ARIC Manuscript Proposal #3880

PC Reviewed: 6/8/21	**Status: _____**	**Priority:2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Can a machine learning derived measure of anatomical dementia risk be considered a measure of brain aging?

b. Abbreviated Title (Length 26 characters): Machine learning & Brain Age

2. Writing Group:

Andrea Anderson
Jamie Justice
Lynne Wagenknecht
Ramon Casanova
Stephen Kritchevsky

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. RC **[please confirm with your initials electronically or in writing]**

First author: Ramon Casanova, PhD
Address: Department of Biostatistics and Data Science
Wake Forest School of Medicine
Winston Salem, NC, 27157
Phone:336 716 8309
Email: casanova@wakehealth.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name: **Lynne Wagenknecht.**
Address: Wake Forest School of Medicine
Division of Public Health Sciences
Wake Forest School of Medicine
Winston Salem, NC, 27157
Phone: Fax:
E-mail: lynne.wagenknecht@wakehealth.edu

3. Timeline:

Expectations are for analyses to be completed within 1 year.

4. Rationale:

The incidence of Alzheimer's disease and related dementias (ADRD) rises exponentially with age. The search for the causes of ADRDs have relatively recently started to give more attention to the age-related changes that appear to be permissive of AD's emergence. In previous work, Dr. Casanova used machine learning to analyze structural MRIs obtained in ARIC to derive an Alzheimer's disease pattern similarity (AD-PS) score[1, 2]. The algorithm behind the scores captures spatial patterns of gray matter brain tissue discriminative of cognitively normal individuals from dementia patients. The scores have been shown to be associated with cognitive status, changes in cognitive function, incident cognitive impairment[3], trajectories of global cognitive function[4] and particulate matter air pollution[5, 6]. More recently, using data from the ARIC cohort, we have shown the scores to be more predictive of incident cognitive impairment than a volumetric composite of regions susceptible to AD available in the ARIC study[7]. Since the AD-PS is not directly assessing specific brain features (e.g. hippocampal, etc.) but rather global and data-driven, we hypothesize it may also be a measure of brain aging able to identify brains that are relatively "old" compared to others of a similar chronological age.

To explore the potential of the AD-PS as a measure of brain aging, we propose related analyses to evaluate thescores both cross-sectionally and longitudinally. We will relate AD-PS scores to total mortality in the ARIC cohort to determine its relationship with age-adjusted total mortality. We will also examine mortality omitting deaths from causes related to CNS pathology (e.g. stroke, Alzheimer's disease). We also plan to associate the AD-PS score with other ARIC measures of biological aging including those derived from proteomic (somologic) data, and a deficit accumulation index. We will examine whether the AD-PS score is associated with a proteomic profile consistent with advanced age. We also hypothesize that the AD-PS will be associated with more frailty and a higher health deficit burden adjusting for age. We hypothesize that higher values of the AD-PS would be associated with worse performance on these measures even in participants without a history of neurologic diseases affecting the central nervous system.

Previously several groups have investigated the potential of sMRI based biomarkers to predict mortality. Kuller et al. reported white matter grade and ventricular volume to be predictors of death[8]. Henneman et al. used MRI scans from 1138 patients to generate visual rating scales for medial temporal lobe atrophy, global cortical atrophy, and white matter hyperintensities (WMH). Number of microbleeds and presence of infarcts were recorded. They found these biomarkers to be predictors of mortality being microbleeds the ones with the strongest associations[9]. More recently a meta-analysis based on 94 studies (N between 14000-16000) found white matter hyperintensities burden, brain infarcts (BI) and microbleeds (MB) to be associated with death[10]. Artificial intelligence methods has been used to estimate brain age which have been shown to predict mortality[11]. We have developed the AD-PS score using machine learning methods[1, 12]. However, none of these approaches has been examined in light of overall organismal aging.

There is a strong interest in the development of techniques to characterize biologic age as a better indication of age-related risk compared to chronologic age. A variety of approaches have been used to identify age-related proteins both in the circulation and in tissues [13, 14]. The studies using the somologic platform have identified over 250 proteins either positively or negatively associated with chronologic age [15-18]. Studies have begun to relate levels of these proteins to various disease outcomes[19, 20]. Machine learning and Artificial intelligence techniques are beginning to be used to produce estimators of chronological age based on proteomic data[21, 22].

It is clear that people develop diseases and pathologic conditions at different rates with age. This is consistent with the idea that people age at different rates. Rockwood and Mitnitsky operationalized this idea through a frailty or deficit accumulation index. The index reflects the proportion of health and functional deficits a person might have from a list of potential deficits. The deficit index rises exponentially with age, and strongly predicts poor outcomes in persons after age-adjustment[23, 24].

5. Main Hypothesis/Study Questions:

Our main study question is to evaluate the potential of the AD-PS score as a brain-focused measure of aging. We propose to relate AD-PS scores to total mortality, and to ARIC measures related to biological age, including a panel of proteins derived from proteomic SOMAscan data and the deficit accumulation index.

Our main hypotheses are:

Hypothesis 1: The AD-PS scores are associated with chronological age.

Hypothesis 1a: Using high-dimensional machine learning methods and the full proteomic panel we will be able to infer an accurate estimator of chronological age (proteomic clock).

Hypothesis 2: The AD-PS scores are associated with measures of biological age.

Hypothesis 2a: The AD-PS scores estimated at visit 5 will be associated with proteins levels from two different panels at visit 5. (Cross-sectional). The first panel will be composed of 32 proteins reported in the literature to be consistently associated with age and which are included in the somologic panels[14]. The second panel will be derived using the top 10 proteins as ranked by a high-dimensional machine learning regression model.

Hypothesis 2b: The AD-PS scores estimated at visit 5 will be associated with the deficit accumulation index based on data collected at visit 5 (Cross-sectional).

Hypothesis 2c: The difference between estimated and chronological age will be strongly associated to the AD-PS scores.

Hypothesis 3: The AD-PS scores estimated at visit 5 will be predictive of total mortality and from causes that are unrelated to neurologic degeneration (i.e., excluding stroke, Parkinson's, ALS, ADRD, etc.). (Longitudinal)

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Design: Cross-sectional and longitudinal study design with follow-up through visit 7.

Outcome sets:

Visits 5-7
Mortality
Cognitive status

Datasets:

Visits 5:

AD-PS scores

Proteomic data

The deficit accumulation score will include 40 health/function items assessed at V5 and which have been included in other published frailty indices including ones used in the SPRINT, Look AHEAD, and the Canadian Longitudinal Study on Aging. Elements include measures of overall health, physical, cognitive and emotional function, diagnosed diseases, clinical laboratories, physiologic, and prevalent diseases.

Demographics

Analyses:

The AD-PS scores were previously estimated for 1857 individuals at visit 5.

Hypothesis 1: **The AD-PS scores are associated with chronological age.**
We will perform analyses to evaluate correlations between age and the AD-PS scores.

Hypothesis 1a: Derivation of proteomic clock.

The precision of estimation of chronological age using high-dimensional machine learning and therefore the feasibility of a **proteomic biological clock** will be evaluated. We will use Random Forests, elastic net and neural networks including the full proteomic panel to estimate chronological age at visit 5. We are going to include proteins with less than 5% missing data. The dataset with proteomic data available (N ~ 6000) will be divided in training and testing. The testing dataset will include participants with the AD-PS scores computed at visit 5 (N = 1857). The rest will compose the training dataset. Estimation performance of the regression models in the testing dataset will be evaluated using correlations between chronological and estimated age, mean squared error and mean absolute deviation.

Hypothesis 2a: **The AD-PS scores will be associated with proteins associated with biological aging.**

Approach 1: For the protein panel referenced above we are going to divide the participants in quartiles according to their AD-PS scores values. Then for each protein we are going to fit logistic regression models using the lowest and upper quartiles groups of participants. Relationships will be adjusted for age, sex, and race. Also Random Forests classification models[25] including the 32 proteins will be fitted to investigate multivariate prediction of the AD-PS scores.

Approach 2: To generate the second panel a we will fit high-dimensional regression models using Random Forests[25] and elastic net (or lasso) regularization[26] to predict chronological age using as predictors the full proteomic somologic panel collected at visit 5. In each case we will build ranks of predictors using variables importance measures available in RF or the absolute value of the coefficients in the case of the regularization based classifiers. We will select the 10 proteins that are common in the top of the ranks produced by the both methods. Once the panel is available, similar association analyses using logistic regression and RF as described above will be performed.

Hypothesis 2b: **The AD-PS scores will be associated with the default accumulation index.**
Linear regression analyses will be performed to evaluate associations adjusting for age, sex and race

Hypothesis 2c: The difference between chronological (CA) and estimated age (EA) will be strongly correlated with the AD-PS scores.
We will perform analyses to evaluate correlations between (CA-EA) and the AD-PS scores.

Hypothesis 3: **The AD-PS scores estimated at visit 5 will be predictive of total mortality**

Analyses will evaluate association of the AD-PS scores with total mortality after visit 5. Participants will be divided in tertiles according to their AD-PS scores values. Cox regression will be used in these analyses adjusted by age, education, sex, race, smoking and hypertension.

Sensitivity and complementary analyses will look omit persons with prevalent neurologic disease at baseline, and causes of death unrelated CNS pathologies.

Limitations/Challenges

7.a. Will the data be used for non-CVD analysis in this manuscript? __X__ Yes ____ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used? ____ Yes ____ No
(This file ICTDER has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? ____ Yes __X__ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"? ____ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

___X___ Yes _______ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

- #3739 - Proteomic age acceleration and cancer incidence: The Atherosclerosis Risk in Communities Study

We contacted the writing group of this paper. It seems there is no conflict. We are not looking into cancer survivors.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? _x___ Yes ___ No

11.b. If yes, is the proposal
_x_ A. primarily the result of an ancillary study (list number* _2008-06_______)

___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* __________ __________ __________)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PubMed Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

References

1. Casanova, R., et al., *Alzheimer's disease risk assessment using large-scale machine learning methods.* PLoS One, 2013. **8**(11): p. e77949.
2. Casanova, R., et al., *High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization.* Front Neuroinform, 2011. **5**: p. 22.
3. Espeland, M.A., et al., *Trajectories of Relative Performance with 2 Measures of Global Cognitive Function.* J Am Geriatr Soc, 2018. **66**(8): p. 1575-1580.
4. Espeland, M.A., et al., *Long Term Effect of Intensive Lifestyle Intervention on Cerebral Blood Flow.* J Am Geriatr Soc, 2018. **66**(1): p. 120-126.
5. Younan, D., et al., *Particulate matter and episodic memory decline mediated by early neuroanatomic biomarkers of Alzheimer's disease.* Brain, 2020. **143**(1): p. 289-302.
6. Younan, D., et al., *PM2.5 associated with gray matter atrophy reflecting increased Alzheimers risk in older women*. Neurology, 2020.
7. Casanova, R., et al., *Comparing data-driven and hypothesis-driven MRI based Predictors of Cognitive Impairment in individuals from the Atherosclerosis Risk in Communities (ARIC) Study.* Alzheimer's & Dementia, 2021. **Under Review**.
8. Kuller, L.H., et al., *White matter grade and ventricular volume on brain MRI as markers of longevity in the cardiovascular health study*. Neurobiol Aging, 2007. **28**(9): p. 1307-15.
9. Henneman, W.J., et al., *MRI biomarkers of vascular damage and atrophy predicting mortality in a memory clinic population.* Stroke, 2009. **40**(2): p. 492-8.
10. Debette, S., et al., *Clinical Significance of Magnetic Resonance Imaging Markers of Vascular Brain Injury: A Systematic Review and Meta-analysis.* JAMA Neurol, 2019. **76**(1): p. 81-94.
11. Cole, J.H., et al., *Brain age predicts mortality*. Mol Psychiatry, 2018. **23**(5): p. 1385-1392.
12. Casanova, R., et al., *High dimensional classification of structural MRI Alzheimer’s disease data based on large scale regularization.* Frontiers of Neuroscience in Neuroinformatics, 2011. **5:22. Epub 2011 Oct 14**.
13. Moaddel, R., et al., *Proteomics in aging research: A roadmap to clinical, translational research*. Aging Cell, 2021. **20**(4): p. e13325.
14. Johnson, A.A., et al., *Systematic review and analysis of human proteomics aging studies unveils a novel proteomic aging clock and identifies key processes that change with age*. Ageing Res Rev, 2020. **60**: p. 101070.
15. Sathyan, S., et al., *Plasma proteomic profile of frailty*. Aging Cell, 2020. **19**(9): p. e13193.

16. Sathyan, S., et al., *Plasma proteomic profile of age, health span, and all-cause mortality in older adults*. Aging Cell, 2020. **19**(11): p. e13250.
17. Tanaka, T., et al., *Plasma proteomic signature of age in healthy humans*. Aging Cell, 2018. **17**(5): p. e12799.
18. Menni, C., et al., *Circulating Proteomic Signatures of Chronological Age*. J Gerontol A Biol Sci Med Sci, 2015. **70**(7): p. 809-16.
19. Tavenier, J., et al., *Longitudinal course of GDF15 levels before acute hospitalization and death in the general population*. Geroscience, 2021.
20. Tavenier, J., et al., *Association of GDF15 with inflammation and physical function during aging and recovery after acute hospitalization: a longitudinal study of older patients and age-matched controls*. J Gerontol A Biol Sci Med Sci, 2021.
21. Galkin, F., et al., *Biohorology and biomarkers of aging: Current state-of-the-art, challenges and opportunities*. Ageing Res Rev, 2020. **60**: p. 101050.
22. Lehallier, B., et al., *Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging*. Aging Cell, 2020. **19**(11): p. e13256.
23. Pajewski, N.M., et al., *Frailty Screening Using the Electronic Health Record Within a Medicare Accountable Care Organization*. J Gerontol A Biol Sci Med Sci, 2019. **74**(11): p. 1771-1777.
24. Mitnitski, A.B., A.J. Mogilner, and K. Rockwood, *Accumulation of deficits as a proxy measure of aging*. ScientificWorldJournal, 2001. **1**: p. 323-36.
25. Breiman, L., *Random Forests*. Machine Learning, 2001. **45**: p. 5-32.
26. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software, 2010. **33**(1): p. 1-22.