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

Authors and affiliations

Kan Z. Gianattasio1, Qiong Wu2, M. Maria Glymour3, Melinda C. Power1

1Department of Epidemiology and Biostatistics, Milken Institute School of Public Health, George Washington University

2Institute of Social Science Survey, Peking University

3Department of Epidemiology and Biostatistics, University of California, San Francisco

**FUNDING**

KZG and MCP were funded by R03 AG055485

**DISCLOSURES**

KZG and MCP have nothing to disclose

**AUTHORSHIP**

**ABSTRACT**

**INTRODUCTION**

Dementia ascertainment is time-consuming and costly, making it difficult to implement in large, representative cohort studies. This fact hinders efforts to use the results from large population surveys to describe and monitor trends and disparities in the prevalence and incidence of cognitive impairment. Recognizing this, several groups of researchers have developed algorithms to use existing data from the large and nationally-representative Health and Retirement Study (HRS) to algorithmically classify dementia status in cohort participants.1–5 HRS provides an ideal setting for algorithm development. First, a strategically selected subset of participants from HRS were evaluated for dementia at four time points between 2001 and 2009 as part of the Aging, Demographics, and Memory Study (ADAMS).6,7 Thus, data from ADAMS provides 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 national level, and to examine the associations of cognitive impairment and dementia with other health factors.8–16

Researchers hoping to use such algorithms in their studies 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. In addition, the algorithm with the best metrics when applied to the data used to create the algorithm may not be the best algorithm to apply widely because of overfitting. Few reports on the existing algorithms provide performance metrics achieved when the algorithm is applied to data other than the training data. Moreover, the algorithm with the best overall performance metrics may be ill-suited to desired applications, particularly efforts to describe disparities, if performance differs across sub-populations. Currently, it is unclear whether there are big differences in the performance of existing algorithms across sub-groups defined by race/ethnicity, age, sex, or education.

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 these algorithms – HRS participants who underwent dementia ascertainment at ADAMS Wave A. We then compared performance metrics across algorithms when applied to data not used for algorithm development – HRS participants who underwent dementia ascertainment at ADAMS Waves B, C, and/or D. At each stage, we also quantified differences in performance metrics across sociodemographic groups. Finally, we conclude with a discussion of when use of each algorithm may be preferred, and recommendations for future algorithm development.

**METHODS**

**Data Sources**

The HRS is a nationally representative, longitudinal study of adults ages 50 and older and their spouses.17 Since enrollment of the original HRS cohort in 1992, several waves of new enrollment have been completed in order to maintain a steady-state cohort of U.S. adults over 50. Since 1998, HRS participants have completed interviews every two years, either by telephone or in person. Relevant to this study, data on sociodemographic and socioeconomic characteristics, functional status, and cognitive status are collected at each interview. The HRS administers cognitive assessments using items from the Telephone Interview for Cognitive Status (TICS) and the Mini-Mental State Examination (MMSE) to those who are sufficiently functional to be interviews. Proxy respondents minimize loss to follow-up and also provide data on cognitive and functional status on participants who are not sufficiently functional to complete the interview.

ADAMS is an HRS sub-study that included systematic dementia ascertainment according to standard criteria.7 HRS participants aged ≥70 years who completed the 2000 or 2002 HRS questionnaires (including those completed via proxy interviews) were sampled for inclusion in ADAMS using a stratified random sampling approach. Ultimately, 856 HRS participants were enrolled and completed initial assessment (Wave A, 2001-2003).18 Three additional waves of data collection that included dementia ascertainment (Waves B, C, and D) were completed through 2009; those diagnosed with dementia at each wave exited the ADAMS.19 The N’s at each wave are indicated in **Figure 1**. Dementia diagnosis was assigned at each wave according to DSM-III-R and DSM-IV criteria and confirmed by a consensus expert panel.7,18,19

**Figure 1: ADAMS sample flow chart**



This study was approved by the Institutional Review Board at the George Washington University.

**Existing Algorithms**

We identified existing algorithms for predicting dementia or cognitive status using core interview data from HRS through a combination of pre-existing knowledge, informal searches of PubMed and Scopus, review of articles cited by each identified manuscript describing an existing algorithm, and review of articles citing identified manuscripts describing an existing algorithm. Ultimately, we identified five existing algorithms that have been used previously in the literature for consideration in this analysis (**Table 1**).1–5 These five algorithms can be categorized into two groups. The first group includes the Herzog & Wallace (H-W)1 and Langa-Kabeto-Weir (L-K-W) algorithms.2,4 Both rely on classifying participants as cognitively impaired if their summary functioning scores fall below a certain cut point. For self-respondents, the summary score is computed exclusively using cognition test scores; for proxy-respondents, the score is computed proxy or interviewer reports on cognition and functional ability. Cut points for each summary score were chosen by the authors to achieve a prevalence of dementia or cognitive impairment similar to the expected population prevalence, derived from external data sources (H-W)1 or ADAMS findings (L-K-W)4. The second group includes the Wu5 , Hurd3 , and Crimmins4 algorithms, which take a regression-based approach to classifying cognitive status, using ADAMS Wave A dementia assessment as the gold-standard diagnosis. These three regression-based algorithms include information on cognitive scores among self-respondents, proxy or interviewer reports on cognitive and functional ability for proxy respondents, and additional socioeconomic, sociodemographic, or physical functioning information. With the exception of Wu et al., who used the missing-indicator method to develop a single algorithm that includes both self- and proxy-respondents, all others developed separate algorithms to predict dementia status for self-respondents vs. proxy-respondents.

Details of the existing algorithms are summarized in **Table 1**.

**Training and Validation Datasets**

We evaluated performance of each of these five algorithms in two samples: (a) a “training” sample comprising HRS/ADAMS participants’ data from Wave A, which were used for training the Wu, Hurd, and Crimmins algorithms, and (b) a “validation” sample comprising HRS/ADAMS participants’ interview and assessment data from ADAMS Waves B, C, and D; though these participants were included in the Wave A assessment, their data from Waves B, C, and D were not previously used in algorithm creation.

The eligibility criteria used to select the original HRS/ADAMS training sample for the Wu, Hurd, and Crimmins algorithms varied slightly across algorithms. In addition, it appears that each algorithm was developed using a slightly different version of the HRS data (e.g., due to use of the RAND version of the HRS data versus use of the core data files, analyst-specific data cleaning choices for common variables, and differences in whether and how missing cognitive scores for persons who refuse to do or complete a cognitive test were imputed). Thus, in order to ensure we provide fair comparisons and mirror likely performance in the hands of future researchers, we created standardized training and validation HRS/ADAMS datasets in which to evaluate performance of the five algorithms.

We used the same criteria for derivation of both or our training and validation HRS/ADAMS datasets. We used the RAND version (version P) of the HRS data for all variables except for proxy cognitive data and interviewer assessment of cognition and Hurd dementia probabilities (which are not included in the RAND data sets). Proxy cognitive data and interviewer assessment of cognition were extracted from the HRS core files, and Hurd algorithm dementia probabilities calculated by the study authors for HRS participants through the 2006 HRS interview wave, which are publicly available on the RAND website. The RAND datasets include imputed cognitive scores for self-respondents with missing cognitive data.20 Whenever available, we used RAND-derived summary variables (e.g. for number of ADL) and we followed the logic used by RAND in creation of the change in ADL limitations variable to create additional variables summarizing change in cognitive variables. 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; observations were dropped if proxy scores from the wave immediately prior were not available.

The HRS/ADAMS training dataset included ADAMS Wave A participants. We used dementia status from ADAMS Wave A and predictor data from the nearest prior HRS interview, either 2000 or 2002 depending on the date of the initial ADAMS assessment. The HRS/ADAMS validation dataset included ADAMS Wave B, C, and D participants who were not previously found to have dementia at a prior ADAMS wave, and included dementia status at each ADAMS wave matched to corresponding HRS data from the nearest prior HRS interview. Note that the ADAMS validation dataset may include up to three records from the same individual because ADAMS participants not identified as having dementia at one wave were asked to return for re-assessment at the next wave of ADAMS repeated measures were treated as an independent observations.

**Statistical Analyses**

In both samples, we restricted our analyses to the subset of observations with complete data on dementia status and variables used in any of the five algorithms, with 760 observations in the training data, and 515 observations (from 375 unique participants) in the validation data. Notably, when creating the ‘standardized’ HRS validation dataset, we excluded all ADAMS Wave D participants whose immediately prior HRS interview occurred in 2008 (N=164) because published Hurd dementia probabilities are not available for HRS interview waves occurring after 2006, and we were unable to apply their published regression22 coefficients to the HRS data due to missing information.

We then classified participants as demented or non-demented according to each algorithm. For the Herzog & Wallace and Langa-Kabeto-Weir algorithms, we simply applied the cut-offs to the relevant summary scores and classified persons as having dementia below the specified cut-offs. For the Wu and Crimmins algorithms, we used the published coefficients to calculate predicted probabilities of class membership, and then classified persons as having dementia if the predicted probability of dementia was greater than 0.5. Though the Crimmins algorithm uses a multinomial logit model predicting probabilities of both cognitive impairment no dementia (CIND) and dementia, we did not consider the predicted CIND probability when classifying dementia status. For the Hurd algorithm, we used the pre-calculated class membership probabilities available for download from the HRS study website and classified persons as having dementia if the predicted probability was greater than 0.5. Finally, we compared the predicted dementia status according to each of the five algorithms to ADAMS-based dementia diagnoses overall and in sociodemographic subgroups. Specifically, we calculated accuracy, sensitivity, and specificity.

To assess the potential impact of alternate cut-points for algorithms that produce a probability of class membership, we plotted receiver-operator curves (ROC) and re-evaluated performance based on a standard set of alternate cut points chosen to achieve (a) 98% sensitivity, (b) 95% sensitivity, (c) 90% sensitivity, (d) 98% specificity, (e) 95% specificity, (f) 90% specificity, and (g) the combination that jointly maximizes sensitivity and specificity.

We conducted several sensitivity analyses.

First, we applied the ADAMS sampling weights to our analyses to be representative of the national population aged 70+, in order to understand the impact of the stratified sampling approach used to select ADAMS participants on performance of each algorithm.

Second, we re-estimated each regression-based algorithm in the standardized dataset, to understand sensitivity of algorithm development to small differences in training sample and data cleaning choices. We then calculated performance metrics based on these new versions of the existing algorithms to consider sensitivity to sample selection choices.

Third, we considered an alternate version of our HRS/ADAMS validation sample that included all HRS/ADAMS participants regardless of whether they previously had a dementia diagnosis in order to understand the usefulness of each algorithm in identifying incident versus prevalent dementia cases. Specifically, we added observations from participants who were known to be alive and demented at the time of Waves B, C, or D but who were not re-evaluated at Waves B, C, or D because of a dementia diagnosis in a prior wave (A, B, or C) . For example, if an ADAMS/HRS participant were diagnosed with dementia at wave A, our alternate validation dataset would now include paired HRS/ADAMS data allowing prediction of dementia at the time of ADAMS Waves B, C and D assessments if they were alive and participated in the nearest prior HRS interview. This alternate validation dataset comprised 1049 observations from 651 unique participants.

Fourth, we repeated our primary analyses only considering persons as having an ADAMS dementia diagnosis at if they were diagnosed with the four most common dementia types in older adults: possible or probable Alzheimer’s disease, possible or probable vascular dementia, frontal lobe dementia, or Lewy body dementia.

We used SAS Version 9.4 for all analyses. We provide annotated code as supplemental material in the spirit of supporting reproducible research and in the hope that it will enable wider use of these algorithms (see Supplemental Methods).

**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 very similar set of self-response cognition test scores. However, there is less consistency in reliance on the proxy cognition, physical function, and demographic variables. The H-W algorithm is the only one that uses the total number of symptoms on the Jorm scale (described in Crimmins4) for proxy-respondents, while the other algorithms use a combination of other proxy- and interviewer- rated cognition scores, including the average IQCODE score also developed by Jorm.21 The Hurd regression is the sole algorithm that accounts for the full set of limitations to basic and instrumental activities of daily living, while the Crimmins and L-K-W approaches each take into account a subset of these functional limitations. Only the Wu algorithm takes into account participant race.

**Table 1: Description of algorithms**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Predictors** | **Herzog-Wallace (1997)** | | **Langa-Kabeto-Weir (2009)** | | **Crimmins (2011)** | | **Hurd (2013)\*** | | **Wu (2013)\*\*** | |
| **Score cutoff** | | **Score cutoff** | | **Multinomial Logit** | | **Ordered probit** | | **Logit** | |
| Self  (0-35) | Proxy  (0-7) | Self  (0-27) | Proxy  (0-11) | Self | Proxy | Self | Proxy | Self | Proxy |
| ***Demographics*** | | | | | | | | | | |
| Age |  |  |  |  | X |  | X | X | X | X |
| Gender |  |  |  |  | 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 | X |
| Delayed word recall | X |  | X |  | X |  | X | X | X | X |
| Serial 7's | X |  | X |  | X |  | X | X | X | X |
| Backward count | X |  | X |  | X |  | X |  | X | X |
| Dates |  |  |  |  | X |  | X | X | X | X |
| Object naming (scissors) | X |  |  |  | X |  | X |  |  |  |
| Object naming (cactus) | X |  |  |  | X |  | X |  | X | X |
| President | X |  |  |  |  |  | X | X | X | X |
| Vice-president | X |  |  |  | X |  | X |  | X | X |
| ***Cognition (proxy)*** | | | | | | | | | | |
| Proxy-rated memory score |  |  |  | X |  | X |  |  | X | X |
| Interviewer assessment |  |  |  | X |  | X |  |  |  |  |
| 16-item Jorm IQCODE |  |  |  |  |  |  |  | X | X | X |
| 7-item Jorm symptoms |  | X |  |  |  | X |  |  |  |  |
| ***Physical functioning (ADL's)*** | | | | | | | | | | |
| Eating |  |  |  |  | X |  | X | X |  |  |
| Bathing |  |  |  |  | X |  | X | X |  |  |
| Dressing |  |  |  |  | X |  | X | X |  |  |
| Transferring |  |  |  |  |  |  | X | X |  |  |
| Walking across room |  |  |  |  |  |  | X | X |  |  |
| ***Physical functioning (IADL's)*** | | | | | | | | | | |
| Using phone |  |  |  | X | X |  | X | X |  |  |
| Taking medication |  |  |  | X |  |  | X | X |  |  |
| Managing money |  |  |  | X | X |  | X | X |  |  |
| Grocery shopping |  |  |  | X |  |  | X | X |  |  |
| Preparing meals |  |  |  | X |  |  | X | X |  |  |
| *\* For predicting dementia status for participants with proxy respondents in the most recent HRS wave, Hurd included an indicator specifying whether they were selfs or also had a proxy two waves prior, and the corresponding cognition assessment scores.* | | | | | | | | | | |
|
| *\*\* Wu used a single algorithm to classify dementia status for selfs 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 sasessments to 0 for selfs, and sets self-cognition assessments to 0 for proxys.* | | | | | | | | | | |
|
|

Summary statistics of each ‘standardized’ dataset are displayed in Table 2. In both the HRS validation and training samples, the large majority of dementia cases are of Alzheimer’s etiology (76%). However, as expected, there are considerably fewer cases of dementia in the HRS validation data (which contains only incident cases) than in the HRS training data (which contains both prevalent and incident cases). Other notable differences are that HRS training data participants are much more likely to have proxy-respondents (22% vs. 6%), and also have greater functional limitations compared to HRS validation data participants (mean 0.97 vs. 0.61 ADL’s and 1.22 vs. 0.53 IADL’s). The overall socio-demographic distributions, as well as cognition test scores are similar across the two samples.

**Table 2: Summary statistics of training and validation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Outcomes and predictors** | **Training data** | | | **Validation data** | | |
| N | Mean | S.D. | N | Mean | S.D. |
| ***Dementia Outcomes*** | | | | | | |
| Dementia Status | 760 | 0.34 | 0.47 | 515 | 0.14 | 0.35 |
| Dementia etiology: Alzheimer's | 760 | 0.26 | 0.44 | 515 | 0.10 | 0.31 |
| Dementia etiology: Vascular | 760 | 0.05 | 0.23 | 515 | 0.02 | 0.14 |
| Dementia etiology: FTD | 760 | 0.00 | 0.04 | 515 | 0.00 | 0.00 |
| Dementia etiology: Lewy Body | 760 | 0.00 | 0.00 | 515 | 0.00 | 0.00 |
| Dementia etiology: Other | 760 | 0.03 | 0.16 | 515 | 0.01 | 0.12 |
| ***Demographics*** | | | | | | |
| Proxy-respondent | 760 | 0.22 | 0.41 | 515 | 0.06 | 0.23 |
| Age | 760 | 80.30 | 6.97 | 515 | 81.20 | 5.76 |
| Female | 760 | 0.59 | 0.49 | 515 | 0.52 | 0.50 |
| Education: LTHS | 760 | 0.49 | 0.50 | 515 | 0.42 | 0.49 |
| Education: HS or GED | 760 | 0.39 | 0.49 | 515 | 0.43 | 0.50 |
| Education: some college or more | 760 | 0.12 | 0.33 | 515 | 0.15 | 0.36 |
| Non-Hispanic White | 760 | 0.69 | 0.46 | 515 | 0.72 | 0.45 |
| Non-Hispanic Black | 760 | 0.18 | 0.39 | 515 | 0.19 | 0.39 |
| Hispanic | 760 | 0.10 | 0.30 | 515 | 0.07 | 0.25 |
| Non-Hispanic Other Race | 760 | 0.02 | 0.15 | 515 | 0.03 | 0.16 |
| ***Cognition (self-response)*** | | | | | | |
| Immediate word recall, 0-10 | 595 | 3.88 | 1.83 | 485 | 4.39 | 1.59 |
| Delayed word recall, 0-10 | 595 | 2.64 | 2.11 | 485 | 3.04 | 1.89 |
| Serial 7's, 0-5 | 595 | 2.41 | 1.91 | 485 | 2.78 | 1.91 |
| Backward count, 0-2\* | 595 | 1.68 | 0.73 | 485 | 1.77 | 0.64 |
| Dates, 0-4 | 595 | 3.35 | 1.01 | 485 | 3.57 | 0.74 |
| Object naming: Cactus, 0-1 | 595 | 0.76 | 0.43 | 485 | 0.83 | 0.37 |
| Object naming: Scissors, 0-1 | 595 | 0.99 | 0.12 | 485 | 0.99 | 0.12 |
| President, 0-1 | 595 | 0.87 | 0.34 | 485 | 0.94 | 0.24 |
| Vice-president, 0-1 | 595 | 0.54 | 0.50 | 485 | 0.66 | 0.48 |
| ***Cognition (proxy)*** | | | | | | |
| Proxy-rated memory score, 1 (excellent) -5 (poor) | 165 | 4.27 | 1.00 | 30 | 3.53 | 1.01 |
| Interviewer assessment , 0-2\*\* | 165 | 4.15 | 0.74 | 30 | 3.39 | 0.50 |
| 16-item IQCODE, 1 (much improved - 5 (much worse) | 165 | 1.51 | 0.77 | 30 | 0.87 | 0.86 |
| Jorm symptoms, prior to 2004 (0-7) | 165 | 2.93 | 2.23 | 30 | 0.80 | 1.13 |
| Jorm symptoms, 2004 onwards, 0-5 | 165 | 1.78 | 1.51 | 30 | 0.63 | 0.96 |
| ***Physical functioning limitations*** | | | | | | |
| Basic activities of daily living (ADL's), 0-5 | 760 | 0.97 | 1.54 | 515 | 0.61 | 1.13 |
| Instrumental activities of daily living (IADL's), 0-5 | 760 | 1.22 | 1.79 | 515 | 0.53 | 1.08 |
| \*0 = Incorrect; 1 = Correct on 2nd attempt; 2 = Correct on 1st attempt | | |  |  |  |  |
| \*\*0 = no cognitive limitations, 1 = some cognitive limitations, 2 = cognitive limitations prevents completion of interview | | | | | | |

As shown in **Table 3**, sensitivity ranged from 53% to 90%, specificity ranged from 79% to 97%, and overall accuracy ranged from 81% to 87% across the five algorithms when applied to the training data. Though overall accuracy was similar in the HRS validation data (range: 79% - 88%), this was largely driven by slightly higher specificities (82% - 98%), as sensitivity was much lower (range: 18% - 62%). Across the two datasets, we also found that the H-W score cutoff-based algorithm had much lower sensitivity, but higher specificity than the three regression-based algorithms. Conversely, with the exception of relative specificity compared to the Crimmins algorithm, the L-K-W score cutoff-based algorithm performed slightly worse across all measures compared to the three regression-based algorithms. We also illustrate the performance of the three regression-based algorithms by plotting ROC curves (**Figure 2**), which clearly demonstrates the higher sensitivities and lower specificities achieved in the training data. Additionally, though the Crimmins algorithm achieved higher sensitivity at a 0.5 probability cutoff (**Table 3**), the ROC curve demonstrates that at higher cut-off points in the training data, the Hurd and Wu algorithms both achieved higher sensitivities.

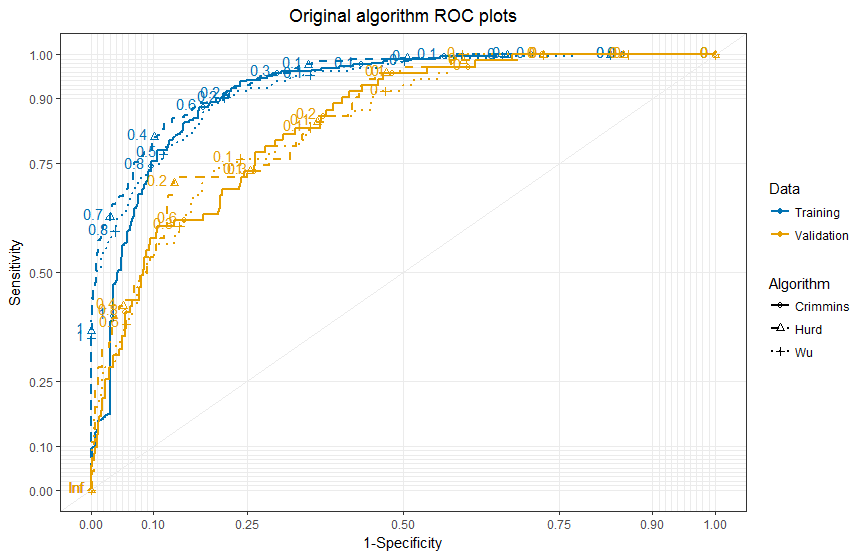
**Table 3: Overall performance metrics**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** |
| ***Training data (N=760)*** | | | |
| Herzog-Wallace (H-W) | 53.49 | 96.61 | 81.97 |
| Langa-Kabeto-Weir (L-K-W) | 75.19 | 83.27 | 80.53 |
| Crimmins | 89.92 | 79.08 | 82.76 |
| Hurd | 76.74 | 91.83 | 86.71 |
| Wu | 77.91 | 88.05 | 84.61 |
| ***Validation data (N=515)*** | | | |
| Herzog-Wallace (H-W) | 18.31 | 97.75 | 86.80 |
| Langa-Kabeto-Weir (L-K-W) | 40.85 | 89.19 | 82.52 |
| Crimmins | 61.97 | 82.21 | 79.42 |
| Hurd | 39.44 | 95.95 | 88.16 |
| Wu | 43.66 | 92.57 | 85.83 |

Overall accuracy was better for all five algorithms and across both datasets in our weighted analyses (**Appendix Table 1a**) primarily due to better specificity in the weighted data. Results were similar when predicting dementia attributable to the four most common forms of age-related dementia (Alzheimer’s disease, vascular dementia, Lewy body dementia, and frontotemporal dementia), who make up over 90% of all dementia cases (**Appendix Table 1b**). All five algorithms had uniformly higher sensitivities when applied to the alternative validation data comprising all participants known to be alive and demented at the time of Waves B, C, or D but who were not re-evaluated at Waves B, C, or D because of a prior dementia (**Appendix Table 1c**).

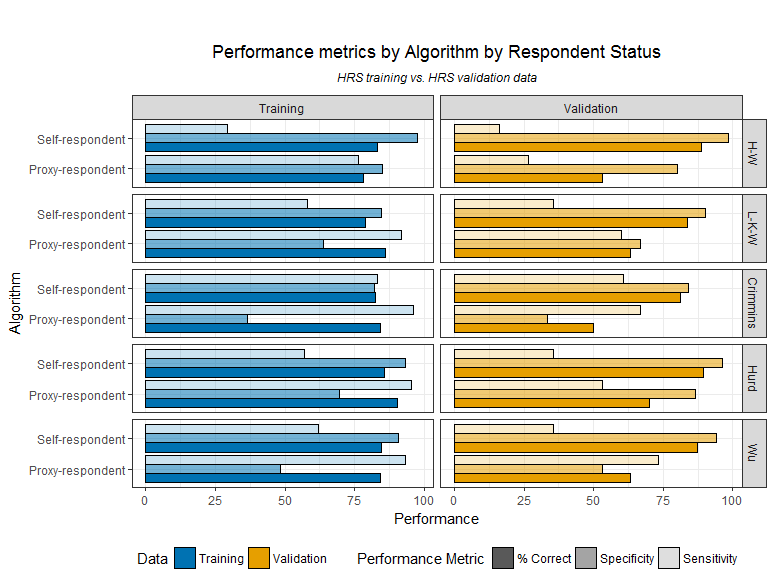
There were large disparities in how the algorithms performed when classifying self-respondents versus proxy-respondents. In both the training and validation data, sensitivity was much higher among proxy-respondents, while specificity was higher among self-respondents across all five algorithms (**Figure 3a)**. Comparing across race/ethnicity groups in the two HRS datasets(**Figure 3b**), specificity and overall classification accuracy was higher for non-Hispanic whites than for non-Hispanic blacks and Hispanics across the board. When applied to the training data, the H-W, L-K-W, Crimmins, and Hurd algorithms all had lower sensitivity for non-Hispanic whites compared to either minority group. This was similarly the case for both score cutoff-based algorithms when applied to the validation data. Conversely, sensitivity was highest among non-Hispanic whites for the Hurd algorithm when applied to the validation data. Interestingly, though Wu et al. excluded all Hispanic participants in the algorithm development, their algorithm had higher sensitivity in predicting dementia status among Hispanics compared to either non-Hispanic group across both datasets. In the training data, the Hurd algorithm had the smallest disparities in overall accuracy across racial/ethnic groups, while in the validation data, the Crimmins and Wu algorithms had the smallest racial/ethnic disparities in overall classification accuracy. Notably, the sensitivity of the Crimmins algorithm across race/ethnicity groups in the validation data was almost equal. Comparing across age groups (**Figure 3c**), we also found uniformly higher specificities and classification accuracies in younger individuals (<80), but generally higher sensitivities among older individuals. The algorithms performed generally better in classifying dementia status among those with at least a high school education, among females, with few exceptions (**Appendix Figure 1**).

**Figure 2: ROC curves for the regression-based algorithms applied to the training and validation data**

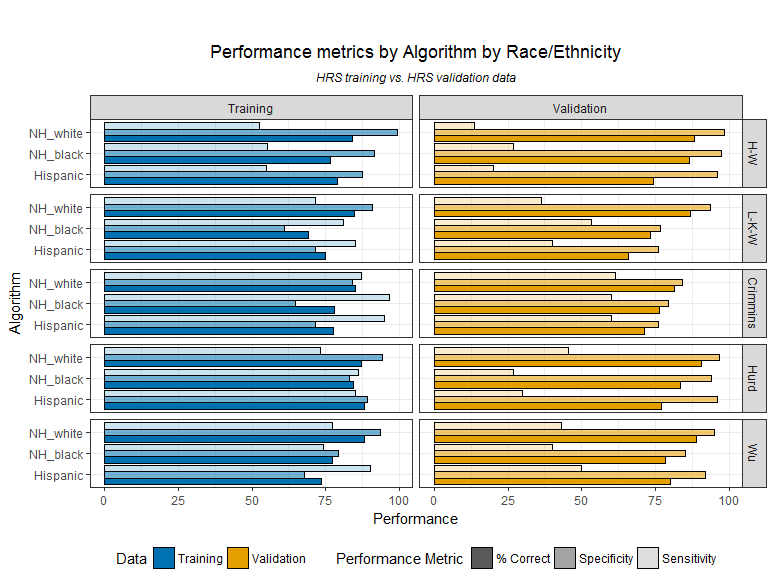


The performance metrics of the three regression-based algorithms at standard alternate cut-points are shown in **Tables 4a – 4c**. At each of the three levels of sensitivity in the training data, the Hurd and Crimmins algorithms had higher specificities and overall accuracy compared to the Wu algorithm, though the difference narrows at lower levels of sensitivity. In the validation data, differences in specificity and overall accuracy were small across the three algorithms, with the Wu algorithm performing best at 98% sensitivity and the Hurd/Crimmins algorithms performing better at 90% sensitivity. **Table 4b** shows that the Hurd algorithm generally outperforms the other two in sensitivity when maximizing specificity in both the training and validation data. Finally, in maximizing XXX, we found that the Hurd algorithm achieved highest overall accuracy, driven primarily by higher specificities, as the Crimmins and Wu algorithms had higher sensitivities (**Table 4c**).

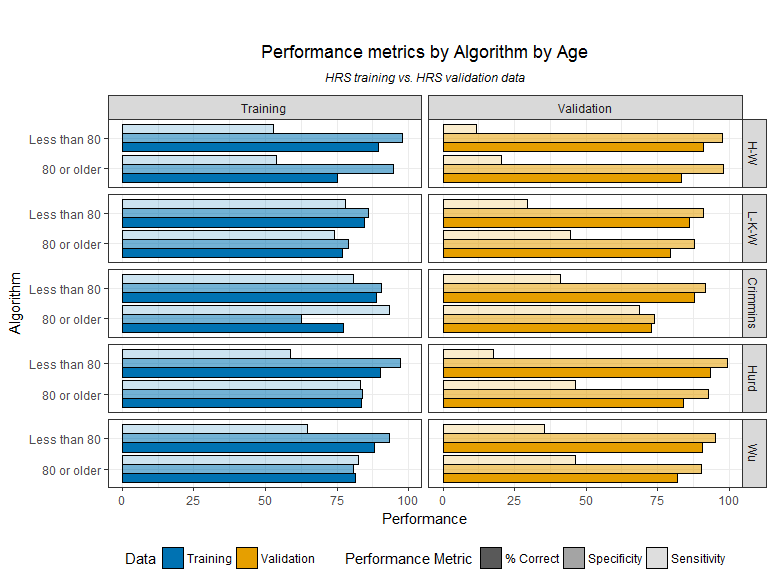
**Figure 3a: Performance metrics by algorithm by respondent status**



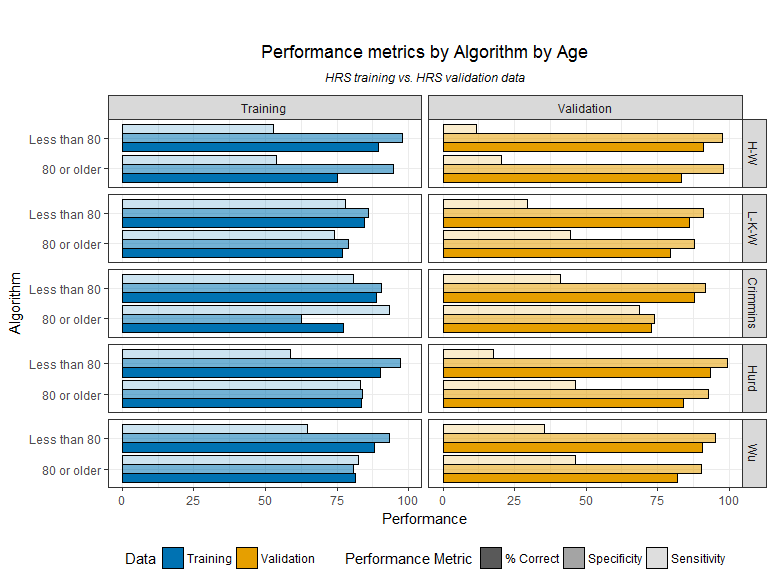
**Figure 3b: Performance metrics by algorithm by race**



**Figure 3c: Performance metrics by algorithm by age**



**Key**

**Table 4a: Cut-points, specificity, and overall accuracy to achieve 98%, 95, and 90% sensitivity**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **98% sensitivity** | | | **95% sensitivity** | | | **90% sensitivity** | | |
| **Algorithm** | **Cut-point** | **Specificity** | **% correct** | **Cut-point** | **Specificity** | **% correct** | **Cut-point** | **Specificity** | **% correct** |
| ***Training data (N=760)*** | | | | | | | | | |
| Crimmins | 0.11 | 55.18 | 69.74 | 0.30 | 71.51 | 79.61 | 0.49 | 79.08 | 82.89 |
| Hurd | 0.10 | 64.54 | 75.92 | 0.14 | 70.72 | 79.08 | 0.22 | 78.88 | 82.76 |
| Wu | 0.04 | 51.20 | 67.11 | 0.10 | 65.74 | 75.79 | 0.20 | 78.29 | 82.37 |
| ***Validation data (N=515)*** | | | | | | | | | |
| Crimmins\* | 0.04 | 38.51 | 46.80 | 0.10 | 53.15 | 59.03 | 0.15 | 59.68 | 63.88 |
| Hurd\* | 0.02 | 42.12 | 49.90 | 0.04 | 53.83 | 59.61 | 0.05 | 56.98 | 61.55 |
| Wu\* | 0.01 | 42.57 | 50.29 | 0.01 | 42.79 | 50.10 | 0.03 | 55.18 | 60.00 |
| \*For each algorithm, closest sensitivity = 98.59 and 95.77 | | | | | | | | | |

**Table 4b: Cut-points, sensitivity and overall accuracy to achieve 98%, 95, and 90% specificity**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **98% specificity** | | | **95% specificity** | | | **90% specificity** | | |
| **Algorithm** | **Cut-point** | **Sensitivity** | **% correct** | **Cut-point** | **Sensitivity** | **% correct** | **Cut-point** | **Sensitivity** | **% correct** |
| ***Training data (N=760)*** | | | | | | | | | |
| Crimmins | 0.99 | 16.28 | 70.39 | 0.94 | 55.81 | 81.71 | 0.80 | 75.58 | 85.13 |
| Hurd | 0.73 | 60.85 | 85.39 | 0.61 | 67.44 | 85.79 | 0.45 | 80.62 | 86.84 |
| Wu | 0.90 | 54.26 | 83.16 | 0.81 | 62.79 | 84.34 | 0.62 | 74.81 | 85.00 |
| ***Validation data (N=513)*** | | | | | | | | | |
| Crimmins | 0.96 | 21.13 | 87.57 | 0.88 | 35.21 | 86.80 | 0.67 | 57.75 | 85.83 |
| Hurd | 0.58 | 32.39 | 89.13 | 0.44 | 42.25 | 87.77 | 0.28 | 53.52 | 85.24 |
| Wu | 0.80 | 26.76 | 88.35 | 0.63 | 36.62 | 86.99 | 0.36 | 53.52 | 85.05 |

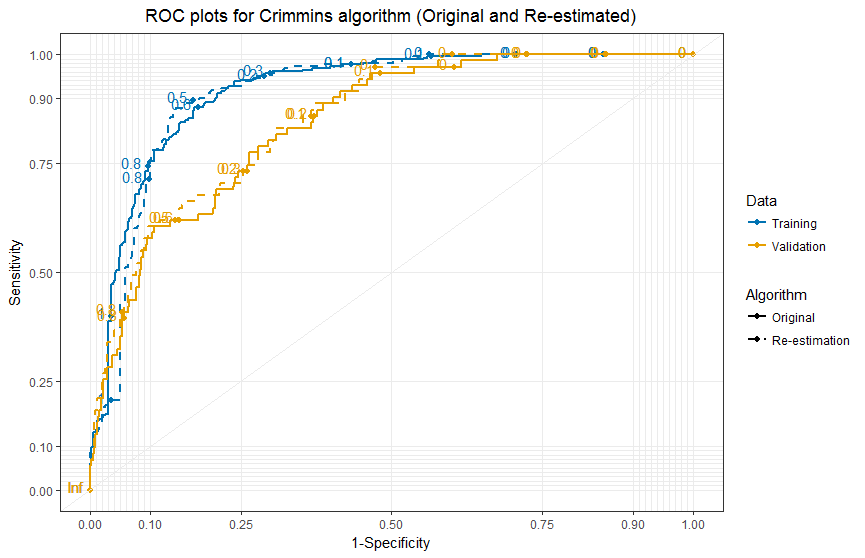
**Table 4c:**

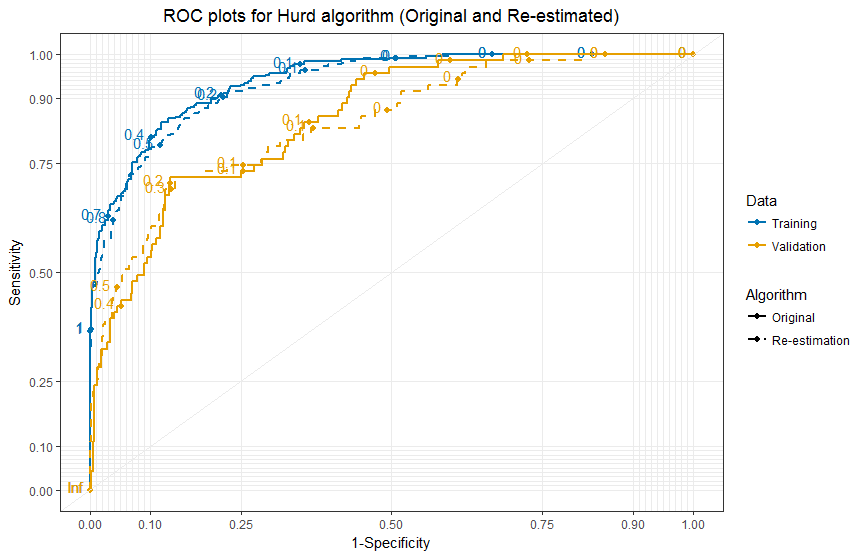
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Maximizing mean sens/spec** | | | |
| **Algorithm** | **Cutpoint** | **Sensitivity** | **Specificity** | **% correct** |
| ***Training data (N=760)*** | | | | |
| Crimmins | 0.58 | 87.98 | 82.67 | 84.47 |
| Hurd | 0.39 | 84.50 | 88.25 | 86.97 |
| Wu | 0.25 | 88.37 | 80.68 | 83.29 |
| ***Validation data (N=515)*** | | | | |
| Crimmins | 0.30 | 77.46 | 73.65 | 74.17 |
| Hurd | 0.19 | 71.83 | 86.49 | 84.47 |
| Wu | 0.13 | 74.65 | 79.50 | 78.84 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Maximizing overall accuracy** | | | |
| **Algorithm** | **Cutpoint** | **Sensitivity** | **Specificity** | **% correct** |
| ***Training data (N=760)*** | | | | |
| Crimmins | 0.78 | 77.91 | 89.44 | 85.53 |
| Hurd | 0.53 | 75.19 | 93.03 | 86.97 |
| Wu | 0.70 | 70.93 | 92.43 | 85.13 |
| ***Validation data (N=515)*** | | | | |
| Crimmins | 0.94 | 25.35 | 97.75 | 87.77 |
| Hurd | 0.66 | 28.17 | 98.87 | 89.13 |
| Wu | 0.80 | 26.76 | 98.20 | 88.35 |

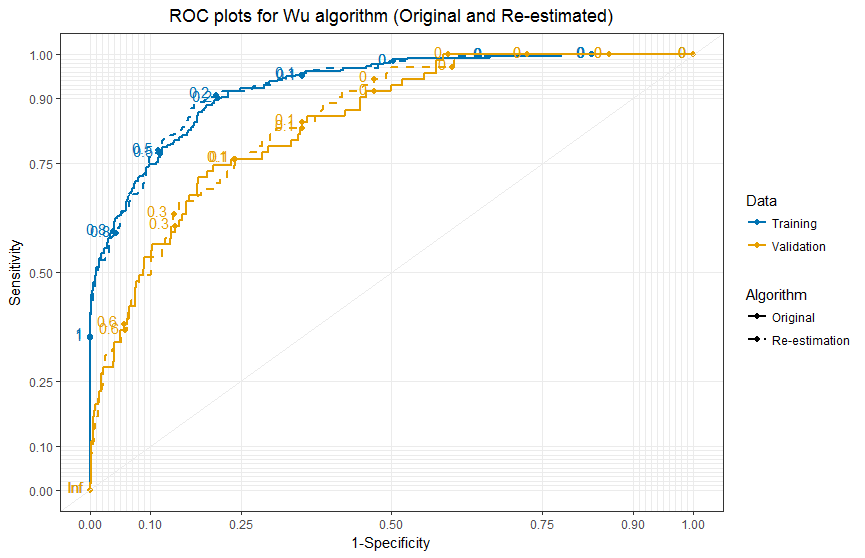
The performance of the three re-estimated regression-based algorithms (**Appendix Table 2**) was largely very similar to that of the original algorithms, with differences across all three performance metrics in the training data falling within three percentage points. In the validation data, the re-estimated Wu and Hurd algorithms achieved much higher sensitivity (both 49%) compared to the original algorithms (respectively 39% and 44%), while the Crimmins algorithm achieved higher specificity (86%) compared to the original algorithm (82%) using a 0.5 cut-point. Evaluating performance at alternate cut-points (**Appendix tables 3a-3c**), we found that the re-estimated Hurd and Crimmins algorithms generally achieved lower specificity at cut-points corresponding to 98% and 95% sensitivity compared to the original algorithms. Similarly, compared to the original algorithms, the Hurd and Wu algorithms achieved lower sensitivity at high levels of specificity when applied to the training data. Conversely, the re-estimated Wu algorithm achieved generally higher specificity at cut-points corresponding to 95% and 90% sensitivity, and the re-estimated Hurd and Crimmins algorithms achieved better sensitivity at 95% and 90% specificity in the validation data compared to the original algorithms. These differences are illustrated in the ROC curves (**Figures 2a-2c**), which further demonstrate that performance differences between the original and re-estimated algorithms tended to be larger in the validation data than in the training data at various cut points.

**Figure 3a: Original and re-estimated Crimmins algorithm**

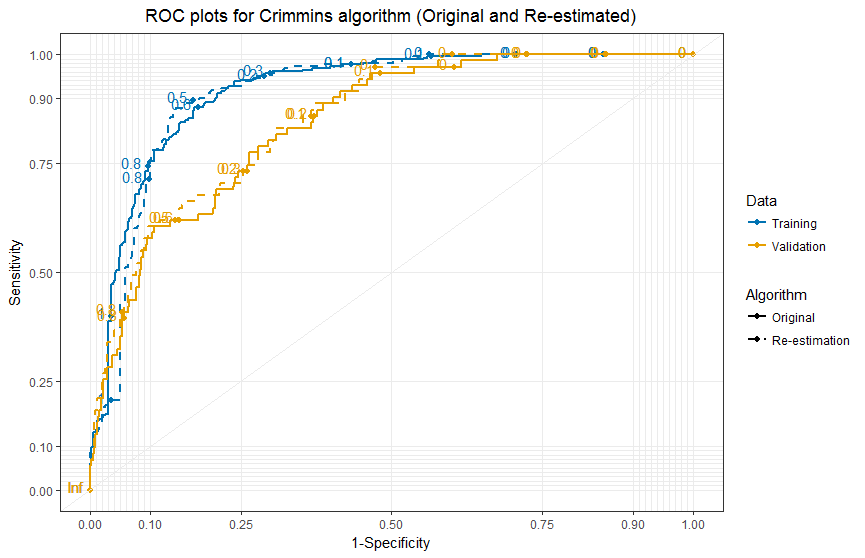
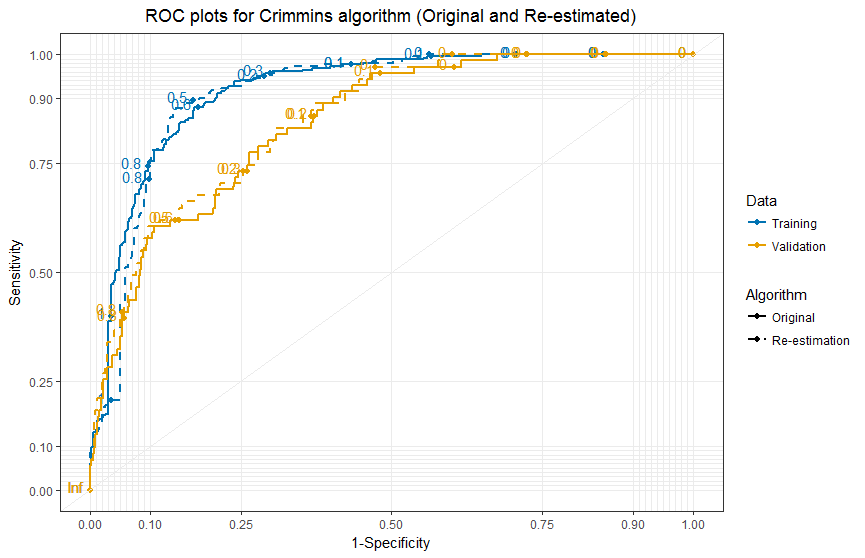
****

**Figure 3b: Original and re-estimated Hurd algorithm**  

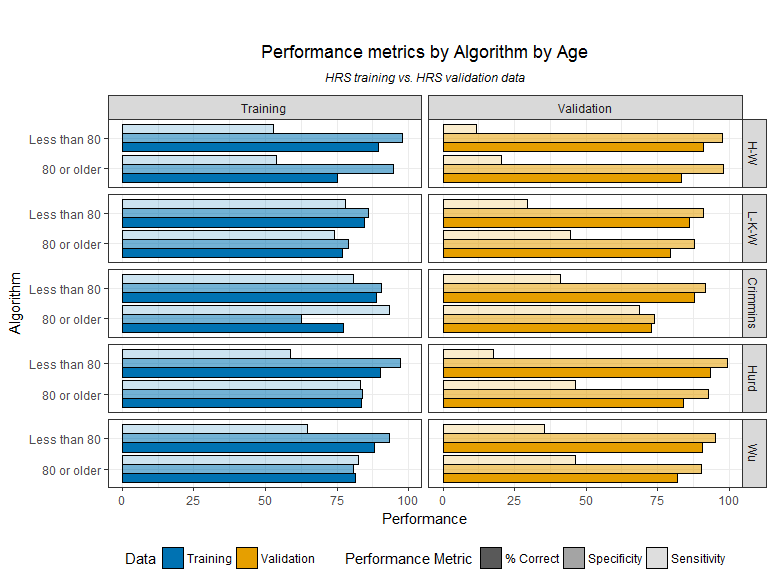
**Figure 3c: Original and re-estimated Wu algorithm**



**KEY**



**Key**

We also examined performance of the re-estimated regression-based algorithms across sociodemographic groups (**Appendix Figures 2a-2e**) and found trends to be largely consistent to those of the original regressions. Proxy respondents tend to perform better insensitivity while self-respondents have higher specificity. Compared to both minority groups, non-Hispanic whites perform better in specificity and overall accuracy in the training data, and in all three performance metrics in the validation data. However, unlike the original Wu algorithm, which achieved highest sensitivity among Hispanics in both the training and validation data, the sensitivity of the re-estimated Wu algorithm among Hispanics in the validation data was marginally lower than that of non-Hispanic whites. Finally, as is the case with the original algorithms, sensitivity was higher among older individuals while specificity was higher among younger individual, and overall performance was generally better among those with at least a high school degree, and among women.

.

**DISCUSSION**

Our findings show that the algorithms perform significantly better in sensitivity (but similarly in specificity) when applied to the training data compared validation data, both overall, and across sociodemographic subgroups, which may suggest overfitting, or that they are better at identifying prevalent than incident ADRD cases. However, the fact that the algorithms perform better when applied to our alternate validation data (comprising all prevalent cases of dementia in addition to incident cases at each wave; **Appendix Table 1c**) than when applied to either the HRS training or original validation data lends greater support to the latter hypothesis. This is encouraging because the usefulness of these algorithms lies in their ability to accurately classify dementia using new data from different sources and years. Across datasets, the algorithms have higher sensitivity among proxy-respondents, and higher specificity for self-respondents, which may be a result of the fact that there is a higher prevalence of dementia among proxies compared to self-respondents (80% vs. 21% in the training data).

Importantly, our head-to-head comparison of the existing algorithms demonstrates the strengths and weaknesses of each, and will help guide future researchers to select the most appropriate algorithm for their purposes. For example, for research purposes where correctly identifying positive cases of dementia is most important, the Crimmins algorithm may be most useful, whereas for research purposes where correctly identifying true negative cases of dementia is most important, the simple cut-off method proposed by Herzog-Wallace would be more useful. Conversely, for cases where consistency in identifying both true positives and true negatives are equally important, the L-K-W or Wu algorithms may be more appropriate for prevalent cases, while the Crimmins algorithm would perform best for incident cases. Finally, it is reassuring that the overall accuracy is not substantially reduced when using only cognitive score cut-offs. Use of these approaches may be appropriate when there is substantial missingness in the covariates needed for other approaches.

The differences in performance across sociodemographic groups must also be considered when selecting an algorithm. Naive applications of these algorithms for disparities research may introduce differential misclassification and bias estimates of disparities, as the accuracy differs across race/ethnicity, educational achievement, and age. For example, taking the point estimates at face value, the performance metrics from the HRS validation data suggest that the Crimmins algorithm performs most consistently across racial groups. Conversely, while the Crimmins algorithm also has the highest sensitivity in classifying dementia cases among more highly educated individuals, it is likely to result in a greater degree of differential misclassification by educational attainment compared to either the Hurd or Wu algorithms. These conclusions are subject to the important caveat that our estimates of sensitivity, specificity, and accuracy are vulnerable to substantial imprecision due to the small sample size. Additionally, they are subject to the fact that we ignored CIND probabilities in classifying dementia for self-respondents using the Crimmins multinomial logit algorithm or the Hurd probit model; accounting for CIND (using the following decision rule: P(dementia) > 0.5 *and* P(dementia) > P(CIND)) substantially lowers its sensitivity and increases its specificity in dementia classification overall, and alters its relative performance to the other algorithms across sociodemographic subgroups **(Appendix tables 4a-4f)**.

The relative ease of applying each algorithm is also an important (and likely the more constraining) factor to consider when selecting which algorithm to use. The regression-based algorithms, while generally performing better, are much more difficult to implement. Apart from the time-consuming task of cleaning and preparing the data for applying the regression coefficients, we had challenges determining specific data cleaning choices for common variables based on the manuscript alone, and relied on access to the original code used to develop the Wu algorithm, as well as correspondence with the authors of the Crimmins algorithm. (Though we also attempted to contact Hurd et al. for clarifications regarding coding decisions, we did not receive a response, and thus made our best educated guess when re-estimating their algorithm.) The cognitive score cutoff-based algorithms were very straightforward to apply, with the L-K-W algorithm being more accessible by requiring fewer predictors. Considering that the sensitivity of the L-K-W algorithm was comparable to (and for certain sociodemographic subgroups, better than) that of the regression-based Wu and Hurd algorithms (as well as the Crimmins algorithm when CIND probabilities are accounted for), it may be the most universally applicable existing algorithm for purposes where identifying positive cases is most important. Furthermore, the L-K-W algorithm is superior to the Crimmins algorithm in cases where users wish to identify individuals with CIND, as it offers relevant cut-off points for both self-respondents *and* proxy-respondents. Comparison of classification performance across the three levels of outcomes between the Crimmins and L-K-W algorithms was outside the scope of this paper, but is an important area of inquiry that should be pursued in the future.

Our study also has limitations. Our methods assume the relationship between our predictors and dementia status is invariant across time. Additionally, the validation dataset used in this evaluation is not ideal. First, it comes from the same study and repeated measures of the same participants from the training data, which limits external validity and renders our conclusions reliant on the assumption that the ADAMS dementia diagnoses are accurate. Second, it captures incident cases of dementia, which is different from prevalent cases of dementia (as captured primarily in the training data). We pursued the possibility of using alternative studies from which to construct out validation dataset, including the ARIC study, HRS sister studies, as well as the Rush MAP, ROS, and MARS cohorts. Unfortunately, none of these studies collected sufficiently similar cognition and physical functioning data to ensure that the algorithms can be applied with a high degree of fidelity. FDespite these limitations, this head-to-head comparison is the first to allow those who wish to use these existing algorithms to make an informed decision regarding which may be most appropriate for their purpose. Additionally, this exercise has revealed that the performance of the algorithms may be substantially different across sociodemographic groups, and that the prediction of prevalent versus incident dementia most likely require different methods and predictors.

Further testing of the existing algorithms should focus on evaluating their performance in predicting prevalent versus incident dementia in external studies and in larger sample sizes to test their external validity. Additionally, comparing the performance of the algorithms in predicting mortality may be worthwhile and potentially shed further insight into the relationship between cognition and mortality in old age. Importantly, we strongly recommend that efforts to develop dementia classification algorithms continue, and that they focus particularly on achieving uniform performance across sociodemographic groups. Based on our head-to-head comparison, it is possible that for summary score cutoff-based algorithms, different thresholds may be needed for different groups, and that for regression-based algorithms, additional sociodemographic indicators may be needed. We also recommend that future efforts to develop algorithms for predicting dementia separate prevalent vs. incident cases in the algorithm training and evaluation processes.

**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. doi:10.1056/NEJMsa1204629.

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.

6. Heeringa SG, Fisher GG, Hurd M, et al. Sample Design, Weighting and Analysis for ADAMS. 2007:1-36.

7. Langa KM, Plassman BL, Wallace RB, et al. The aging, demographics, and memory study: Study design and methods. *Neuroepidemiology*. 2005;25(4):181-191. doi:10.1159/000087448.

8. Langa KM, Chernew ME, Kabeto MU, et al. National estimates of the quality and cost of informal caregiving for the elderly with dementia. *J Gen Intern Med*. 2001;16:770-778.

9. 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? *Alzheimer’s Dement*. 2008;4(2):134-144.

10. 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(1):51.

11. Gure TR, Blaum CS, Giordani B, et al. Prevalence of cognitive impairment in older adults with heart failure. *J Am Geriatr Soc*. 2012;60(9):1724-1729.

12. 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(5):564-573.

13. Suemoto CK, Gilsanz P, Mayeda ER, Glymour MM. Body mass index and cognitive function: The potential for reverse causation. *Int J Obes*. 2015;39(9):1383-1389.

14. Rist PM, Capistrant BD, Wu Q, Marden JR, Glymour MM. Dementia and dependence: do modifiable risk factors delay disability? *Neurology*. 2014;82(17):1543-1550.

15. 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.

16. 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(37):235-243.

17. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JWR, Weir DR. Cohort profile: The Health and Retirement Study (HRS). *Int J Epidemiol*. 2014;43(2):576-585.

18. 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.

19. Plassman BL, Langa KM, McCammon RJ, et al. Incidence of dementia and cognitive impairment, not dementia in the united states. *Ann Neurol*. 2011;70(3):418-426. doi:10.1002/ana.22362.

20. Fisher GG, Hassan H, Faul JD, Rodgers WL, Weir DR. *Health and Retirement Study Imputation of Cognitive Functioning Measures: 1992-2014 (Final Release Version)*. Anne Arbor, MI; 2017.

21. 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 https://www.alz.org/documents\_custom/shortiqcode\_english.pdf.

22. Hurd MD, Martorell P, Delavande A, Mullen KJ, Langa KM. Monetary Costs of Dementia in the United States: On-Line Supplement. *N Engl J Med*. 2013:1-23. http://www.nejm.org/doi/suppl/10.1056/NEJMsa1204629/suppl\_file/nejmsa1204629\_appendix.pdf.

**APPENDICES**

**Table 1a: Sensitivity analyses – weighted performance metrics**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** |
| ***Training data (N=22,471,962)*** | | | |
| Herzog-Wallace | 42.21 | 99.26 | 92.07 |
| Langa-Kabeto-Weir | 56.52 | 96.60 | 91.55 |
| Crimmins | 77.92 | 92.69 | 90.83 |
| Hurd | 65.37 | 98.32 | 94.17 |
| Wu | 63.79 | 97.55 | 93.29 |
| ***HRS validation data (N=24,199,101)*** | | | |
| Herzog-Wallace | 13.35 | 99.69 | 92.61 |
| Langa-Kabeto-Weir | 24.02 | 97.63 | 91.58 |
| Crimmins | 38.98 | 91.15 | 86.87 |
| Hurd | 25.84 | 99.00 | 93.00 |
| Wu | 34.82 | 98.26 | 93.05 |

**Table 1b: Performance predicting four most common age-related dementias (AD, VD, FTD, Lewy Bodies)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** |
| ***Training data (N=760)*** | | | |
| Herzog-Wallace | 53.16 | 94.46 | 81.58 |
| Langa-Kabeto-Weir | 73.42 | 80.11 | 78.03 |
| Crimmins | 90.30 | 76.48 | 80.79 |
| Hurd | 77.22 | 89.29 | 85.53 |
| Wu | 78.48 | 85.66 | 83.42 |
| ***Validation data (N=515)*** | | | |
| Herzog-Wallace | 17.19 | 97.34 | 87.38 |
| Langa-Kabeto-Weir | 39.06 | 88.47 | 82.33 |
| Crimmins | 59.38 | 81.15 | 78.45 |
| Hurd | 37.50 | 95.12 | 87.96 |
| Wu | 42.19 | 91.80 | 85.63 |

**Table 1c: Performance metrics in alternative validation dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** |
| ***Alternative Validation data (N=1,049)*** | | | |
| Herzog-Wallace | 57.41 | 97.58 | 74.83 |
| Langa-Kabeto-Weir | 78.79 | 88.79 | 83.13 |
| Crimmins | 91.08 | 81.98 | 87.13 |
| Hurd | 82.15 | 95.60 | 87.99 |
| Wu | 80.47 | 92.31 | 85.61 |

**Table 2: Overall performance metrics for re-estimated regressions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** |
| ***Training data (N=760)*** | |  |  |
| Crimmins | 89.53 | 82.07 | 84.61 |
| Hurd | 76.74 | 90.44 | 85.79 |
| Wu | 77.91 | 88.65 | 85.00 |
| ***Validation data (N=515)*** | | |  |
| Crimmins | 61.97 | 86.49 | 83.11 |
| Hurd | 49.30 | 94.37 | 88.16 |
| Wu | 49.30 | 91.67 | 85.83 |

**Table 3a: Cut-points, specificity, and overall accuracy to achieve 98%, 95, and 90% sensitivity**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **98% sensitivity** | | | **95% sensitivity** | | | **90% sensitivity** | | |
| **Algorithm** | **Cut-point** | **Specificity** | **% correct** | **Cut-point** | **Specificity** | **% correct** | **Cut-point** | **Specificity** | **% correct** |
|  | ***Training data (N=760)*** | | | | | | | | |
| Crimmins | 0.06 | 53.39 | 68.55 | 0.23 | 70.12 | 78.68 | 0.49 | 81.87 | 84.61 |
| Hurd | 0.06 | 57.97 | 71.58 | 0.10 | 66.93 | 76.45 | 0.20 | 78.09 | 82.24 |
| Wu | 0.04 | 51.20 | 67.11 | 0.11 | 64.54 | 75.00 | 0.25 | 79.88 | 83.42 |
|  | ***Validation data (N=515)*** | | | | | | | | |
| Crimmins\*~ | 0.02 | 42.34 | 50.10 | 0.05 | 53.60 | 59.42 | 0.07 | 57.66 | 62.14 |
| Hurd\*~ | 0.02 | 34.23 | 43.11 | NA | NA | NA | 0.04 | 48.42 | 54.17 |
| Wu^~ | NA | NA | NA | 0.03 | 50.68 | 56.89 | 0.05 | 59.23 | 63.50 |
| \*Sensitivity = 98.59 and 95.78 | | | | | | | | | |
| ^98% sensitivity not available 0.001 increments; Sensitivity increases from 97% to 100 at 0.009 cutoff; | | | | | | | | | |
| ~Crimmins/Wu sensitivity = 95.77; Hurd sensitivity increases from 94% to 97% at 0.022 cutoff | | | | | | | | | |

**Table 3b: Cut-points, sensitivity and overall accuracy to achieve 98%, 95, and 90% specificity**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **98% specificity** | | | **95% specificity** | | | **90% specificity** | | |
| **Algorithm** | **Cut-point** | **Specificity** | **% correct** | **Cut-point** | **Specificity** | **% correct** | **Cut-point** | **Specificity** | **% correct** |
|  | ***HRS training data (N=760)*** | | | | | | | | |
| Crimmins | 0.99 | 18.22 | 70.92 | 0.96 | 42.25 | 77.11 | 0.78 | 75.97 | 85.26 |
| Hurd | 0.89 | 53.88 | 83.03 | 0.73 | 65.89 | 85.13 | 0.48 | 78.29 | 86.05 |
| Wu | 0.88 | 52.71 | 82.63 | 0.79 | 61.63 | 83.68 | 0.54 | 76.74 | 85.53 |
|  | ***HRS validation data (N=515)*** | | | | | | | | |
| Crimmins | 0.97 | 22.54 | 87.77 | 0.83 | 39.44 | 87.38 | 0.58 | 60.56 | 86.02 |
| Hurd | 0.74 | 30.99 | 88.93 | 0.52 | 46.48 | 88.35 | 0.33 | 59.15 | 85.83 |
| Wu | 0.80 | 22.54 | 87.77 | 0.63 | 33.80 | 86.60 | 0.41 | 49.30 | 84.47 |

**Table 3c:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Cut-point** | **Sensitivity** | **Specificity** | **% correct** |
| ***HRS training data (N=760)*** | | | | |
| Crimmins | 0.61 | 87.60 | 86.06 | 86.58 |
| Hurd | 0.31 | 85.27 | 84.46 | 84.74 |
| Wu | 0.29 | 89.15 | 82.47 | 84.74 |
| ***HRS validation data (N=515)*** | | | | |
| Crimmins | 0.15 | 83.10 | 69.14 | 71.07 |
| Hurd | 0.26 | 71.83 | 85.59 | 83.69 |
| Wu | 0.09 | 83.10 | 68.92 | 70.87 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Cut-point** | **Sensitivity** | **Specificity** | **% correct** |
| ***HRS training data (N=760)*** | | | | |
| Crimmins | 0.61 | 87.60 | 86.06 | 86.58 |
| Hurd | 0.59 | 72.48 | 93.43 | 86.32 |
| Wu | 0.54 | 76.74 | 90.04 | 85.53 |
| ***HRS validation data (N=515)*** | | | | |
| Crimmins | 0.92 | 33.80 | 97.07 | 88.35 |
| Hurd | 0.69 | 38.03 | 97.75 | 89.51 |
| Wu | 0.71 | 30.99 | 97.52 | 88.35 |

**Table 4a: Overall performance accuracy, with alternative Crimmins classification (using decision rule P(dementia) > 0.5 *and* P(dementia) > P(CIND))**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** |
| ***HRS training data (N=760)*** | | | |
| Herzog-Wallace | 53.49 | 96.61 | 81.97 |
| Langa-Kabeto-Weir | 75.19 | 83.27 | 80.53 |
| Crimmins | 89.92 | 79.08 | 82.76 |
| **Crimmins (accounting CIND)** | **81.01** | **87.05** | **85.00** |
| Hurd | 76.74 | 91.83 | 86.71 |
| Wu | 77.91 | 88.05 | 84.61 |
| ***HRS validation data (N=515)*** | | | |
| Herzog-Wallace | 18.31 | 97.75 | 86.80 |
| Langa-Kabeto-Weir | 40.85 | 89.19 | 82.52 |
| Crimmins | 61.97 | 82.21 | 79.42 |
| **Crimmins (accounting CIND)** | **46.48** | **89.19** | **83.30** |
| Hurd | 39.44 | 95.95 | 88.16 |
| Wu | 43.66 | 92.57 | 85.83 |

**Table 4b: Performance by respondent status, with alternative Crimmins classification (using decision rule P(dementia) > 0.5 *and* P(dementia) > P(CIND))**

Crimmins et al. only used a multinomial logit model for classifying cognition outcomes for self-respondents, and used a standard logit model (with outcomes normal vs. demented) for classifying outcomes for proxy-respondents. Thus performance metrics for proxies across the two classification methods.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** | **Sensitivity** | **Specificity** | **% correct** |
| **Self-respondent** | | | **Proxy-respondent** | | |
| ***HRS training data*** | **N=595** | | | **N=165** | | |
| Herzog-Wallace | 29.37 | 97.44 | 83.03 | 76.52 | 84.85 | 78.18 |
| Langa-Kabeto-Weir | 57.94 | 84.65 | 78.99 | 91.67 | 63.64 | 86.06 |
| Crimmins | 83.33 | 82.09 | 82.35 | 96.21 | 36.36 | 84.24 |
| **Crimmins (accounting CIND)** | **65.08** | **90.62** | **85.21** | **96.21** | **36.36** | **84.24** |
| Hurd | 57.14 | 93.39 | 85.71 | 95.45 | 69.70 | 90.30 |
| Wu | 61.90 | 90.83 | 84.71 | 93.18 | 48.48 | 84.24 |
| ***HRS validation data*** | **N=485** | | | **N=30** | | |
| Herzog-Wallace | 16.07 | 98.37 | 88.87 | 26.67 | 80.00 | 53.33 |
| Langa-Kabeto-Weir | 35.71 | 89.98 | 83.71 | 60.00 | 66.67 | 63.33 |
| Crimmins | 60.71 | 83.92 | 81.24 | 66.67 | 33.33 | 50.00 |
| **Crimmins (accounting CIND)** | **41.07** | **91.14** | **85.36** | **66.67** | **33.33** | **50.00** |
| Hurd | 35.71 | 96.27 | 89.28 | 53.33 | 86.67 | 70.00 |
| Wu | 35.71 | 93.94 | 87.22 | 73.33 | 53.33 | 63.33 |

**Table 4c: Performance by race/ethnicity, with alternative Crimmins classification (using decision rule P(dementia) > 0.5 *and* P(dementia) > P(CIND))**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Sens** | **Spec** | **% correct** | **Sens** | **Spec** | **% correct** | **Sens** | **Spec** | **% correct** |
| **Non-Hispanic White** | | | **Non-Hispanic Black** | | | **Hispanic** | | |
| ***HRS training data*** | **N=526** | | | **N=140** | | | **N=76** | | |
| Herzog-Wallace | 52.33 | 99.44 | 84.03 | 55.17 | 91.46 | 76.43 | 55.00 | 87.50 | 78.95 |
| Langa-Kabeto-Weir | 71.51 | 90.96 | 84.60 | 81.03 | 60.98 | 69.29 | 85.00 | 71.43 | 75.00 |
| Crimmins | 87.21 | 83.90 | 84.98 | 96.55 | 64.63 | 77.86 | 95.00 | 71.43 | 77.63 |
| **Crimmins (accounting CIND)** | **82.56** | **89.55** | **87.26** | **77.59** | **79.27** | **78.57** | **80.00** | **83.93** | **82.89** |
| Hurd | 73.26 | 94.07 | 87.26 | 86.21 | 82.93 | 84.29 | 85.00 | 89.29 | 88.16 |
| Wu | 77.33 | 93.50 | 88.21 | 74.14 | 79.27 | 77.14 | 90.00 | 67.86 | 73.68 |
| ***HRS validation data*** | **N=369** | | | **N=97** | | | **N=35** | | |
| Herzog-Wallace | 13.64 | 98.46 | 88.35 | 26.67 | 97.56 | 86.60 | 20.00 | 96.00 | 74.29 |
| Langa-Kabeto-Weir | 36.36 | 93.85 | 86.99 | 53.33 | 76.83 | 73.20 | 40.00 | 76.00 | 65.71 |
| Crimmins | 61.36 | 84.31 | 81.57 | 60.00 | 79.27 | 76.29 | 60.00 | 76.00 | 71.43 |
| **Crimmins (accounting CIND)** | **52.27** | **91.08** | **86.45** | **33.33** | **86.59** | **78.35** | **30.00** | **84.00** | **68.57** |
| Hurd | 45.45 | 96.62 | 90.51 | 26.67 | 93.90 | 83.51 | 30.00 | 96.00 | 77.14 |
| Wu | 43.18 | 95.08 | 88.89 | 40.00 | 85.37 | 78.35 | 50.00 | 92.00 | 80.00 |

**Table 4d: Performance by age, with alternative Crimmins classification (using decision rule P(dementia) > 0.5 *and* P(dementia) > P(CIND))**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** | **Sensitivity** | **Specificity** | **% correct** |
| **Under 80 years** | | | **80 years or older** | | |
| ***HRS training data*** | **N=364** | | | **N=396** | | |
| Herzog-Wallace | 52.94 | 97.97 | 89.56 | 53.68 | 94.66 | 75.00 |
| Langa-Kabeto-Weir | 77.94 | 86.15 | 84.62 | 74.21 | 79.13 | 76.77 |
| Crimmins | 80.88 | 90.54 | 88.74 | 93.16 | 62.62 | 77.27 |
| **Crimmins (accounting CIND)** | **66.18** | **93.58** | **88.46** | **86.32** | **77.67** | **81.82** |
| Hurd | 58.82 | 97.30 | 90.11 | 83.16 | 83.98 | 83.59 |
| Wu | 64.71 | 93.24 | 87.91 | 82.63 | 80.58 | 81.57 |
| ***HRS validation data*** | **N=225** | | | **N=290** | | |
| Herzog-Wallace | 11.76 | 97.60 | 91.11 | 20.37 | 97.88 | 83.45 |
| Langa-Kabeto-Weir | 29.41 | 90.87 | 86.22 | 44.44 | 87.71 | 79.66 |
| Crimmins | 41.18 | 91.83 | 88.00 | 68.52 | 73.73 | 72.76 |
| **Crimmins (accounting CIND)** | **35.29** | **95.19** | **90.67** | **50.00** | **83.90** | **77.59** |
| Hurd | 17.65 | 99.52 | 93.33 | 46.30 | 92.80 | 84.14 |
| Wu | 35.29 | 95.19 | 90.67 | 46.30 | 90.25 | 82.07 |

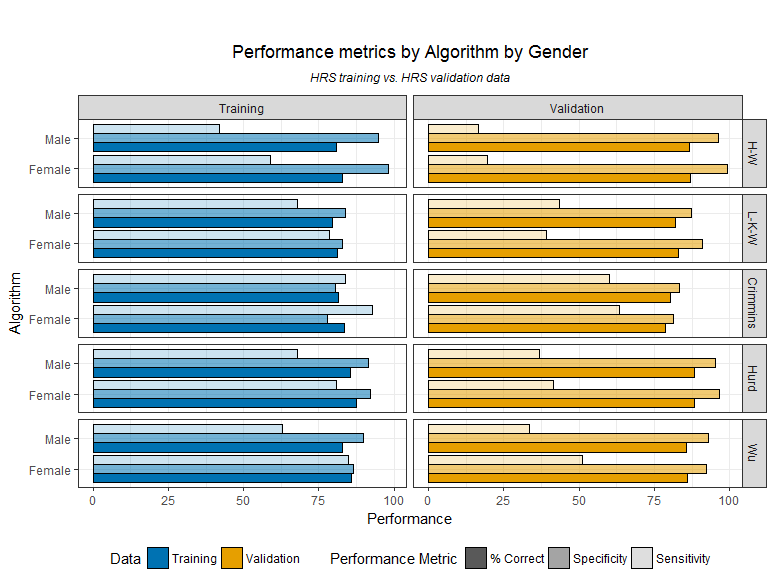
**Table 4e: Performance by gender, with alternative Crimmins classification (using decision rule P(dementia) > 0.5 *and* P(dementia) > P(CIND))**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** | **Sensitivity** | **Specificity** | **% correct** |
| **Male** | | | **Female** | | |
| ***HRS training data*** | **N=308** | | | **N=452** | | |
| Herzog-Wallace | 41.98 | 94.71 | 80.84 | 58.76 | 98.18 | 82.74 |
| Langa-Kabeto-Weir | 67.90 | 83.70 | 79.55 | 78.53 | 82.91 | 81.19 |
| Crimmins | 83.95 | 80.62 | 81.49 | 92.66 | 77.82 | 83.63 |
| **Crimmins (accounting CIND)** | **72.84** | **88.11** | **84.09** | **84.75** | **86.18** | **85.62** |
| Hurd | 67.90 | 91.63 | 85.39 | 80.79 | 92.00 | 87.61 |
| Wu | 62.96 | 89.87 | 82.79 | 84.75 | 86.55 | 85.84 |
| ***HRS validation data*** | **N=245** | | | **N=270** | | |
| Herzog-Wallace | 16.67 | 96.28 | 86.53 | 19.51 | 99.13 | 87.04 |
| Langa-Kabeto-Weir | 43.33 | 87.44 | 82.04 | 39.02 | 90.83 | 82.96 |
| Crimmins | 60.00 | 83.26 | 80.41 | 63.41 | 81.22 | 78.52 |
| **Crimmins (accounting CIND)** | **46.67** | **91.16** | **85.71** | **46.34** | **87.34** | **81.11** |
| Hurd | 36.67 | 95.35 | 88.16 | 41.46 | 96.51 | 88.15 |
| Wu | 33.33 | 93.02 | 85.71 | 51.22 | 92.14 | 85.93 |

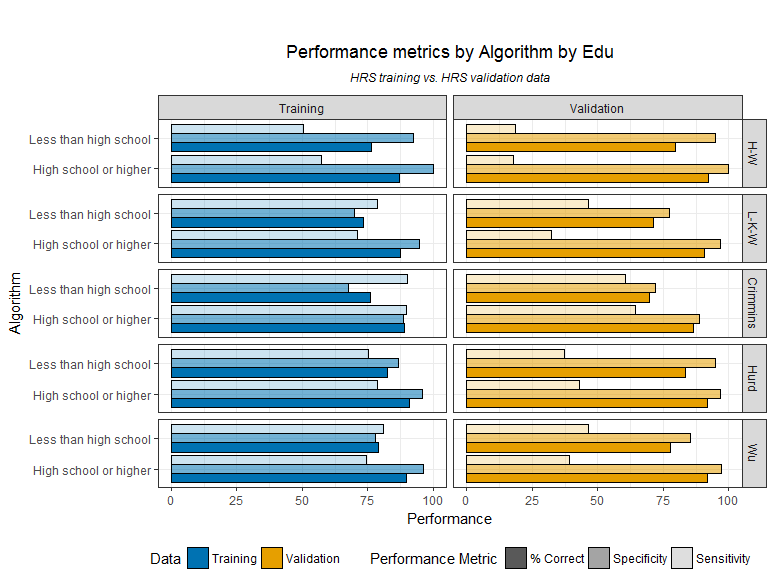
**Table 4e: Performance by gender, with alternative Crimmins classification (using decision rule P(dementia) > 0.5 *and* P(dementia) > P(CIND))**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Sensitivity** | **Specificity** | **% correct** | **Sensitivity** | **Specificity** | **% correct** |
| **Self-respondent** | | | **Proxy** | | |
| ***HRS training data*** | **N=595** | | | **N=165** | | |
| Herzog-Wallace | 29.37 | 97.44 | 83.03 | 76.52 | 84.85 | 78.18 |
| Langa-Kabeto-Weir | 57.94 | 84.65 | 78.99 | 91.67 | 63.64 | 86.06 |
| Crimmins | 83.33 | 82.09 | 82.35 | 96.21 | 36.36 | 84.24 |
| **Crimmins (accounting CIND)** | **65.08** | **90.62** | **85.21** | **96.21** | **36.36** | **84.24** |
| Hurd | 57.14 | 93.39 | 85.71 | 95.45 | 69.70 | 90.30 |
| Wu | 61.90 | 90.83 | 84.71 | 93.18 | 48.48 | 84.24 |
| ***HRS validation data*** | **N=485** | | | **N=30** | | |
| Herzog-Wallace | 16.07 | 98.37 | 88.87 | 26.67 | 80.00 | 53.33 |
| Langa-Kabeto-Weir | 35.71 | 89.98 | 83.71 | 60.00 | 66.67 | 63.33 |
| Crimmins | 60.71 | 83.92 | 81.24 | 66.67 | 33.33 | 50.00 |
| **Crimmins (accounting CIND)** | **41.07** | **91.14** | **85.36** | **66.67** | **33.33** | **50.00** |
| Hurd | 35.71 | 96.27 | 89.28 | 53.33 | 86.67 | 70.00 |
| Wu | 35.71 | 93.94 | 87.22 | 73.33 | 53.33 | 63.33 |

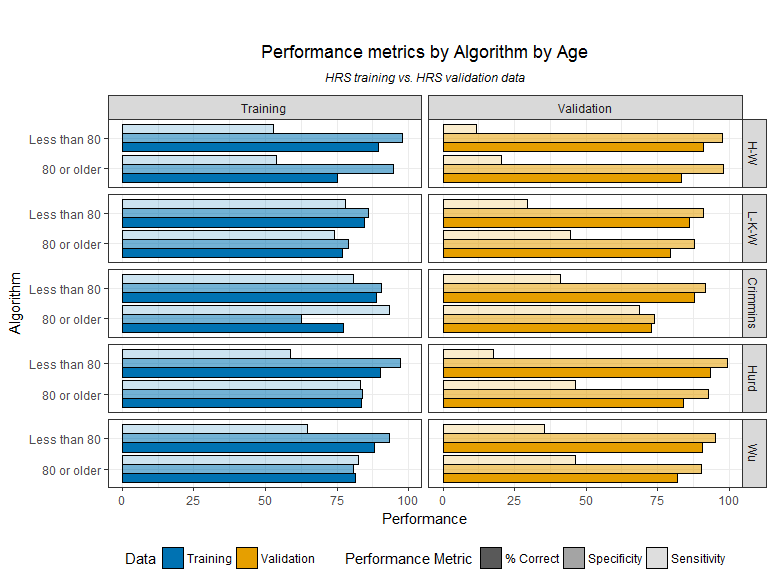
**Figure 1a: Performance metrics by algorithm by gender**



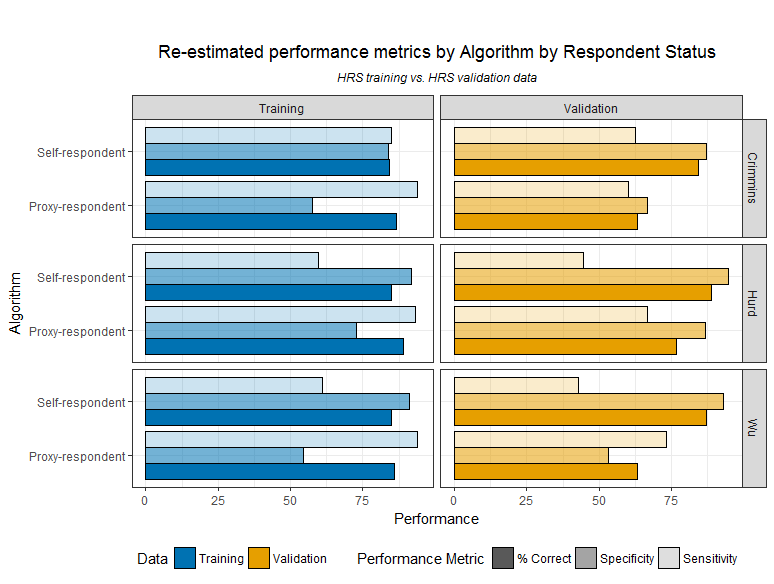
**Figure 1b: Performance metrics by algorithm by education**



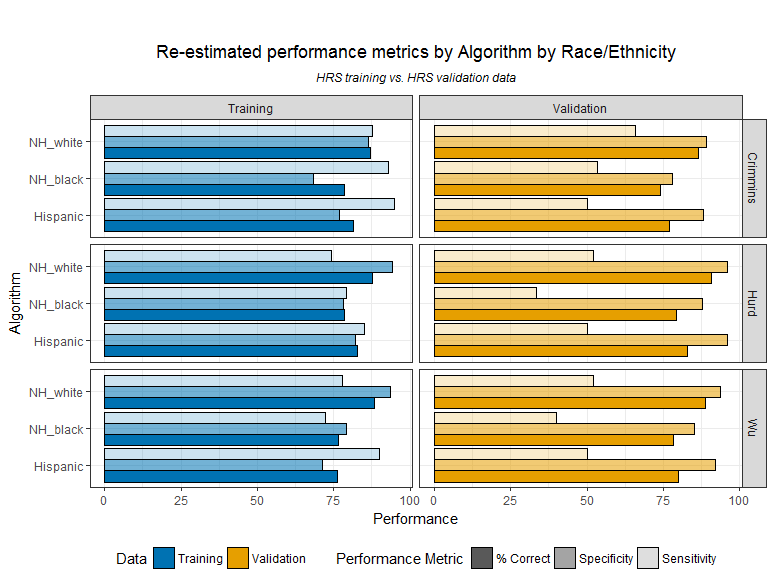
**Key**

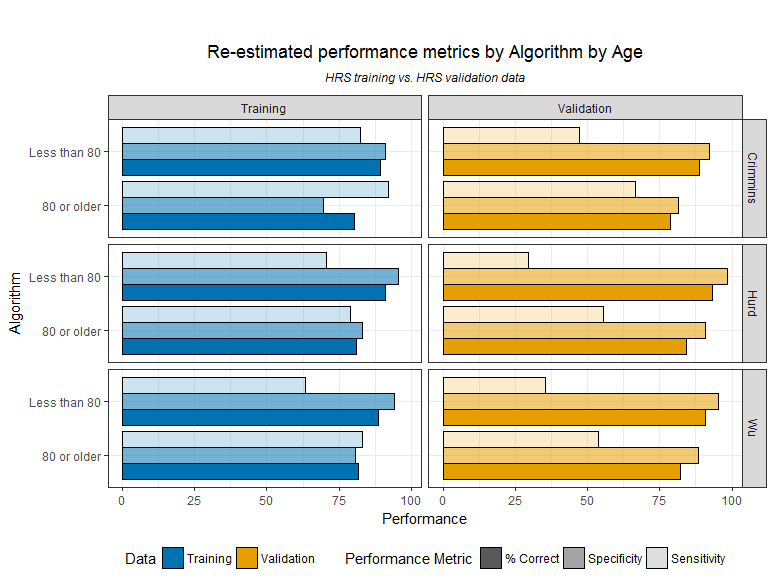
**Figure 2a: Re-estimated performance metrics by algorithm by respondent status**



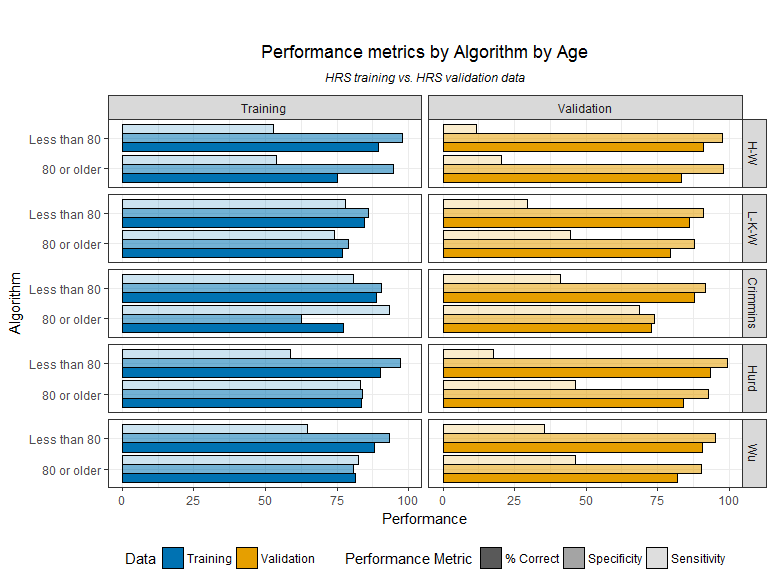
**Figure 3b: Re-estimated performance metrics by algorithm by race**



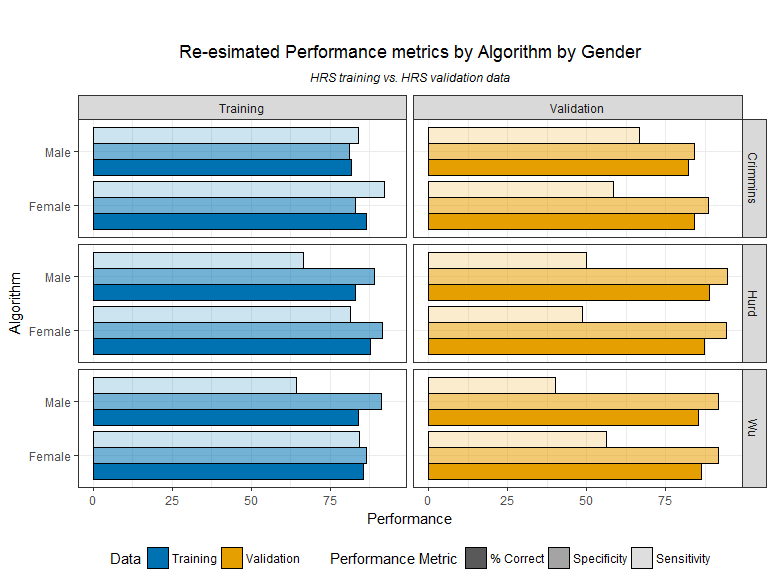
**Figure 2c: Re-estimated performance metrics by algorithm by age**



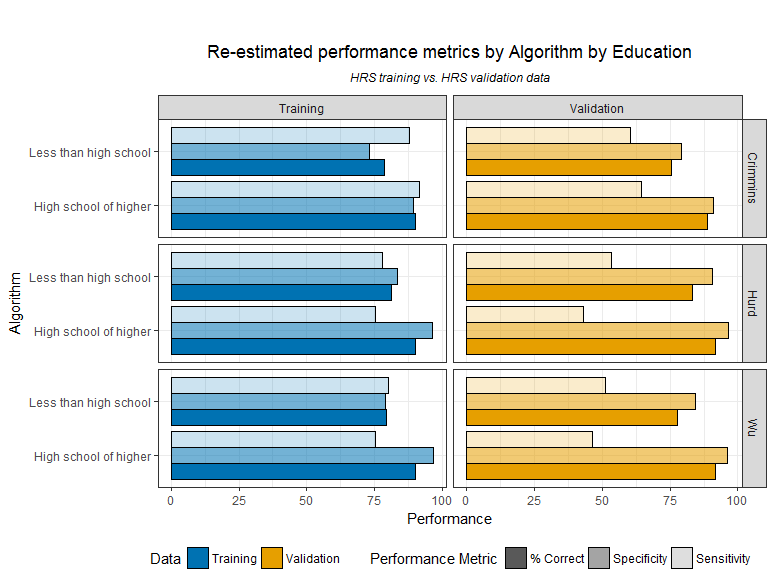
**Key**

**Figure 2d: Re-estimated performance metrics by algorithm by gender**



**Figure 2e: Re-estimated performance metrics by algorithm by education**



**Key**

