Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: https://www.journals.elsevier.com/eclinicalmedicine

Research Paper

Ensemble machine learning classification of daily living abilities among older people with HIV

Robert Paul[a,b,*], Torie Tsuei[c], Kyu Cho[a], Andrew Belden[a], Benedetta Milanini[c], Jacob Bolzenius[a], Shireen Javandel[c], Joseph McBride[a], Lucette Cysique[d], Samantha Lesinski[a], Victor Valcour[c,e]

[a] *Missouri Institute of Mental Health, University of Missouri-St. Louis, St. Louis, MO 63121-4400, United States*
[b] *Department of Psychological Sciences, University of Missouri-St. Louis, St. Louis, MO, United States*
[c] *Memory and Aging Center, Department of Neurology, University of California San Francisco, San Francisco, CA, United States*
[d] *School of Psychology, University of New South Wales, Sydney, Australia*
[e] *Global Brain Health Institute, University of California San Francisco, San Francisco, CA, United States*

ARTICLE INFO

Article History:
Received 22 October 2020
Revised 11 March 2021
Accepted 17 March 2021
Available online xxx

Keywords:
ADLs
HIV
Aging
Machine learning

ABSTRACT

Background: clinically relevant methods to identify individuals at risk for impaired daily living abilities secondary to neurocognitive impairment (ADLs) remain elusive. This is especially true for complex clinical conditions such as HIV-Associated Neurocognitive Disorders (HAND). The aim of this study was to identify novel and modifiable factors that have potential to improve diagnostic accuracy of ADL risk, with the long-term goal of guiding future interventions to minimize ADL disruption.

Methods: study participants included 79 people with HIV (PWH; mean age = 63; range = 55–80) enrolled in neuroHIV studies at University California San Francisco (UCSF) between 2016 and 2019. All participants were virally suppressed and exhibited objective evidence of neurocognitive impairment. ADL status was defined as either normative (n = 39) or at risk (n = 40) based on a task-based protocol. Gradient boosted multivariate regression (GBM) was employed to identify the combination of variables that differentiated ADL subgroup classification. Predictor variables included demographic factors, HIV disease severity indices, brain white matter integrity quantified using diffusion tensor imaging, cognitive test performance, and health co-morbidities. Model performance was examined using average Area Under the Curve (AUC) with repeated five-fold cross validation.

Findings: the univariate GBM yielded an average AUC of 83% using Wide Range Achievement test 4 (WRAT-4) reading score, self-reported thought confusion and difficulty reading, radial diffusivity (RD) in the left external capsule, fractional anisotropy (FA) in the left cingulate gyrus, and Stroop performance. The model allowing for two-way interactions modestly improved classification performance (AUC of 88%) and revealed synergies between race, reading ability, cognitive performance, and neuroimaging metrics in the genu and uncinate fasciculus. Conversion of Neuropsychological Assessment Battery Daily Living Module (NAB-DLM) performance from raw scores into T scores amplified differences between White and non-White study participants.

Interpretation: demographic and sociocultural factors are critical determinants of ADL risk status among older PWH who meet diagnostic criteria for neurocognitive impairment. Task-based ADL assessment that relies heavily on reading proficiency may artificially inflate the frequency/severity of ADL impairment among diverse clinical populations. Culturally relevant measures of ADL status are needed for individuals with acquired neurocognitive disorders, including HAND.

1. Introduction

Prevailing diagnostic nosologies for acquired neurocognitive disorders (e.g., Diagnostic and Statistical Manual for Mental Disorders-5th edition) [1] are based on the premise that altered brain structure impairs cognitive function relative to premorbid abilities based on

[*] Corresponding author at: Missouri Institute of Mental Health, University of Missouri-St. Louis, St. Louis, MO 63121-4400, United States.
E-mail address: paulro@umsl.edu (R. Paul).

https://doi.org/10.1016/j.eclinm.2021.100845

Research in Context

Evidence before this study

We searched PubMed, Google Scholar, and Google through February 2021, using the terms: "activities of daily living (ADLs)" and "HIV" and any combination of the following: "neuroimaging", "machine learning", "cognition", "race", and "HAND". Multiple studies have examined psychosocial and neuroimaging correlates of cognitive impairment in people with HIV (PWH). Several studies report modest associations between alterations in brain integrity and ADLs in PWH, but none have integrated psychosocial, cognitive, and neuroimaging metrics into a single explanatory model of ADL risk.

Added value of this study

This is the first study to establish a data-driven predictive model of ADL risk status derived from multimodal bio behavioral and neurological outcomes to identify key interactions among predictors that point toward more specific mechanisms that explain individual differences in task-based ADL performance. Results from the study highlight the importance of structural brain integrity quantified via neuroimaging and performance on tests of executive function. Additionally, Black race and low reading ability were important factors that distinguished individuals according to ADL risk group.

Implications of all the available evidence

Study findings emphasize the importance of brain structure and cognitive function as predictors of ADL risk, consistent with existing diagnostic criteria for neurocognitive disorders. Importantly, the findings also highlight challenges associated with existing ADL paradigms and the potential for task-based methods to misclassify racially diverse individuals as impaired. Development of culturally relevant methods to ascertain ADL status represents a clinical and research priority for clinical populations with neurocognitive disorders.

subjective report or performance on cognitive tests. Mild levels of cognitive impairment have minimal impact on an individual's ability to complete real world activities, referred to as activities of daily living (ADLs). By contrast, moderate to severe cognitive impairment results in demonstrable inability to complete instrumental skills (financial management, medication adherence, etc.) and eventually, basic skills (bathing, grooming, etc.) [2].

In practice, predicting risk of ADL impairment is a significant clinical challenge. This is particularly true when ascertaining functional capacity among individuals with complex clinical conditions that involve confounding factors (psychosocial, co-morbid health concerns, etc.) that potentially explain below average performance on standardized testing. Both of these concerns complicate the assessment of ADL risk among the increasingly large population of people with HIV (PWH) who are now reaching advanced age due to the success of suppressive antiretroviral therapy (ART) [3–5].

Numerous studies report modest associations between white matter alterations on diffusion tensor imaging (DTI) and cognitive dysfunction among PWH [6,7], as well as associations between cognitive dysfunction and ADL impairment [8–10], but less is known about the interplay across these dimensions among older PWH who have achieved viral suppression. Further, prior studies of ADL risk among older PWH have not examined interactions between ADL risk, brain structure and function, and social determinants of health that likely account for a significant amount of variance in ADL risk. Here, we employed a novel combination of data driven and traditional inferential methods to develop a predictive model of ADL risk in a sample of older PWH who completed standardized, multi-dimensional assessments. We predicted that specific features extracted from neuroimaging, psychosocial, and HIV disease severity metrics would classify individuals according to ADL status. We also hypothesized that interactions across predictor variables would improve classification performance and provide novel insights into mechanisms that differentiate individuals according to ADL subgroup designation. The ultimate goal of this work is to identify novel and modifiable factors that have potential to improve diagnostic accuracy of neurocognitive disorders and, ideally, to define clinical targets to guide interventions capable of minimizing ADL disruption.

2. Methods

2.1. Study design, setting, and participants

Study participants included 79 virally suppressed (plasma viral load <50 copies/mL) older PWH (mean age 63, range 55–80). Inclusion criteria were: (1) outpatient, ambulatory status; (2) 55 years of age or older; (3) English as the primary language; (4) objective evidence of neurocognitive impairment; (5) self-reported ADL decline documented using the Patient's Assessment of Own Functioning Inventory (PAOFI) [11]; and (5) capacity to provide informed consent. Exclusion criteria: (1) neurologic injury other than HIV; (2) recreational cocaine or methamphetamine use in the six months before study participation or use of alcohol or other drugs deemed by case conference to confound HAND determination; and (3) untreated major psychiatric disorder (e.g., schizophrenia or bipolar disorder). The study was approved by the affiliated Institutional Review Boards (IRBs). Participants provided written consent following a thorough explanation of study procedures. Individuals were compensated for their time and transportation Table 1.

2.2. Procedure

Potential study participants were identified via IRB-approved phone screen. Interested individuals who met general inclusion/exclusion criteria were scheduled for an initial clinic visit to confirm eligibility. A structured clinical interview was completed by a neurologist to identify neurologic injury other than HIV (e.g., stroke). Once enrolled, individuals completed all procedures within a four-month window; the exact sequence of assessments was dependent on scheduling (e.g., scanner availability). Research psychometricians completed annual training and re-certification for cognitive testing, and participants were provided rest breaks as needed to ensure high quality data acquisition. HAND diagnosis was determined through case conference review of demographic, clinical, and cognitive test performance, with at least three HIV experts involved in the consensus diagnosis procedure.

2.3. Main outcome variable

2.3.1. ADL subgroup designation

The typical approach to ascertain ADL status involves the use of subjective rating scales, which are inherently prone to bias from under- and over-reporting [12]. Information from collateral sources (e.g., family, spouse, partner) may not be available or sufficient to determine ADL status [13]. Performance-based measures offer an alternative approach to determine ADL capacity using measures intended to represent real-world living skills. Prior studies of task-based ADL performance in the HIV field have typically relied on very specific aspects of instrumental ADLs, such as the Medication Management Test [14]. However, this approach is time-intensive and the

Table 1
Sociodemographic and clinical characteristics for the total sample, by race, and by ADL subgroup status.

	Total Sample (N = 79)	White Participants (n = 59)	Black Participants (n = 13)	Non-White/ Non-Black (n = 7)	ADL Non-Risk (n = 40)	ADL Risk Group (n = 39)
Age in years: Mean (SD)	63·1 (5·1)	63·3 (4·9)	64·1 (5·9)	59·7 (3·7)	64·3 (5·1)	61·9 (4·8)
Education in years: Mean (SD)	15·8 (2·5)	16·3 (2·3)	14·5 (2·4)	13·7 (2·1)	16·5 (2·5)	15·0 (2·2)
% Male	95%	98%	77%	100%	95%	95%
% MSM	79%	88%	31%	86%	83%	74%
Nadir CD4 T-cell count: Median (IQR)	183 (219)	180 (159)	217 (296)	100 (288)	199 (176)	128 (262)
Current CD4+ T-cell count: Median (IQR)	620 (382)	578 (321)	731 (555)	781 (902)	643 (333)	540 (438)
Plasma HIV RNA suppression	100%	100%	100%	100%	100%	100%
WRAT-4 T-score: Mean (SD)	50 (10)	51·50 (8·86)	42 (14·98)	48·81 (2·43)	52·8 (9·9)	46·7 (9·2)
Geriatric Depression Scale: Mean (SD)	9·4 (6·3)	9·9 (6·7)	7·8 (6·0)	9·0 (4·0)	9·5 (6·1)	9·4 (6·7)

MSM = Men who have sex with men. SD = Standard deviation. IQR = Interquartile Range. One participant reported a history of learning difficulty, and one participant reported current use of an opioid medication for chronic pain. Non-White/Non-Black subgroup was comprised of the following: Native Hawaiian or Pacific Islander (n = 2), Asian (n = 1), and (n = 4) endorsed “other”.

results do not address a broad range of ADLs important to patients, families, and care providers (e.g., driving, financial management).

The current study examined task-based ADL performance using the Neuropsychological Assessment Battery Daily Living Module (NAB-DLM) [15]. The NAB-DLM includes five subtests: (1) Daily Living Immediate and Delayed Memory; (2) Bill Payment; (3) Judgment; (4) Map Reading; and (5) Driving Scenes. The Daily Living Memory subtest required individuals to learn and remember information related to medication instructions as well as a fictitious name, address, and phone number. Bill Payment required individuals to demonstrate the steps involved in payment and record keeping of a fictitious utility bill; a blank check, check ledger, and envelope were provided. The Judgment subtest required participants to answer questions pertaining to home safety, health, and medical issues. On Map Reading, participants were shown a fictitious map depicting highways, boulevards with street names, and directional markers and then were asked questions about the information (e.g., number of miles between points). On the Driving Scenes subtest, participants were shown a line drawing depiction of a two-lane road viewed from the perspective of the driver. Subsequently, modified scenes were presented, and participants were required to identify the specific modifications. Raw scores were converted to demographically adjusted T-scores using normative data from the NAB manual.

We defined two ADL subgroups based on the total T-score for the NAB-DLM module. A T-score of 45 was employed to differentiate normative performance (T-score of $\geq$45) vs. at risk performance (T-score $<$45). We used the full-scale total score rather than individual domain subscales of the NAB-DLM as recommended by the NAB manual. The designated cutoff score of 45 represents performance that is one-half standard deviation below normative expectations, which is recognized as a salient threshold for detecting clinically relevant effects [16]. The cutoff also yielded similar sample sizes between the normative vs. at risk subgroups, a design consideration that supports reliability of the machine learning approach [17].

2.4. Predictor variables

2.4.1. Imaging acquisition

Diffusion images were acquired using a 3T Siemens Prisma Fit MRI with a 64-channel head coil. A single shot spin- echo planar imaging (SE-EPI) sequence was acquired for 69 axial slices providing whole brain coverage, using 2·0 mm isotropic voxels, 85/180 flip/refocusing angle. An integrated parallel acquisition technique (iPAT) acceleration factor of two and a multi-band acceleration factor of three was utilized with an EPI factor of 110. Two B0 volumes with opposite phase encoding (AP/PA) were acquired with TE=7080, TR=72·2 ms. Diffusion-weighted parameters were as follows: TE=2420 and TR=72·2 ms used to three multi-shells (10 vol) with 96 non-collinear diffusion sensitization directions at b = 2500 s/mm^2, 48 directions at b = 1000 s/mm^2, and 30 directions at 500 s/mm^2.

Scans were visually inspected and denoised to remove artifacts. Diffusion-weighted images were registered to the first B0 vol using MCFLIRT in FSL [18]. Relative and absolute displacement thresholds were set at 1 mm. Brain tissue was extracted by applying a median Otsu function [19] to create a mask by applying a 4 mm radius with four iterations to the B0 acquisition. Images were then corrected for eddy currents and fit to tensor eigenvalues using the Diffusion Imaging in Python Package [20]. A non-linear least squares approach was utilized to derive fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD), and radial diffusivity (RD) maps normalized and warped to group template space using Diffusion Tensor Imaging—Tool Kit (DTI-TK). Skeletonized FA maps eliminated anatomical variability by projecting voxels of maximum anisotropy onto a group template skeleton for registered statistical comparison.

Five regions of interest (ROIs) were derived from the Johns Hopkins University (JHU) ICBM-DTI-81 atlas. The JHU atlas space was transformed to DTI-TK space using warping parameters. Advanced Normalization Tools (ANTs; http://stnava.github.io/ANTs/) was utilized to extract FA, MD, AD, and RD from averaged, transformed group maps. The five ROIs included the genu of the corpus callosum, uncinate fasciculus, superior longitudinal fasciculus, corticospinal tract, and the cingulate gyrus portion of the cingulum bundle. These tracts were selected *a priori* based on results from prior studies demonstrating relevance to cognitive performance in PWH, including alterations in brain connectivity observed in older PWH [6]. FA, MD, RD, and AD from each hemisphere were included in the analyses.

2.4.2. Neurocognitive assessment

Participants completed neurocognitive tests with known sensitivity to HIV and aging (see Lezak et al. [21] for additional information on each test and relevant citations). The cognitive battery is provided below by domain. However, the raw scores from each test (rather than aggregated domain scores) were included in the analysis. *Attention*: Stroop Color Naming, California Verbal Learning Test-II (CVLT-II) Trial 1, and Digit Span Forward; *Psychomotor/Motor*: Trail Making A, Finger Tapping (dominant and non-dominant hand), and Grooved Pegboard (dominant and nondominant hands); *Executive Function:* Trail Making B, Stroop Interference Test, Lexical Fluency (D words), Digit Span Backward, and Design Fluency; *Learning and Memory:* total recall across the learning trials of the CVLT-II and recall on the long delay trial. Visual learning and memory were examined using the Benson Figure Recall; *Visuospatial:* Benson Figure Copy and Visual Object and Space Perception (VOSP); *Language:* Category Fluency and Boston Naming Test-Short Form (BNT-SF).

Subjective ratings of cognition and ADL status were recorded using two self-report questionnaires. As noted above, the PAOFI [11] was administered to document functional decline believed to be associated with cognitive difficulties. The PAOFI is a multi-dimensional assessment of cognitive and daily living skills comprised of 33 questions. Responses are rated using a 6-point scale ("almost never" to "almost always"). Item level responses were included in the primary analyses to define specific components of self-reported functional decline that corresponded to ADL risk from the perspective of the study participant. Participants also completed a self-report measure of basic activities of daily living (BADLs) and instrumental activities of daily living (IADLs) using a measure based on scales developed by Katz [22] and Lawton and Brody [23] The questionnaire included 16 questions pertaining to ADL function. Participants rated their level of performance as either independent or in need of at least some assistance (to full dependence). Unlike the PAOFI, this additional measure of ADL status did not differentiate between cognitive and physical etiologies to the ADL decline. Item level responses were included in the machine learning analysis.

2.4.3. Sociodemographic variables

Age, number of years of education, and race were included. Race was included as a three-level variable that included Black participants, White participants, and non-Black/non-White participants. The non-Black/non-White group was made up of the following: Native Hawaiian or Pacific Islander (*n* = 2), Asian (*n* = 1), and "other" (*n* = 4). The at-risk ADL subgroup (*n* = 39) consisted of 24 White participants, 10 Black participants, 1 Asian participant, 1 Native Hawaiian or Pacific Islander participant, and 3 "other" race individuals, while the normative ADL subgroup (*n* = 40) consisted of 35 White participants, 3 Black participants, 1 Native Hawaiian or Pacific Islander participant, and 1 "other" race individuals. Participants completed the Wide Range Achievement Test-4 Word Reading test (WRAT-4) [24], a measure of word reading performance that requires individuals to pronounce words spelled with either typical or atypical phonetics. The WRAT-4 has been utilized in prior studies of PWH as a proxy for educational quality and cognitive reserve [25,26].

2.4.4. Mood assessment

Participants completed the Geriatric Depression Scale-II (GDS-II) [27]. Total score served as the dependent variable.

2.4.5. HIV variables

Plasma CD4 T-cell count was quantified using standard laboratory procedures from blood acquired at study enrollment. Nadir plasma CD4 T-cell count was documented from medical records. To be eligible for this study, all participants were defined as undetectable based on having no history of two or more consecutive clinical visits (approximately three months apart) with detectable plasma viral load (i.e., >50 copies/mL). Viral blips (defined as a single event of detectability) were not exclusionary.

2.5. Statistical analysis

Demographic and HIV clinical variables were summarized using descriptive methods. Categorical comparisons (e.g., race) were examined using Chi Square. The classification model of ADL subgroup designation (normative vs. at risk) was built using gradient boosted multivariate regression (GBM). A more detailed review of GBM is provided by Miller et al. [28] In brief, GBM combines the outcomes of two or more individual machine learning prediction models to establish a final composite model that benefits from the errors generated by the preliminary base models. The ultimate goal is to reduce the variance component of the prediction error by adding bias (i.e., in the context of the bias-variance trade-off) during the learning trials. Conceptually, GBM benefits from a "wisdom of crowds" approach, with prediction performance similar to meta analytic strategies such as Super Learner [29]. GBM has been successfully employed to identify novel explanatory mechanisms of complex psychiatric disorders [30−33], and neuroimaging substrates of Alzheimer's disease (AD) [34]. Additionally, we recently employed GBM to classify older PWH who meet clinical criteria for frailty [35], and to predict neurodevelopmental outcomes in children infected with HIV perinatally [36].

2.5.1. Feature selection and model evaluation

Feature selection was conducted using an in-house program based on SciKit and PDPBox [35,36]. Class labels were determined using a probability score based on the sigmoid function (1/(1 + *e*^(-x))). A 0·5 decision boundary and gradient descent were implemented to minimize prediction error. Highly correlated features (*r*>0·65) were managed via retention of the feature with the highest mutual criterion information. Missing data are permissible using GBM given no greater than 20% missing data per variable, a rule that was exceeded in this data set by over 90% of entered variables (out of over 200). Missing data for the remaining cases were imputed based on median values per variable. The final classification model was based on the six input features with the highest mutual criterion information. We have utilized this approach in past studies to reduce the risk of overfitting as well as facilitate clinical interpretation of the final classification model [37,38]. Additionally, we tested whether a larger number of features resulted in meaningful gain in classification accuracy, defined by an increase in average area under the Curve (AUC) >1SD from the base model. This approach ensured that the final algorithm was comprised of the most salient features. For the current analysis, model saturation was achieved with six input features. Separate algorithms evaluated univariate features and two-way interactions. To minimize overfitting, the stability of the models was evaluated using five-fold cross validation with five repeated trials (25 validation trials). The algorithms were trained using each iteration of the folds. The final metric of model performance was AUC, averaged across the repeated validation trials.

2.5.2. Comparison to logistic regression

Results from the univariate and interactive GBM models were compared to logistic regression, a traditional classification approach used for binary outcomes. To facilitate this comparison, we implemented a two-step process to build the logistic regression models. The first step ranked the predictive features by relative association strength from all pairwise correlation coefficients. The second step involved selection of the top six features defined in step one to examine the predictive relevance in the logistic regression. The first step increases model performance beyond what can be achieved through traditional regression by exploring and selecting the most salient correlates, but the method allows for a more direct comparison to model performance relative to the GBM. Logistic regression is more susceptible to missing data than GBM, and therefore missing data were removed in a list wise fashion.

2.6. Role of the funding source

The funding source had no role in the study design, collection, analysis, or interpretation of the data, writing of the report, or decision to submit the paper for publication. Dr Robert Paul had access to the dataset and responsible for the decision to submit the manuscript for publication.

3. Results

3.1. GBM classification of ADL subgroup status

The base GBM model with univariate features yielded an average AUC of 83% (Accuracy: 74%; Precision: 75%; Recall: 74%; F1 score: 75%; Fig. 1, panel a). This model indicated that (1) lower WRAT-4 scores; (2) greater frequency of thought confusion (PAOFI item #25); (3) lower RD in the left external capsule; (4) difficulty understanding

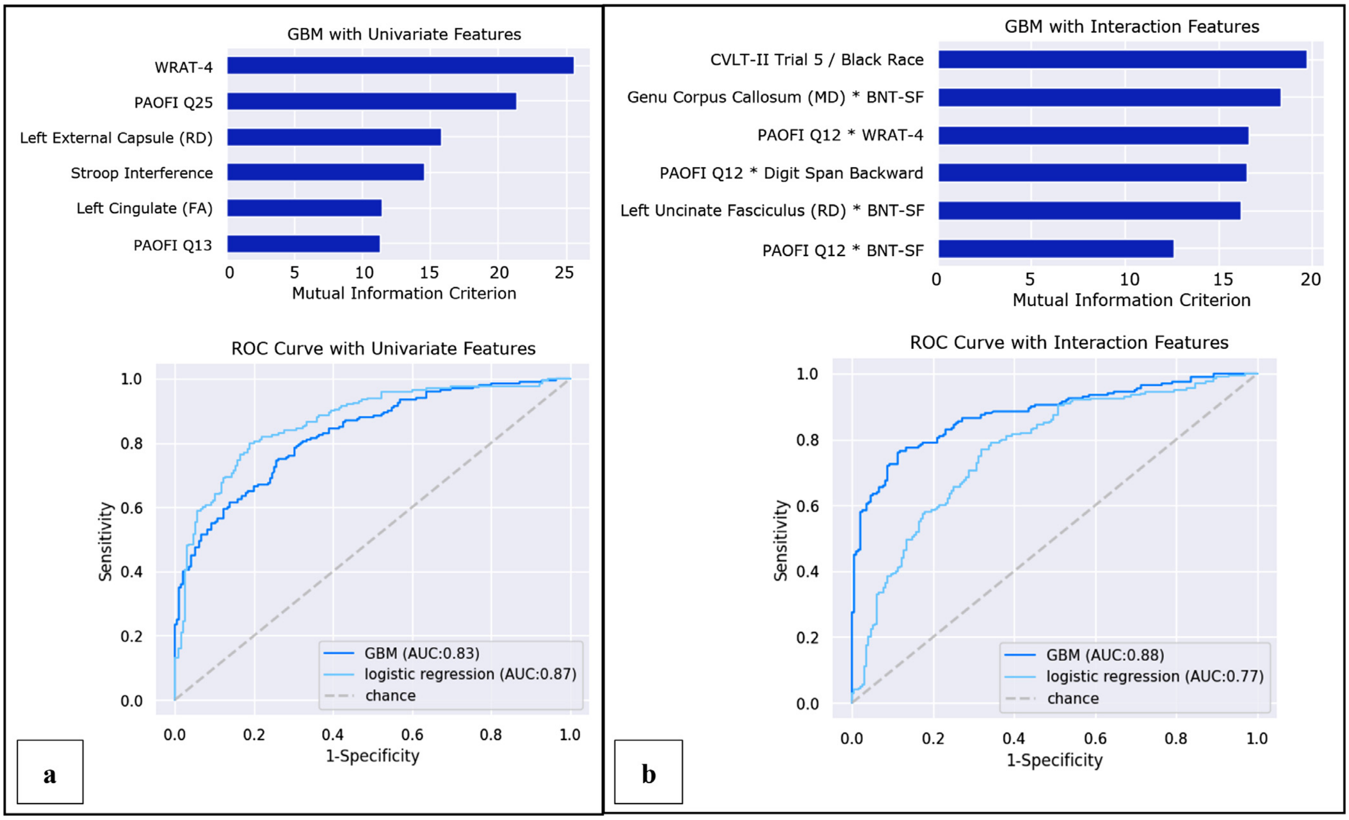

Fig. 1. Predictors of ADL risk groups defined by the univariate and interactive GBM models.

reading (PAOFI item #13); (5) lower performance on the Stroop interference task; and (6) higher FA in the left cingulate gyrus predicted membership in the at-risk ADL subgroup (Fig. 2, panel a).

3.2. *GBM classification of ADL subgroup status with two-way interaction features*

The GBM model allowing for two-way interactions yielded an average AUC of 88% (Accuracy: 80%; Precision: 82%; Recall: 79%; F1 score: 80%; Fig. 1, panel b) using the following features: (1) identification of Black race and lower CVLT-II trial 5 score; (2) frequent difficulty recognizing/identifying printed words (PAOFI item #12) and lower WRAT-4 score; (3) higher MD in the genu of the corpus callosum and lower BNT-SF score; (4) lower left hemisphere RD in the uncinate fasciculus and lower BNT-SF score; (5) self-reported difficulty recognizing/identifying printed words (PAOFI item #12) and lower performance on the Digit Span Backward test; (6) self-reported difficulties recognizing/identifying printed words (PAOFI item #12) and lower score on the BNT-SF (Fig. 2, panel b).

3.3. *Classification of ADL subgroup status using logistic regression*

Ranked by regression coefficient, the base logistic regression model yielded an average AUC of 87% (Accuracy: 80%; Precision: 79%; Recall: 82%; F1 score: 80%; not shown) based on (1) difficulty understanding reading material (PAOFI item #13); (2) lower BNT-SF score; (3) difficulties losing things or remembering where things are placed (PAOFI item #9); (4) self-reported difficulty getting in or out of a bed or chair; (5) confused or illogical thoughts (PAOFI item #25); and (6) lower score on Digits Span Forward.

The logistic regression allowing for two-way interactions yielded an average AUC of 77% (Accuracy: 72%; Precision: 70%; Recall: 78%; F1 score: 74%; not shown) based on: (1) higher RD in the left uncinate fasciculus with worse BNT-SF score; (2) self-reported difficulty with bathing and taking medication; (3) worse performance on Stroop Color Naming; (4) poorer performance on the Benson Figure Recall with worse BNT-SF; (5) difficulty understanding reading material (PAOFI item #13) and self-reported preference to be shown versus told how to do things (PAOFI item #14); and (6) self-reported difficulty with dressing oneself.

3.4. *Post-hoc analysis*

Given the relevance of race defined in the GBM models, we examined variability in NAB-DLM scores as a function of self-identified race. Average performance on the NAB Daily Living Memory Recall test (*T* = 36) and Bill Payment test (*T* = 35) was more than 1 SD below norms for Black participants. Similarly, the average score on the Bill Payment test was more than 1·5 SD (*T* = 34) below the mean for individuals who identified as neither White nor Black. Ridgeline plots (Fig. 3) depict significant race-related discrepancies between raw scores and corresponding T scores on the NAB-DLM. A two-point difference in raw scores was sufficient to drive the T scores for Black participants into the impaired range after converting raw scores into T scores using the norms in the NAB-DLM manual. See Supplemental Materials Table 2.

4. Discussion

Study findings provide important insights related to ADL risk assessment for older PWH. Consistent with our primary hypothesis, our machine learning approach yielded a robust classifier of ADL status. We also demonstrate that psychosocial factors (e.g., racial identity and reading proficiency) contribute to the prediction model for ADL risk subgroup status. This observation is important as existing diagnostic schemes for neurocognitive disorders do not provide

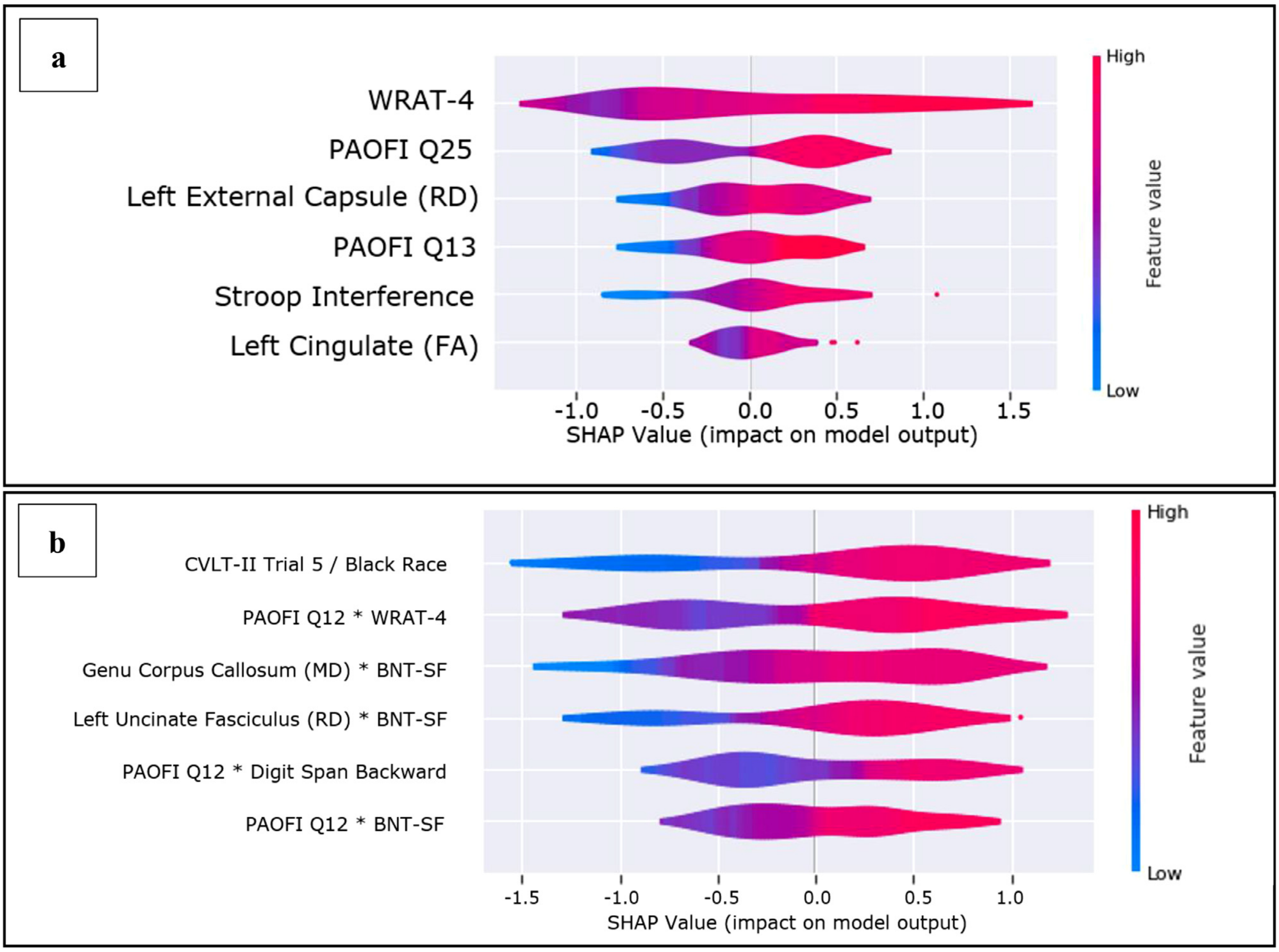

Fig. 2. Directionality of predictors of NAB-DLM ADL risk subgroup.

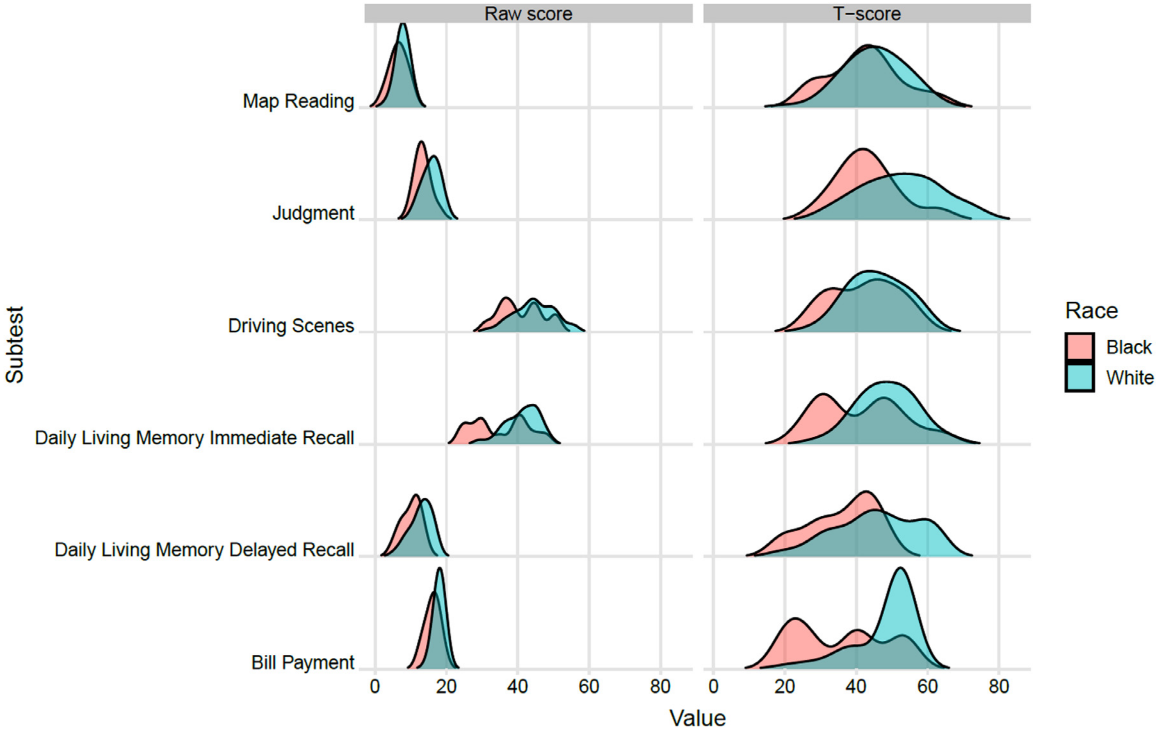

Fig. 3. Ridgeline plot displaying distributions of NAB-DLM raw and *T*-scores by race.

Table 2
NAB-DLM *T*-scores for the total sample, by race and by ADL subgroup status.

	Total Sample (*N* = 79)	White Participants (*n* = 59)	Black Participants (*n* = 13)	Non-White/ Non-Black (*n* = 7)	ADL Non-Risk (*n* = 40)	ADL Risk Group (*n* = 39)
Daily Living Memory Immediate Recall	46·77 (9·36)	48·49 (8·31)	41·31 (12·02)	42·43 (8·26)	52·08 (7·36)	41·33 (8·03)
Daily Living Memory Delayed Recall	43·91 (11·96)	45·81 (11·75)	36·00 (9·43)*	42·57 (13·15)	50·25 (10·60)	37·41 (9·62)*
Bill Payment	45·66 (10·53)	48·02 (8·49)	35·54 (12·99)*	34·01 (11·41)*	50·28 (6·11)	40·92 (11·99)
Judgment	50·28 (11·44)	52·58 (11·25)	43·08 (8·77)	44·29 (10·80)	56·08 (10·73)	44·33 (8·88)
Map Reading	44·54 (9·03)	45·19 (8·78)	42·46 (10·49)	43·00 (8·87)	47·28 (9·44)	41·74 (7·75)
Driving Scenes	44·95 (8·42)	45·81 (8·03)	42·15 (9·55)	42·86 (9·34)	48·90 (7·69)	40·90 (7·17)
Overall NAB-DLM score	46·02 (5·93)	47·65 (5·10)	40·09 (6·40)	43·29 (3·60)	50·81 (3·19)	41·11 (3·57)

* Average T-score below clinical threshold for impairment ($T<40$).

guidance related to social determinants of ADL risk. Finally, we developed a parsimonious classification algorithm that leveraged explanatory power from two-way interactions, with significant contributions from brain white matter tracts that support executive (e.g., uncinate fasciculus, genu of the corpus callosum) and psychomotor (e.g., external capsule) functions.

Consistent with prior studies demonstrating associations between executive performance and ADL disruption [39–41], the GBM model selected two measures of executive function/working memory (Stroop and Digits Backward) as top classifiers of ADL subgroup status. By contrast, delayed memory performance did not emerge as a relevant classifier of ADL subgroup designation in this sample of older PWH, a finding that reinforces the need to attend to etiological mechanisms that disrupt brain regions supporting executive functions rather than degenerative mechanisms such as Alzheimer's disease.

It should be noted that the brain regions implicated in the current study are also relevant to mood/emotional dysfunction [42,43]. After the introduction of ART in 2005, the relevance of mood dysfunction on HIV-related clinical outcomes focused on the potential negative effects on suboptimal adherence to ART. However, work conducted in the pre-ART era describes a direct link between mood and immune dysregulation and disease progression in PWH [44,45]. In alignment with these studies, recent work by our team reveals a plausible pathway that links mood dysregulation and neuroimmune activation to chronic CD4/CD8 T-cell inversion and the development of *T*-cell exhaustion, both of which increase the risk of significant health complications for PWH despite otherwise successful ART [45]. Longitudinal studies are needed to more completely delineate the intersection between white matter microstructural alterations, mood dysregulation and cognitive impairment as a causal pathway toward ADL disruption among PWH who are advancing into older ages.

Importantly, our results highlight the importance of reading ability as a precursor to performance on task-based ADL protocols such as the NAB-DLM. In the present study, the WRAT-4 Reading subtest and the Boston Naming Test-Short Form emerged as important classifiers in the univariate model and features in the two-way model. The observation that language-centric cognitive measures predicted performance on the NAB-DLM is most likely explained by the emphasis of language-based components of everyday living in the subtests of the NAB-DLM [15]. To this end, the associations between reading proficiency and racial identity observed in the present study may represent autocorrelations.

Differences on the NAB-DLM between White and non-White study participants were most prominent on the Bill Payment, Map Reading, and the Driving Scenes subtests. It is probable that differences in exposure/experience conducting these activities contribute to the performance differences, rather than acquired cognitive impairment. Ethnic minorities in the US are more likely to utilize alternative financial service providers (check cashing, money orders, etc.) [46,47] compared to commercial banking institutions and personal checking accounts. As such, task-based methods that require individuals to apply skills related to the management of a personal checking account (e.g., balancing a fictitious checkbook) are likely to amplify performance differences. A similar concern exists for the Map Reading and Driving Scenes subtests given that minorities and individuals with lower income are more likely to utilize public transportation versus drive personal vehicles [48]. While the relatively limited number of Non-White individuals in this sample is recognized as a methodological limitation, our results suggest that clinicians should consider the degree to which sub-average performance on task-based ADL measures reflect sociocultural factors.

Another source of potential bias relates to cultural representativeness in published normative data. The NAB-DLM normative sample included a disproportionately low number of older Black individuals when compared to US census data at the time. As consequence, small differences in performance on the NAB-DLM between Black and White study participants translated into pronounced differences when raw scores were converted to standardized *T*-scores. On three NAB-DLM subtests, raw score differences between Black and White participants of approximately two points became ≥10-point gaps after conversion to *T*-scores. As such, Non-White participants with raw scores that differed by only one or two questions from White participants were likely to be defined as impaired. Misclassification of cognitive impairment is not a new problem [49], but few studies have focused on social determinants that have potential to artificially inflate impairment in ADLs among individuals from diverse ethnic and racial backgrounds.

Limitations of the study merit discussion. The dimensional characterization of the sample allowed for the most complete phenotyping of ADL status of virally suppressed older adults conducted to date, yet the total sample size was restricted. As such, it is possible that extreme scores on a given variable (particularly the WRAT-4) could skew the results of the machine learning analysis. We do not believe this is a likely outcome as the reading and cognitive test scores in this sample are consistent with the results reported in other studies [50,51]. Given the older age of the study participants, all of whom were virally suppressed, the results may not generalize to younger PWH and/or those with uncontrolled viremia. Additionally, while we utilized a number of strategies to minimize overfitting, including five-fold cross validation with multiple repeat trials, focused inclusion of input features in the algorithm, etc., additional studies are needed to examine the clinical relevance of the classification algorithm in larger, more racially and ethnically diverse samples who are followed longitudinally. Additionally, studies that examine predictive features of the individual subtests of the NAB would be of interest.

Results from the current study provide an important foundation to establish clinically relevant predictive models of ADL risk among older PWH with neurocognitive impairment. Our study findings highlight the importance of sociocultural factors, in addition to brain structure and function, as predictors of ADL status when measured using existing task-based protocols. Development of culturally relevant methods to ascertain ADL status, including methods that are specific to HIV, represents a clinical and research priority to prevent errors that could propagate health disparities among diverse populations.

Declaration of Competing Interest

Dr. Paul reports grants from National Institute of Mental Health during the conduct of the study. Dr. Valcour reports other relevant activities from Southeast Tennessee AIDS Education & Training Center and from International AIDS Society - USA, outside the submitted work. Dr. Cysique reports grants, personal fees and other relevant activities from Missouri Institute of Mental Health during the conduct of the study. All other authors report no conflicts of interest.

Acknowledgments

The authors gratefully acknowledge the contribution of the study participants.

Author contributions

RP, BM, and VV contributed to the conception and design of the study. SJ was responsible for data collection. TT, KC, AB, and JB completed data analyses. RP, TT, AB, JB, JM, LC, SL, and VV contributed to the drafting of the manuscript. RP verified the underlying data. All authors reviewed and approved of the final version of the manuscript.

Funding

This work was supported by the National Institutes of Mental Health.

Data sharing statement

Data collected for the study, including deidentified participant data, data dictionary, study protocol, and informed consent form, will be made available with publication. Individuals will be required to submit a data access agreement, including proposed analyses, in order to access the data. Upon approval of the data access agreement, the data will be shared via a HIPAA-compliant, cloud storage such as Box.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.eclinm.2021.100845.

References

[1] APA. Diagnostic and statistical manual of mental disorders (DSM-5®). Washington, D.C.: American Psychiatric Association; 2013.
[2] Antinori A, Arendt G, Becker JT, et al. Updated research nosology for HIV-associated neurocognitive disorders. Neurology 2007;69(18):1789–99.
[3] Samji H, Cescon A, Hogg RS, et al. Closing the gap: increases in life expectancy among treated HIV-positive individuals in the United States and Canada. PLoS ONE 2013;8(12):e81355.
[4] Wandeler G, Johnson LF, Egger M. Trends in life expectancy of HIV-positive adults on antiretroviral therapy across the globe: comparisons with general population. Curr Opin HIV AIDS 2016;11(5):492–500.
[5] Cohort Antiretroviral Therapy. Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. Lancet HIV 2017;4(8):e349–56.
[6] Chang K, Premeaux TA, Cobigo Y, et al. Plasma inflammatory biomarkers link to diffusion tensor imaging metrics in virally suppressed HIV-infected individuals. AIDS 2020;34(2):203–13.
[7] Chang L, Wong V, Nakama H, et al. Greater than age-related changes in brain diffusion of HIV patients after 1 year. J Neuroimmune Pharmacol 2008;3(4):265–74.
[8] Boyle PA, Malloy PF, Salloway S, Cahn-Weiner DA, Cohen R, Cummings JL. Executive dysfunction and apathy predict functional impairment in Alzheimer disease. Am J Geriatr Psychiatry 2003;11(2):214–21.
[9] Hosaka KRJ, Greene M, Premeaux TA, et al. Geriatric syndromes in older adults living with HIV and cognitive impairment. J Am Geriatr Soc 2019;67(9):1913–6.
[10] Morgello S, Gensler G, Sherman S, et al. Frailty in medically complex individuals with chronic HIV. AIDS 2019;33(10):1603–11.
[11] Chelune GJ, Heaton RK, Lehman RAW. Neuropsychological and personality correlates of patients' complaints of disability. In: Goldstein G, Tarter RE, editors. Advances in clinical neuropsychology. Boston, MAUS: Springer; 1986. p. 95–126.
[12] Blackstone K, Moore DJ, Heaton RK, et al. Diagnosing symptomatic HIV-associated neurocognitive disorders: self-report versus performance-based assessment of everyday functioning. J Int Neuropsychol Soc 2012;18(1):79–88.
[13] Yasuda N ZS, Hawkes WG, Gruber-Baldini AL, Hebel JR, Magaziner J. Concordance of proxy-perceived change and measured change in multiple domains of function in older persons. J Am Geriatr Soc 2004;52(7):1157–62.
[14] Cooley SA, Paul RH, Ances BM. Medication management abilities are reduced in older persons living with HIV compared with healthy older HIV- controls. J Neurovirol 2020;26(2):264–9.
[15] Stern RA, White T. NAB, neuropsychological assessment battery: administration, scoring, and interpretation manual. Lutz, FL: Psychological Assessment Resources; 2003.
[16] Norman GR, Sloan JA, Wyrwich KW. The truly remarkable universality of half a standard deviation: confirmation through another look. Expert Rev Pharmacoecon Outcomes Res 2004;4(5):581–5.
[17] Mehrabi N., Morstatter F., Saxena N., Lerman K., Galstyan A. A survey on bias and fairness in machine learning. arXiv preprint arXiv:190809635. 2019.
[18] Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 2002;17(2):825–41.
[19] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 1979;9(1):62–6.
[20] Garyfallidis E, Brett M, Amirbekian B, et al. Dipy, a library for the analysis of diffusion MRI data. Front Neuroinform 2014;8(8).
[21] Lezak MD, Howieson DB, Loring DW, Fischer JS. Neuropsychological assessment. Fourth ed USA: Oxford University Press; 2004.
[22] Katz S. Assessing self-maintenance: activities of daily living, mobility and instrumental activities of daily living. J Am Geriatr Soc 1983;31(12):721–7.
[23] Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist 1969;9(3):179–86.
[24] Wilkinson GS, Robertson GJ. Wide range achievement test 4 (WRAT4). Lutz, FL: Psychological Assessment Resources; 2006.
[25] Rivera Mindt M, Byrd D, Saez P, Manly J. Increasing culturally competent neuropsychological services for ethnic minority populations: a call to action. Clin Neuropsychol 2010;24(3):429–53.
[26] Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y. Reading level attenuates differences in neuropsychological test performance between African American and White elders. J Int Neuropsychol Soc 2002;8(3):341–8.
[27] Yesavage JA. Geriatric depression scale. Psychopharmacol Bull 1988;24(4):709–11.
[28] Miller PJ, Lubke GH, McArtor DB, Bergeman C. Finding structure in data using multivariate tree boosting. Psychol Methods 2016;21(4):583.
[29] Polley EC, Rose S, van der Laan MJ. Super Learning. In M.J. van der Laan and S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data, Chapter 3. New York, Springer 2011.
[30] Papini S, Pisner D, Shumake J, et al. Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. J Anxiety Disord 2018;60:35–42.
[31] Yang H, Liu J, Sui J, Pearlson G, Calhoun VD. A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. Front Hum Neurosci 2010;4:192.
[32] Blankers M, van der Post LFM, Dekker JJM. Predicting hospitalization following psychiatric crisis care using machine learning. BMC Med Inform Decis Mak 2020;20(1):332.
[33] Chen Q, Zhang-James Y, Barnett EJ, et al. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. PLoS Med 2020;17(11):e1003416.
[34] Riedel BC, Daianu M, Ver Steeg G, et al. Uncovering biologically coherent peripheral signatures of health and risk for Alzheimer's disease in the aging brain. Front Aging Neurosci 2018;10:390.
[35] Pedregosa FVG, Michel V, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[36] Jiangchun L. SauceCat/PDPbox. https://github.com/SauceCat/PDPbox. Published 2019. Accessed September 10, 2020.
[37] Paul RH, Cho KS, Luckett P, et al. Machine learning analysis reveals novel neuroimaging and clinical signatures of frailty in HIV. J Acquir Immune Defic Syndr 2020;84(4):414–21.
[38] Paul RH, Cho KS, Belden AC, et al. Machine-learning classification of neurocognitive performance in children with perinatal HIV initiating de novo antiretroviral therapy. AIDS 2020;34(5):737–48.
[39] Watson CW, Sundermann EE, Hussain MA, et al. Effects of trauma, economic hardship, and stress on neurocognition and everyday function in HIV. Health Psychol 2019;38(1):33–42.
[40] Heaton RK, Marcotte TD, Mindt MR, et al. The impact of HIV-associated neuropsychological impairment on everyday functioning. J Int Neuropsychol Soc 2004;10 (3):317–31.
[41] Woods SP, Iudicello JE, Morgan EE, et al. Household everyday functioning in the internet age: online shopping and banking skills are affected in HIV-associated neurocognitive disorders. J Int Neuropsychol Soc 2017;23 (7):605–15.
[42] Masuda Y, Okada G, Takamura M, et al. White matter abnormalities and cognitive function in euthymic patients with bipolar disorder and major depressive disorder. Brain Behav 2020;10(12):e01868.

[43] Cyprien F, de Champfleur NM, Deverdun J, et al. Corpus callosum integrity is affected by mood disorders and also by the suicide attempt history: a diffusion tensor imaging study. J Affect Disord 2016;206:115–24.

[44] Ickovics JR, Hamburger ME, Vlahov D, et al. Mortality, CD4 cell count decline, and depressive symptoms among HIV-seropositive women: longitudinal analysis from the HIV epidemiology research study. JAMA 2001;285(11):1466–74.

[45] Farinpour R, Miller EN, Satz P, et al. Psychosocial risk factors of HIV morbidity and mortality: findings from the multicenter AIDS cohort study (MACS). J Clin Exp Neuropsychol 2003;25(5):654–70.

[46] Kim KT, Lee J, Lee JM. Exploring racial/ethnic disparities in the use of alternative financial services: the moderating role of financial knowledge. Race Soc Probl 2019;11(2):149–60.

[47] Fowler CS, Cover JK, Kleit RG. The geography of fringe banking. J Reg Sci 2014;54 (4):688–710.

[48] Sanchez TW, Stolz R, Ma JS. Moving to equity: addressing inequitable effects of transportation policies on minorities. Cambridge, MA: The Civil Rights Project at Harvard University; 2003.

[49] Brooks BL, Iverson GL, Holdnack JA, Feldman HH. Potential for misclassification of mild cognitive impairment: a study of memory scores on the Wechsler Memory Scale-III in healthy older adults. J Int Neuropsychol Soc 2008;14 (3):463–78.

[50] Casaletto KB, Cattie J, Franklin DR, et al. The Wide Range Achievement Test-4 Reading subtest "holds" in HIV-infected individuals. J Clin Exp Neuropsychol 2014;36(9):992–1001.

[51] Yu B, Pasipanodya E, Montoya JL, et al. Metabolic syndrome and neurocognitive deficits in HIV infection. J Acquir Immune Defic Syndr 2019;81(1):95–101.