CLIN-SUMM: Incremental Longitudinal Summarization of Clinical Notes Enables Scalable Representation and Early Disease Prediction

Valentina D'Souza, MSc[1*], Danielle F. Pace, PhD[1*], Alaleh Azhir, MD[1,2], Arash Nargesi, MD, MPH[1,2], Erik B. Holbrook, MD, MSc[1,2], Wei He[2], Tristan Naumann, PhD[3], Samuel Friedman, PhD[1], Steven J. Atlas, MD[2,4], Christopher D. Anderson, MD, MSc[2], Judy Hung, MD[2], and Mahnaz Maddah, PhD[1†]

[1] Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2] Massachusetts General Brigham, Boston, MA, USA
[3] Microsoft Research, Cambridge, MA, USA
[4] Harvard Medical School, Boston, MA, USA

*Authors contributed equally to this project

[†]Corresponding Author:
Mahnaz Maddah, PhD
Director, Machine Learning for Health
Broad Institute of MIT and Harvard
415 Main Street, Cambridge, MA 02142
Email: maddah@broadinstitute.org
Phone: (617) 714-7000

Abstract

Electronic health records contain years of longitudinal clinical notes rich in evolving patient information, yet their volume, redundancy, and fragmentation limit clinical usability and scalable modeling. We present CLIN-SUMM (Clinical Longitudinal Insight from Notes using Summarization), a framework that restructures summarization as a longitudinal representation problem. Rather than collapsing histories into static summaries, CLIN-SUMM incrementally constructs structured, categorized, date-partitioned patient representations, summarizing only newly documented information at each encounter while preserving temporal fidelity without access to future data. This standardized representation can be computed once and reused across downstream tasks, thereby decoupling narrative processing from prediction. Across 12,356 Massachusetts General Hospital patients, CLIN-SUMM achieved 70% token

reduction while maintaining high clinician-rated correctness and completeness. Using dementia as a case study, fine-tuning Clinical ModernBERT on CLIN-SUMM summaries yielded AUROC 0.86 for diagnosis and 0.81 for 3-year risk prediction, with longitudinal analyses demonstrating progressive risk separation years before a formal diagnosis. CLIN-SUMM summaries also enabled efficient extraction of longitudinal medication trajectories, improving medication capture compared to structured EHR data while maintaining high dosage agreement. CLIN-SUMM provides a scalable representation layer for clinical review and longitudinal machine learning, enhancing disease modeling, risk prediction, and other longitudinal reasoning tasks.

Introduction

Electronic health records (EHRs) contain years of longitudinal clinical notes documenting the evolving trajectory of a patient's medical history. These narratives capture symptoms, diagnoses, treatment decisions, responses to therapy, lifestyle factors, labs, imaging, and contextual clinical reasoning that are often incompletely represented in the structured codes typically used for machine learning [1–3]. Clinical notes, therefore, are likely to remain narrative documents containing essential clinical information that can support a range of downstream tasks, including automated diagnosis, risk prediction, phenotyping, trajectory analysis, and clinical question answering.

Despite their richness, longitudinal notes remain difficult to use effectively at scale, whether by clinicians doing chart review or by machine learning models attempting to learn temporal patterns. The most obvious challenge is sheer volume: records of individual patients frequently accumulate hundreds to thousands of notes over many years. There is substantial redundancy due to use of note templates and copy-forward practices, with some estimating that over 50% of text is duplicated over encounters [4]. The most critical clinical updates, such as new diagnoses, medication adjustments, or symptom progression, are often buried within thousands of lines of unstructured narrative text [5]. Fragmentation compounds the problem: clinical observations are

scattered across encounters and specialties, and individual notes rarely integrate longitudinal context into a coherent temporal account. Consequently, both clinicians and algorithms face substantial barriers when attempting to summarize care over unstructured narratives across time. Even long context models that aim to process larger portions of the patient record still contend with low signal-to-noise ratios, redundancy, heterogeneous formatting, and the "lost in the middle" phenomenon, limiting their ability to extract coherent longitudinal insight [6, 7].

However, most large language model (LLM) applications addressing clinical note summarization, whether extractive [8] or abstractive [9], typically operate only at the level of individual notes [10] or episodes (e.g., a hospitalization) [11], or collapse multi-visit histories into static summaries that obscure temporal evolution. Multi-modal approaches combining structured EHR data with clinical notes similarly operate only at the note or episode level [12], while methods that derive structured concepts from unstructured text lack the nuance contained within clinical narratives while remaining repetitive and unorganized [13, 14].

The central challenge is therefore not simply summarization, but representation: how can we reliably transform large volumes of longitudinal clinical text into an organized, concise format that preserves the evolving clinical context, while remaining clinically navigable for chart review and computationally tractable for downstream modeling?

We introduce CLIN-SUMM (Clinical Longitudinal Insight from Notes using Summarization), a framework designed to generate structured, categorized, date-partitioned patient-level summaries through sequential summarization of multi-visit clinical notes (**Figure 1**). CLIN-SUMM does not simply "squash" a patient's history into a paragraph, as in conventional summarization. At each encounter, CLIN-SUMM populates predefined clinical sections (e.g., "Diagnosis", "Medications & Allergies", "Vital signs or Lab/Imaging Findings" and more) with summaries of only the <u>new</u> information arising from that specific visit and appends date-stamped updates to explicitly capture changes over time.

The resulting representation grows cumulatively as care continues, forming a longitudinal, structured summary of the patient's health trajectory. In this sense, CLIN-SUMM encodes computationally how clinicians mentally update their understanding of a patient: new information refines and extends prior knowledge rather than rewriting history. Notes are processed strictly in chronological order. At every step the model updates the patient-level summary using only information available up to that time point, and the system never has access to "future" documentation when summarizing a given encounter. This preserves temporal fidelity and causal direction, aligning the representation with real-world clinical decision-making. Because CLIN-SUMM updates patient summaries after each encounter, a hospital system could feasibly generate and maintain longitudinal summaries as part of routine documentation in near real-time.

Conceptually, CLIN-SUMM functions as a representation learning layer between raw narrative text and downstream modeling tasks. Rather than training task-specific models directly on the original highly variable and repetitive clinical notes, the same durable longitudinal patient summary can be computed once and reused across endpoints, whether for manual clinical review, medication or vital sign extraction, symptom trend analysis, development of any number of machine learning models for disease classification or risk prediction, or other targeted tasks. This decouples representation construction from prediction, reducing repeated work during model development. Computational expense is also dramatically reduced, both because the summaries must only be computed once, and because the substantially compressed summaries require a fraction of tokens.

CLIN-SUMM uses a two-prompt architecture (one for baseline synthesis and another for incremental updates), automated redundancy filtering, and a sliding window for context management. The method is agnostic to the exact choice of LLM, and here we demonstrate CLIN-SUMM implemented with the open-source, low-cost Qwen3 [15] model, which ensures reproducibility and privacy-preserving local deployment. We evaluate the framework using longitudinal primary care data from a cohort of more than 12,000 patients, and evaluate the

summaries according to examples of their intended use, rather than relying solely on surface-level similarity to a reference text [16,17]. We demonstrate that CLIN-SUMM summaries (1) achieve high compression efficiency, (2) maintain clinical fidelity under structured clinician review, with high scores for correctness and completeness, (3) underpin strong downstream performance in dementia diagnosis and 3-year prediction modeling, and (4) facilitate automated medication extraction. Together, these results suggest that incremental, structured longitudinal summarization can provide a scalable and reusable mechanism for unlocking the value of unstructured clinical notes for both patient care and predictive modeling.

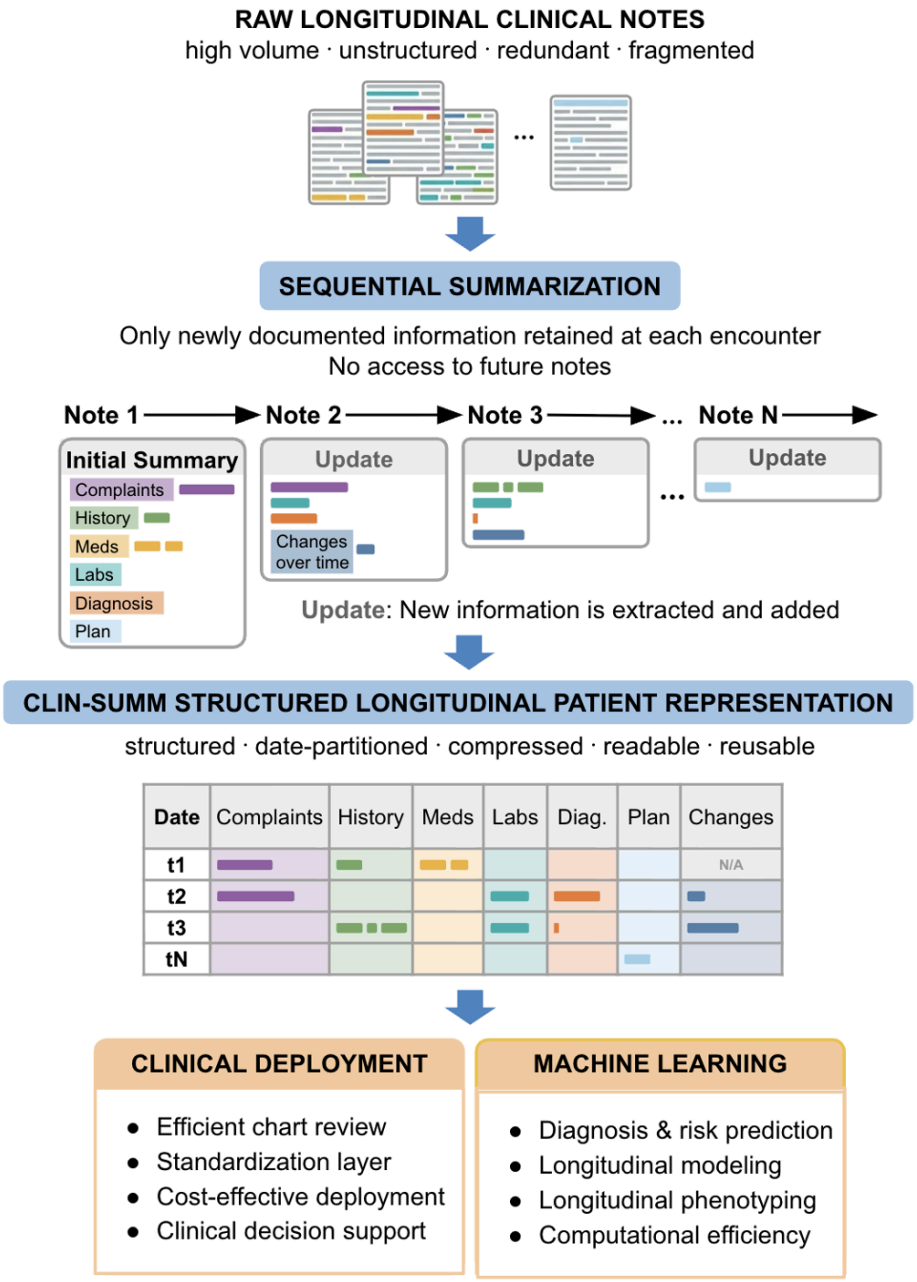


Figure 1: Overview of the CLIN-SUMM framework. The resulting patient-level data for each patient includes a summary of the first clinical note followed by summaries of the new updates described in subsequent notes, where each section captures how a patient's clinical state evolves across encounters with corresponding dates. Patient visit notes are processed using LLM-powered

summarization prompts, redundancy filtering, and sliding-window management to generate a longitudinal, date-partitioned clinical summary.

2 Results

2.1 Summary generation and experimental setting

CLIN-SUMM longitudinally summarizes fragmented and repetitive EHR patient visit notes into a coherent, organized, efficient and date-partitioned summary representation for each patient that grows over time as clinical notes accumulate from encounters with the healthcare system. CLIN-SUMM uses a two-prompt architecture to convert clinical notes to compressed summaries using an LLM, as shown in **Figure 2**, with full prompt templates and technical implementation details provided in **Supplementary Figure 1** and **Supplementary Section 5 (Prompt Optimization and Parameter Selection)**. The first initial summary prompt uses an LLM to organize all of the text from the first clinical note into seven sections ("Chief Complaints", "History of Present Illness", "Past Medical History", "Medications & Allergies", "Vital signs or Lab/Imaging Findings", "Diagnosis", and "Treatment Plan and Follow-up"). The second, change-focused update prompt operates on each subsequent note. This prompt inputs the patient's previously generated summaries so that the LLM adds only novel information from that specific encounter to each section and also populates a "Changes over Time" section that summarizes key temporal updates such as medication adjustments, new diagnoses, or changes in lab/imaging findings. CLIN-SUMM includes redundancy control via a Jaccard similarity [18] filter that avoids preprocessing near-identical notes, uses a sliding window mechanism to handle patients with long encounter histories without exceeding LLM context limits, and has a generalizable design that can be applied to a wide range of phenotypes and downstream tasks. Our setup is generic, and we demonstrate our main results using the Qwen3 [15] LLM.

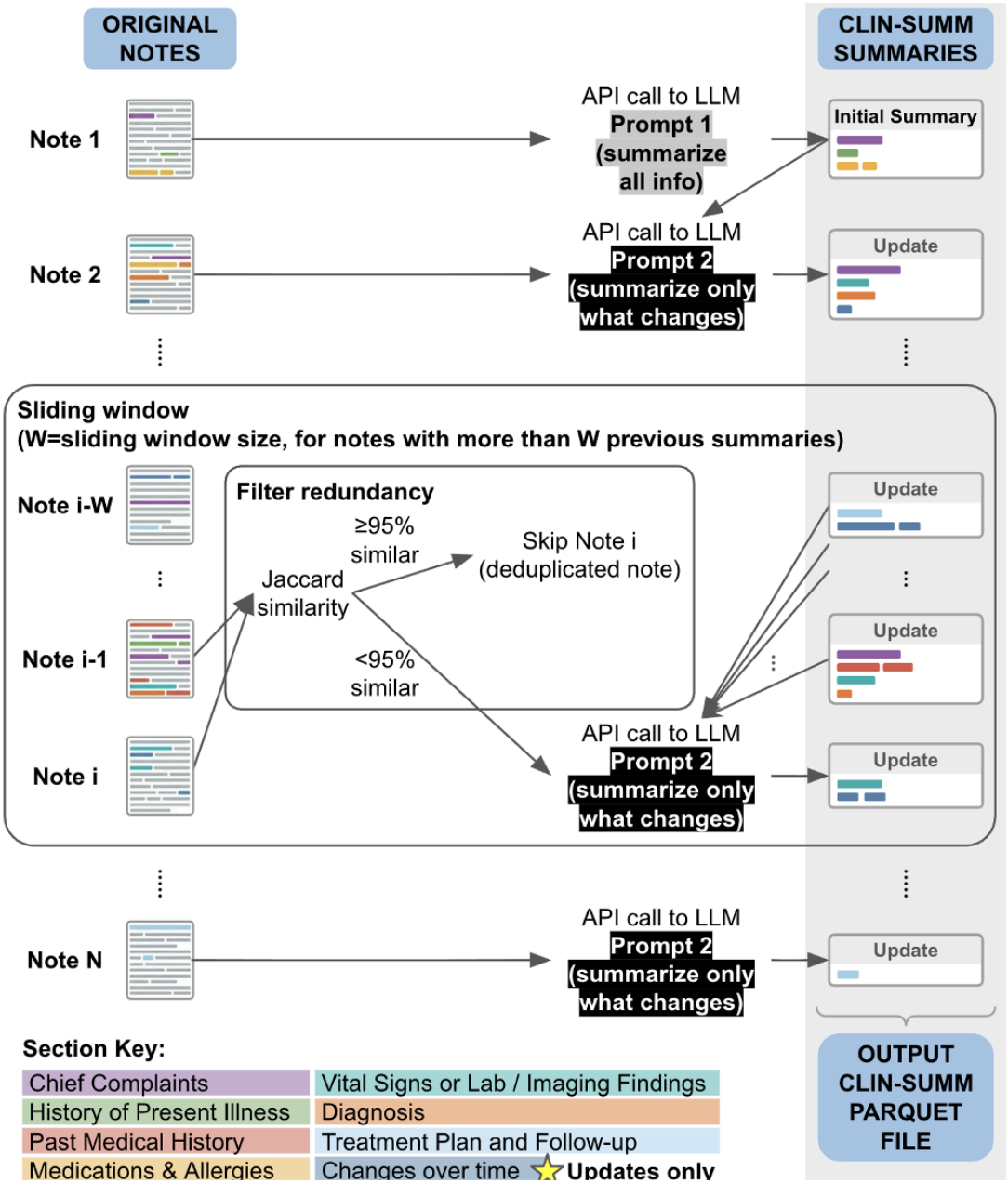


Figure 2: Raw longitudinal clinical notes are converted into structured patient timelines through sequential LLM-based summarization. The first note is summarized to create an initial patient summary, while subsequent notes are processed to extract only newly documented information. Redundant notes are filtered via a Jaccard similarity threshold, and a sliding-window mechanism limits context length for long note histories. The resulting updates are appended to a structured CLIN-SUMM

representation, yielding a compressed, organized, date-partitioned longitudinal clinical summary for each patient.

We developed the methods, models and analyses in this study using clinical notes from individuals receiving longitudinal care at Massachusetts General Hospital (MGH) primary care practices between 2005-2023. Our cohort includes 6,178 dementia patients and 6,178 age- and sex-matched controls, with a total of 2,599,333 notes with an average of 15 years of follow-up per subject. Each subject was assigned a censoring date. For dementia patients, this was defined as 30 days prior to the recorded dementia diagnosis date. For matched controls, the censoring date was determined using the diagnosis age of the matched dementia case. The date at which the control reached that same age was calculated from the control's date of birth, and 30 days were subtracted to maintain a comparable pre-diagnosis window. After restricting notes to those recorded prior to the censoring dates, the dataset includes 1,974,755 notes with an average of 13 years of follow-up per subject. **Supplementary Table 1** describes the basic participant characteristics of the population. We used a pilot set of 1500 subjects (750 randomly selected dementia patients aged 50-75 at diagnosis and their corresponding 750 age- and sex-matched controls) for initial development (65% training, 25% validation, 10% testing). We subsequently used the entire cohort to evaluate the clinical utility of CLIN-SUMM summaries (21% training, 6.5% validation and 72.5% testing, by combining 5% internal and 67.5% external test sets for evaluation).

2.2 CLIN-SUMM enables significant compression

CLIN-SUMM summarization substantially compressed the clinical narratives in the full cohort, reducing 2,717,908,027 words (3,623,877,369 tokens) to 819,060,671 words (1,092,080,895 tokens), yielding an overall space savings of 69.86% (space savings is defined as the relative reduction in word count between the original clinical notes and their corresponding summaries). Analyzing compression at the patient level (which reduces the impact of patients with

disproportionately numerous or lengthy notes) yielded a space savings of 52.6 ± 31.2%, indicating consistent compression across patients.

Early notes typically showed lower compression due to higher content density, while later notes were more highly compressed as CLIN-SUMM removed repeated information (**Supplementary Figure 2**). Compression was further evaluated at the patient level by analyzing the distribution of space-saving percentages. To account for multiple notes per day, notes were aggregated at the day level before computing average space savings per individual. While most patients showed positive compression, a small subset exhibited negative compression, more commonly in cases with very short notes (<200 words) or minimal new information. In these instances, the summarization process sometimes erroneously incorporated information from prior summaries, leading to longer outputs. These cases were often associated with low information note types (e.g., phone calls, emails, addenda), and 56% of notes in this group were under 200 words, suggesting limited information content as a key driver (**Supplementary Figure 3**).

2.3 CLIN-SUMM summaries achieve high correctness and completeness scores

A clinician-led review of CLIN-SUMM summaries from 30 patients (698 notes) sampled from the pilot set (n = 1,500, described above) using a structured questionnaire (**Supplementary Table 2**) revealed high scores for correctness and completeness and low hallucination rates. An independent evaluation by a second clinician of summaries from 10 patients (255 notes) revealed high inter-rater agreement (**Supplementary Table 3**).

Correctness scores (assessing alignment of the summary with the key points in the original notes without factual errors or misinterpretations) averaged 4.69 ± 1.02 out of 5 with an inter-rater agreement of 91.30% [95% CI 87.70 - 94.8], with Cohen's $\boldsymbol{K}$ = 0.39. Completeness scores (assessing the level of information coverage in the summary compared to the original notes) averaged 4.65 ± 1.02 out of 5 with an inter-rater agreement of 91.90% [95% CI 88.5 - 95.3], with Cohen's $\boldsymbol{K}$ = 0.47.

Fewer than 4% of summaries exhibited hallucinations. Qualitatively, these mostly involved (1) inferring diagnoses that were not explicitly documented based on presenting symptoms, (2) misinterpreting vitals or laboratory values due to inconsistent formatting, and (3) incorrectly attributing lack of treatment to unavailability. Representative examples are provided in **Supplementary Section 1 (Error Analysis of CLIN-SUMM Summaries)**. These error types reflect known limitations of current LLMs and are likely to diminish as models continue to improve in reasoning, factual consistency, and clinical grounding.

2.4 CLIN-SUMM summaries can form the basis for machine-learning modeling for early disease screening

2.4.1 Dementia diagnosis and 3-year prediction

The rich textual data contained in patients' historical clinical notes is summarized in the CLIN-SUMM longitudinal dataset and can be leveraged for machine learning. We demonstrated this by training transformer models for dementia diagnosis and 3-year prediction that input CLIN-SUMM summaries (and not the original, much longer clinical notes). The two models are based on Clinical ModernBERT [19], and are predictive, encoder-only models for binary classification of dementia cases versus controls. Note that the controls were not healthy: many had multiple comorbidities and overlapping clinical risk factors, making the task of distinguishing dementia cases versus controls more challenging.

The diagnosis model aims to improve dementia case identification by complementing International Classification of Diseases (ICD) and Systematized Nomenclature of Medicine (SNOMED)-based definitions, using notes to detect individuals who may not have formal codes. The 3-year prediction model aims to identify patients at high risk of developing dementia, supporting opportunities for early, proactive intervention through preventative measures, lifestyle changes, and therapeutic strategies, potentially transforming how chronic diseases are managed.

The diagnosis model was trained using all of each patient's CLIN-SUMM summaries up to 30 days before formal dementia diagnosis, and the 3-year prediction model was trained using all of each patient's CLIN-SUMM summaries up to 3 years before formal dementia diagnosis (**Figure 3A**). For controls, we used 30 days or 3 years before the censoring date, respectively. For very long summary histories, we truncated the input to the most recent part, corresponding to the period closest to the diagnosis date (or censoring date for controls), that could fit within Clinical ModernBERT's context window (8,192 tokens). For the diagnosis model, this design excludes periods likely to contain explicit diagnostic determinations (e.g., confirmed dementia diagnoses), while still retaining earlier, less definitive references (e.g., concerns or evaluations for cognitive decline), thereby encouraging the model to rely on dementia-consistent clinical language patterns rather than direct statements of disease presence. In contrast, the 3-year prediction model aims to capture subtle early signals that might signal disease onset well before symptoms emerge.

Clinical ModernBERT models using CLIN-SUMM summaries to identify dementia cases under the diagnosis and 3-year prediction scenarios consistently outperformed multifactorial dementia risk models [20] derived from the CAIDE dementia risk model [21] (**Figures 3B and 3C**, with detailed results in **Supplementary Table 4**). For the dementia diagnosis task, performance was strong (Area Under the Receiver Operating Characteristic (AUROC) 0.86, Area under the Precision-Recall Curve (AUPRC) 0.86, F1 score 0.77, precision 0.78, recall 0.76), reflecting high discrimination and a balanced trade-off between sensitivity and specificity. As expected, the dementia 3-year prediction model had slightly lower performance (given that the indicators of future dementia diagnosis are less specific the further one goes back in time and may be prone to misclassification) but retained substantial discriminatory power (AUROC 0.81, AUPRC 0.78, F1 score 0.76, precision 0.73, recall 0.79). For context, logistic baselines incorporating age, sex, prevalent comorbidities and vitals (CAIDE dementia risk model), showed more limited predictive ability in the subset of test subjects with vitals (diagnosis: AUROC 0.67 vs. 0.85; 3-year prediction: AUROC 0.63 vs. 0.79), reflecting the insufficiency of structured risk factors alone in this setting.

These limitations may in part be attributable to incomplete capture of clinical nuance (e.g., symptom progression, cognitive complaints, and clinician assessments), variability in data availability (e.g., missing or irregularly recorded vitals), and constraints in representing longitudinal context using structured features (e.g., temporal ordering and evolution of clinical events).

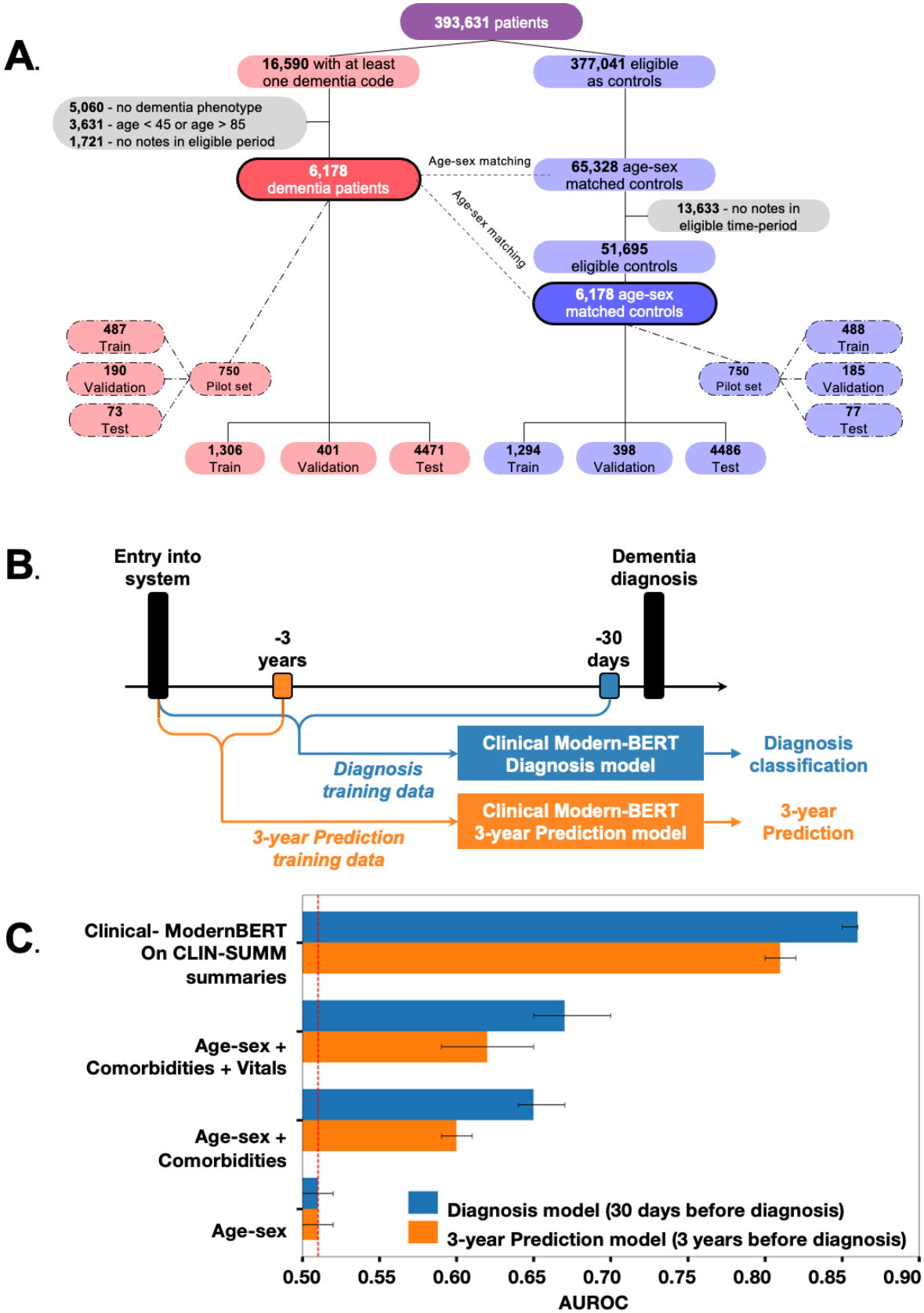

A.
393,631 patients
16,590 with at least one dementia code
377,041 eligible as controls
5,060 - no dementia phenotype
3,631 - age < 45 or age > 85
1,721 - no notes in eligible period
6,178 dementia patients
Age-sex matching
65,328 age-sex matched controls
13,633 - no notes in eligible time-period
Age-sex matching
51,695 eligible controls
6,178 age-sex matched controls
487 Train
190 Validation
73 Test
750 Pilot set
750 Pilot set
488 Train
185 Validation
77 Test
1,306 Train
401 Validation
4471 Test
1,294 Train
398 Validation
4486 Test
B.
Entry into system
Dementia diagnosis
-3 years
-30 days
Clinical Modern-BERT Diagnosis model
Diagnosis classification
Diagnosis training data
Clinical Modern-BERT 3-year Prediction model
3-year Prediction
3-year Prediction training data
C.
Clinical- ModernBERT On CLIN-SUMM summaries
Age-sex + Comorbidities + Vitals
Age-sex + Comorbidities
Age-sex
Diagnosis model (30 days before diagnosis)
3-year Prediction model (3 years before diagnosis)
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
AUROC

Figure 3: Overview of study design and modeling framework for dementia diagnosis and 3-year prediction. (A) Cohort construction and study flow diagram, including age-sex matching between dementia cases and controls. **(B)** Temporal framing of the prediction tasks, with diagnosis prediction evaluated 30 days before diagnosis and 3-year prediction evaluated three years prior to diagnosis. **(C)** Comparison of model performance across approaches. Clinical ModernBERT trained on CLIN-SUMM summaries achieves the highest performance for both diagnosis and 3-year prediction tasks, substantially outperforming traditional feature-based models. Bars show AUROC, with the dashed red line indicating the age-sex baseline.

2.4.2 Early detection time horizon

To evaluate the model's clinical utility and early-detection capabilities, we conducted a longitudinal lead-time analysis. We applied the frozen diagnosis model (originally trained using all CLIN-SUMM summaries up to 30 days before diagnosis) to incrementally expanded temporal windows, simulating deployment in a real clinical scenario in which a risk assessment model is repeatedly applied as a patient's clinical trajectory evolves (**Figure 4A**). Detailed longitudinal evaluation procedures are provided in **Supplementary Section 2 (Early detection time horizon).**

When applied to test sets right-truncated at progressively earlier time points (-5 years, -3 years, -1 year, -6 months before diagnosis), the diagnosis model generalizes to an early prediction tool (**Figure 4B**, with detailed results in **Supplementary Table 5)**. The model maintained meaningful discriminatory power even years before diagnosis, with an AUROC of 0.84 at -6 months and 0.82 at -1 year. Notably, the 30-day diagnosis model achieved an AUROC of 0.79 on data truncated at -3 years, closely matching the 3-year prediction model (AUROC 0.81) despite significant differences in the training data.

Longitudinal patterns in mean predicted dementia risk, computed over progressively longer CLIN-SUMM summaries in one-year increments, revealed significant differences in the dementia vs control groups and suggest that model predictions are grounded in clinically meaningful cues that evolve over time (**Figure 4C**). Many years prior to dementia diagnosis (or for controls, the censor date), the dementia and control groups had similar mean predicted dementia risk, both of which were below 0.5. The mean dementia risk steadily increased in the dementia group as the clinical histories became progressively richer and the formal dementia diagnosis dates neared, while the mean dementia risk in the control group was stable or decreasing. A visible separation between the distributions of predicted dementia risk for the dementia versus controls groups is clear between 3 to 8 years before diagnosis, consistent with our previous findings of high AUROC for binary classification. Plotting mean dementia risk on a two-dimensional grid as a function of

both the number of years prior to endpoint and subject age at the time (**Supplementary Figure 4**) reveals that dementia risk generally increases with age so long as the endpoint is sufficiently near (e.g., a maximum of 10 years away), but that differences in mean predicted dementia risk between the two groups are highest near clinical endpoints regardless of age.

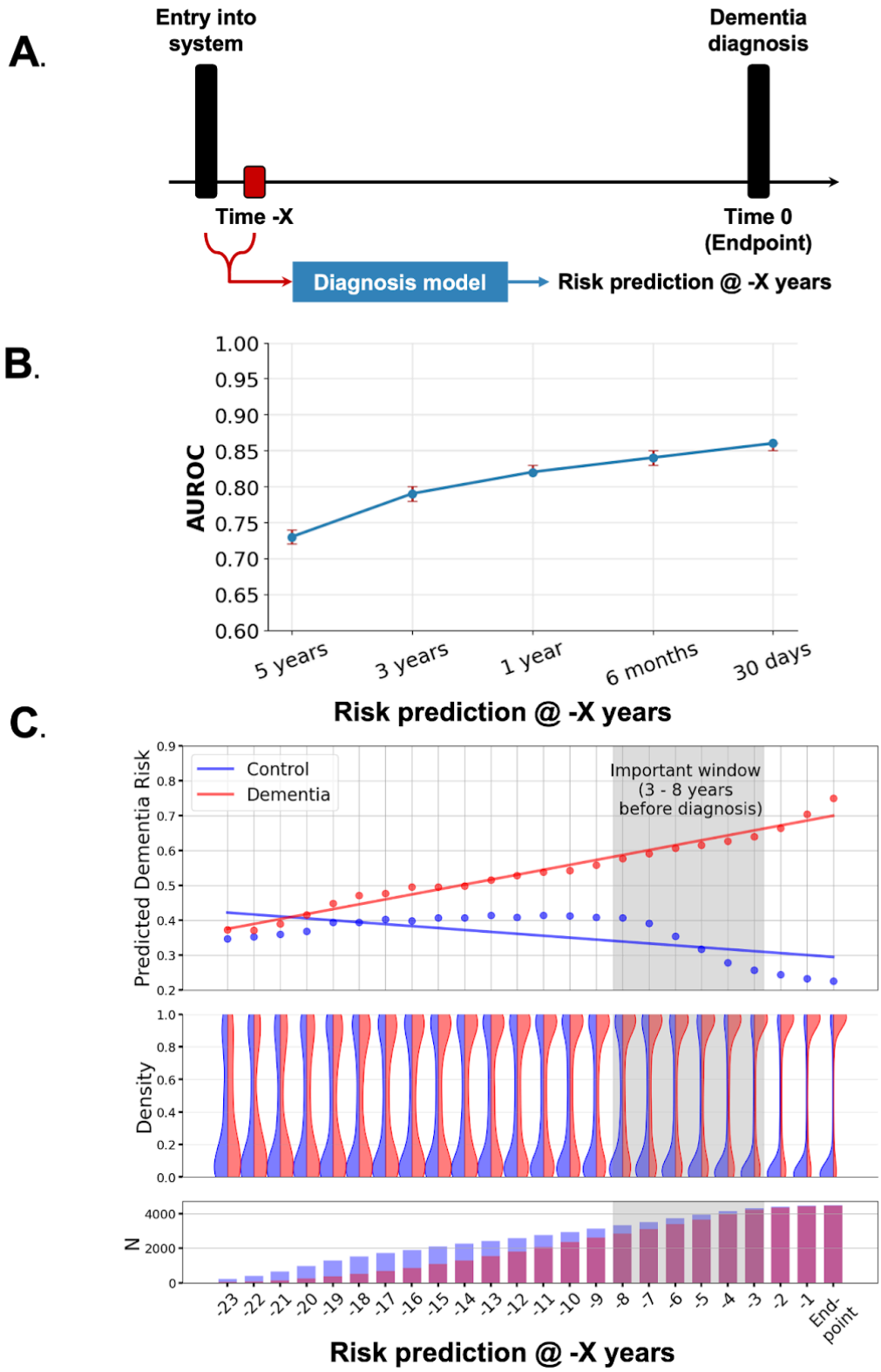

A.
Entry into system
Dementia diagnosis
Time -X
Time 0 (Endpoint)
Diagnosis model
Risk prediction @ -X years
B.
AUROC
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
5 years
3 years
1 year
6 months
30 days
Risk prediction @ -X years
C.
Predicted Dementia Risk
Control
Dementia
Important window (3 - 8 years before diagnosis)
Density
N
-23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 End-point
Risk prediction @ -X years

Figure 4: Longitudinal dementia risk prediction across time prior to diagnosis. (A) Temporal prediction framework illustrates how the diagnosis model estimates dementia risk at different time horizons before the diagnostic endpoint (Time 0). **(B)** Diagnosis model performance across prediction windows, showing AUROC for risk prediction evaluated at 5 years, 3 years, 1 year, 6 months, and 30 days prior to diagnosis, with performance increasing as predictions approach the diagnosis date. **(C)** Predicted dementia risk across time windows for dementia cases and controls. Mean predicted risk diverges between dementia cases and controls as the endpoint approaches, with a pronounced divergence **3–8** years before diagnosis. The middle panel shows the distribution of predicted risks for both groups across time, and the bottom panel shows the number of samples available at each prediction window.

2.4.3 Interpretability

To understand which linguistic features drive model predictions, we analyzed word-level attribution scores derived from LIME [22] explanations of the fine-tuned Clinical ModernBERT dementia diagnosis model (**Figure 5**). Across semantic categories, words associated with cognitive impairment and neurological symptoms were among the words with the highest positive attribution scores, indicating strong positive contributions to dementia predictions. These included terms such as ‘cognitive’, ‘memory’, ‘impairment’, ‘confusion’, ‘decline’, ‘deficits’ and ‘neurological’ [23,24]. Additional promoting signals included functional or symptom descriptors such as ‘gait’, ‘fall’, ‘dizziness’ and ‘difficulty’, which are commonly documented in patients with neurocognitive disorders. Given that LIME operates on whitespace-tokenized words, certain multi-word clinical concepts (e.g., ‘hearing loss’) may appear as separate tokens (e.g., ‘hearing’ and ‘loss’, both of which were identified as promoting words). In exploratory analyses, co-occurrence patterns identified via simple regex-based matching showed that ‘hearing’ and ‘loss’ appeared together within the same context in 20 of 36 individuals (55.6%) among those in whom either term was identified. This suggests that LIME attributions may reflect clinically meaningful concepts, such as hearing loss, a known risk factor for cognitive decline and dementia [25]. Additional words associated with established risk factors and comorbidities, including hyperglycemia, depression and hypertension, also showed positive attribution, consistent with their links to increased dementia risk [25-27]. In contrast, words with negative attribution scores, contributing toward the control class, were largely associated with routine clinical documentation, stable clinical status and general health descriptions. These included terms such as ‘appointment’, ‘follow-up’, ‘visit’, ‘annual’, ‘well’ and ‘mild’. Overall, the salience patterns demonstrate that the model prioritizes clinically interpretable signals consistent with known manifestations of dementia, including cognitive impairment, neurological symptoms and functional decline. These findings suggest that

the model's predictions are driven by medically relevant language rather than spurious textual artifacts.

← Protecting (lower risk)
Promoting (higher risk) →
Lifestyle/ General Health
physical
exercise
oxygen
food
medical
risk
undergoing
recurrence
hearing
concerns
illness
advised
resolved
Clinical Procedures/ Tests
colonoscopy
panel
testing
imaging
screening
blood
cuff
exam
lipid
performed
treatment
evaluation
labimaging
Clinical Documentation
annual
history
identified
visit
presented
management
followup
patient
dr
note
findings
documented
reports
chief
referral
appointment
Body Parts/ Anatomy
mouth
prostate
prostatic
knee
eye
urinary
abdominal
left
ophthalmic
bladder
extremities
right
Demographics
male
yearold
female
postmenopausal
family
social
Conditions/ Diagnoses
gastroesophageal
cancer
allergy
obesity
allergies
disease
obstructive
hyperglycemia
dementia
disorder
depression
osteopenia
hypothyroidism
hypertension
hyperlipidemia
allergies
incontinence
Symptoms/ Signs
pain
swelling
cough
rash
breath
tenderness
dysphagia
decline
dizziness
loss
fall
bruits
gait
difficulty
confusion
complaints
Neurocognitive
cognitive
memory
oriented
impairment
cerebral
pupils
speech
neurological
deficits
cognition
Drugs/ Medications
tablet
calcium
metformin
omeprazole
fluticasone
cream
atorvastatin
propionate
medications
cholecalciferol
fluoxetine
metoprolol
warfarin
aspirin
discontinued
medication
tylenol
tab
Descriptors/ Qualifiers
vital
chronic
new
newly
soft
essential
currently
current
lower
mild
poor
clear
possible
slightly
acute
mild
continued
recent
previous
well
Word size ∝ attribution score
Word color intensity ∝ prevalence
-0.10 -0.08 -0.06 -0.04 -0.02 0.00 0.003 0.006 0.009 0.012 0.015
Mean attribution score

Figure 5: Word-level attribution scores derived from **LIME** highlight influential features contributing to dementia risk in the diagnosis model. Word size and position reflect attribution magnitude, and color intensity reflects prevalence in the dataset. Blue terms correspond to features associated with lower predicted risk, while red terms indicate features associated with higher predicted risk. **Neurocognitive** terms and **clinical diagnoses** show strong positive attribution, whereas general health and procedural terms tend to show negative attribution.

2.4.4 Additional analyses of model behavior

Additional analyses examining model behavior, including error patterns and evaluation of predictions in relation to mild cognitive impairment (MCI), are provided in **Supplementary Section 3 (Additional analyses of model behavior) and Supplementary Figure 5**. Overall, misclassifications were primarily associated with ambiguous or incomplete clinical context, while analyses of MCI-related subgroups showed graded risk patterns consistent with disease progression.

2.5 CLIN-SUMM summaries augment medication extraction

As medication information is consolidated into a dedicated section for each patient summary, efficient extraction is possible by querying a single structured segment of the longitudinal record, eliminating the need to scan thousands of raw clinical notes. To assess the completeness of medication extraction from CLIN-SUMM summaries, we compared LLM-extracted Donepezil medication events with medication records available in structured EHR tables (Snowflake) (**Figure 6A**). This comparison served both as a validation step and as a test of the hypothesis that summary-based extraction can reduce apparent missingness in structured data sources. CLIN-SUMM extraction identified substantially more Donepezil medication instances than structured tables alone (2,725 vs. 1,312 in the dementia cohort). While many events were detected by both sources, a large number were identified exclusively in summaries, suggesting that clinically relevant medication information often appears in clinical notes but was absent or fragmented in the retrieved structured medication-containing tables. Discrepancies where events appeared only in structured tables were relatively few (n=73) and typically occurred during time periods with missing clinical notes, limiting their presence in CLIN-SUMM summaries. To further evaluate concordance between sources, we matched each CLIN-SUMM-extracted medication record to the closest corresponding entry in the structured EHR data based on ordering date.

Dosage agreement between CLIN-SUMM extractions and structured records was high, with 96.2% events showing exact agreement. Discrepancies were infrequent and largely explained by clinically interpretable factors, including dose titration (2.3%), where notes captured gradual dose escalation while structured records reflected the final prescribed dose, and split or fractional dosing (0.4%), where notes documented detailed administration patterns (e.g., ¼ tablet taken twice daily), while structured records capture only the prescribed medication without encoding these instructions. In 0.5% of cases, the CLIN-SUMM extracted dose aligned with the structured 'description field', while the structured 'dose quantity' field was inaccurate, highlighting inconsistencies within structured medication data. True mismatches were observed in 0.6% of cases, although the raw clinical notes supported the CLIN-SUMM extracted dose, indicating inconsistencies in structured records. Together, these results suggest that CLIN-SUMM summaries may improve medication capture while maintaining strong agreement with the clinically administered dosage.

Medication trajectories reconstructed from CLIN-SUMM summaries revealed clinically meaningful treatment patterns for dementia-related medications. For example, Donepezil use increased substantially in the years preceding diagnosis among dementia patients, with the number of individuals receiving the medication rising sharply near their diagnosis dates (**Figure 6B**). In contrast, Donepezil use remained rare among controls across the observation period. This likely reflects prescriptions for suspected cognitive impairment or early memory complaints that did not lead to a recorded dementia diagnosis during the study period, or individuals who may have received a diagnosis after the study cutoff date, or lagging documentation practices following diagnosis of dementia. Dosage distributions (**Figure 6B**) also showed increasing variability closer to diagnosis in the dementia cohort, consistent with treatment initiation and adjustment during disease progression, while trajectories for individual patients show trends over time (**Figure 6C**).

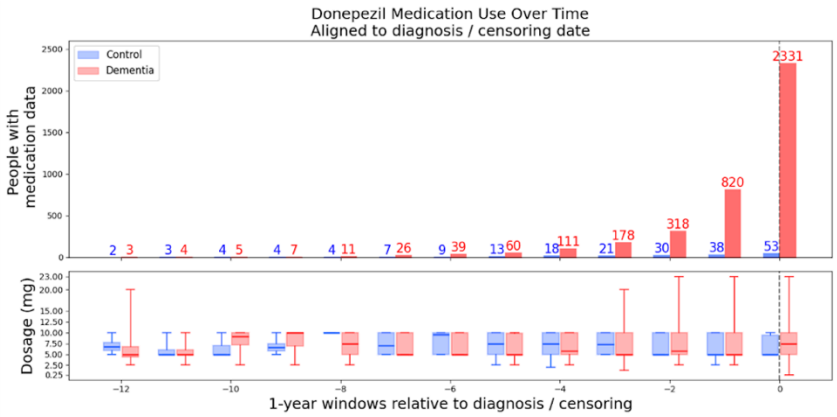
Together, these results demonstrate that CLIN-SUMM summaries can serve as an important source for reconstructing longitudinal medication trajectories and may help mitigate missingness,

inconsistencies, fragmentation or lag in structured EHR medication data with reduced computational cost compared to processing the original notes.

A.

	Controls	Dementia
Extraction counts		
Structured EHR (Snowflake)	35	1312
CLIN-SUMM	59	2725
Agreement between sources		
Found in both	30	1243
Only in Structured EHR	5	69
Only in CLIN-SUMM	29	1482

B.



C.

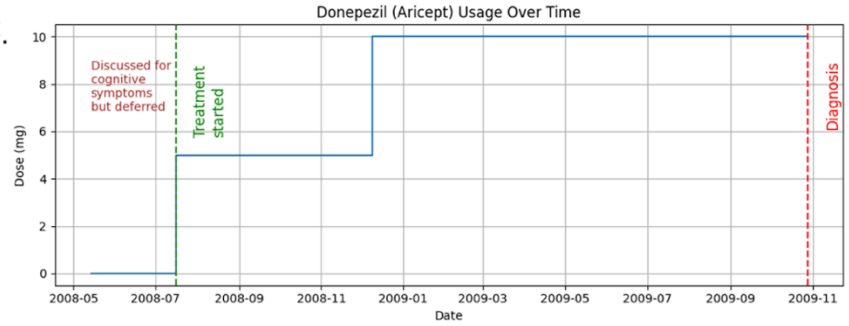


Figure 6: Medication extraction and longitudinal treatment patterns identified from CLIN-SUMM summaries. (A) Comparison of medication extraction counts from structured EHR data and CLIN-SUMM summaries, showing increased detection of medications from summarized clinical notes. **(B)** Population-level trends in **Donepezil** use across dementia patients and controls, with

corresponding dosage, aligned relative to dementia diagnosis or censoring date. **(C)** Example patient trajectory illustrating medication initiation and dose changes over time.

2.6 Evaluation of a Commercial Frontier LLM

Although CLIN-SUMM is designed to operate with any LLM, and Qwen3 [15] was ultimately used for summarization in this study, we explored a state-of-the-art proprietary LLM as a reference point given that such models are widely regarded as benchmarks for high-quality text generation. Specifically, we generated CLIN-SUMM summaries using the GPT-4o [28] LLM in the pilot set. We then evaluated whether comparable performance could be achieved with Qwen3 to support cost-efficient and scalable deployment across the full cohort. Qwen3 produced stable and faithful clinical summaries, with a mean BERTScore recall of 0.91 between Qwen- and GPT-generated text (**Supplementary Table 6)**. To assess the downstream impact of summarization differences, we trained Clinical ModernBERT dementia diagnosis and 3-year prediction models separately on GPT-4o- and Qwen3-generated summaries. Performance was comparable across models, with slightly higher AUROC for models trained on GPT-4o summaries compared to Qwen3 (diagnosis: AUROC 0.81 versus 0.79; 3-year prediction: AUROC 0.75 vs. 0.72; **Supplementary Table 7,** with no statistically significant differences observed (DeLong tests: p-value = 0.37 for diagnosis; p-value = 0.33 for 3-year prediction)). Other metrics, including AUPRC and F1, showed similarly small differences. These findings demonstrate that open-source alternatives such as Qwen3 can achieve near-parity with proprietary LLMs for generating useful longitudinal summaries of clinical notes, while substantially reducing computational and financial costs (e.g., $250 with GPT vs. $35 with Qwen3 for summarizing 100 individuals with an average of 100 notes each), and mitigating key barriers to deployment, including limited scalability and restricted accessibility associated with application programming interface (API)-based models. Moreover, since Qwen3 can be deployed entirely within a secure, institution-controlled environment, it eliminates data-transfer risks and

enhances compliance with privacy regulations, including HIPAA (Health Insurance Portability and Accountability Act), an essential requirement for clinical translation.

2.7 Additional Applications

Additional pilot applications illustrating the versatility of CLIN-SUMM summaries, including applications in disease progression analysis and longitudinal symptom trend characterization, are presented in **Supplementary Section 4 (Additional Applications of CLIN-SUMM Summaries)**.

3 Discussion

In this study, we introduce CLIN-SUMM, a longitudinal summarization framework designed to condense clinical notes and create a reusable representation of narrative EHR data that preserves temporal evolution while reducing redundancy. Rather than repeatedly processing raw, unstructured notes for each downstream task, CLIN-SUMM produces a structured, date-partitioned representation that captures incremental clinical updates over time. This representation retains the nuance, versatility, and comprehensibility of raw clinical text while overcoming its inherent redundancy and lack of organization, and can be incrementally updated as new encounters occur, avoiding the need to regenerate summaries from scratch in clinical workflows. In doing so, CLIN-SUMM defines a new intermediate layer between raw clinical text and downstream applications, offering a more tractable and scalable way to reason over longitudinal patient trajectories. Compared to existing approaches, this representation offers distinct advantages over structured codes [1–3] (which often suffer from missingness and coding errors, and lack nuance), original notes [13,14] (which are lengthy, repetitive and unorganized), hierarchical textual representations [29] (which are derived from notes and structured data and retain many of their problems) and conventional clinical summaries [8–11] (which typically lack an explicit temporal dimension, although recent research has highlighted the benefits of summarizing on a daily basis [30]).

From a clinical perspective, this representation enables a more efficient and interpretable view of patient history and how it may evolve over time. By isolating only, the “new” clinical signal at each encounter, CLIN-SUMM reduces the cognitive burden associated with manual chart review and allows clinicians to rapidly grasp longitudinal changes. The structured, temporally organized summaries support tracking of evolving problems and medications, facilitate near-real-time maintenance of patient summaries, and provide a coherent view of disease progression. High clinician-rated scores for correctness and completeness, alongside a low hallucination rate, suggest that these summaries can serve as a reliable cognitive scaffold in clinical workflows, for example by providing an update of relevant clinical activity since a patient was last seen by a clinician. Importantly, CLIN-SUMM can be implemented using open-source models in a local, privacy-preserving setting, making it feasible for institutional deployment. The medication extraction analysis further highlights its practical value: while some medication events overlap with structured EHR tables, many are captured only within note-derived summaries, indicating that CLIN-SUMM complements and augments structured data sources.

Beyond clinical applications, CLIN-SUMM introduces a new paradigm for machine learning on clinical text by decoupling raw note processing from prediction. As an application-agnostic representation, CLIN-SUMM summaries can be generated once at the system level and reused across multiple downstream tasks, eliminating the need to repeatedly process raw notes and improving computational efficiency. For example, the same set of summaries could support models for predicting or diagnosing a range of conditions, such as diabetes, cardiovascular disease, stroke, and dementia, without requiring task-specific reprocessing of the underlying clinical notes. The substantial reduction in token count further enables more scalable training and inference while preserving clinically meaningful signals. Given the heterogeneity of raw clinical text and variability in documentation practices, this abstraction layer represents a key step toward robust and generalizable machine learning systems. In this sense, the primary contribution of CLIN-SUMM extends beyond summarization to establishing a reusable infrastructure for

longitudinal clinical reasoning and learning. Compression with clinical fidelity is not merely a technical metric, but a bridge to deployment: reducing information burden while preserving clinical accuracy makes CLIN-SUMM both usable and trustworthy in real-world settings. Moreover, by leveraging compressed representations instead of raw clinical notes, CLIN-SUMM reduces computational and financial costs, an important consideration for deployment at scale across large healthcare systems.

We evaluated this framework using diagnosis and future risk prediction of dementia as a proof-of-value application, motivated by the inherent difficulty of these predictive tasks due to the long prodromal phase of this disease and reliance on subtle, distributed signals in clinical notes. In our matched cohort, models based solely on demographic features (age and sex) performed near chance, indicating that signal from demographic factors alone is limited. Despite achieving substantial compression, CLIN-SUMM summaries supported robust classification performance for diagnosis and enabled meaningful future risk prediction up to three years in advance. This predictive performance compares favorably to established multifactorial models like CAIDE [21] and ANU-ADRI [31], yet it operates directly on existing notes without requiring structured variables that are often missing. These findings demonstrate that compression is not destructive: CLIN-SUMM preserves clinically relevant signals while removing redundancy, enabling effective prediction without reliance on raw notes or structured-only approaches.

Despite being trained on notes within 30 days of diagnosis, our dementia diagnosis model accurately distinguished dementia patients from controls up to 3–8 years in advance. Specifically, the model assigned approximately the same predicted dementia risk to dementia cases and controls many years prior to diagnosis (as early records likely contain fewer explicit disease-related signals and less frequent or subtler cognitive, neurological or behavioral indicators), and a steady separation between the two groups over time (as later notes capture evolving symptoms and correlated diagnoses). This capability is particularly vital given that dementia is insidious in onset, and formal ICD codes are often either never recorded or lag

significantly. Indeed, in our cohort dementia medications were sometimes prescribed up to two years before a formal code appeared, yet the model accurately distinguished dementia patients from controls well before this lead-time. Importantly, intermediate-risk patterns observed in patients with MCI further support that the model is capturing a graded, longitudinal disease trajectory rather than simply detecting the presence or absence of explicit diagnostic factors.

Interpretability analyses suggest that the model is learning a clinically coherent signature of dementia rather than relying solely on heuristic cues or explicit diagnostic language. Notably, CLIN-SUMM appears to preserve clinically relevant information in a way that allows these signals to be effectively leveraged by the model. Beyond expected neurocognitive terms such as *memory* and *cognitive*, the model assigns importance to hearing-related language, dizziness, gait and fall-related terms, mood changes, depression, and vascular or metabolic comorbidities, aligning with prior literature [25-27]. These signals are best interpreted as evidence of clinical face validity rather than causal relationships, and they support the conclusion that CLIN-SUMM captures a distributed and realistic prodromal signature embedded in routine clinical documentation.

Error analysis further strengthens confidence in the model by revealing systematic and clinically interpretable failure modes. False positives are often observed in patients whose records contain dense cognitive-related vocabulary, while false negatives are more common in cases with sparse documentation or fragmented care across specialties. Importantly, discrepancies between model predictions and ICD labels often suggest potential label noise and delayed coding. In many such cases, model predictions appeared clinically plausible, supporting the reliability of the model in capturing meaningful clinical patterns.

While dementia serves as a compelling use case, the utility of CLIN-SUMM extends well beyond a single disease prediction task. The framework facilitates reconstruction of medication trajectories, tracking of symptom evolution, and summarization of disease progression over time. These capabilities support a broad range of applications, including longitudinal phenotyping and clinical decision support. We anticipate that CLIN-SUMM will be particularly valuable for conditions

characterized by gradual and complex trajectories, such as autoimmune diseases, chronic kidney disease, and oncology treatment response, where relevant signals accumulate over time. This positions CLIN-SUMM as a general-purpose infrastructure rather than a task-specific solution.

CLIN-SUMM achieves strong performance using an open-source model (Qwen3 [15]), with results comparable to those obtained using proprietary frontier models. This demonstrates that the framework's utility is not dependent on closed, high-cost systems and supports its feasibility for local, privacy-preserving deployment. By enabling on-premise inference and reducing reliance on external APIs, CLIN-SUMM aligns with institutional requirements for data security and governance, further reinforcing its practicality and scalability.

Several limitations warrant consideration. First, although CLIN-SUMM demonstrated low hallucination rates, even rare factual errors carry clinical risk, and the sequential summarization framework may propagate or amplify errors over time. While prompt-level safeguards aim to constrain outputs to verifiable information, future work will focus on strengthening reliability through systematic error analysis, incorporation of source attribution to link summaries to original notes, and additional validation and monitoring strategies to ensure safe clinical integration. Technical refinements, such as optimized redundancy filtering, improved windowing strategies, and the pre-pruning of uninformative notes, could further enhance both summary accuracy and computational efficiency. Also, it is worth noting that the accuracy of the generated summaries and the cost required to compute them will improve alongside new developments in LLM technology. Second, even though training was restricted to notes prior to 30 days before diagnosis, potential prediagnostic signal leakage cannot be fully excluded. However, our longitudinal analyses and findings related to MCI suggest that model predictions are not solely driven by explicit diagnostic terms, supporting the presence of meaningful prediagnostic patterns. Third, direct comparisons with models trained on raw clinical notes are needed to isolate the incremental value of summarization. Models trained on raw notes may be sensitive to variability in note length, style, and documentation practices; we hypothesize that CLIN-SUMM, by

standardizing note structure into guided sections, may mitigate this heterogeneity. However, this requires systematic evaluation. Fourth, clinician evaluation was conducted on a limited sample due to the time-intensive nature of manual review, which requires clinicians to read, interpret, and reconcile information across numerous lengthy, technically complex clinical notes per patient and to verify alignment with the generated summaries. Expanded validation, either through larger-scale clinician assessment or complementary evaluation using LLM-based approaches, is needed for more robust and comprehensive assessment. Finally, external validation across diverse healthcare systems is necessary to assess generalizability and robustness across diverse patient populations, clinical settings, and documentation practices. In addition, the clinical context in which models trained on CLIN-SUMM summaries are applied may influence performance; for example, applying these models to patients seen primarily in specialty care rather than longitudinal primary care may increase the risk of false negatives due to more limited or focused documentation of certain clinical conditions.

Despite these limitations, CLIN-SUMM provides a practical pathway for transforming longitudinal clinical notes into a reusable infrastructure layer for both clinical reasoning and machine learning. By substantially reducing narrative volume while preserving temporal context and clinically meaningful signal, CLIN-SUMM enables more efficient chart review and supports robust downstream prediction, including early disease risk stratification years prior to diagnosis. As an application-agnostic representation, it can be computed once and reused across tasks, offering a scalable approach to leveraging unstructured clinical data. These findings position CLIN-SUMM as a promising foundation for integrating longitudinal clinical narratives into real-world analytical and decision-support systems.

4 Methods

4.1 Study Cohort and Data

We developed the methods, models, and analyses in this study using EHR data from individuals receiving longitudinal care at MGB primary care practices. The study population was derived from the **Primary Care Longitudinal Cohort** (PCLC), which includes adult patients linked to the Massachusetts General Hospital (MGH) primary care network between 2005 and 2023. The PCLC retrospectively identifies patients seen in primary care within the prior three years and assigns them to a primary care provider (PCP) or practice based on rules incorporating patient age, time since last visit, and visit patterns. Although longitudinal clinical data were available as early as 1976, patient–PCP linkage was assessed beginning in 2005.

The original dataset comprises 393,631 patients and includes both structured and unstructured data. This includes patient demographics, diagnosis codes, procedure codes, medications, notes from health care encounters and annual ambulatory visit notes. Prior to downstream processing, all clinical notes were processed for Protected Health Information (PHI) minimization, including the removal or masking of identifiable information such as patient and provider names. As the machine learning models developed in this study focus on dementia diagnosis and 3-year prediction, we conducted a retrospective case-control study (**Figure 3A**). Subjects diagnosed with dementia were identified based on the presence of at least two relevant ICD or Systematized Nomenclature of Medicine (SNOMED) codes, which must have been recorded on different dates within a 12-month period (**Supplementary Tables 8 and 9**). In order to be included, subjects must have been aged 45-85 at their initial dementia diagnosis. To ensure a 3-year lead-time, we only included subjects who had notes documented prior to the 3-years window preceding the first dementia diagnosis code, yielding a final set of 6,178 dementia subjects. From this group, we randomly selected a pilot cohort of 750 subjects aged 50-75 years at their initial dementia diagnosis. An age-and sex-matched control group was selected from patients without any dementia-related codes, using birth dates, such that the sampling maintained a balanced 1:1 case-to-control distribution for both the pilot cohort and the entire cohort. In total, the dataset

included 1,605,524 ambulatory and visit notes from the dementia group (1,034,564 recorded prior to the dementia diagnosis date) and 993,809 ambulatory and visit notes from the control group (940,191 recorded prior to the censor date). The cohort spanned 870 primary care providers (PCPs) across 12,356 patients and included diverse note types and documentation templates (>100,000 templates, capturing variation across PCPs, nurses, and other care settings), reflecting substantial heterogeneity in clinical practice and documentation, even within a single-institution setting. Detailed demographic characteristics are described in **Supplementary Table 1**.

4.2 CLIN-SUMM Summarization

4.2.1 Summarization Workflow

CLIN-SUMM uses two distinct prompts to iteratively summarize longitudinal clinical notes (**Supplementary Figure 1**). The initial prompt generates a comprehensive summary of the patient's inaugural entry note into the system, with explicit sections encompassing "Chief Complaints", "History of Present Illness", "Past Medical History", "Medications & Allergies", "Vital signs or Lab/Imaging Findings", "Diagnosis", and "Treatment Plan and Follow-up". The second prompt explicitly incorporates the previously generated summaries and aims to expand the patient-level summary with new information from each subsequent note, focusing exclusively on summarizing only novel content not encompassed within all of the preceding note-level summaries, and appending a "Changes Over Time" section that summarizes key temporal updates across encounters, such as medication adjustments, new diagnoses, or changes in lab/imaging findings, to track the patient's evolving clinical state. Each summarized entry is timestamped using the original note's date and time, preserving the chronological order of events.

CLIN-SUMM output is stored as a structured Parquet file, linking each patient's notes with the note date and time, brief note-level descriptions (such as 'patient note', 'wellness annual', 'lab results', 'medication-management', 'consult' or ''neurology', provided by the EHR), and the

generated summaries. This format preserves longitudinal detail while enabling flexible aggregation of summaries by month, year, clinical event, or section.

4.2.2 Prompt Optimization and Parameter Selection

Prompts were iteratively refined following best practices [32], to produce structured, clinically faithful summaries. The CLIN-SUMM framework was implemented using the Qwen3 model [15] deployed via the vLLM inference framework [33]. The approach incorporates context management and redundancy control mechanisms (via Jaccard similarity evaluation [18]) to support stable longitudinal summarization. Full prompt templates, parameter settings (**Supplementary Table 10**), and implementation details are provided in the **Supplementary Section 5 (Prompt Optimization and Parameter Selection)**.

4.3 Compression Efficiency

We measured compression efficiency in 2,599,333 notes from the 12,356 patients in the full cohort. For each note, we compared the length of the original clinical text with the length of the corresponding CLIN-SUMM summary. Space savings is computed as the percent reduction in the number of words via

$$Space\ Savings = \left(1 - \frac{\#\ Words\ in\ Summary}{\#\ Words\ in\ Original\ Note}\right) * 100\%.$$

Compression was evaluated at multiple aggregation levels. First, overall space savings was computed across the entire cohort by comparing the total number of words in all original notes versus the generated summaries. Second, patient-level compression was computed by calculating space savings for each patient individually and summarizing the distribution across patients. Because multiple clinical notes may be generated on the same day, notes were aggregated at the

day level before evaluating the distribution of space savings. This aggregation reduces the influence of fragmented documentation patterns (e.g., multiple short addenda on the same date).

4.4 Clinical Evaluation of Summaries

We randomly selected 30 patients from the pilot set. A clinician (A.A.) evaluated all CLIN-SUMM summaries from these subjects, encompassing summaries from 698 notes, using a validation questionnaire. A second clinician (A.N.) independently assessed the summaries from 10 of these subjects, encompassing summaries from 255 notes.

The questionnaire comprised structured five-point scales (1=poor, 5=excellent) and yes/no questions to evaluate correctness, completeness, and hallucinations (**Supplementary Table 2**). Correctness verifies if the summary fully aligns with the key points in the notes without factual errors or misinterpretations. Completeness assesses the level of information coverage compared to the original notes and identifies any omissions. Finally, a yes/no question assesses whether the summary includes hallucinations or invented content not supported by the original notes.

4.5 Dementia Diagnosis and 3-year Prediction Models

4.5.1 Data splits

We randomly stratified the pilot cohort of 1,500 CLIN-SUMM patient-level CLIN-SUMM summaries at the patient level into training (N=975, 65%), validation (N=375, 25%), and test (N=150, 10%) sets. Splits of the pilot cohort were used for initial development and evaluation, after which subsequent experiments were performed using the entire population data. The full cohort was partitioned into 2,600 training cases (21%), 799 validation cases (6.5%), and 601 test cases (5%), with an additional 8,356 cases (67.5%) reserved as an external test set for further evaluation. The initial 4,000-case subset was therefore used to derive and test the scaled model. For the final evaluation, results were reported on the combined test cohort (N = 8,957), consisting of the 601 internal test cases and the 8,356 external cases (**Supplementary Table 4, 5**).

4.5.2 Data, architecture and training

We trained models for two binary classification tasks: development of a dementia diagnosis model for dementia case identification from notes, and a dementia 3-year prediction model for early clinical risk prediction. We fine-tuned the Clinical ModernBERT model [19], a transformer-based language model derived from BERT [34]. BERT is a bidirectional transformer-based language model that learns contextual representations of text through self-attention. Clinical ModernBERT is an updated transformer architecture optimized for efficiency and longer context lengths, and further pre-trained on clinical text to better capture medical terminology and documentation patterns. We trained the dementia diagnosis model using CLIN-SUMM summaries for all notes up to 30 days before the first formal dementia diagnosis (or for controls, 30 days before the censor date). We trained the dementia 3-year prediction model using CLIN-SUMM summaries for all notes up to 3 years before the first formal dementia diagnosis (or for controls, 3 years before the censor date). Notes were preprocessed by removing punctuation and special characters and the cleaned text was subsequently tokenized and converted to tensors. The Clinical ModernBERT model accepted sequences up to 8,192 tokens; for longer summaries, only the most recent summaries were used. Both the diagnosis and 3-year prediction models were trained using a binary cross-entropy loss with a learning rate of 5×10^{-7}, batch size of 4, and a weight decay of 0.02 for up to 60 epochs. A scheduler reduced the rate by 0.5 if validation loss plateaued for five epochs, and early stopping triggered after 15 stagnant epochs. The best-performing checkpoint was selected as the final model.

4.5.3 Age-sex baseline

As a reference model, we trained a logistic regression classifier using only age and sex as predictors. For the diagnosis and 3-year prediction models, age was defined as the patient's age 30 days or 3 years prior to the diagnosis/censor date, respectively. Age was standardized prior to

model training. Because the difference between 30 days and 3 years represents a relatively small-time shift, the standardized age values, and therefore the resulting model performance metrics, were nearly identical between the diagnosis and 3-year prediction models. This baseline therefore provides a minimal demographic benchmark for comparison with models incorporating additional clinical information.

4.5.4 Age-sex-comorbidities baseline

We next trained a logistic regression baseline incorporating age, sex, and prevalent comorbidities derived from ICD and SNOMED coded diagnoses. Comorbidities included anemia, arthritis, depression, delirium, osteoporosis, dizziness, syncope, hyperlipidemia, dyspnea, Parkinson's disease, hearing impairment, visual impairment, anorexia, anxiety, alcohol use disorder, hypertension, and history of falls (**Supplementary Table 11**). The selected conditions were based on established dementia risk factors and clinical variables included in previously published dementia risk scores, including the CAIDE [21] dementia risk model. Each condition was represented as a binary indicator denoting the presence of a corresponding condition code 30 days or 3 years prior to the diagnosis/censor date, respectively.

4.5.5 Age-sex-comorbidities-vitals baseline

To assess whether routinely collected physiologic measurements improve predictive performance, we extended the age-sex-comorbidities baseline to include vital signs and lab values commonly used in dementia risk models, including those included in the CAIDE dementia risk score. Specifically, we incorporated body mass index (BMI), total cholesterol, systolic blood pressure and diastolic blood pressure (**Supplementary Table 12**). For the diagnostic task, measurements were obtained from records within the period spanning 5 years to 30 days prior to the diagnosis/censor date, with the most recent available value used for each variable. For the 3-year prediction task, measurements were obtained from the period spanning 8 years to 3 years prior to the

diagnosis/censor date, again using the most recent available value. Because these variables were not available for all individuals, inclusion of vital signs and lab values resulted in a reduced sample size due to missing data.

4.5.6 Interpretability

To interpret the predictions of the fine-tuned Clinical ModernBERT dementia classification model, we applied Local Interpretable Model-agnostic Explanations (LIME) [22] to quantify the contribution of individual words toward model predictions. For each patient summary, LIME generates a set of perturbed versions of the original text by randomly masking subsets of words. The fine-tuned Clinical ModernBERT model is evaluated on each perturbed version to obtain predicted probabilities for the dementia class. LIME then fits a locally weighted linear surrogate model that approximates the behavior of the neural network around the original document.

$$f(z_j) \approx \beta_0 + \sum_{w=1}^{K} \beta_w z_{jw}$$

where z_j represents a perturbed document expressed as a binary word-presence vector, $z_{jw} = 1$ indicates that word w is present in perturbation j, and β_w represents the local attribution coefficient for word w. Perturbations closer to the original document receive higher weights using an exponential kernel based on cosine distance, ensuring the surrogate model approximates the neural network locally.

For each patient we retained the top 100 words selected by LIME according to their contribution magnitude. These coefficients quantify the direction and strength of each word's contribution to the model prediction. Positive coefficients indicate words that increase the predicted probability of dementia (risk-promoting words), while negative coefficients indicate words that decrease the predicted probability and support the control class (protective words).

To identify consistent linguistic patterns across individuals, word-level attribution coefficients were aggregated across patients. Interpretability analysis was conducted on 100 test cases, consisting of 50 high-risk individuals, randomly selected from the top 5% of predicted dementia risk (≥ 95th percentile), and 50 low-risk individuals, randomly selected from the bottom 5% of predicted risk (≤ 5th percentile). Selection was stratified by age and sex to ensure comparable demographic distributions between groups. For each word w, we computed the conditional attribution score, representing the average attribution coefficient across patients in whose LIME explanation the word was identified as an informative feature:

$$\mu_w^{cond} = \frac{1}{n_w} \sum_{i=1}^{n_w} \beta_{w,i}$$

where n_w denotes the number of patients in whose explanation word w appears, and $\beta_{w,i}$ represents the attribution coefficient of word w for patient i.

Word attribution results were visualized using a category-structured text plot in which words were arranged by semantic category and positioned along a horizontal axis according to their attribution score. Words positioned to the right of zero represent terms that increase the predicted probability of dementia (promoting words). Words positioned to the left of zero represent terms associated with the healthy class (protective words). Font size scales with the magnitude of the attribution coefficient, allowing visually prominent representation of strongly influential words. Words were grouped into predefined categories (Neurocognitive, Symptoms/Signs, Conditions/Diagnoses, Drugs/Medications, Clinical Documentation, Clinical Procedures/Tests, Descriptors/Qualifiers, Demographics, Body Parts/Anatomy, Lifestyle/General Health, Other, Dosage/Administration, Units/Measurements), with distinct colors assigned to each category. Intensity of color represents the prevalence of the word across the cohort. Within each category, words were ranked by the absolute value of their attribution score, and the 10 top-ranked terms

within each category were displayed for both positive and negative contributions. Categories such as Other, Dosage/Administration and Units/Measurements were excluded from the final visualization because the associated terms were largely non-informative and did not meaningfully contribute to model interpretation.

4.6 Medication Extraction

Medication usage events were extracted from the CLIN-SUMM patient summaries. Each patient summary contains a dedicated section describing medication history aggregated across clinical notes. Using these condensed summaries enables efficient downstream processing, as CLIN-SUMM summaries provide a compact representation of clinically relevant information while preserving temporal context across visits.

Medication extraction was performed using GPT-4o. For each patient summary, the model was prompted to identify medication usage events for a predefined set of dementia-related and neuropsychiatric medications relevant to our cohort: Donepezil, Galantamine, Memantine, Rivastigmine, Lecanemab, Donanemab, Suvorexant, and Brexpiprazole. Brand and generic names were treated as equivalent. The model was instructed to extract all medication events mentioned in the summary and return them in structured JSON format, including the event date (ISO format, when available) and the current dose in milligrams if specified. If a medication was discontinued, the dose was recorded as 0 mg on the corresponding date.

The following prompt was used:

> *Given the following summary of patient notes, extract all mentioned medication usage events for Donepezil, Galantamine, Memantine, Rivastigmine, Lecanemab, Donanemab, Suvorexant, and BrexpiprazoleTreat brand and generic names as equivalent. For each medication, return a list of events including the date (YYYY-MM-DD if possible) and the current dose in mg if mentioned. If*

discontinued on a particular date, assign 0 mg to the current dose. Return the result strictly as valid JSON using the structure {"MED 1":[{"date":"YYYY-MM-DD","dose_mg":<value>}], ... }.

Using medication events extracted from CLIN-SUMM summaries, we reconstructed longitudinal medication trajectories for dementia-related treatments. To validate this approach and evaluate its ability to recover medication information not captured in structured EHR data, we compared LLM-extracted medication events from CLIN-SUMM summaries with medication records obtained from Snowflake tables.

4.7 Evaluation of a Commercial Frontier LLM

To evaluate the CLIN-SUMM framework across different model families, we implemented the pipeline using a commercial frontier model (GPT-4o) accessed via the OpenAI API in the pilot subset. This evaluation was designed to provide a reference point for summarization quality and to enable comparison with an open-source alternative (Qwen3) used for large-scale deployment. Because frontier models differ in context capacity, instruction following, and API constraints, several implementation adjustments were made. GPT-4o required fewer prompt constraints to achieve structured output. For example, explicit enumeration of laboratory test names and detailed examples used in the Qwen prompt were no longer necessary. The pipeline incorporated a sliding context window to ensure that longitudinal summaries remained within the model's input limits. For GPT-4o, a window size of 50 prior summaries was used, representing the largest configuration that maintained stable API performance while avoiding request size limits and latency constraints. Specifically, it avoids intermittent rate-limit and timeout errors that stemmed not from token overflows but from API-level constraints, including transient throughput throttling, payload-size serialization overhead, and server-side latency caps on long requests, leading to slower response times and occasional failed generations. Overall, the frontier model produced stable and well-structured summaries with minimal prompt modification.

The GPT-4o evaluation was conducted on the pilot dataset to allow direct comparison with summaries generated by Qwen3 under identical conditions. To assess textual similarity between summaries generated by GPT-4o and Qwen3, we computed BERTScore [35], which measures semantic similarity between two texts using contextual embeddings. BERTScore evaluates token-level alignment between candidate and reference texts in embedding space and reports precision, recall, and F1 scores. Precision measures how well tokens in the candidate summary match semantically relevant tokens in the reference summary, recall measures how well the candidate summary captures the semantic content of the reference summary, and the F1 score provides a balanced measure of both coverage and fidelity. These metrics were computed between GPT-4o-generated summaries and Qwen3-generated summaries for the same patient records. In addition, to evaluate the downstream predictive utility of the summaries, we trained the Clinical ModernBERT dementia diagnosis and 3-year prediction models using GPT-4o-generated summaries from the pilot subset, following the same training procedure described above. For comparison, we trained corresponding models on Qwen3-generated summaries using the same pilot data. These pilot-based Qwen3 models were used only for comparison with GPT-4o, whereas the primary diagnosis and 3-year prediction model results reported elsewhere in the study were obtained from models trained on the full dataset (2600 training cases, 799 validation cases, and 601 test cases, with an additional 8,356 cases reserved as an external test set). Performance was then compared between the two.

5 Data availability

The Massachusetts General Brigham source data are not publicly available because they are electronic health records. Making the data publicly available without additional consent or ethics could compromise privacy.

6 Code availability

The LLM prompts are provided in their entirety in **Supplementary Figure 1**. Code for the sliding window, Jaccard similarity evaluation and Clinical ModernBERT fine-tuning for disease diagnosis and 3-year prediction will be made publicly available via a public repository upon publication.

7 Acknowledgments

M.M. is partially supported by grants from the National Institutes of Health (3OT2OD035404-01S3, 1R01NS134597, and 1UG3HG014379-01) and from the American Heart Association (961045). C.D.A. reports funding by NIH U01NS069673, RF1NS139183, R01NS093870, AHA 21SFRN812095, and the Global Brain Care Coalition.

8 Author contributions

V.D. contributed to methodology, data curation, formal analysis, visualization, and writing. D.F.P. contributed to methodology, visualization, and writing. A.A., A.N., and E.B.H. contributed to clinical validation and investigation. W.H. contributed to data curation. T.N. and S.F. contributed to methodology. S.J.A. contributed to data curation, clinical conceptualization and interpretation. C.D.A. contributed to clinical conceptualization and interpretation. J.H. contributed to conceptualization and funding acquisition. M.M. contributed to conceptualization, methodology, funding acquisition, and supervision. All authors contributed to review and editing of the final manuscript.

9 Competing interests

C.D.A. has received sponsored research support from Bayer AG and Massachusetts General Hospital and has consulted for ApoPharma and MPM BioImpact. S.J.A. receives sponsored research support from Bristol Myers Squibb / Pfizer. He has consulted for Bristol Myers Squibb, Fitbit (Google), Pfizer, and Premier. The remaining authors declare no competing interests.

10 Human Ethics and Consent to Participate

This study protocol was approved by the Institutional Review Board (IRB) of Massachusetts General Brigham (#2023P001077) and complied with the Declaration of Helsinki. The requirement for informed consent was waived by the Institutional Review Board given the low-risk retrospective nature of the study and privacy-protecting measures implemented by the investigators.

References

[1] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. Scientific Reports, 10:7155, April 2020.

[2] Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. arXiv preprint arXiv:2405.14567, 2024.

[3] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. arXiv preprint arXiv:2412.16178, December 2024.

[4] Jackson Steinkamp, Jacob J. Kantrowitz, and Subha Airan-Javia. Prevalence and sources of duplicate information in the electronic medical record. JAMA Network Open, 5(9):e2233348, September 2022.

[5] Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. Computers in Biology and Medicine, 155:106649, March 2023.

[6] OpenAI. New models and developer products announced at DevDay. https://openai.com/index/new-models-and-dev eloper-products-announced-at-devday/, 2023. [Online; accessed August 2025].

[7] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024.

[8] Neel Kanwal and Giuseppe Rizzo. Attention-based clinical note summarization. In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, volume abs/2104.08942, page 813–820. Association for Computing Machinery, 2022.

[9] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. Learning to summarize radiology findings. In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis (EMNLP-LOUHI), 2018.

[10] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine, 30:1134–1142, February 2024.

[11] Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. What's in a summary? laying the groundwork for advances in hospital-course summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 4794–4811. Association for Computational Linguistics, June 2021.

[12] Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for

interpretable in-hospital mortality prediction. In Proceedings of the AMIA Annual Symposium, pages 719–728, 2023.

[13] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfie Baston, Jack Ross, Esther Idowu, James T Teo, and Richard J Dobson. Foresight – generative pretrained transformer (gpt) for modelling of patient timelines using ehrs, 2023.

[14] Zeljko Kraljevic, Joshua Au Yeung, Daniel Bean, James Teo, and Richard J. Dobson. Large language models for medical forecasting – foresight 2, 2024.

[15] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv preprint arXiv:2505.09388, May 2025.

[16] Emma Croxford et al. Current and future state of evaluation of large language models for medical summarization tasks. NPJ health systems, 2(1):6, 2025.

[17] Liyan Tang et al. Evaluating large language models on medical evidence summarization. NPJ Digital Medicine, 6(1):158, 2023.

[18] Wikipedia contributors. Jaccard index — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Jaccar d_index. [Online; accessed October 2024].

[19] Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text. arXiv preprint arXiv:2504.03964, April 2025.

[20] Mika Kivimäki, Gill Livingston, Archana Singh-Manoux, Nina Mars, Joni V. Lindbohm, Jaana Pentti, Solja T. Nyberg, Matti Pirinen, Emma L. Anderson, Aroon D. Hingorani, Pyry N. Sipilä, and et al. Estimating dementia risk using multifactorial prediction models. JAMA Network Open, 6(6):e2318132, June 2023.

[21] Shireen Sindi, Elisabeth Calov, Jasmine Fokkens, Tiia Ngandu, Hilkka Soininen, Jaakko Tuomilehto, and Miia Kivipelto. The caide dementia risk score app: The development of an evidence-based mobile application to predict the risk of dementia. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1(3):328–333, July 2015.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, August 2016.

[23] C. R. Jack, D. A. Bennett, K. Blennow, et al. Nia-aa research framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14:535–562, 2018.

[24] Louise M. Allan, Clive G. Ballard, Elise N. Rowan, and Rose Anne Kenny. Incidence and prediction of falls in dementia: A prospective study in older people. PLoS ONE, 4(5):e5521, 2009.

[25] AARP. What are the modifiable risk factors for dementia?, 2024. Accessed: 2026-03-20.

[26] S. Hay, K. Ong, D. Santomauro, et al. Burden of 375 diseases and injuries, risk-attributable burden of 88 risk factors and healthy life expectancy in 204 countries and territories including 660 subnational locations, 1990–2023: a systematic analysis for the global burden of disease study 2023. The Lancet, 406:1873–1922, 2025.

[27] J. Manuello, J. Min, P. McCarthy, et al. The effects of genetic and modifiable risk factors on brain regions vulnerable to aging and disease. Nature Communications, 15:2576, 2024.

[28] Hurst, A., Lerer, A., Goucher, A. P., and et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, October 2024.

[29] Pellegrini, C., Özsoy, E., Bani-Harouni, D., Keicher, M., & Navab, N. (2025). From ehrs to patient pathways: Scalable modeling of longitudinal health trajectories with llms. *arXiv preprint arXiv:2506.04831*.

[30] Hartman V, Campion TR. A Day-to-Day Approach for Automating the Hospital Course Section of the Discharge Summary. AMIA Jt Summits Transl Sci Proc. 2022 May 23;2022:216-225. PMID: 35854728; PMCID: PMC9285173.

[31] Kaarin J. Anstey, Nicolas Cherbuin, Pushpani M. Herath, Chengxuan Qiu, Lewis H. Kuller, Oscar L. Lopez, Robert S. Wilson, and Laura Fratiglioni. A self-report risk index to predict occurrence of dementia in three independent cohorts of older adults: the anu-adri. PLoS One, 9(1):e86141, January 2014.

[32] OpenAI. Best practices for prompt engineering with OpenAI API. https://help.openai.com/en/articles/6654000-b est-practices-for-prompt-engineering-with-openai-api, 2023. [Online; accessed October 2024].

[33] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Yonghao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with paged attention. Association for Computing Machinery, September 2023.

[34] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.

[35] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *ArXiv, abs/1904.09675*.