1 and weight decay $1e^{-2}$; classification head with dropout 0.1 on the [CLS] token. Training runs for 3 epochs with early stopping on validation macro-F1, evaluated every 200 steps. Parameters: $batch\_size = 64$, $lr = 2e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

5 Results

5.1 Model Performance on Classification Tasks

We first evaluate model performance on binary and multi-class classification tasks under the training configurations described in the Experiments section. As shown in Table 2, augmenting the gold training set with DualAlign (full) yields the strongest binary classification performance, reaching an F1 of 0.84 and accuracy of 0.95—substantially surpassing the gold-only baseline. Gold + DualAlign (matched) also outperforms Gold + Bronze, underscoring the value of demographic and symptom-guided alignment. Even when used alone, DualAlign (full) achieves an accuracy of 0.82, demonstrating that well-designed synthetic data can provide meaningful signal in low-resource settings despite a lower F1 of 0.47.

In the sentence-level multi-class classification task (Table 3), DualAlign (full) augmentation improves performance across all five symptom categories. The largest relative gains appear in more challenging categories such as Requires Assistance

Training Set (Positive)	Precision	Recall	F1	Accuracy
Gold Only	0.87	0.61	0.72	0.87
+Bronze	0.89	0.69	0.77	0.91
+DualAlign (matched)	0.90	0.72	0.79	0.93
+DualAlign (full)	**0.92**	**0.77**	**0.84**	**0.95**
DualAlign (full) only	0.45	0.49	0.47	0.82

Table 2: Binary classification performance on the gold test set, with negative samples selected from real world notes at a 5:1 ratio.

Train Set	Acc.	Cog.	Conc.	Req.	Phys.	Neuro.
Gold	0.70	0.77	0.46	0.57	0.62	0.71
+Bronze	0.75	0.79	0.53	0.67	0.74	0.78
+DA-M	0.77	0.81	0.54	0.70	0.77	0.81
+DA-F	**0.80**	**0.82**	**0.60**	**0.72**	**0.78**	**0.86**
DA-F only	0.53	0.60	0.22	0.39	0.48	0.62

Table 3: Multi-class sentence classification of AD symptoms. **Acc.** = overall accuracy; DA-M = DualAlign (matched), DA-F = DualAlign (full). Category abbreviations: **Cog.** = Cognitive impairment, **Conc.** = Concerns by others, **Req.** = Requires assistance, **Phys.** = Physiological changes, **Neuro.** = Neuropsychiatric symptoms.

(F1 from 0.57 to 0.72) and Concerns by Others (F1 from 0.46 to 0.60), with substantial improvements also observed for Physiological Changes and Neuropsychiatric Symptoms. These results indicate that DualAlign enhances coverage of underrepresented and semantically nuanced categories, leading to more robust classification.

We further examined the effect of synthetic data volume by incrementally adding DualAlign-generated data to the gold training set in 10% increments. As shown in Figure 3, both binary F1 (left) and multi-class accuracy (right) improve steadily with additional synthetic data, plateauing around the 40% mark. Beyond this threshold, performance gains diminish, suggesting that a moderate amount of well-aligned synthetic data suffices to approach peak performance.

Together, these findings demonstrate that DualAlign is an effective complement to limited gold data, capable of scaling coverage while maintaining clinical plausibility. DualAlign's demographic and semantic alignment produces richer training signals, yielding improvements that generalize across diverse symptom types.

5.2 Human Evaluation

Two clinical experts assessed the quality of DualAlign-generated data along two dimensions:

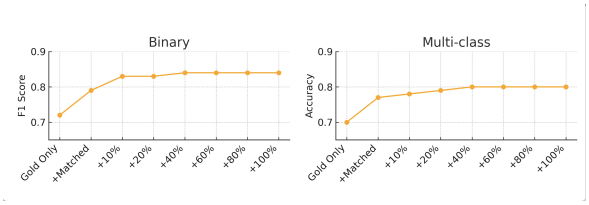

Figure 3: Performance with incremental addition of DualAlign-generated data. Left: Binary F1 score. Right: Multi-class accuracy.

(1) Sentence-level quality — realism, linguistic diversity, clinical complexity, and labeling accuracy — on 500 sentences from DualAlign and 500 from an unconstrained LLM baseline ("Bronze data"); and (2) Patient-level quality — narrative coherence and temporal plausibility across three full synthetic patient trajectories. We chose 500 sentences per system to balance category coverage with the feasibility of detailed expert review. Experts also provided qualitative feedback on symptom plausibility, disease progression consistency and annotation correctness.

5.2.1 Sentence-level Evaluation

Experts reported that DualAlign-generated sentences were markedly richer and more clinically informative than those from the unconstrained LLM baseline ("bronze data"). DualAlign outputs more frequently used precise clinical terminology, incorporated structured findings such as MMSE or MoCA scores with clear interpretation, and described functional or behavioral changes in a way that was contextually grounded rather than vague or repetitive. Compared to bronze data, reviewers observed fewer boilerplate or overly generic statements and greater variation in symptom presentation. Factual accuracy improved relative to the bronze baseline but was not perfect. Reviewers still found occasional clinical misstatements (e.g., confusion between *anosmia* and *hyposmia*) and some inconsistencies in how symptoms were mapped to disease stages, especially when interpreting cognitive test scores. Annotation accuracy was estimated at roughly 85%, with most residual errors concentrated in categories such as concerns raised by others and requires assistance. Common error types included incorrect handling of negation, semantic ambiguity, and overgeneralization of functional or behavioral limitations.

5.2.2 Longitudinal Evaluation

Generating longitudinal clinical narratives spanning many years remains challenging for current LLMs, due to their limitations in modeling long-range dependencies and the slow, complex progression of Alzheimer's disease. At the patient level, expert reviewers found that DualAlign-generated notes provided strong symptom coverage and clinically grounded structure, although they do not yet fully match real-world EHRs in temporal realism. Reviewers emphasized that, compared with prior synthetic datasets, DualAlign represents a clear advance in narrative breadth, organization, and clinical plausibility.

Nonetheless, several limitations were noted. Disease trajectories were occasionally temporally compressed, with cognitive transitions (e.g., from mild cognitive impairment to dementia) lacking nuance or consistency across visits. Reviewers also observed sporadic mismatches in clinical staging and timeline alignment. These challenges reflect well-known limitations of current LLMs in generating long-horizon, temporally coherent clinical narratives, rather than deficiencies specific to the DualAlign framework.

We also conducted a topic modeling analysis comparing synthetic and gold-standard data. We applied Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to both datasets to extract latent topics. The overlap rate of the top 20 topics over the 0–9 years prior to AD ranges from 55% to 100%, indicating a moderate but variable degree of alignment between the two data sources.

Further examination of the topic words reveals that the DualAlign-generated data aligns well with real data at the beginning and end of the studied period. This suggests that the proposed system is capable of capturing characteristic early and late-stage signals. However, its ability to model disease progression over time is more limited. This limitation is expected, as the progression of AD can vary significantly across individuals. While the observed trends reflect population-level patterns, the model struggles to capture patient-specific variability when applied at the individual level. These findings highlight the inherent challenges in generating realistic longitudinal patient profiles.

We also visualize the extracted topics using t-SNE; additional details are provided in the Appendix A.8.

6 Discussion

DualAlign addresses key limitations of existing LLM-based synthetic EHR generation, which often produces repetitive, context-poor note fragments with limited demographic realism and clinical grounding. By decomposing generation into two explicit and controllable stages—persona generation and symptom-trajectory alignment—DualAlign anchors narratives in realistic patient profiles while conditioning content on empirically observed, stage-specific symptom patterns. This design yields synthetic patients and narratives that are both demographically calibrated and clinically coherent, reducing generic repetition and improving contextual fidelity.

Evaluation demonstrates that DualAlign substantially improves the quality and downstream utility of synthetic data for AD research. Compared with unconstrained prompt-based generation, DualAlign produces symptom descriptions that are more informative, lexically diverse, and clinically plausible, as supported by diversity metrics, expert review, and consistent gains in both binary and multi-class classification tasks. Notably, clinicians highlighted improvements in contextual specificity and variation for semantically nuanced and underrepresented symptom categories.

Although demonstrated here for AD, DualAlign is inherently disease-agnostic. Persona generation can incorporate condition-specific risk factors and comorbidities, while symptom-trajectory alignment can be adapted to empirical progression patterns in other chronic or progressive diseases. We provide a guidance on how to apply it to other chronic diseases in the appendix as a reference A.9. The framework is also compatible with multilingual corpora and heterogeneous EHR systems, making it well suited for low-resource or privacy-constrained settings where access to high-quality annotated clinical data is limited.

At the same time, our findings underscore a persistent gap between synthetic and real longitudinal EHR data. Even with strong grounding in real-world distributions, current LLM-based approaches to long-form clinical text generation do not yet match the richness or temporal fidelity of real patient records. This gap reflects broader challenges in modeling long-range clinical trajectories, including limitations in temporal reasoning and difficulty capturing latent patient state beyond explicitly documented symptoms—challenges intrinsic to cur-

rent generative models rather than deficiencies of DualAlign itself.

As a step toward more realistic synthetic longitudinal EHR generation, DualAlign provides both a practical resource and a methodological foundation. Most importantly, it highlights longitudinal coherence—the consistent maintenance of patient context and disease progression across extended timelines—as a central open challenge. Addressing this limitation will require more advanced, clinically informed generation methods with stronger long-horizon reasoning capabilities, which we view as a critical direction for future work.

7 Conclusion

DualAlign introduces a framework for generating synthetic clinical narratives by combining demographic conditioning with symptom-trajectory alignment. Human evaluation verifies the effectiveness of the proposed method. Experiments on AD symptom classification show that DualAlign-generated data improves performance under limited supervision, reducing reliance on large volumes of human annotation and outperforming unguided synthetic baselines.

8 Limitations

Despite these advances, several limitations remain. First, achieving fully realistic longitudinal coherence over multi-year horizons remains challenging for current LLMs. While DualAlign generates clinically plausible trajectories with broad symptom coverage, expert review indicates that symptom progression can be temporally compressed and stage transitions may lack real-world granularity. These issues reflect limitations of LLMs in modeling long-range dependencies and complex temporal structure. Future work could incorporate planning-based generation, external memory, or clinician-in-the-loop refinement to improve temporal fidelity.

Second, annotation accuracy—though guided by structured protocols—showed an error rate of roughly 15%, particularly in semantically subtle categories (e.g., requires assistance, concerns raised by others), often due to ambiguity, negation, and overgeneralization. More advanced semantic parsing or post-hoc correction could improve reliability. While DualAlign data provides strong training signals, models trained on synthetic data should be externally validated on real-world cohorts, especially under distribution shifts. Synthetic data should therefore be viewed as a complement to, rather than a replacement for, real annotated EHRs.

Third, although DualAlign is designed to be disease-agnostic, our empirical evaluation focused exclusively on AD. Additional validation across other chronic or progressive conditions will be necessary to assess the generality of the framework.

Fourth, our real-world grounding relies on EHR data from a single healthcare system, which may reflect institution-specific documentation practices. While our demographic reweighting and alignment strategy aim to mitigate single-site bias, future work should examine performance when grounding on data from multiple health systems. Moreover, human evaluation—though conducted by clinicians—was limited in scale; broader expert review will be important to further assess realism, temporal fidelity, and annotation quality.

Finally, we do not include conventional ablation studies (e.g., persona-only vs. symptom-only generation). In the context of longitudinal LLM-based synthetic EHR generation, such ablations are difficult to interpret: prompt-level modifications do not correspond to isolatable model components, and independently removing demographic context or symptom trajectories would break clinical coherence rather than reveal causal contributions. Moreover, generating sufficiently large longitudinal cohorts for each variant is not practical. We therefore rely on downstream task performance, expert review, and diversity and distributional analyses, which more directly assess synthetic data fidelity and utility.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) under Award Number 5R01AG080670. This study uses data from the U.S. Veterans Health Administration (VHA) and was conducted using resources and facilities at the Center for Health Organization and Implementation Research (CHOIR), VA Bedford Healthcare System, Bedford, MA, USA. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and VHA. The funding agencies had no role in the design of the study; the collection, analysis, and interpretation of data.

References

Austin A Barr, Joshua Quan, Eddie Guo, and Emre Sezgin. 2025. Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data. *Frontiers in Artificial Intelligence*, 8:1533508.

MAPPING A Better. 2023. Alzheimer's disease facts and figures. *Alzheimers Dement*, 19(4):1598–1695.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Matthias Ganzinger, Nicola Kunz, Pascal Fuchs, Cornelia K Lyu, Martin Loos, Martin Dugas, and Thomas M Pausch. 2025. Automated generation of discharge summaries: leveraging large language models with clinical data. *Scientific Reports*, 15(1):16466.

Mauro Giuffrè and Dennis L Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, 6(1):186.

Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082.

Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45.

Aaron Jones, Muhammad Usman Ali, Alexandra Mayhew, Komal Aryal, Rebecca H Correia, Darly Dash, Derek R Manis, Atiya Rehman, Megan E O'Connell, Vanessa Taler, and 1 others. 2025. Environmental risk factors for all-cause dementia, alzheimer's disease dementia, vascular dementia, and mild cognitive impairment: An umbrella review and meta-analysis. *Environmental research*, 270:121007.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, and 1 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Qian Li, Xi Yang, Jie Xu, Yi Guo, Xing He, Hui Hu, Tianchen Lyu, David Marra, Amber Miller, Glenn Smith, and 1 others. 2023a. Early prediction of alzheimer's disease and related dementias using real-world electronic health records. *Alzheimer's & Dementia*, 19(8):3506–3518.

Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. 2025. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541.

Rumeng Li, Xun Wang, and Hong Yu. 2023b. Two directions for clinical data generation with large language models: data-to-label and label-to-data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 7129.

JL Liss, S Seleri Assunção, Jeffrey Cummings, A Atri, DS Geldmacher, SF Candela, DP Devanand, HM Fillit, J Susman, J Mintzer, and 1 others. 2021. Practical recommendations for timely, accurate diagnosis of symptomatic alzheimer's disease (mci and dementia) in primary care: a review and synthesis. *Journal of internal medicine*, 290(2):310–334.

Fang Liu and Demosthenes Panagiotakos. 2022. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, 22(1):287.

Yintong Liu, U Rajendra Acharya, and Jen Hong Tan. 2025. Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation. *Computer Methods and Programs in Biomedicine*, 260:108571.

Gill Livingston, Jonathan Huntley, Kathy Y Liu, Sergi G Costafreda, Geir Selbæk, Suvarna Alladi, David Ames, Sube Banerjee, Alistair Burns, Carol Brayne, and 1 others. 2024. Dementia prevention, intervention, and care: 2024 report of the lancet standing commission. *The Lancet*, 404(10452):572–628.

Mohammad Loni, Fatemeh Poursalim, Mehdi Asadi, and Arash Gharehbaghi. 2025. A review on generative ai models for synthetic medical text, time series, and longitudinal data. *npj Digital Medicine*, 8(1):281.

Mohamad Homam Mawaldi and Martin Mladenov. 2024. Synthetic data generation using large language models evaluating the utility of synthetic clinical text generated via fine-tuning llama-2 on mimic-iii data when used as training data for clinical named entity recognition.

Meta AI. 2024. Introducing Llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/. Accessed: 2025-08-01.

Graciela Muniz-Terrera, Ofer Mendelevitch, Rodrigo Barnes, and Michael D Lesh. 2021. Virtual cohorts and synthetic data in dementia: an illustration of their potential to advance research. *Frontiers in Artificial Intelligence*, 4:613956.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, and 1 others. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Aziel Alejandro Peralta Ramirez, Sergio Trujillo López, Gonzalo Armando Navarro Armendariz, Sayil Alejandra De la Torre Othón, Marcial Ramin Sierra Cervantes, and Juan Antonio Medina Aguirre. 2025. Clinical simulation with chatgpt: A revolution in medical education? *Journal of CME*, 14(1):2525615.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, and 1 others. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119.

Susanne Röhr, Alexander Pabst, Ronny Baber, Christoph Engel, Heide Glaesmer, Andreas Hinz, Matthias L Schroeter, A Veronica Witte, Samira Zeynalova, Arno Villringer, and 1 others. 2022. Social determinants and lifestyle factors for brain health: implications for risk reduction of cognitive decline and dementia. *Scientific Reports*, 12(1):12965.

Miguel Rujas, Rodrigo Martín Gómez del Moral Herranz, Giuseppe Fico, and Beatriz Merino-Barbancho. 2025. Synthetic data generation in healthcare: A scoping review of reviews on domains, motivations, and future applications. *International Journal of Medical Informatics*, 195:105763.

Muhammad Sajjad, Farheen Ramzan, Muhammad Usman Ghani Khan, Amjad Rehman, Mahyar Kolivand, Suliman Mohamed Fati, and Saeed Ali Bahaj. 2021. Deep convolutional generative adversarial network for alzheimer's disease classification using positron emission tomography (pet) and synthetic data augmentation. *Microscopy Research and Technique*, 84(12):3023–3034.

Savyasachi V Shah. 2024. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953–e2425953.

Daniel Smolyak, Margrét V Bjarnadóttir, Kenyon Crowley, and Ritu Agarwal. 2024. Large language models and synthetic health data: progress and prospects. *JAMIA open*, 7(4):ooae114.

Shana D Stites, Sharnita Midgett, Dawn Mechanic-Hamilton, Megan Zuelsdorff, Crystal M Glover, David X Marquez, Joyce E Balls-Berry, Marissa L Streitz, Ganesh Babulal, Jean-Francois Trani, and 1 others. 2022. Establishing a framework for gathering structural and social determinants of health in alzheimer's disease research centers. *The Gerontologist*, 62(5):694–703.

Andrea Taloni, Giulia Coco, Marco Pellegrini, Matthias Wjst, Niccolò Salgari, Giovanna Carnovale-Scalzo, Vincenzo Scorcia, Massimo Busin, and Giuseppe Giannaccare. 2025. Exploring detection methods for synthetic medical datasets created with a large language model. *JAMA ophthalmology*, 143(6):517–522.

Donna Tjandra, Raymond Q Migrino, Bruno Giordani, and Jenna Wiens. 2020. Cohort discovery and risk stratification for alzheimer's disease: an electronic health record-based approach. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 6(1):e12035.

Wiesje M van der Flier, Marjolein E de Vugt, Ellen MA Smets, Marco Blom, and Charlotte E Teunissen. 2023. Towards a future where alzheimer's disease pathology is stopped before the onset of dementia. *Nature aging*, 3(5):494–505.

Liqin Wang, John Laurentiev, Jie Yang, Ying-Chih Lo, Rebecca E Amariglio, Deborah Blacker, Reisa A Sperling, Gad A Marshall, and Li Zhou. 2021. Development and validation of a deep learning model for earlier detection of cognitive decline from clinical notes in electronic health records. *JAMA network open*, 4(11):e2135174–e2135174.

Christopher YK Williams, Charumathi Raghu Subramanian, Syed Salman Ali, Michael Apolinario, Elisabeth Askin, Peter Barish, Monica Cheng, W James Deardorff, Nisha Donthi, Smitha Ganeshan, and 1 others. 2025. Physician-and large language model–generated hospital discharge summaries. *JAMA Internal Medicine*.

Jie Xu, Fei Wang, Zhenxing Xu, Prakash Adekkanattu, Pascal Brandt, Guoqian Jiang, Richard C Kiefer, Yuan Luo, Chengsheng Mao, Jennifer A Pacheco, and 1 others. 2020. Data-driven discovery of probable alzheimer's disease and related dementia subphenotypes using electronic health records. *Learning Health Systems*, 4(4):e10246.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. 2024. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15496–15523.

A Appendix

A.1 AD Risk Factors

Factor	Category	Prevalence (%)
1. DEMOGRAPHIC & SOCIOECONOMIC FACTORS		
Age	<65 (early-onset)	8.0%
Age	65–74	22.0%
Age	75–84	45.0%
Age	≥85	25.0%
Gender	Male	42.0%
Gender	Female	56.0%
Gender	Non-binary/Other	2.0%
Race	White	58.0%
Race	Black or African American	22.0%
Race	Asian	8.0%
Race	American Indian or Alaska Native	0.5%
Race	Native Hawaiian or Other Pacific Islander	0.5%
Race	Mixed/Multiracial	6.0%
Race	Others/unknown	5.0%
Ethnicity	Hispanic/Latino	15.0%
Ethnicity	Non-Hispanic/Latino	80.0%
Ethnicity	Others/unknown	5.0%
Geographic Location	Urban	55.0%
Geographic Location	Suburban	30.0%
Geographic Location	Rural	15.0%
Education Level	No formal education	3.0%
Education Level	Primary	25.0%
Education Level	Secondary	45.0%
Education Level	College	20.0%
Education Level	Postgraduate	7.0%
Financial Status	Low income	38.0%
Financial Status	Middle income	55.0%
Financial Status	High income	7.0%
Employment/Occupation	Retired	65.0%
Employment/Occupation	Manual labor	20.0%
Employment/Occupation	Professional	10.0%
Employment/Occupation	Unemployed	5.0%
Health Insurance	None	5.0%
Health Insurance	Public (e.g., Medicare/Medicaid)	75.0%
Health Insurance	Private	20.0%
Health Literacy	Low	35.0%
Health Literacy	Moderate	50.0%
Health Literacy	High	15.0%
Housing Instability	Stable	82.0%
Housing Instability	Unstable (eviction/foreclosure)	15.0%
Housing Instability	Homeless	3.0%

Factor	Category	Prevalence (%)
2. MEDICAL & BIOLOGICAL FACTORS		
Family History of AD	Yes	28.0%
Family History of AD	No	72.0%
Hypertension	Yes	68.0%
Hypertension	No	32.0%
Diabetes	Yes	34.0%
Diabetes	No	66.0%
Cardiovascular Disease	Yes	45.0%
Cardiovascular Disease	No	55.0%
Obesity	Yes	41.0%
Obesity	No	59.0%
Stroke History	Yes	18.0%
Stroke History	No	82.0%
Autoimmune Disorders	Yes	12.0%
Autoimmune Disorders	No	88.0%
Traumatic Brain Injury (TBI)	Yes	9.0%
Traumatic Brain Injury (TBI)	No	91.0%
Epilepsy	Yes	4.0%
Epilepsy	No	96.0%
Chronic Inflammation	Yes	27.0%
Chronic Inflammation	No	73.0%
Depression Diagnosis	Diagnosed	22.0%
Depression Diagnosis	Undiagnosed	15.0%
Depression Diagnosis	Untreated	8.0%
Anxiety Diagnosis	Diagnosed	18.0%
Anxiety Diagnosis	Undiagnosed	12.0%
Anxiety Diagnosis	Untreated	7.0%
Bipolar Disorder Diagnosis	Diagnosed	4.0%
Bipolar Disorder Diagnosis	Undiagnosed	2.0%
Bipolar Disorder Diagnosis	Untreated	1.0%
Schizophrenia Diagnosis	Diagnosed	3.0%
Schizophrenia Diagnosis	Undiagnosed	1.0%
Schizophrenia Diagnosis	Untreated	0.5%
PTSD	Yes	11.0%
PTSD	No	89.0%
Hearing Loss Severity	None	45.0%
Hearing Loss Severity	Mild	35.0%
Hearing Loss Severity	Moderate	15.0%
Hearing Loss Severity	Severe	5.0%
Vision Loss Severity	None	50.0%
Vision Loss Severity	Mild	30.0%
Vision Loss Severity	Moderate	15.0%
Vision Loss Severity	Severe	5.0%

Factor	Category	Prevalence (%)
Chronic Pain	Yes	39.0%
Chronic Pain	No	61.0%
Acute Pain	Yes	25.0%
Acute Pain	No	75.0%
Physical Disability	Yes	33.0%
Physical Disability	No	67.0%
Cognitive Disability	Yes	28.0%
Cognitive Disability	No	72.0%
3. LIFESTYLE & ENVIRONMENTAL FACTORS		
Diet Type	Balanced	48.0%
Diet Type	Poor (high processed foods)	52.0%
Substance Abuse (legal/illicit)	Yes	17.0%
Substance Abuse (legal/illicit)	No	83.0%
Smoking Status	Never	45.0%
Smoking Status	Former	35.0%
Smoking Status	Current	20.0%
Alcohol Use	None	40.0%
Alcohol Use	Moderate	50.0%
Alcohol Use	Heavy	10.0%
Physical Activity Level	Sedentary	55.0%
Physical Activity Level	Moderate	35.0%
Physical Activity Level	Active	10.0%
Sleep Patterns	Regular	60.0%
Sleep Patterns	Irregular	40.0%
Air Pollution Exposure	Yes	35.0%
Air Pollution Exposure	No	65.0%
4. PSYCHOSOCIAL & STRESS-RELATED FACTORS		
Physical Abuse	Yes	7.0%
Physical Abuse	No	93.0%
Emotional Abuse	Yes	15.0%
Emotional Abuse	No	85.0%
Sexual Abuse	Yes	4.0%
Sexual Abuse	No	96.0%
Combat Exposure	Yes	6.0%
Combat Exposure	No	94.0%
Racism/Discrimination	Yes	22.0%
Racism/Discrimination	No	78.0%
Legal Problems	Yes	9.0%
Legal Problems	No	91.0%
Cultural Stigma Around AD	Yes	31.0%
Cultural Stigma Around AD	No	69.0%

Factor	Category	Prevalence (%)
Internalized Shame/Guilt	Yes	19.0%
Internalized Shame/Guilt	No	81.0%
Social Engagement	High (regular social interaction)	35.0%
Social Engagement	Moderate	45.0%
Social Engagement	Isolated	20.0%
Marital Status	Single	15.0%
Marital Status	Married	50.0%
Marital Status	Divorced	25.0%
Marital Status	Widowed	10.0%
Caregiver Availability	Family	65.0%
Caregiver Availability	Professional caregiver	25.0%
Caregiver Availability	None	10.0%
Stress Levels	Low	25.0%
Stress Levels	Moderate	50.0%
Stress Levels	High	25.0%
5. ACCESS TO CARE & STRUCTURAL BARRIERS		
Proximity to Healthcare	Easy access	60.0%
Proximity to Healthcare	Limited access	30.0%
Proximity to Healthcare	Hard	10.0%
Public Transport Access	Easy access	55.0%
Public Transport Access	Limited access	30.0%
Public Transport Access	No access	15.0%
Primary Language	English	82.0%
Primary Language	Spanish	12.0%
Primary Language	Other	6.0%
6. DEVELOPMENTAL & LIFECOURSE FACTORS		
Childhood Trauma	Yes	13.0%
Childhood Trauma	No	87.0%
Undocumented Immigrant Status	Yes	4.0%
Undocumented Immigrant Status	No	96.0%

A.2 Note Statistics

Table 5 summarizes the average number of notes per patient by note type and time window before formal Alzheimer's disease diagnosis. These statistics describe the empirical visit patterns used to guide our synthetic note generation process.

A.3 Keyword Distribution

Tables 6 and 7 summarize the real-world distribution of AD-related keywords that guided our synthetic note generation. Table 6 shows the average number of keywords mentioned per note across different years before formal diagnosis, capturing temporal trends in symptom documentation. Table 7 reports the relative frequency of keywords across major symptom categories (normalized to memory = 1), providing category-level priors for keyword sampling.

Table 5: Average Number of Notes per Patient by Note Type and Time Window Before Diagnosis

Note Type	**10–7 Years Before**	**6–4 Years Before**	**3–2 Years Before**	**1 Year Before**
Primary care	2.54	2.59	2.95	5.01
Neurology	0.31	0.74	1.18	2.51
Memory clinic	0.31	0.74	1.18	2.51
Neuropsychology	0.31	0.74	1.18	2.51
Geriatrics	1.02	0.74	0.59	1.67
Psychiatry/Mental health	0.51	0.74	0.89	1.67
Emergency visits	1.02	1.11	1.18	2.51
Home-based primary care (HBPC)	0.00	0.00	0.59	1.67

Table 6: Average Keyword Count per Note by Years Before Diagnosis

Years Before Diagnosis	**Avg. Keywords per Note**
10	2.745
9	2.874
8	2.993
7	3.101
6	3.272
5	3.384
4	3.508
3	3.678
2	3.829
1	4.160

Table 7: Average Keyword Proportion by Category (Relative to Memory = 1)

Keyword Category	**Proportion**
Speech/language	2.746
Memory	1.000
Learning/Perception	1.733
Assistance Needed	1.531
Physiological Changes	8.766
Neuropsychiatric Symptoms	4.399

A.4 Example Note

Clinical Note

Patient Name: [redacted] **Date**: [redacted]
Patient ID: [redacted] **DOB**: [redacted]
Provider: Dr. [redacted] **Visit Type**: Primary Care

Subjective:
The patient is a 70-year-old African American female, currently retired, living in urban conditions with unstable housing. She reports moderate levels of stress and is experiencing increased forgetfulness, which she notices mainly in her daily activities, such as misplacing household items and missing appointments. Her daughter, who accompanied her, mentioned the patient has been asking the same questions repeatedly. The patient is also experiencing difficulty with balance and has had minor falls at home. She reports mild vision loss but denies any significant hearing loss (HoH). Though she has no serious auditory issues, she sometimes misses parts of conversations. The patient complains of chronic pain exacerbated by her physical disability. Sleep patterns are irregular, and she describes symptoms consistent with insomnia. The daughter reports some issues with smell (hyposmia) but denies complete anosmia. The patient also notes occasional difficulty swallowing, though no significant dysphagia is observed. Recent challenges with incontinence have been causing distress.

Objective:
- **Vital Signs**: BP 150/95 mmHg (hypertensive), HR 82 bpm, Temp 98.6°F
- **Physical Exam**: No acute distress, gait appears slightly unsteady, mild vision loss confirmed. Hearing (auditory) screening shows within normal range, but sporadic minor misses in auditory cues reported.
- **Neurological Examination**: Balance slightly impaired; Mini-Mental State Examination (MMSE) reveals a score of 26/30, indicating possible early cognitive decline.
- **Recent labs**: Elevated cholesterol levels consistent with cardiovascular disease; no hyperglycemia present.
- **Cognitive testing**: Reflects mild cognitive impairment, consistent with early signs of Alzheimer's disease.

Assessment:
The patient presents with early prodromal signs that could be consistent with developing Alzheimer's disease, characterized by forgetfulness, imbalance issues, and mild cognitive impairment. Her cardiovascular disease and lifestyle, including poor diet and sedentary activity, could contribute to cognitive decline. Chronic inflammation due to her physical conditions and comorbid depression and bipolar disorder could also be influencing factors. Differential diagnoses include:

1. Early Alzheimer's disease
2. Vascular dementia related to her significant cardiovascular disease
3. Mild Cognitive Impairment (MCI) secondary to depression and lifestyle factors

Plan:

1. Encourage lifestyle modifications, including improved diet and regular physical activity, to address cardiovascular risks and potentially slow cognitive decline.
2. Referral to a neurologist for further evaluation and management of cognitive decline.
3. Adjust current antihypertensive regimen to better manage high blood pressure, possibly improving cognitive and balance issues.
4. Recommend a sleep hygiene program and consider non-pharmacological interventions for insomnia.
5. Monitor any worsening of dysphagia or balance issues and consider referral to specialists if

symptoms persist or worsen.
6. Counseling and support for depression and bipolar disorder, possibly adjusting psychiatric medications if necessary.
7. Discuss social support and housing stability solutions to reduce stressors that may exacerbate symptoms.

Follow-up: Scheduled in 3 months for reassessment and ongoing management of symptoms.

Provider Signature: Dr. [redacted] **Date**: [redacted]

A.5 Temporal Context

Year before AD: AD development stage
10: "Early prodromal stage",
9: "Early prodromal stage",
8: "Early prodromal stage",
7: "Early prodromal stage",
6: "Mild cognitive impairment stage",
5: "Mild cognitive impairment stage",
4: "Mild dementia stage",
3: "Mild dementia stage",
2: "Moderate dementia stage",
1: "Moderate dementia stage"

A.6 Keywords for Each Category

Speech/language
- communication, speech, speaking
- word-finding, word-retrieval, naming, encoding, phonemic
- aphasia, paraphasia, anomia, dysnomia
- fluency, perseveration, repetition
- language, linguistic
- comprehend, understand, alexia

Memory
- memory, amnesia, amnestic
- remembering, recognizing, recall, recount, retain
- forget, lapse

Learning/Perception
- attention, concentration, focus
- learning, abstraction, problem-solving
- executive function, cognitive, neurocognitive, thinking, processing
- visuospatial, multidomain, global, agnosia
- getting lost, trouble finding, disoriented, confusion
- handwriting deterioration

Assistance Needed
- ADLs: eating, dressing, grooming, toileting, bathing, mobility
- iADLs: cooking, housekeeping, cleaning, laundry, shopping
- phone use, computer use
- managing medications, managing bills, managing finances
- driving, transportation
- medical and legal decision-making
- healthcare proxy, HPOA, guardian, guardianship
- supervision required

Physiological Changes
- hearing, auditory, SNHL, HoH
- vision
- smell, anosmia, hyposmia
- swallowing, dysphagia
- gait, balance
- sleep, insomnia
- pain
- incontinence

Neuropsychiatric Symptoms
- mood, affect, behavior, apathy
- personality
- depressed, anhedonia

- anxiety, anxious, agitation, hypervigilance, restless, overwhelmed
- insight, judgment, impulsive, anosognosia
- anger, short-tempered, irritable, aggressive, shouting
- erratic, rummaging
- wandering
- thought disorder
- delusion, hallucination, paranoia, psychosis

A.7 Annotation Guideline

Class 1: Cognitive impairment
Cognitive impairment is when a person has trouble remembering, learning, concentrating, or making decisions that affect their everyday life, including patient subjective statements as well as Dr. statements.

Class 2: Notice/concern by others (not by provider; refers to friends/family/neighbors)
Family complains of X (may be related to any class including cognitive decline, functional impairment, neuropsychiatric and physiological changes).

Class 3: Requires assistance / functional impairment
Needs help with or loss of ability with ADLs/i-ADLs, difficulty with self-care, trouble managing belongings; may include need for supervision.

Class 4: Physiological changes
Senses (vision/hearing/smell), sleep changes, swallowing issues, movement/gait/balance. (Focus on early possible associations; very late physiological changes may be skipped.)

Class 5: Neuropsychiatric symptoms
Mood/behavior changes (depression, irritability, aggression, anxiety, apathy), personality change, paranoia/delusions/hallucinations/psychosis.

A.8 Longitudinal Evaluation of DualAlign

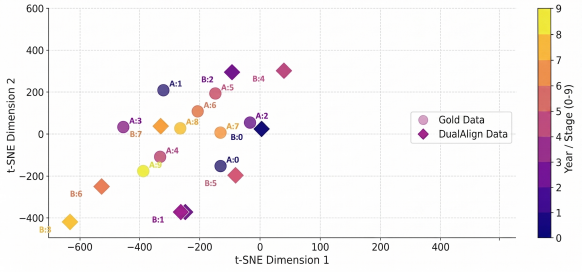

Figure 4: Plotting Topics of Gold and DualAlign Data for 9 to 0 Years prior to AD

Years Prior to AD	Overlap Rate (%)
0	95
1	55
2	95
3	65
4	60
5	55
6	60
7	60
8	75
9	100

Table 8: Topic Overlap Rate between DualAlign and Gold Data

A.9 Guidance on Applying to Other Diseases

To extend DualAlign beyond Alzheimer's disease, adapt each alignment component to disease-specific clinical structure:

1. Define disease-specific personas
   - Replace AD risk factors with condition-relevant variables (e.g., for diabetes: BMI, HbA1c, lifestyle; for COPD: smoking history, environmental exposure).
   - Use epidemiological data to sample realistic demographic and comorbidity distributions.
2. Model longitudinal trajectories
   - Identify key disease stages (e.g., onset → progression → complications).
   - Derive data like time-dependent patterns of visits, symptom evolution, from real-world data.
   - Align note frequency and type with typical disease management pathways.
3. Build a symptom/event ontology
   - Curate clinician-validated lexicons covering core domains (symptoms, functional impact, etc.).
   - Preserve category proportions and temporal trends.
4. Align temporal symptom patterns
   - Estimate how symptoms emerge over time.
   - Use stratified sampling to reflect progression dynamics (e.g., gradual worsening vs. episodic flares).
5. Adapt prompt templates
   - Encode disease context: patient profile, stage, care setting, and required symptom anchors.
6. Customize annotation schema
   - Define clinically meaningful categories aligned with downstream tasks (e.g., complications, severity, treatment response).
   - Focus on observable signals in routine care notes.

7. Validate and calibrate

• Perform expert review for clinical plausibility and temporal coherence.

• Benchmark utility on disease-relevant tasks (e.g., risk prediction, phenotype classification).

Key principle: Maintain the dual alignment structure—population-grounded personas + empirically derived longitudinal patterns—while replacing AD-specific components with disease-specific clinical knowledge.