Vantage Point: Algorithmic dementia classification via data integration and latent truth modeling

Crystal Shaw, M.S.[1,2], Elizabeth Rose Mayeda, Ph.D.[2], Thomas R. Belin, Ph.D.[1]
UCLA Department of Biostatistics[1], UCLA Department of Epidemiology[2]; Los Angeles, California, USA

Email: c.shaw@ucla.edu

UCLA Fielding School of Public Health

MOTIVATION: ALGORITHMIC DEMENTIA ASCERTAINMENT IN LARGE COHORT STUDIES

Large, representative cohort studies help us understand trends in population health and risk factors for diseases

Objective: estimate the societal burden of dementia-- currently estimated that 5.8 million Americans are living with Alzheimer's dementia [1]

Gold-standard dementia ascertainment (expert judgement) is time-intensive, expensive, and nearly impossible in population-representative studies

Algorithmic dementia classification can help us close the research gap, but supervised learning techniques require adjudicated dementia cases

By combining **multiple data sources** and modeling **source reliability**, Bayesian data integration via the latent truth model provides an **unsupervised learning** alternative to dementia classification with **high specificity** compared to supervised learning algorithms used in practice.

DATA SOURCE: HEALTH AND RETIREMENT STUDY

Health and Retirement Study (HRS)

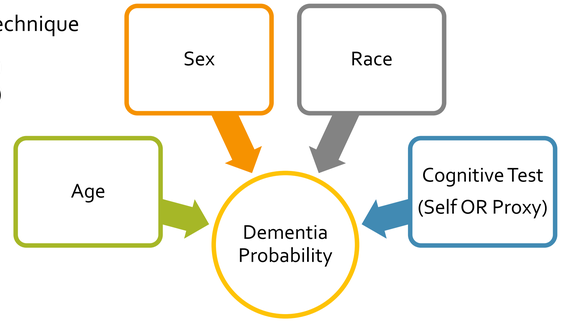
- Nationally representative, longitudinal study of over 37,000 adults age 50 or older (and their spouses)
- Wave 1 in 1992, biennial interviews since + new enrollment (13 waves of data available)
- Studies predictors of successful aging, behavior choices, and life transitions

Aging, Demographics, and Memory Study (ADAMS)
*used to validate classification algorithms

- HRS participants age 70+ (n = 856)
- Gold-standard dementia ascertainment
- Wave A: prevalent dementia (2001-2003)
- Waves B – D: incident dementia (2003-2009)

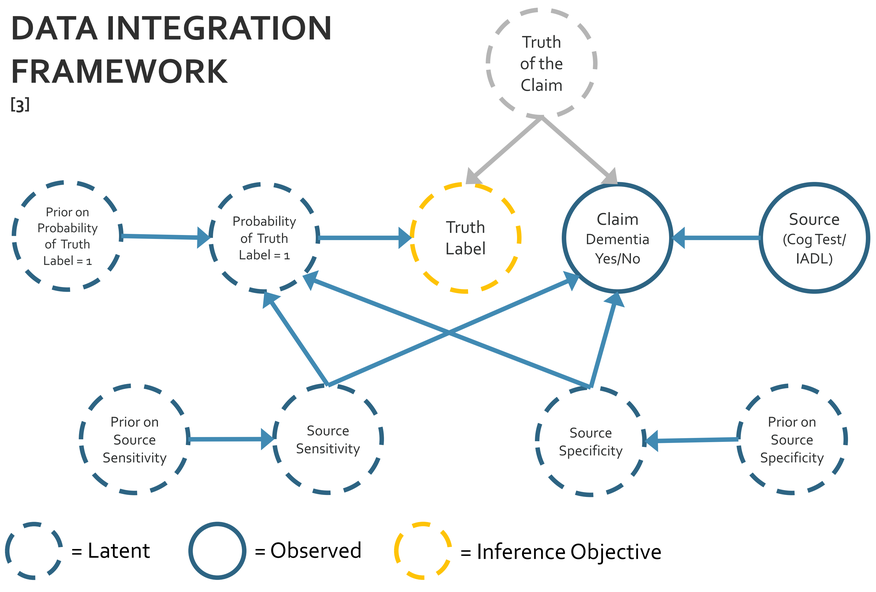
COMPARISON ALGORITHM: WU LOGIT MODEL

- Supervised learning technique
- Sensitivity: 44 (32, 55)
  Specificity: 93 (90, 95)
  [2]



DATA INTEGRATION FRAMEWORK

[3]



METHODS

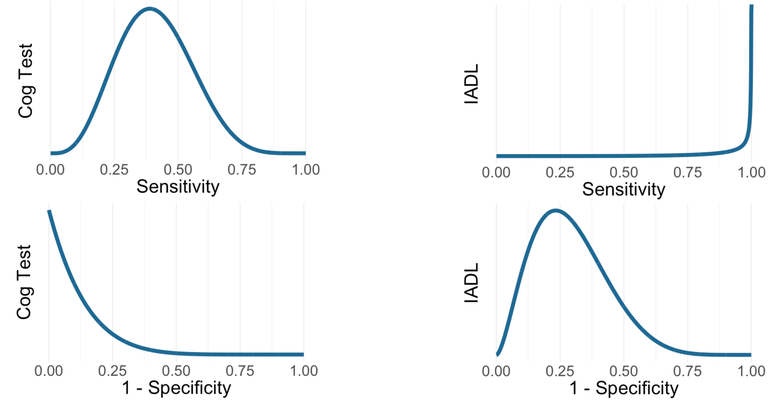
Sample Selection
- Randomly sample n = 25,000 observations from HRS waves 3-12
- "observation" refers to an individual at an interview wave
- An individual could appear more than once in the sample

Sources Considered
- Summary of individual's wave-specific cognitive test performance
- Summary of individual's wave-specific instrumental activities of daily living (IADL) limitations

Specify Source Priors
- Cog Test Sensitivity: Beta(4.5, 6.5)
- Cog Test 1 - Specificity: Beta(1, 7.8)
- IADL Sensitivity: Beta(4.4, 0.3)
- IADL 1 - Specificity: Beta(2.7, 6.6)
- Truth Label: Beta(2, 2)

Determine Truth Labels
- Run Collapsed Gibbs Sampler to determine source-specific P(truth label = 1)
- Average source-specific P(truth label = 1) for each observation
- If average P(truth label = 1) ≥ 0.5, truth label = 1

SENSITIVITY & SPECIFICITY PRIORS



PROOF OF CONCEPT

Compare dementia classification based on data integration with classification based on the Wu model

Results
Sensitivity: 61.6%
Specificity: 93.2%

FUTURE WORK

- Include covariates such as age, sex, race/ethnicity, and education when specifying priors
- Include proxy measures as well as direct measures as sources
- Use ADAMS data as the gold standard for algorithm validation

REFERENCES
[1] Herbert et al., *Neurology* (2013)
[2] Gianatassio et al., *Epidemiology* (2019)
[3] Zhao et al., *Proceedings of the VLDB Endowment* (2012)