Comparison of methods for algorithmic classification of dementia status in the Health and Retirement Study

Kan Z. Gianattasio[1], Qiong Wu[2], M. Maria Glymour[3], Melinda C. Power[1]

[1]Department of Epidemiology and Biostatistics, Milken Institute School of Public Health, George Washington University; [2]Institute of Social Science Survey, Peking University; [3]Department of Epidemiology and Biostatistics, University of California, San Francisco

FUNDING ACKNOWLEDGMENT: This work was funded by R03 AG055485

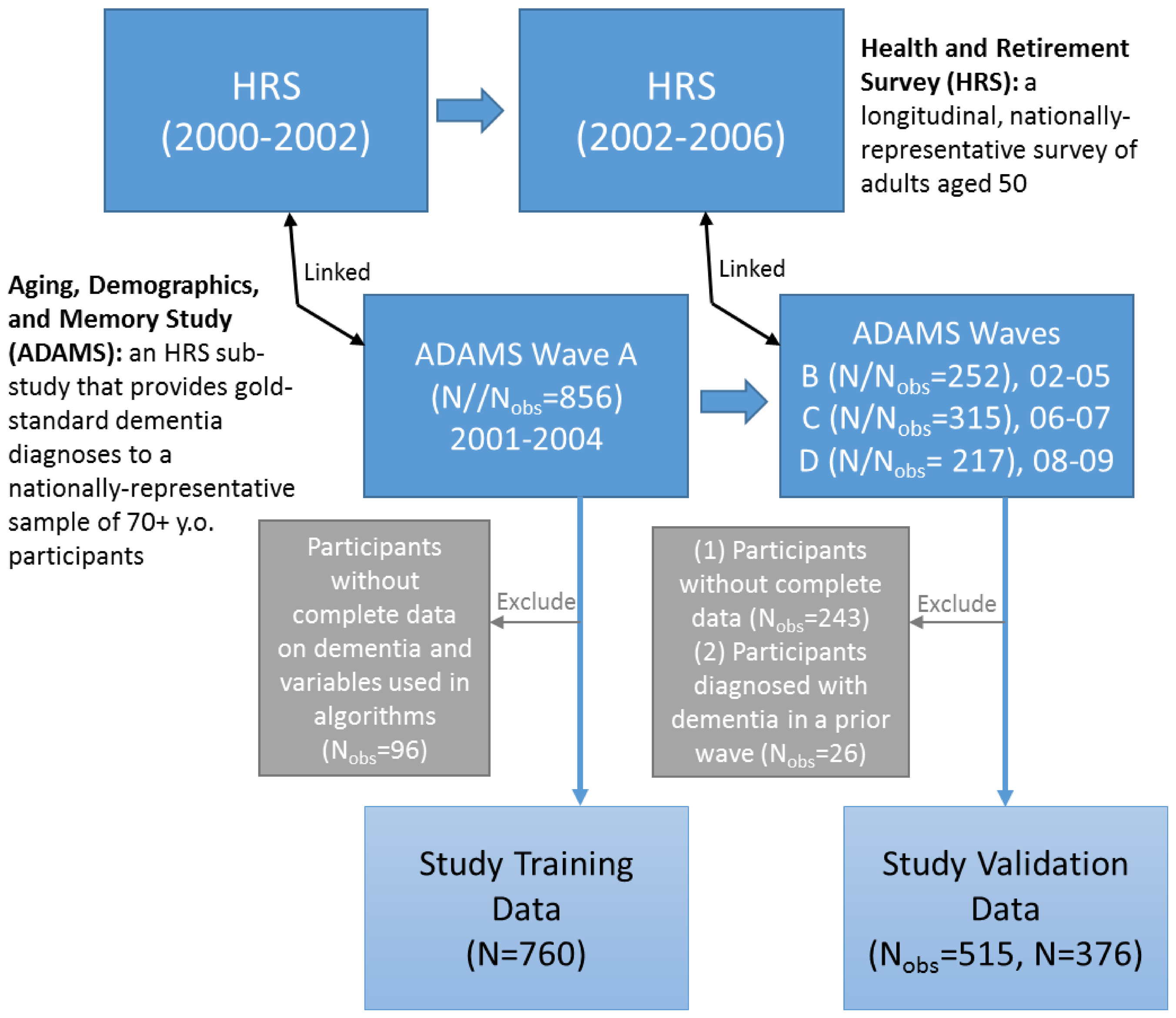
GW | Public Health

Introduction and Objectives

- Dementia ascertainment is time-consuming and costly, thus it is difficult to describe and monitor trends and disparities in the prevalence and incidence of cognitive impairment.
- Researchers have independently developed algorithms to use existing data from the *Health and Retirement Study (HRS)* to algorithmically classify dementia status in cohort participants,[1–5] but reporting of performance metrics is inconsistent.
- The **objective** of this study is to conduct a head-to-head comparison of performance of *5* existing algorithms for algorithmic classification of dementia in the HRS.

Data and Methods

Data:



Statistical Analyses:

- We applied each algorithm (**Table 1**) to the training and validation data and computed (a) sensitivity, (b) specificity, and (c) overall accuracy.
- We performed various sensitivity and robustness checks:
  - re-evaluating performance on alternate validation sample that included participants with previously diagnosed dementia known to be alive at waves B, C, D;
  - bootstrapping all analyses to obtain 95% confidence intervals;
  - and others.

References

1. Herzog AR, Wallace RB. Measures of cognitive functioning in the AHEAD Study. *J Gerontol B Psychol Sci Soc Sci*. 1997;52 Spec No:37-48.

2. Alzheimer's Association. 2010 Alzheimer's disease facts and figures. *Alzheimer's Dement*. 2010;6(2):158-194.

3. Hurd MD, Martorell P, Delavande A, Mullen KJ, Langa KM. Monetary Costs of Dementia in the United States. *N Engl J Med*. 2013;368(14):1326-1334.

4.Crimmins EM, Kim JK, Langa KM, Weir DR. Assessment of Cognition Using Surveys and Neuropsychological Assessment: The Health and Retirement Study and the Aging, Demographics, and Memory Study. *Journals Gerontol Ser B Psychol Sci Soc Sci*. 2011;66B(Supplement 1):i162-i171.

5. Wu Q, Tchetgen Tchetgen EJ, Osypuk TL, White K, Mujahid M, Maria Glymour M. Combining direct and proxy assessments to reduce attrition bias in a longitudinal study. *Alzheimer Dis Assoc Disord*. 2013;27(3):207-212.

Results

Sample Descriptions: There was a higher portion of dementia cases (34% vs. 15%), and proxy-respondents (22% vs. 6%) in the training data compared to the validation data. While training data participants also had more physical functioning limitations, the two groups were similar in sociodemographics and cognitive functioning. ***Algorithm Descriptions:*** Separate algorithms were used for self-respondents versus proxy-respondents by all authors except Wu et al, who used the missing-indicator method to combine self- and proxy- respondents into a single algorithm and setting non-applicable items to 0 (**Table 1**).

Table 1: Details of model choice and variables included by algorithm

Model	Herzog-Wallace (1997)		Langa-Kabeto-Weir, (2009)		Crimmins (2011)		Hurd (2013)		Wu (2013)	
	Score cutoff		Score cutoff		Multinom. Logit		Ordered probit		Logit	
Predictors	Self (35)	Proxy (7)	Self (27)	Proxy (11)	Self	Proxy	Self	Proxy	Self	Proxy
Demographics										
Age, Gender					X		X	X	X	X
Education					X		X	X		
Race									X	X
Cognition (self-response)										
Word recall	X		X		X		X	X	X	
Serial 7's	X		X		X		X	X	X	
Backward count	X		X		X		X		X	
Dates					X		X	X	X	
Object naming (2)	X				X		X			
President	X						X	X	X	
Vice-president	X				X		X		X	
Cognition (proxy)										
Proxy-rated memory				X		X				X
Interviewer assess.				X		X				
16-item IQCODE								X		X
7-item Jorm symps		X				X				
Physical Functioning										
ADLS's					X		X	X		
IADL's				X	X		X	X		

Table 2: Overall Performance metrics for each data sample (0.5 cut-point)

Algorithm	*Training (N=760)*			*Validation (N=515)*		
	Sens	Spec	Acc	Sens	Spec	Acc
Summary score cutoff-based algorithms						
Herzog-Wallace	53.5	96.6	82.0	18.3	97.8	86.8
Langa-Kabeto-Weir	75.2	83.3	80.5	40.9	89.2	82.5
Regression-based algorithms						
Crimmins	89.9	79.1	82.8	62.0	82.2	79.4
Hurd	76.7	91.8	86.7	39.4	96.0	88.2
Wu	77.9	88.1	84.6	43.7	92.6	85.8

- Sensitivity was higher in the training data; specificity and accuracy were similar in the training and validation data across algorithms (**Table 2**).
- Of the three regression-based algorithms, Hurd achieved highest specificity when maximizing sensitivity, as well as highest sensitivity when maximizing specificity (**Figure 1**).
- Specificity and overall accuracy was higher for non-Hispanic whites compared to both minority groups across the board (**Figure 2**).
- Sensitivity was higher among proxy and older (80+) respondents
- Specificity was uniformly better among self- and younger respondents.
- The algorithms performed generally better in classifying dementia for those with at least a high school education and for females, with few exceptions.

Figure 1: ROC curves for each regression-based algorithm in the training and validation data

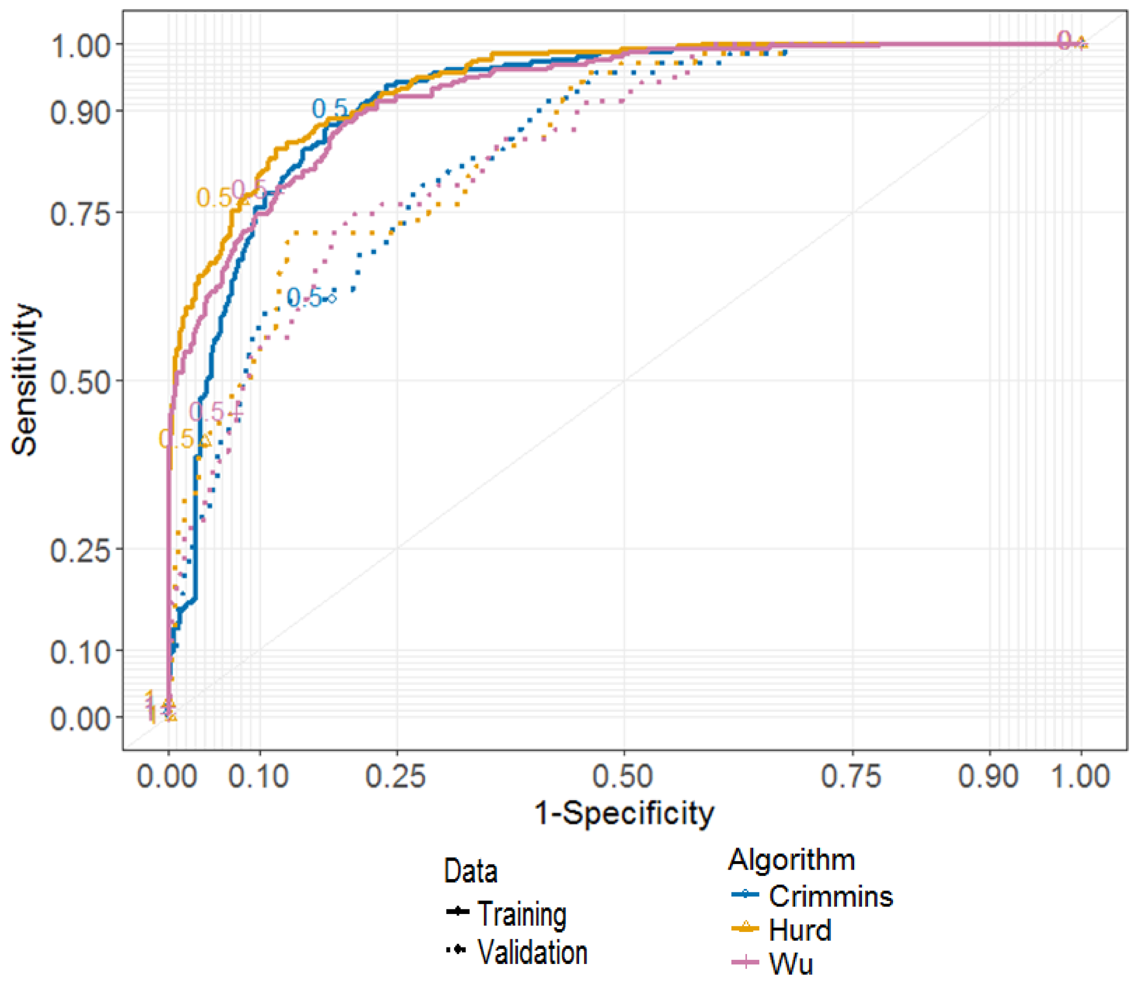


Figure 2: Performance metrics for each algorithm, by race/ethnicity, by sample

Conclusions and Discussion

- Higher sensitivity in the training and alternate validation data suggest that existing algorithms are better at predicting prevalent than incident dementia.
- The usefulness of each algorithm will be determined by the purpose:
  - At cut-point = 0.5, Crimmins provides highest sensitivity and Herzog-Wallace provides highest specificity, while Hurd offers highest overall accuracy. Hurd also minimizes race/ethnic disparities in prevalent cases, while Wu/Crimmins minimize these disparities in incident dementia.
- The relative ease of applying these algorithms will also be a key factor to consider: regression-based algorithms are much more difficult and time-consuming to implement.

Limitations:

- We assume a time-invariant relationship between predictors and dementia.
- Validation and training data drawn from same study -> limits external validity.
- Validation data includes only incident cases, which are not ideal for evaluating algorithms developed with prevalent cases.
- Small N's limit conclusiveness of sub-group differences.

Future Directions:

- Further testing of existing algorithms using external data sources, separately for prevalent and incident dementia.
- Developing improved algorithms for classifying dementia using variables commonly collected in large population surveys, with a particular focus on achieving uniform performance across subgroups.