OPEN

# Comparison of Methods for Algorithmic Classification of Dementia Status in the Health and Retirement Study

*Kan Z. Gianattasio,*[a] *Qiong Wu,*[b] *M. Maria Glymour,*[c] *and Melinda C. Power*[a]

**Background:** Dementia ascertainment is time-consuming and costly. Several algorithms use existing data from the US-representative Health and Retirement Study (HRS) to algorithmically identify dementia. However, relative performance of these algorithms remains unknown.

**Methods:** We compared performance across five algorithms (Herzog–Wallace, Langa–Kabeto–Weir, Crimmins, Hurd, Wu) overall and within sociodemographic subgroups in participants in HRS and Wave A of the Aging, Demographics, and Memory Study (ADAMS, 2000–2002), an HRS substudy including in-person dementia ascertainment. We then compared algorithmic performance in an internal (time-split) validation dataset including participants of HRS and ADAMS Waves B, C, and/or D (2002–2009).

**Results:** In the unweighted training data, sensitivity ranged from 53% to 90%, specificity ranged from 79% to 97%, and overall accuracy ranged from 81% to 87%. Though sensitivity was lower in the unweighted validation data (range: 18%–62%), overall accuracy was similar (range: 79%–88%) due to higher specificities (range: 82%–98%). In analyses weighted to represent the age-eligible US population, accuracy ranged from 91% to 94% in the training data and 87% to 94% in the validation data. Using a 0.5 probability cutoff, Crimmins maximized sensitivity, Herzog–Wallace maximized specificity, and Wu and Hurd maximized accuracy. Accuracy was higher among younger, highly-educated, and non-Hispanic white participants versus their complements in both weighted and unweighted analyses.

**Conclusion:** Algorithmic diagnoses provide a cost-effective way to conduct dementia research. However, naïve use of existing algorithms in disparities or risk factor research may induce nonconservative bias. Algorithms with more comparable performance across relevant subgroups are needed.

**Keywords:** Algorithms; Alzheimer's disease; Dementia; Health and Retirement Study

(*Epidemiology* 2019;30: 291–302)

Clinical diagnosis of dementia requires evidence of substantial decline in one or more domains of cognitive function and cognitive impairment that interferes with activities of daily living, as well as lack of evidence that findings are attributable to another disorder. Study-based adjudication procedures typically involve neuropsychological testing and an informant interview, and may also consider additional information from medical history, laboratory testing, neuroimaging findings, and neurological examination, especially when etiologic diagnoses are desired.[1–3] Final diagnosis is made by a consensus panel of clinicians and other experts after review of the available data. Thus dementia ascertainment is time-consuming and costly, making it difficult to implement in large, representative cohort studies. This hinders efforts to use large population surveys to monitor trends and disparities in the prevalence and incidence of dementia or to conduct risk factor analyses in representative populations.

Recognizing this, several groups of researchers have developed algorithms using existing data from the large, nationally representative Health and Retirement Study (HRS) to algorithmically classify dementia status in cohort participants.[4–8] The HRS provides an ideal setting for algorithm development. First, a strategically selected subset of HRS participants was evaluated for dementia up to four times between 2001 and 2009 as part of the Aging, Demographics, and Memory Study (ADAMS).[9,10] Thus, data from ADAMS provide gold-standard dementia diagnoses against which to train and evaluate algorithms. Second, as HRS is nationally representative, algorithmic diagnoses in HRS can be used to monitor trends in cognitive impairment and dementia at the

Submitted June 12, 2018; accepted November 13, 2018.

From the [a]Department of Epidemiology and Biostatistics, Milken Institute School of Public Health, George Washington University, Washington, DC; [b]Institute of Social Science Survey, Peking University, Beijing; and [c]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA.

The results herein correspond to specific aims of Grant R03 AG055485 to M.C.P. from National Institutes of Health. M.M.G. was also supported by Grant R01 AG051170 from National Institutes of Health. The Health and Retirement Study data are sponsored by the National Institute on Aging (Grant Number U01AG009740) and was conducted by the University of Michigan.

The authors report no conflicts of interest.

The data used in this study are available on the Health and Retirement Study website (http://hrsonline.isr.umich.edu/). SAS code for reproducing our datasets and assigning algorithmic diagnoses are available on https://github.com/powerepilab/AD_algorithm_comparison.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Kan Z. Gianattasio, 950 New Hampshire Ave, 5th Floor, Washington DC 20052. E-mail: kzhang0316@gwu.edu.



ISSN: 1044-3983/19/3002-0291
DOI: 10.1097/EDE.0000000000000945

national level,[11–19] and would allow for disparities or risk factor analyses in a representative population.

Researchers hoping to use existing algorithms face a difficult choice. Each algorithm was developed independently, often in the context of other objectives. Thus, reporting of performance metrics is inconsistent, and whether there are substantial differences in performance remains unknown. Moreover, few reports provide performance achieved when algorithms are applied to data other than that used to develop the algorithm. Finally, the algorithm with the best overall performance metrics may be ill-suited to efforts to describe disparities or evaluate risk factor associations, if performance differs across subpopulations.

The objective of this study was to conduct a head-to-head comparison of existing algorithms for algorithmic classification of dementia in HRS. We first compared overall performance metrics across algorithms within a sample commonly used to develop existing algorithms—HRS/ADAMS participants who underwent dementia ascertainment at ADAMS Wave A. We then compared performance metrics across algorithms when applied to HRS/ADAMS data points that had not been used for algorithm development—data from HRS/ADAMS participants who underwent dementia ascertainment at ADAMS Waves B, C, and/or D. At each stage, we also quantified differences in performance metrics across subpopulations. We conclude with a discussion of when use of one or more of these algorithms may be appropriate.

## METHODS

### Data Sources

The HRS is a nationally representative, longitudinal study of adults >50 years of age and their spouses.[20,21] Since enrollment of the original cohort in 1992, HRS has enrolled several waves of new participants to maintain a steady-state cohort. Relevant to this study, data on sociodemographic characteristics, functional status (activities of daily living [ADLs]; instrumental activities of daily living [IADLs]), and cognitive status are collected at each interview, which participants have completed biennially since 1998. For respondents who are able and willing to be interviewed, HRS administers direct cognitive assessments using items from the Telephone Interview for Cognitive Status (TICS)[22] and the Mini-Mental State Examination (MMSE).[23] For those not able or willing to be interviewed, HRS collects relevant data from proxy respondents through proxy-rated assessment of memory and functional status, interviewer perceptions of cognitive status, proxy-reported Jorm symptoms of cognitive impairment,[24] and the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE).[25,26]

ADAMS is an HRS substudy that conducted systematic dementia ascertainment.[10] HRS self- and proxy-respondent participants ≥70 years of age who contributed data to the 2000 or 2002 waves were sampled for inclusion in ADAMS using stratified random sampling. Ultimately, 856 HRS participants were enrolled and completed initial assessment for prevalent dementia (Wave A, 2001–2003).[27] Three additional waves of data collection (Waves B, C, and D) were completed through 2009 among those without a dementia diagnosis in a previous wave (exclusive of a small number of repeat assessments for confirmation of prior diagnoses).[28] ADAMS evaluations were conducted in-person by a nurse and neuropsychology technician and included both respondent and proxy interviews. Respondents completed a battery of cognitive tests, neurologic examination, depression screening, and blood pressure measurement. Proxies provided information on neuropsychiatric symptoms, medical and cognitive history, medications, functional impairment, family history of memory problems, and caregiving. When agreed to, ADAMS investigators also collected data from participant medical records. Initial diagnoses based on the Diagnostic and Statistical Manual of Mental Disorders (editions 3 [DSM-III-R] and 4 [DSM-IV]) criteria were made by study experts based on all data collected at the ADAMS evaluation. These initial diagnoses were reviewed and revised by a geropsychiatrist in light of additional medical information obtained from the medical records when available. Final dementia diagnoses were confirmed by a consensus expert panel.[10,27,28]

### Existing Algorithms

We identified five[4–8] existing algorithms developed to predict dementia or significant cognitive impairment in HRS participants through pre-existing knowledge, informal searches of PubMed and Scopus, and review of articles cited by and citing identified manuscripts. Each algorithm is described in detail elsewhere.[4–8] Briefly, two of the algorithms—Herzog & Wallace (H–W)[4] and Langa–Kabeto–Weir (L–K–W)[5,7]—apply cutpoints to derived scores summarizing cognitive and/or functional data from the HRS interview to identify persons with severe cognitive impairment or dementia. For self-respondents, the summary scores reflect overall cognitive test performance. For proxy respondents, the H–W score is a sum of Jorm symptoms of cognitive impairment, whereas the L–K–W score sums proxy-rated memory, interviewer-perceived cognition, and IADLs. Cut points for summary scores were chosen to achieve a prevalence of dementia or cognitive impairment similar to the expected population prevalence, derived from external data sources (H–W)[4] or ADAMS findings (L–K–W)[7] (Table 1). The remaining three algorithms take varying regression-based approaches to predict the ADAMS Wave A cognitive status using HRS interview data: Wu et al.[8] applies a single logistic model to both self- and proxy-respondents, using an indicator for proxy respondents (similar to a missing indicator method); Hurd et al.[6] applies separate ordered probit models to self- and proxy-respondents; and Crimmins et al.[7] applies a multinomial logistic model to self-respondents and a logistic model to proxy respondents. Both Hurd and Crimmins predict a three-level outcome (dementia, cognitive impairment no dementia, normal); we focus

**TABLE 1. Predictors Used to Classify Dementia Status of HRS Participants in Each of the Five Algorithms**

| | Classification Algorithm | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Herzog–Wallace (1997)[a] Score Cutoff | | Langa–Kabeto–Weir (2009)[b] Score Cutoff | | Crimmins (2011)[c] Multinomial Logit | | Hurd (2013)[c,d] Ordered probit | | Wu (2013)[c,e] Logit | |
| | Self | Proxy | Self | Proxy | Self | Proxy | Self | Proxy | Self | Proxy |
| Predictors | | | | | | | | | | |
| Demographics | | | | | | | | | | |
| Age | — | — | — | — | X | — | X | X | X | X |
| Sex | — | — | — | — | X | — | X | X | X | X |
| Education | — | — | — | — | X | — | X | X | — | — |
| Race | — | — | — | — | — | — | — | — | X | X |
| Proxy indicator | — | — | — | — | — | — | — | X | X | X |
| Cognition (self-response) | | | | | | | | | | |
| Immediate word recall | X | — | X | — | X | — | X | X | X | — |
| Delayed 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 | X | — |
| Object naming (scissors) | X | — | — | — | X | — | X | — | — | — |
| Object naming (cactus) | X | — | — | — | X | — | X | — | X | — |
| President | X | — | — | — | X | — | X | X | X | — |
| Vice-president | X | — | — | — | X | — | — | — | X | — |
| Cognition (proxy) | — | — | — | — | — | — | — | — | — | — |
| Proxy-rated memory | — | — | — | X | — | X | — | — | — | X |
| Interviewer assessment | — | — | — | X | — | X | — | — | — | — |
| 16-item Jorm IQCODE | — | — | — | — | — | — | — | X | — | X |
| Jorm symptoms of cognitive impairment | — | X | — | — | — | — | — | — | — | — |
| Physical functioning (ADLs) | | | | | | | | | | |
| Eating | — | — | — | — | X | — | X | X | — | — |
| Bathing | — | — | — | — | X | — | X | X | — | — |
| Dressing | — | — | — | — | X | — | X | X | — | — |
| Transferring | — | — | — | — | — | — | X | X | — | — |
| Walking across room | — | — | — | — | — | — | X | X | — | — |
| Physical functioning (IADLs) | | | | | | | | | | |
| Using phone | — | — | — | X | X | X | X | X | — | — |
| Taking medication | — | — | — | X | — | X | X | X | — | — |
| Managing money | — | — | — | X | X | X | X | X | — | — |
| Grocery shopping | — | — | — | X | — | X | X | X | — | — |
| Preparing meals | — | — | — | X | — | X | X | X | — | — |

[a]For self-respondents, a score of 8 or lower out of 35 is classified as demented; a score of 8 is 2.70 SDs below the mean in the weighted training sample, and 3.02 SDs below the mean in the weighted validation sample. For proxy respondents, having two or more Jorm symptoms are classified as demented.

[b]For self-respondents, a score of 6 or lower out of 27 is classified as demented; a score of 6 is 1.85 SDs below the mean in the weighted training sample, and 2.02 SDs below the mean in the weighted validation sample. For proxy respondents, a score of 6 or higher out of 11 is classified as demented.

[c]A predicted probability greater than 0.5 is classified as demented for all three regression-based algorithms in our primary analyses. Among self-respondents, a 0.5 probability corresponds to the 89.7 percentile (Crimmins) and 95.9 percentile (Hurd and Wu) in the weighted training sample, and to the 90.2 percentile (Crimmins), 97.6 percentile (Hurd), and 97.7 percentile (Wu) in the weighted validation sample. Among proxy respondents, a 0.5 probability corresponds to the 28.9 percentile (Crimmins), 35.9 percentile (Hurd) and 27.3 percentile (Wu) in the weighted training sample, and to the 43.8 percentile (Crimmins), 77.5 percentile (Hurd) and 38.5 percentile (Wu) in the weighted validation sample.

[d]For predicting dementia status for proxy-respondent participants at a given HRS wave, Hurd included an indicator specifying whether they were self-respondents during the prior HRS wave, and if so, the algorithm uses data on cognitive assessments completed as a self-respondents two waves prior. If they also had a proxy respondent two waves prior, the algorithm uses change in proxy cognition scores.

[e]Wu used a single algorithm to classify dementia status for selves and participants who had a proxy in the most recent HRS wave using the missing-indicator method. The algorithm includes a binary proxy indicator, sets proxy cognition assessments to 0 for selves, and sets self-cognition assessments to 0 for proxies. As such, the Wu algorithm uses the variables under "proxy" heading when the participant is a proxy respondent and uses the variables under the "self" heading when the participant is a self-respondent.

only on the ability of the algorithms to identify dementia and consider participants classified as cognitive impairment no dementia or normal as "not demented."

## Training and Validation Datasets

Each algorithm was developed using a slightly different version of the HRS data (e.g., due to use of RAND versus core HRS data files, data-cleaning choices, and differences in dealing with missing data and eligibility criteria). To provide a fair comparison, we created standardized "training" and "validation" datasets containing HRS interview data and ADAMS diagnosis data in which to evaluate algorithm performance. Note that the validation dataset is an internal time-split validation dataset. Although we recognize that an external validation sample is preferable, we were unable to identify an external data source with sufficient overlap in measures. We used the RAND HRS data (Version P) for all variables except for proxy- and interviewer-reported cognitive data, which were not available in the RAND datasets and were extracted from the HRS core data files. Whenever available, we used RAND-derived summary variables (e.g., for ADLs) and we followed the RAND logic for computing change in ADL limitations to create variables summarizing change in cognition. The RAND datasets include imputed cognitive scores for self-respondents with missing cognitive data.[29] To address missing data in HRS proxy cognition measures for proxy respondents, we replaced missing HRS proxy cognition data with proxy scores from the HRS wave immediately prior, when available (affecting <0.1% of observations). We were unable to directly calculate the published Hurd formula due to missing information in the published description[6]; thus, we used Hurd dementia probabilities calculated by the study authors for HRS participants through the 2006 interview (publicly-available on the RAND web site).

Our standardized training dataset included ADAMS Wave A dementia diagnoses and the corresponding nearest prior HRS interview data from the 2000 or 2002 interviews from ADAMS Wave A participants; these data were commonly used across authors for algorithm development. Our standardized validation dataset included ADAMS Wave B, C, and D dementia diagnoses and corresponding nearest prior HRS interview data from the 2002, 2004, 2006, and/or 2008 HRS interviews from ADAMS participants who were evaluated at Waves B, C, or D. This is an internal validation dataset (i.e., a time-split sample); all of these participants were also Wave A participants, even though their data from Waves B, C, and D and the corresponding HRS interview data were not previously used in algorithm creation. Unfortunately, we were unable to identify a separate study that had sufficient coverage of the data to construct an external validation dataset on which to test the algorithms. Note that the ADAMS internal validation dataset may include up to three records from the same individual because ADAMS participants were followed longitudinally until dementia diagnosis or censoring. To ensure a fair comparison, we excluded observations that were missing data on any variable included in any algorithm, including missing precalculated Hurd probabilities; we additionally excluded observations (N = 26) from the validation dataset contributed by ADAMS participants who underwent repeat dementia ascertainment to confirm a previous diagnosis of dementia from a prior ADAMS wave (Figure 1). Of note, while the Wu algorithm was developed on a dataset that excluded Hispanic participants, we included Hispanic participants in our data and applied the Wu algorithm without alteration.

## Statistical Analyses

We applied each published algorithm to both the training and validation datasets to classify HRS participants as having or not having dementia proximal to a given HRS interview. For the H–W and L–K–W algorithms, we calculated the relevant summary scores and classified persons with scores outside the specified cutoffs as having dementia (Table 1). For the Wu and Crimmins algorithms, we used the published coefficients to calculate predicted probabilities of class membership, classifying those with a predicted dementia probability >0.5 as having dementia. For the Hurd algorithm, we classified persons as having dementia if the precalculated predicted probability of dementia was >0.5. These classifications were then compared with the ADAMS dementia diagnoses to compute classification accuracy (% correctly classified), sensitivity, and specificity, as well as the proportion of true positives, false positives, true negatives, and false negatives. We derived these measures separately in the training and validation datasets, overall and within prespecified subpopulations defined by age, sex, education, race/ethnicity, and respondent type (self- or proxy-respondent). We report both unweighted results, reflective of performance in the raw ADAMS sample (which oversampled cognitively impaired participants), and weighted results, reflective of performance in a nationally representative sample of the US population ≥70 years of age.[9] In addition, recognizing that the arbitrary selection of a 0.5 probability cutoff for the three regression-based algorithms affects sensitivity, specificity, and accuracy, we plotted receiver-operator curves (ROC) and computed the area under the curve (AUC).

We conducted several sensitivity analyses in the unweighted data. First, to investigate the susceptibility of algorithm development to small differences in eligibility criteria and data-cleaning choices, we re-estimated each regression-based algorithm in the standardized training dataset and then used the resulting predicted probabilities to classify dementia status and evaluate algorithm performance. Second, we used leave-one-out cross-validation in the training data when re-estimating each regression-based algorithm to estimate out-of-sample performance in identifying prevalent dementia. Third, we evaluated the performance of each algorithm in an alternate version of our validation sample allowing the contribution of data from prevalent dementia cases. Specifically, we included

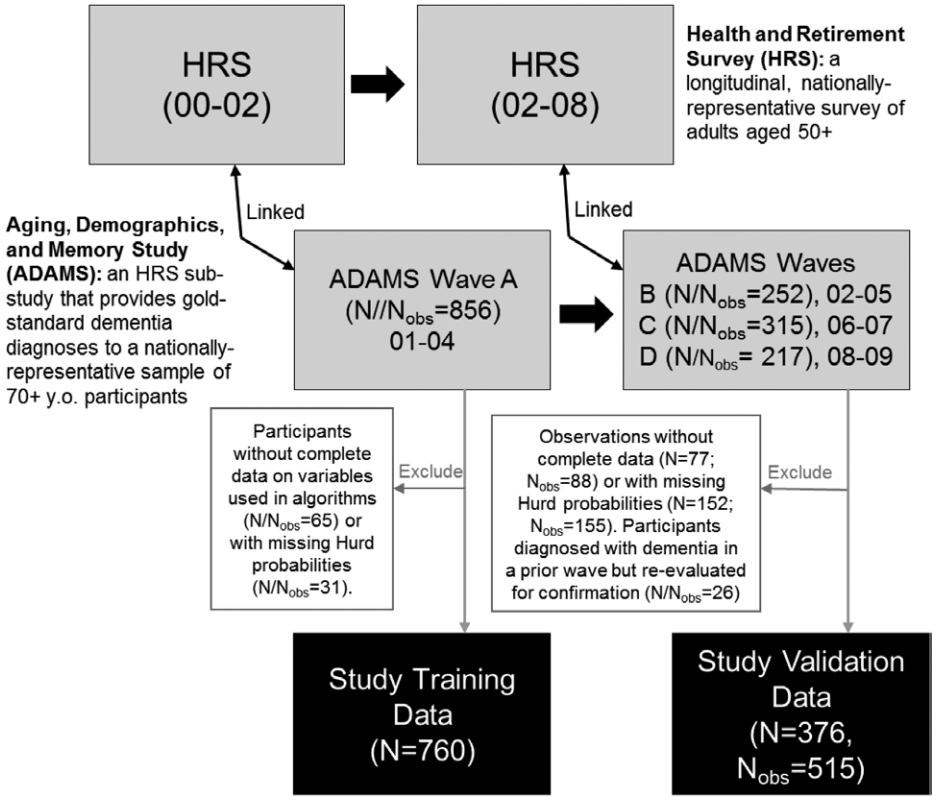


**FIGURE 1.** Flowchart showing derivation of our standardized training and validation HRS/ADAMS datasets.

observations from participants who were known to be alive at the time of ADAMS Waves B, C, or D, but who did not contribute to our primary validation data at those waves due to a dementia diagnosis in a prior wave (A, B, or C; see eFigure 1; http://links.lww.com/EDE/B433). This analysis addresses questions about whether differences in algorithm performance across the training and validation datasets were attributable to differences in ability to identify incident versus prevalent dementia cases. Fourth, to illustrate the potential impact of alternate cutpoints for the three regression-based algorithms that produce a probability of class membership, we re-evaluated performance based on alternate cut points chosen to achieve (1) 98%, 95%, or 90% sensitivity, (2) 98%, 95%, or 90% specificity, and (3) maximal accuracy. Finally, we used an alternate classification rule for the Crimmins algorithm whereby we only classify persons as having dementia if they had an estimated dementia probability greater than 50% and greater than estimated probability of cognitive impairment-no dementia.

We bootstrapped all of our analyses using 10,000 bootstrap samples, with the exception of the sensitivity analyses using leave-one-out cross-validation. We used the bootstrap percentile method to obtain point estimates (i.e., the median of the bootstrap statistic distribution) and 95% confidence intervals (i.e., the 2.5th and 97.5th percentiles of the bootstrap statistic distribution) for all performance metrics. We used a simple random selection process with replacement to construct training data bootstrap samples. For the validation data, we constructed bootstrap samples by selecting with replacement on unique individuals rather than observations to account for repeated measures.

This study was approved by the George Washington University Institutional Review Board. HRS and ADAMS participants provided informed consent at the time of data collection. We used SAS Version 9.4 and R Version 3.4.4. Code allowing recreation of our training and algorithmic datasets and assigning algorithmic diagnoses is available on GitHub (https://github.com/powerepilab/AD_algorithm_comparison).

## RESULTS

The predictors used in each algorithm are shown in Table 1. With the exception of the L–K–W algorithm, the algorithms rely on a similar set of self-response cognitive test scores. Proxy-respondent cognitive and functional data predictors vary by algorithm. Although the three regression-based algorithms all include age and sex, only Crimmins and Hurd include education, and only Wu includes race.

Our training dataset was larger than our validation dataset (training $N/N_{obs}$ = 760; validation N = 376/$N_{obs}$ = 515, Figure 1). Demographic, cognitive, and functional characteristics of each sample are provided in Table 2. Reflecting the design of ADAMS, which identified prevalent dementia at Wave A and incident dementia at Waves B, C, and D, use of proxy respondents, functional limitations, and dementia diagnosis is less common and cognitive tests scores are generally higher in the validation sample than in the training sample.

**TABLE 2.** Distribution of ADAMS Dementia Status and HRS Interview Predictor Variables in the Training and Validation Datasets

| Outcomes and Predictors | Mean (SD) or N (%) | |
|---|---|---|
| | Training Data ($N/N_{obs}$ = 760) | Validation Data ($N_{obs}$ = 515) |
| Dementia outcomes | | |
| Dementia status, N (%) | 258 (34) | 71 (14) |
| Dementia etiology: Alzheimer's Vascular, Lewy body, FTD, N (%) | 237 (31) | 64 (12) |
| Dementia etiology: other, N (%) | 21 (3) | 7 (1) |
| Demographics | | |
| Age, mean (SD) | 80.3 (7.0) | 81.2 (5.8) |
| Proxy respondent, N (%) | 165 (22) | 30 (6) |
| Female, N (%) | 452 (59) | 270 (52) |
| Education | | |
| Less than high school, N (%) | 369 (49) | 215 (42) |
| High school or GED, N (%) | 298 (39) | 222 (43) |
| Some college or more, N (%) | 93 (12) | 78 (15) |
| Race/ethnicity | | |
| Non-Hispanic White, N (%) | 526 (69) | 369 (72) |
| Non-Hispanic Black, N (%) | 140 (18) | 97 (19) |
| Hispanic, N (%) | 76 (10) | 35 (7) |
| Non-Hispanic other race, N (%) | 18 (2) | 14 (3) |
| Cognition (self-response) | | |
| Immediate word recall, 0–10, mean (SD) | 3.9 (1.8) | 4.4 (1.6) |
| Delayed word recall, 0–10, mean (SD) | 2.6 (2.1) | 3.0 (1.9) |
| Serial 7's, 0–5, mean (SD) | 2.4 (1.9) | 2.8 (1.9) |
| Dates, 0–4, mean (SD) | 3.4 (1.0) | 3.6 (0.7) |
| Object naming: cactus, N (%) | 451 (76) | 404 (83) |
| Object naming: scissors, N (%) | 587 (99) | 478 (99) |
| President, N (%) | 518 (87) | 454 (94) |
| Vice-president, N (%) | 319 (54) | 318 (66) |
| Backwards counting | | |
| Incorrect (0), N (%) | 92 (15) | 55 (11) |
| Correct on first or second attempt (1 or 2), N (%) | 503 (85) | 430 (88) |
| Cognition (proxy) | | |
| Interviewer assessment of cognitive limitations | | |
| No cognitive limitations, N (%) | 28 (17) | 13 (43) |
| Some cognitive limitations, N (%) | 25 (15) | 8 (27) |
| Cognitive limitations prevents completion of interview, N (%) | 112 (68) | 9 (30) |
| Proxy-rated memory score, 1 (excellent)–5 (poor), mean (SD) | 4.3 (1.0) | 3.5 (1.0) |
| 16-item Jorm IQCODE, 1 (much improved)–5 (much worse), mean (SD) | 4.2 (0.7) | 3.4 (0.5) |
| Jorm symptoms, 2004 onwards, 0–5, mean (SD) | 1.8 (1.5) | 0.6 (1.0) |
| Physical functioning limitations | | |
| ADLs, 0–5, mean (SD) | 1.0 (1.5) | 0.6 (1.1) |
| IADLs, 0–5, mean (SD) | 1.2 (1.8) | 0.5 (1.1) |

FTD indicates frontotemporal dementia; GED, general educational development.

In the unweighted training data, overall sensitivity ranged from 53% to 90%, specificity ranged from 79% to 97%, and overall accuracy ranged from 81% to 87% across the five algorithms (Table 3). Overall accuracy was similar in the unweighted validation data (range: 79%–88%); however, this was largely driven by slightly higher specificities (82%–98%), as sensitivity was much lower (range: 18%–62%). In both the training and validation datasets, the H–W algorithm had the highest specificity, the Crimmins algorithm had the highest sensitivity, and the Wu and Hurd algorithms had the highest overall accuracy based on point estimates; however, confidence intervals for these metrics frequently overlap across

**TABLE 3.** Performance Metrics in the Training and Validation Data, Overall and by Subgroups

| | Training Data | | | Validation Data | | |
|---|---|---|---|---|---|---|
| **Algorithm** | **Sensitivity % (95% CI)** | **Specificity % (95% CI)** | **Accuracy % (95% CI)** | **Sensitivity % (95% CI)** | **Specificity % (95% CI)** | **Accuracy % (95% CI)** |
| Overall | | $N/N_{obs}$ = 760 | | | N = 376, $N_{obs}$ = 515 | |
| H–W | 53 (47, 60) | 97 (95, 98) | 82 (79, 85) | 18 (10, 28) | 98 (96, 99) | 87 (84, 90) |
| L–K–W | 75 (70, 80) | 83 (80, 87) | 81 (78, 83) | 41 (30, 52) | 89 (86, 92) | 83 (79, 86) |
| Crimmins | 90 (86, 93) | 79 (75, 83) | 83 (80, 85) | 62 (51, 73) | 82 (78, 86) | 79 (75, 83) |
| Hurd | 77 (72, 82) | 92 (89, 94) | 87 (84, 89) | 39 (28, 51) | 96 (94, 98) | 88 (85, 91) |
| Wu | 78 (73, 83) | 88 (85, 91) | 85 (82, 87) | 44 (32, 55) | 93 (90, 95) | 86 (83, 89) |
| By respondent status | | | | | | |
| Self | | $N/N_{obs}$ = 595 | | | N = 351, $N_{obs}$ = 485 | |
| H–W | 29 (21, 38) | 97 (96, 99) | 83 (80, 86) | 16 (7, 26) | 98 (97, 100) | 89 (86, 92) |
| L–K–W | 58 (49, 67) | 85 (81, 88) | 79 (76, 82) | 36 (23, 49) | 90 (87, 93) | 84 (80, 87) |
| Crimmins | 83 (77, 90) | 82 (78, 85) | 82 (79, 85) | 61 (48, 73) | 84 (80, 88) | 81 (77, 85) |
| Hurd | 57 (48, 66) | 93 (91, 96) | 86 (83, 88) | 36 (23, 48) | 96 (94, 98) | 89 (86, 92) |
| Wu | 62 (53, 70) | 91 (88, 93) | 85 (82, 87) | 36 (23, 49) | 94 (91, 96) | 87 (84, 90) |
| Proxy | | $N/N_{obs}$ = 165 | | | N = 25, $N_{obs}$ = 30 | |
| H–W | 77 (69, 83) | 85 (72, 96) | 78 (72, 84) | 26 (6, 50) | 81 (53, 100) | 53 (35, 72) |
| L–K–W | 92 (87, 96) | 64 (47, 80) | 86 (81, 91) | 60 (33, 85) | 67 (40, 93) | 64 (45, 81) |
| Crimmins | 96 (93, 99) | 36 (20, 53) | 84 (78, 90) | 67 (41, 90) | 33 (9, 62) | 50 (32, 69) |
| Hurd | 96 (92, 99) | 70 (53, 85) | 90 (86, 95) | 54 (27, 80) | 88 (69, 100) | 70 (55, 85) |
| Wu | 93 (89, 97) | 48 (31, 66) | 84 (79, 90) | 74 (50, 94) | 54 (27, 82) | 64 (45, 81) |
| By race/ethnicity | | | | | | |
| NH White | | $N/N_{obs}$ = 526 | | | N = 271, $N_{obs}$ = 369 | |
| H–W | 52 (45, 60) | 99(99, 100) | 84 (81, 87) | 13 (4, 24) | 99 (97, 100) | 88 (85, 92) |
| L–K–W | 72 (65, 78) | 91 (88, 94) | 85 (81, 88) | 36 (22, 51) | 94 (91, 97) | 87 (83, 91) |
| Crimmins | 87 (82, 92) | 84 (80, 88) | 85 (82, 88) | 62 (47, 76) | 84 (79, 89) | 82 (77, 86) |
| Hurd | 73 (67, 80) | 94 (91, 96) | 87 (84, 90) | 45 (31, 60) | 97 (95, 98) | 91 (87, 93) |
| Wu | 77 (71, 83) | 94 (91, 96) | 88 (85, 91) | 43 (29, 58) | 95 (92, 98) | 89 (85, 92) |
| NH Black | | $N/N_{obs}$ = 140 | | | N = 65, $N_{obs}$ = 97 | |
| H–W | 55 (42, 68) | 92 (85, 97) | 77 (69, 83) | 26 (6, 50) | 98 (94, 100) | 87 (79, 93) |
| L–K–W | 81 (71, 91) | 61 (50, 71) | 69 (61, 77) | 53 (27, 79) | 77 (66, 87) | 73 (64, 82) |
| Crimmins | 97 (91, 100) | 65 (54, 75) | 78 (71, 84) | 60 (33, 85) | 79 (69, 89) | 76 (67, 85) |
| Hurd | 87 (77, 95) | 83 (74, 91) | 84 (78, 90) | 26 (6, 50) | 94 (87, 99) | 84 (75, 91) |
| Wu | 74 (63, 85) | 79 (70, 88) | 77 (70, 84) | 40 (15, 67) | 86 (76, 94) | 79 (68, 88) |
| Hispanic | | $N/N_{obs}$ = 76 | | | N = 31, $N_{obs}$ = 35 | |
| H–W | 55 (33, 77) | 88 (78, 96) | 79 (70, 88) | 20 (0, 50) | 96 (86, 100) | 74 (57, 89) |
| L–K–W | 86 (67, 100) | 71 (59, 83) | 75 (65, 84) | 40 (10, 75) | 76 (61, 91) | 66 (50, 80) |
| Crimmins | 95 (83, 100) | 71 (59, 83) | 78 (68, 86) | 60 (27, 91) | 76 (59, 92) | 71 (56, 86) |
| Hurd | 86 (67, 100) | 89 (81, 97) | 88 (80, 95) | 29 (0, 63) | 96 (86, 100) | 77 (61, 91) |
| Wu | 90 (74, 100) | 68 (55, 80) | 74 (63, 84) | 50 (17, 83) | 93 (81, 100) | 80 (66, 92) |
| By sex | | | | | | |
| Male | | $N/N_{obs}$ = 308 | | | N = 182, $N_{obs}$ = 255 | |
| H–W | 42 (31, 53) | 95 (92, 97) | 81 (76, 85) | 17 (4, 31) | 96 (93, 99) | 87 (82, 91) |
| L–K–W | 68 (57, 78) | 84 (79, 88) | 80 (75, 84) | 43 (26, 62) | 88 (82, 92) | 82 (77, 87) |
| Crimmins | 84 (76, 92) | 81 (75, 86) | 82 (77, 86) | 60 (42, 78) | 83 (77, 89) | 80 (75, 86) |
| Hurd | 68 (57, 78) | 92 (88, 95) | 85 (81, 89) | 36 (20, 55) | 95 (92, 98) | 88 (84, 92) |
| Wu | 63 (52, 73) | 90 (86, 94) | 83 (78, 87) | 33 (17, 51) | 93 (89, 97) | 86 (81, 90) |
| Female | | $N/N_{obs}$ = 452 | | | N = 194, $N_{obs}$ = 270 | |
| H–W | 59 (52, 66) | 98 (96, 100) | 83 (79, 86) | 19 (8, 33) | 99 (98, 100) | 87 (83, 91) |
| L–K–W | 79 (72, 84) | 83 (78, 87) | 81 (78, 85) | 39 (25, 55) | 91 (87, 95) | 83 (78, 88) |
| Crimmins | 93 (89, 96) | 78 (73, 83) | 84 (80, 87) | 64 (48, 78) | 81 (75, 87) | 79 (73, 84) |

(*Continued*)

**TABLE 3.** (*Continued*)

| Algorithm | Training Data | | | Validation Data | | |
|---|---|---|---|---|---|---|
| | Sensitivity % (95% CI) | Specificity % (95% CI) | Accuracy % (95% CI) | Sensitivity % (95% CI) | Specificity % (95% CI) | Accuracy % (95% CI) |
| Hurd | 81 (75, 86) | 92 (89, 95) | 88 (85, 91) | 41 (27, 57) | 97 (94, 99) | 88 (84, 92) |
| Wu | 85 (79, 90) | 87 (82, 90) | 86 (83, 89) | 51 (36, 67) | 92 (88, 96) | 86 (81, 90) |
| By age | | | | | | |
| <80 | | $N/N_{obs}$ = 364 | | | N = 170, $N_{obs}$ = 225 | |
| H–W | 53 (41, 65) | 98 (96, 99) | 90 (86, 93) | 11 (0, 29) | 98 (95, 100) | 91 (87, 95) |
| L–K–W | 78 (68, 88) | 86 (82, 90) | 85 (81, 88) | 29 (8, 54) | 91 (86, 95) | 86 (81, 91) |
| Crimmins | 81 (71, 90) | 91 (87, 94) | 89 (85, 92) | 41 (18, 67) | 92 (88, 95) | 88 (83, 92) |
| Hurd | 59 (47, 71) | 97 (95, 99) | 90 (87, 93) | 17 (0, 38) | 100 (98, 100) | 93 (90, 96) |
| Wu | 65 (53, 76) | 93 (90, 96) | 88 (84, 91) | 35 (13, 60) | 95 (92, 98) | 91 (87, 94) |
| 80+ | | $N/N_{obs}$ = 396 | | | N = 206, $N_{obs}$ = 290 | |
| H–W | 54 (47, 61) | 95 (91, 98) | 75 (71, 79) | 20 (10, 32) | 98 (96, 100) | 84 (79, 88) |
| L–K–W | 74 (68, 80) | 79 (73, 84) | 77 (72, 81) | 44 (31, 58) | 88 (83, 92) | 80 (75, 84) |
| Crimmins | 93 (89, 97) | 63 (56, 69) | 77 (73, 81) | 69 (56, 81) | 74 (67, 80) | 73 (67, 78) |
| Hurd | 83 (78, 88) | 84 (79, 89) | 84 (80, 87) | 46 (33, 60) | 93 (89, 96) | 84 (80, 88) |
| Wu | 83 (77, 88) | 81 (75, 86) | 82 (78, 85) | 46 (33, 60) | 90 (86, 94) | 82 (77, 87) |
| By education | | | | | | |
| <High school | | $N/N_{obs}$ = 369 | | | N = 157, $N_{obs}$ = 215 | |
| H–W | 50 (42, 59) | 93 (89, 96) | 76 (72, 81) | 18 (8, 31) | 95 (91, 98) | 80 (74, 85) |
| L–K–W | 79 (72, 85) | 70 (64, 76) | 73 (69, 78) | 47 (32, 62) | 77 (70, 84) | 71 (64, 78) |
| Crimmins | 90 (85, 95) | 68 (61, 74) | 76 (72, 80) | 60 (46, 75) | 72 (64, 80) | 70 (63, 77) |
| Hurd | 75 (68, 82) | 87 (82, 91) | 82 (78, 86) | 37 (23, 52) | 95 (91, 98) | 83 (78, 88) |
| Wu | 81 (74, 87) | 78 (72, 83) | 79 (75, 83) | 47 (31, 62) | 86 (79, 91) | 78 (71, 84) |
| High school + | | $N/N_{obs}$ = 391 | | | N = 219, $N_{obs}$ = 300 | |
| H–W | 57 (48, 66) | 100 (100, 100) | 87 (84, 90) | 17 (4, 33) | 100 (99, 100) | 92 (89, 95) |
| L–K–W | 71 (63, 79) | 95 (92, 97) | 87 (84, 91) | 32 (15, 50) | 97 (94, 99) | 91 (87, 94) |
| Crimmins | 90 (84, 95) | 89 (85, 92) | 89 (86, 92) | 64 (46, 82) | 89 (84, 93) | 86 (82, 90) |
| Hurd | 79 (71, 86) | 96 (94, 98) | 91 (88, 94) | 43 (24, 62) | 97 (94, 99) | 92 (88, 95) |
| Wu | 74 (66, 82) | 96 (94, 98) | 90 (87, 93) | 39 (21, 58) | 97 (95, 99) | 92 (88, 95) |

algorithms. Figure 2 details the proportions of correctly and incorrectly classified persons by ADAMS dementia status, algorithm, and dataset.

Though the overall pattern of performance across algorithms observed above largely held in our unweighted analyses after stratifying by respondent status, sex, age, education, and race/ethnicity, performance appeared to differ substantially across these subgroups, although confidence intervals often overlap (Table 3). Proxy respondents typically had higher sensitivity and uniformly lower specificity compared with self-respondents in both datasets. In the validation sample, proxy respondents also had substantially lower overall accuracy compared with self-respondents. Non-Hispanic black and Hispanic participants had similar or higher sensitivity, but lower specificity and overall accuracy in the training data compared with non-Hispanic whites. Differences in sensitivity and specificity by race/ethnicity were less consistent in the validation data; however, overall accuracy generally remained lower in minorities compared with that in non-Hispanic whites. With the exception of lower sensitivities for men in the training data, performance metrics were similar across genders. Those >80 years of age had higher sensitivity, but lower specificity and overall accuracy compared with younger participants in the three regression-based algorithms in both datasets. For H–W and L–K–W, classification accuracy of older adults was lower than that for younger adults in both datasets, driven by higher frequency of dementia rather than big differences in sensitivity and specificity. Specificity and overall accuracy were higher within the group with at least a high school education when compared with those without in both datasets.

After weighing the sample to obtain estimates applicable to the US age-eligible population, the relative patterns of performance across algorithms (Table 4) and across subgroups (eTable 1; http://links.lww.com/EDE/B433) were generally consistent with the patterns observed in the unweighted data. However, overall accuracy was higher across algorithms after weighting due to both a lower frequency of dementia and higher achieved specificity in the weighted data.

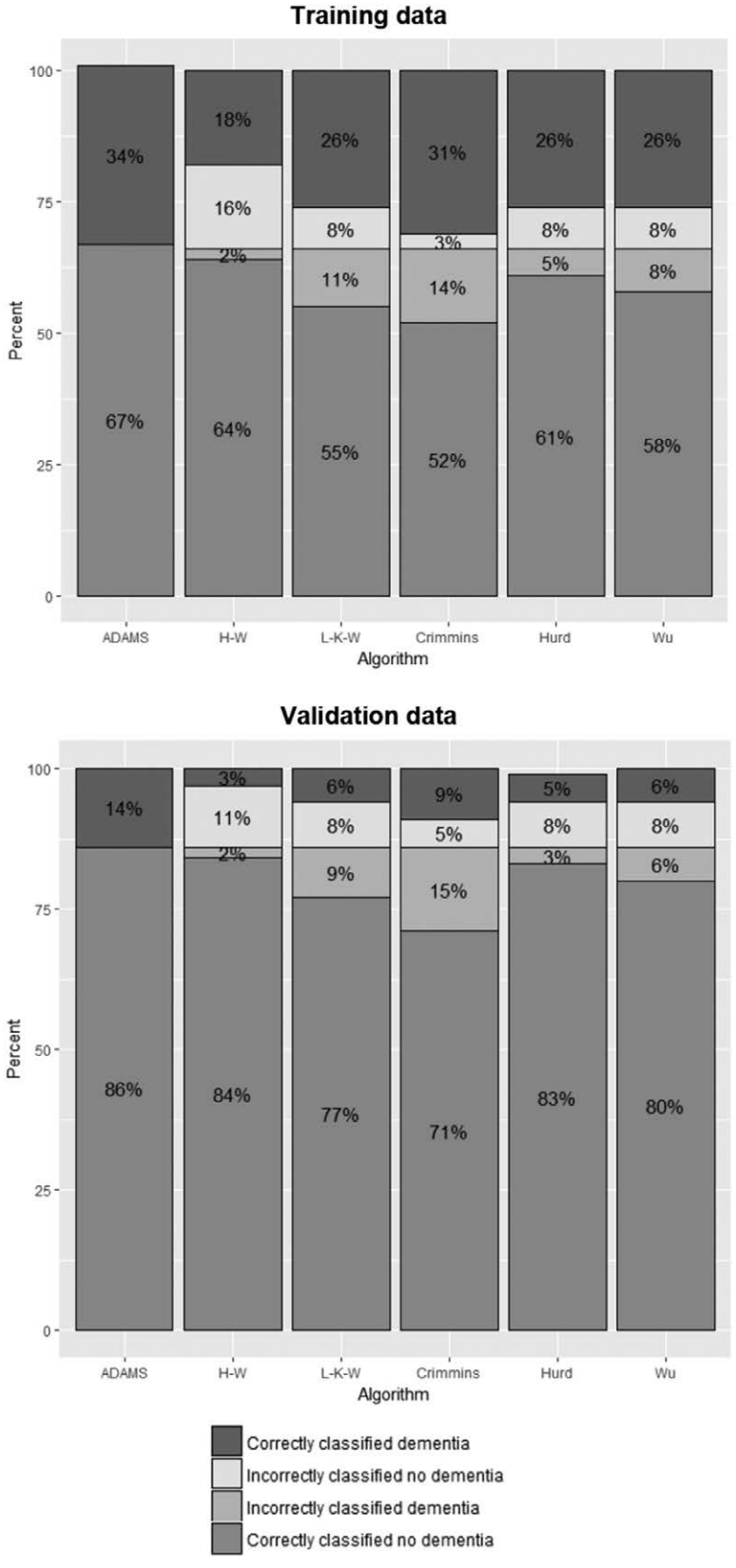


**FIGURE 2.** Proportion of correctly and incorrectly classified observations by ADAMS dementia status, algorithm, and dataset in unweighted analyses.

**TABLE 4.** Performance Metrics in the Training and Validation Data Overall, After Weighting to Obtain Estimates Applicable to the US Age-eligible Population

| | Training Data | | | Validation Data | | |
|---|---|---|---|---|---|---|
| **Algorithm** | **Sensitivity % (95% CI)** | **Specificity % (95% CI)** | **Accuracy % (95% CI)** | **Sensitivity % (95% CI)** | **Specificity % (95% CI)** | **Accuracy % (95% CI)** |
| Overall | | $N/N_{obs}$ = 760 | | | N = 376, $N_{obs}$ = 515 | |
| H–W | 42 (33, 52) | 99 (98, 100) | 92 (90, 94) | 13 (3, 26) | 100 (99, 100) | 93 (91, 95) |
| L–K–W | 57 (47, 67) | 97 (95, 97) | 91 (89, 93) | 24 (13, 38) | 98 (97, 99) | 92 (89, 94) |
| Crimmins | 78 (68, 87) | 93 (91, 95) | 91 (89, 93) | 39 (25, 56) | 91 (87, 95) | 87 (83, 91) |
| Hurd | 65 (56, 74) | 98 (98, 99) | 94 (92, 96) | 26 (15, 40) | 99 (98, 100) | 93 (91, 96) |
| Wu | 64 (54, 73) | 97 (96, 98) | 93 (91, 95) | 35 (20, 51) | 98 (97, 99) | 94 (91, 96) |

The performance of the three re-estimated regression-based algorithms was consistent with our primary analyses (eTable 2; http://links.lww.com/EDE/B433), suggesting that differences in eligibility criteria or variable coding had minimal effect. Reassuringly, performance remained roughly similar in our training-data leave-one-out cross-validation analyses (eTable 3; http://links.lww.com/EDE/B433), which estimate the expected out-of-sample performance, suggesting that overfitting is not a substantial issue. Performance in the alternative validation data including prevalent dementia cases was generally consistent with performance in the training data, suggesting differences in ability of the algorithms to identify prevalent versus incident cases drives differences in performance across the training and validation datasets (eTable 4; http://links.lww.com/EDE/B433).

Although there are substantial differences in sensitivity, specificity, and accuracy across the three regression-based algorithms in each dataset when using an arbitrary 0.5 probability cutoff, the AUCs from ROC analyses of each algorithm are comparable (Table 5). The ROC curves for the three algorithms in eFigure 2 (http://links.lww.com/EDE/B433) also demonstrate this, as the curves for the three algorithms are close within each dataset, but the use of a 0.5 cutpoint selects for different sensitivities and specificities for each algorithm. Sensitivity was lower whereas specificity was higher for the Crimmins algorithm when applying our alternate classification rule (eTable 5; http://links.lww.com/EDE/B433). Alternate cutpoints that achieve prespecified metrics (i.e., sensitivity or specificity of 90%, 95%, or 98%, or maximal accuracy), as well as the corresponding performance metrics associated with the use of these cutoffs in the unweighted data, are presented in eTable 6 (http://links.lww.com/EDE/B433). Due to the small number of dementia cases in the validation data (N = 71), we were unable to attain very high sensitivities (98% and 95%) with precision. Although Hurd generally performs best at these prespecified cutoffs when considering point estimates, the advantage is small and confidence intervals for the corresponding sensitivities and specificities across algorithms frequently overlap. Cutoffs producing maximal accuracy in the unweighted validation data (range: 88%–89%) uniformly maximize specificity (97%–99%) at the expense of sensitivity (27%–28%).

**TABLE 5.** ROC AUCs of Regression-based Algorithms

| | **Training ($N/N_{obs}$ = 760) AUC (95% CI)** | **Validation Data (N = 376, $N_{obs}$ = 515) AUC (95% CI)** |
|---|---|---|
| Crimmins | 0.92 (0.90, 0.94) | 0.84 (0.80, 0.88) |
| Hurd | 0.94 (0.93, 0.96) | 0.85 (0.81, 0.89) |
| Wu | 0.93 (0.91, 0.94) | 0.84 (0s.79, 0.88) |

## DISCUSSION

This article provides a head-to-head comparison of existing algorithms for classifying dementia status in HRS participants. Generally, H–W maximized specificity, Crimmins maximized sensitivity, and Wu and Hurd maximized accuracy. As expected, the most sensitive algorithms tended to be the least specific, resulting in a relatively narrow range for overall accuracy in the full samples. However, performance varied substantially across subgroups. Overall accuracy of algorithmic diagnoses was uniformly worse among race/ethnic minorities, older adults, and less-educated participants when compared with their complements, driven by substantial differences in sensitivity and specificity across subgroups. Overall accuracy also varied by respondent type, which has implications for estimating differences in dementia prevalence or incidence across groups with unequal proportions of proxy respondents. Accuracy was higher after weighting the sample to be representative of the age-eligible US population, due to higher specificity from up-weighting those with better cognition and a lower overall frequency of dementia in the weighted sample. However, even in the weighted analyses, substantial differences in performance across subgroups remained. When considering alternate cutoffs chosen to achieve high sensitivity, specificity, or overall accuracy, the three regression-based algorithms performed similarly.

While overall accuracy achieved in our weighted samples may sufficiently justify further use of these algorithms in HRS or other age-eligible US-representative samples to estimate overall incidence or prevalence of dementia, the substantial differences in performance across subgroups argue against

the naïve use in disparities or risk factor studies. Of most concern, when there are differences in classification accuracy in dementia by the exposure of interest, these algorithms may introduce nonconservative differential misclassification. We have demonstrated this to be the case in disparities research, which requires identifying and understanding differences in dementia incidence or prevalence across sociodemographic characteristics. We caution that similar differences in performance metrics are also likely to be observed across subgroups defined by other risk factors of interest and that these differences may persist even after adjustment for sociodemographic characteristics, thereby introducing issues of bias. Thus, further analysis-specific methodologic work to address this known source of bias is necessary for using any of these algorithms in disparities or risk factor epidemiology.

Another feature to note is that algorithm performance differs in identifying prevalent dementia (i.e., as in the training and alternate validation data) versus incident dementia (i.e., as in the validation data). While overall accuracy was similar, sensitivity was uniformly lower and specificity was often higher in identifying incident dementia. Dementia diagnosis is not the result of a dramatic change, but rather passing a threshold on a continuum of cognitive and functional ability; thus, it should not come as a surprise that it is more difficult to correctly identify incident dementia (i.e., new and less severe) than prevalent dementia (i.e., established and more severe).

The algorithms differ in practical ways. The cognitive score cutoff-based algorithms were the most straightforward to apply. Variable missingness varies across predictors in the full HRS sample, thus restricting to those with available data for any given algorithm may limit sample size, impair generalizability, and possibly induce selection bias. Finally, different algorithms may be preferred if the goal is to maximize sensitivity versus to maximize specificity or accuracy.

Our study has several strengths. To our knowledge, it is the first to provide a head-to-head comparison of existing algorithms created to classify dementia status in the nationally representative HRS cohort. We consider performance in both a training and validation dataset and conducted sensitivity analyses to provide context for observed differences. Our study also has limitations. The conclusiveness of our subgroup analyses is limited given the large overlapping confidence intervals resulting from the small number of observations in some strata, particularly in the validation dataset. Our methods assume no error in the ADAMS dementia diagnoses and that the relation between our predictors and dementia status is invariant across time. The training dataset included prevalent cases whereas the validation dataset included incident cases. Our validation dataset is comprised of repeated measures from the same participants used in the training data. Although we attempted to identify an external validation sample (e.g., the Atherosclerosis in Risk [ARIC] study, and Rush University Memory and Aging Project, Minority Aging Research Study, and Religious Orders Study [MAP, MARS, and ROS] cohorts), we were unable to find a study with sufficient coverage of the data needed to apply the existing algorithms; thus, we proceeded without an external dataset to avoid false conclusions related to our inability to differentiate between differences attributable to the algorithms from those created by using questionable surrogate data. However, HRS sister studies and the HRS-linked Healthy Cognitive Aging Project (HCAP) may offer future opportunities to assess algorithm performance. Given the lack of overlap in predictor variables across existing cohort studies with dementia diagnosis, algorithm development will likely need to remain cohort specific in the absence of comparable data and harmonization across cohorts.

Currently, only a handful of cohorts have systematic dementia ascertainment. Notably, race/ethnic minorities, those who live in rural areas, and those of lower socioeconomic status are under-represented in the available data. Thus, algorithmic assessment may provide a cost-effective way to expand the number of data sources that can meaningfully contribute to dementia research. However, algorithms developed for other purposes may not be ideal for disparities or risk factor research. Efforts to develop new algorithms may want to prioritize comparable performance across subgroups of interest.

## REFERENCES


1. Bennett DA, Schneider JA, Aggarwal NT, et al. Decision rules guiding the clinical diagnosis of Alzheimer's disease in two community-based cohort studies compared to standard practice in a clinic-based cohort study. *Neuroepidemiology*. 2006;27:169–176.
2. Bachman DL, Wolf PA, Linn R, et al. Prevalence of dementia and probable senile dementia of the Alzheimer type in the Framingham Study. *Neurology*. 1992;42:115–119.
3. Gottesman RF, Albert MS, Alonso A, et al. Associations between midlife vascular risk factors and 25-year incident dementia in the Atherosclerosis Risk in Communities (ARIC) cohort. *JAMA Neurol*. 2017;74:1246–1254.
4. 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.
5. Alzheimer's Association. 2010 Alzheimer's disease facts and figures. *Alzheimer's Dement*. 2010;6:158–194.
6. Hurd MD, Martorell P, Delavande A, Mullen KJ, Langa KM. Monetary costs of dementia in the United States. *N Engl J Med*. 2013;368: 1326–1334.
7. 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. *J Gerontol B Psychol Sci Soc Sci*. 2011;66(suppl 1):i162–i171.
8. 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: 207–212.
9. Heeringa SG, Fisher GG, Hurd M, et al. Aging, Demographics and Memory Study (ADAMS): sample design, weighting and analysis for ADAMS. http://hrsonline.isr.umich.edu/sitedocs/userg/ADAMSSampleWeights_Jun2009.pdf. Published 2009. Accessed 2 January 2018.
10. Langa KM, Plassman BL, Wallace RB, et al. The Aging, Demographics, and Memory Study: study design and methods. *Neuroepidemiology*. 2005;25:181–191.
11. Langa KM, Chernew ME, Kabeto MU, et al. National estimates of the quantity and cost of informal caregiving for the elderly with dementia. *J Gen Intern Med*. 2001;16:770–778.

12. Langa KM, Larson EB, Karlawish JH, et al. Trends in the prevalence and mortality of cognitive impairment in the United States: is there evidence of a compression of cognitive morbidity? *Alzheimers Dement*. 2008;4:134–144.
13. Langa KM, Larson EB, Crimmins EM, et al. A comparison of the prevalence of dementia in the United States in 2000 and 2012. *JAMA Intern Med*. 2017;177:51–58.
14. Gure TR, Blaum CS, Giordani B, et al. Prevalence of cognitive impairment in older adults with heart failure. *J Am Geriatr Soc*. 2012;60:1724–1729.
15. Donovan NJ, Wu Q, Rentz DM, Sperling RA, Marshall GA, Glymour MM. Loneliness, depression and cognitive function in older U.S. adults. *Int J Geriatr Psychiatry*. 2017;32:564–573.
16. Suemoto CK, Gilsanz P, Mayeda ER, Glymour MM. Body mass index and cognitive function: the potential for reverse causation. *Int J Obes (Lond)*. 2015;39:1383–1389.
17. Rist PM, Capistrant BD, Wu Q, Marden JR, Glymour MM. Dementia and dependence: do modifiable risk factors delay disability? *Neurology*. 2014;82:1543–1550.
18. Wu Q, Tchetgen Tchetgen EJ, Osypuk T, et al. Estimating the cognitive effects of prevalent diabetes, recent onset diabetes, and the duration of diabetes among older adults. *Dement Geriatr Cogn Disord*. 2015;39:239–249.
19. Wang Q, Mejía-Guevara I, Rist PM, Walter S, Capistrant BD, Glymour MM. Changes in memory before and after stroke differ by age and sex, but not by race. *Cerebrovasc Dis*. 2014;37:235–243.
20. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort Profile: the Health and Retirement Study (HRS). *Int J Epidemiol*. 2014;43:576–585.
21. Health and Retirement Study. Produced and Distributed by the University of Michigan with Funding from the National Institute on Aging (Grant Number U01AG009740). Ann Arbor, MI.
22. Brandt J, Spencer M, Folstein M. The telephone interview for cognitive status. *Neuropsychiatry Neuropsychol Behav Neurol*. 1998;1:111–117.
23. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*. 1975;12:189–198.
24. Jorm AF. Disability in dementia: assessment, prevention, and rehabilitation. *Disabil Rehabil*. 1994;16:98–109.
25. Jorm AF. *Short Form of the Informant Questionnaire on Cognitive Decline in the Elderly (Short IQCODE*). Centre for Mental Health Research, The Australian National University, Canberra: Australia. Available at: https://www.alz.org/documents_custom/shortiqcode_english.pdf. Accessed 9 February 2018.
26. Jorm AF. The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): a review. *Int Psychogeriatrics*. 2004;16:1–19.
27. Plassman BL, Langa KM, Fisher GG, et al. Prevalence of cognitive impairment without dementia in the United States. *Ann Intern Med*. 2008;148:427–434.
28. Plassman BL, Langa KM, McCammon RJ, et al. Incidence of dementia and cognitive impairment, not dementia in the United States. *Ann Neurol*. 2011;70:418–426.
29. Fisher GG, Hassan H, Faul JD, Rodgers WL, Weir DR. *Health and Retirement Study Imputation of Cognitive Functioning Measures: 1992–2014 (Final Release Version)*. Ann Arbor, MI; 2017.