DOI: 10.1002/alz.095607

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

DEMENTIA CARE RESEARCH AND PSYCHOSOCIAL FACTORS

POSTER PRESENTATION

DEMENTIA CARE RESEARCH (RESEARCH PROJECTS; NONPHARMACOLOGICAL)

Validation of Large Language Models to Identify Dementia Cases in Electronic Health Records

Richard Yang | **Liqin Wang**

Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Correspondence
Liqin Wang, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
Email: lwang@bwh.harvard.edu

Abstract

Background: Dementia diagnosis presents global healthcare challenges due to its elusive nature and inconsistent documentation in Electronic Health Records (EHRs). Approaches relying on diagnostic codes and rules may overlook undiagnosed cases or mistakenly identify others, as these codes are frequently used for purposes beyond disease diagnosis. This highlights the need for more sophisticated diagnostic tools. The advent of artificial intelligence (AI) and large language models (LLMs), with their advanced natural language understanding, promises more accurate detection. This study evaluates LLMs' performance in identifying dementia cases using aggregated EHR data.

Method: The study utilized the EHR from Mass General Brigham (MGB) to identify potential dementia patients through keywords and codes indicating cognitive decline. From this pool, 200 patients underwent independent chart reviews by two experts, with discrepancies resolved by a third. We employed two LLMs, GPT-3.5 and GPT-4, and two data preparation approaches for dementia detection: daily medical record aggregation and patient record aggregation. The first approach concatenated daily records related to cognitive decline and analyzed the data chronologically with zero-shot prompting using both LLMs. Conversely, the second approach concatenated entire patient medical records related to cognitive decline and applied GPT-4 with few-shot prompting. For comparison, we included a baseline that requires two separate dementia ICD diagnoses in different years, at least 30 days apart. We assessed each approach using accuracy, positive predictive value (PPV), sensitivity, specificity, and F1 score.

Result: The patient record aggregation plus few-shot prompting strategy with GPT-4 had the best performance, achieving an accuracy of 0.86 and an F1 score of 0.80. In the daily medical record aggregation approach, GPT-4's zero-shot prompting outperformed GPT-3.5, with F1 scores of 0.78 and 0.57, respectively. The baseline

rule-based approach scored a good F1 of 0.74 but showed lower sensitivity, suggesting underdiagnosis with traditional methods.

Conclusion: This study demonstrates that LLMs can significantly enhance dementia diagnosis using EHRs. Our findings reveal that the aggregated patient record and few-shot prompting strategy with GPT-4 outperforms traditional methods, offering a more comprehensive and accurate evaluation of dementia.

Figure 1. A flowchart for Evaluating LLMs for Identifying Dementia Cases in EHRs.

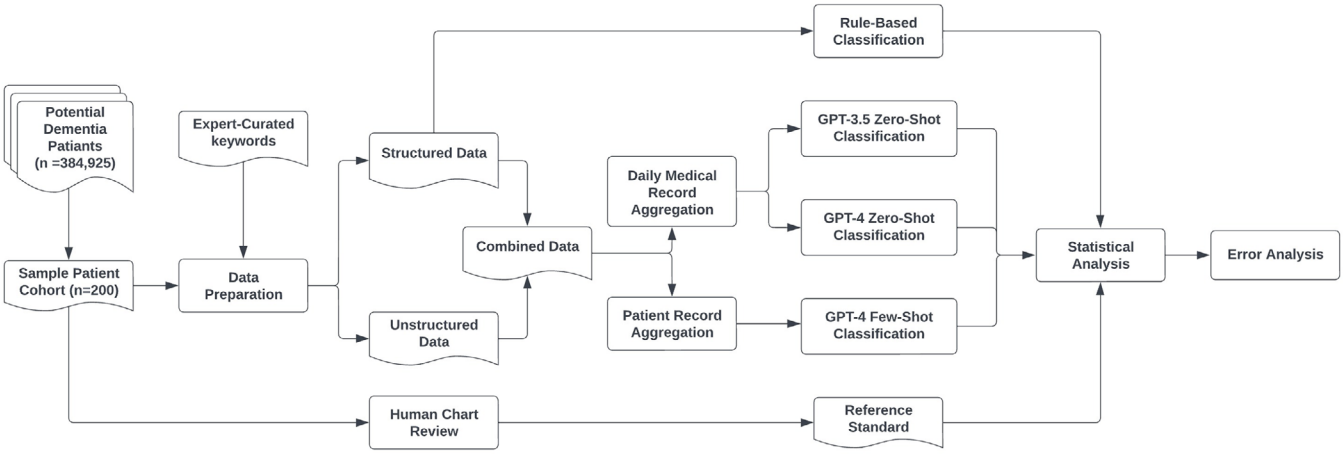

Table 1. Diagnostic Rules and LLM Prompts for Dementia Diagnosis

Approach	Description
Rule-Based Algorithm	2 codes in 2 years separated by at least 30 days ICD-9: 46.1, 290.0, 290.1, 290.2, 290.3, 290.4, 294.x, 331.0, 331.1, 331.5, 331.82 ICD-10: F00.x, F01.x, F02.x, F03.x, G30.x
Zero-Shot Prompt	Context: "Act as a specialist in neurology." Request: "Based on the medical summary, determine the dementia status of this patient (not his/her family member). You are strictly looking for dementia, not less severe conditions like mild cognitive impairment. Do not consider medications. Return your answer in JSON format. For the 'Dementia_Status' field: 'YES' if the patient has dementia, 'NO' if not. For the 'Reason' field, explain your rationale in one sentence."
Few-Shot Prompt	Context: "Act as a specialist in neurology. Review all medical summaries of the patient and determine dementia status." Initial Review of Each Medical Summary: Request: "Review this medical summary and note any indicators of dementia. Do not make a final decision yet." Final Decision and Rationale: Request: "Now that all summaries have been reviewed, please provide your final assessment of the patient's dementia status (not his/her family member). You are strictly looking for dementia, not less severe conditions like mild cognitive impairment. Do not consider medications. Return your answer in JSON format. For the 'Dementia_Status' field: 'YES' if the patient has dementia, 'NO' if not. For the 'Reason' field, explain your rationale in one sentence. "

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

Table 2. Performance of Dementia Identification Approaches

Approach	LLM Model	Accuracy	PPV	Sensitivity	Specificity	F1
Rule-Based Approach	N/A	0.85	0.75	0.74	0.90	0.74
Daily Medical Record Aggregation, Zero-Shot	GPT-3.5	0.57	0.41	0.98	0.40	0.57
	GPT-4	0.84	0.65	1.0	0.78	0.78
Patient Record Aggregation, Few-shot	GPT-4	0.86	0.69	0.95	0.82	0.80

Abbreviations: LLM, large language model; PPV, positive predictive value