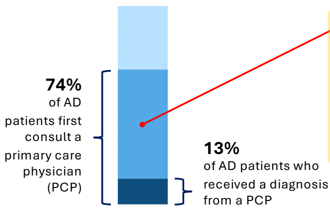
# WATCH-SS: Developing a Trustworthy and Explainable Modular Framework for Detecting Cognitive Impairment from Spontaneous Speech

Sydney Pugh, PhD[1], Matthew Hill[1], Sy Hwang, MS[1], Rachel Wu[1], Kuk Jang[1,2], Stacy Iannone, DHSc, MS[1], Karen O'Connor, MS[1], Kyra O'Brien, MS, MSHP[1], Eric Eaton, PhD[2], and Kevin Johnson, MD, MS[1,2]

[1] Perelman School of Medicine, University of Pennsylvania, [2] School of Engineering and Applied Science, University of Pennsylvania

## Introduction

- 7.2 million Americans age 65 and older are estimated to be living with Alzheimer's disease (AD) in 2025[1]
- Over 50% of patients with Alzheimer's disease and related dementias (ADRD) are **undiagnosed** or **unaware of diagnosis**[2]
- Disparities in AD diagnosis and treatment disproportionately affect underrepresented racial, ethnic, and socioeconomic groups[3]
  - E.g., AD is almost twice as prevalent in Black individuals than White individuals (~19% vs 10%), yet Black individuals comprise < 3% of participants in two pivotal new medication trials
- Primary care is an optimal setting for early detection of ADRD
  - Often the first point of contact for emerging health concerns
  - Long-standing relationship with patient may reveal subtle signs (e.g., medication or appointment adherence)
- Key challenges to ADRD diagnosis in primary care:
  - Time/Competing priorities
  - Lack of expertise
  - Lack of comfort with diagnosis or providing follow-up care
  - Lack of support (e.g., access to neurologists)

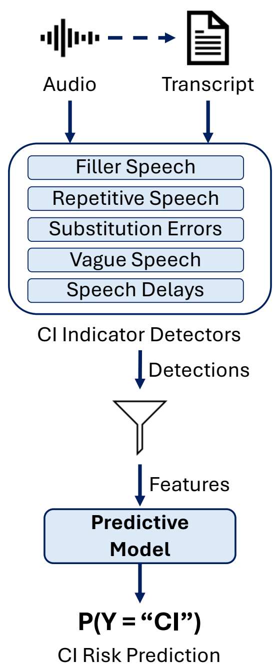



**Goal**: To close this gap with **WATCH-SS**, a screening tool that **passively analyzes patient speech** to alert PCPs to patients who would benefit from a **formal cognitive assessment**.

[1] 2025 Alzheimer's Disease Facts and Figures. Alzheimer's Association. 2025.
[2] H. Amjad, et al. Journal of General Internal Medicine. 2018.
[3] B. Cavedoni and K. O'Brien. Practical Neurology. 2025.

## Methods

### The Warning Assessment and Alerting Tool for Cognitive Health using Spontaneous Speech (WATCH-SS) Framework



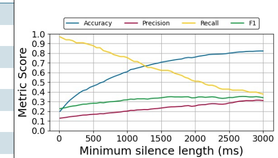

- DementiaBank ADReSS dataset: recordings of subjects performing a standardized picture description task
- For linguistic indicators, we compared two approaches:
  1. **Traditional NLP** (e.g., keyword search, n-gram analysis)
  2. **Large Language Models (LLMs)** (zero- or few-shot prompting with GPT-4o)
- For speech delays, we use a **silence detector** on the audio waveform
- Detections aggregated into **clinically interpretable set of summary features** provided to a **LightGBM** model to produce the final risk prediction

## Results

### Detector Performance

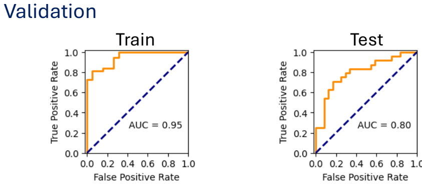
- Simple NLP baselines achieve best performance for lexically-defined tasks like filler and repetitive speech
- LLMs were superior for more semantically complex tasks
- The silence detector for speech delays achieved a peak F1-score of 35%

| Indicator | Detector | Precision | Recall | F1 |
|---|---|---|---|---|
| Filler Speech | Keywords | **0.941** | 0.935 | **0.938** |
| | LLM | 0.623 | **0.941** | 0.750 |
| Repetitive Speech | Unigrams | **0.557** | **0.957** | **0.704** |
| | LLM | 0.407 | **0.957** | 0.571 |
| Substitution Errors | MLM | 0.049 | **0.720** | 0.093 |
| | LLM | **0.107** | 0.640 | **0.184** |
| Vague Speech | Keywords | 0.032 | **0.875** | 0.061 |
| | LLM | **0.061** | **0.875** | **0.115** |




### Model Performance

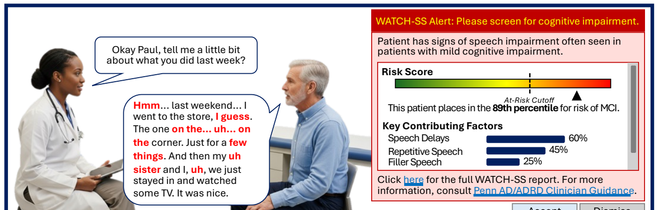
#### Internal Validation




#### External Validation

- On a set of 27 clinic visit recordings for patients 65+ from the OBSERVER Repository, WATCH-SS yielded lower predictive performance (AUC = 0.63), highlighting the challenge of using fragmented patient speech samples common in primary care

## Clinical Use Case




## Conclusion

- WATCH-SS demonstrates that a modular, feature-based approach can achieve strong predictive performance (AUC=80%) while maintaining the interpretability required for a trustworthy, clinically-usable screening tool for cognitive health
- **Future Work**: (i) Refine and expand the set of detectors, (ii) retrain the predictive model on larger, more diverse datasets, and (iii) larger-scale validation study using real clinic visits

## Acknowledgements

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number P30AG073105, and by the University of Pennsylvania ASSET Center under the Collaborative Research in Trustworthy AI for Medicine grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.