ARIC Manuscript Proposal # 3327

PC Reviewed: 1/8/19  Status: _____  Priority: 2  
SC Reviewed: _________  Status: _____  Priority: _____

1.a. Full Title: A proteomic analysis of incident dementia: The ARIC Study

b. Abbreviated Title (Length 26 characters):

2. Writing Group:
   Writing group members: Keenan Walker; Aozhou Wu; Adrienne Tin; Thomas Mosley; Myriam Fornage; David Knopman; Christie Ballantyne; Eric Boerwinkle; Rebecca Gottesman; Josef Coresh (last author): others welcome

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

First author: Keenan Walker, PhD
Address: Johns Hopkins Hospital, Department of Neurology
         Phipps 446
         600 North Wolfe Street
         Baltimore, MD 21287

         Phone: 626-840-6216  Fax: 410-955-0672
         E-mail: kwalke26@jhmi.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: Josef Coresh
Address: Johns Hopkins Bloomberg School of Public Health and Welch Center for Prevention, Epidemiology, and Clinical Research
         2024 E. Monument St.,
         Suite 2-600
         Baltimore, MD 21287

         Phone: 410-995-0495  Fax: 410-955-0476
         E-mail: coresh@jhu.edu

3. Timeline: 6-9 months; manuscript submission in summer of fall of 2019
4. **Rationale:**
There is an urgent need for reliable non-invasive biomarkers that can be used to identify individuals with preclinical Alzheimer’s disease and related dementias. While a number of previous studies have examined the degree to which blood protein levels are abnormal among persons living with dementia, the vast majority of these studies have used a targeted approach to examine only a small number of proteins. Although a number of biologically relevant proteins have been associated with Alzheimer’s disease, many of these associations have been inconsistent, and as such, no blood-based biomarker has progressed beyond the discovery phase.\(^1\) Given the complexity of Alzheimer’s disease pathogenesis and the considerable size of the human proteome, there is likely a wide range of to-be-discovered protein changes that occur during the multi-decade prodromal phase of Alzheimer’s disease. Although several case-control studies have evaluated abnormalities in blood protein levels among individuals with Alzheimer’s disease, these studies have had little or no follow-up, relatively small sample sizes, and have assayed only a small portion of the human proteome.\(^2,3,12,13,4–11\) Proteomic studies of Alzheimer’s disease to date have been almost exclusively cross-sectional and therefore have been unable to determine which proteins are abnormally expressed in non-demented individuals who eventually progress to dementia.

The recent development of high-throughput technology for the characterization of the human proteome has enabled the simultaneous assessment of approximately 5,200 proteins in an accurate and reliable manner using a small amount of blood.\(^14,15\) Although gene expression studies have been informative in identifying transcriptional changes that co-occur with dementia, the limited relationship between mRNA-protein pairs, with regard to expression levels, underscores the need to examine protein levels directly.\(^16–18\) By conducting a large-scale prospective analysis of altered plasma protein levels in non-demented individuals who progress to dementia, we will take a data-driven approach to identify blood-based protein biomarkers that can be used for early disease detection and to better understand the molecular pathways that may be altered in the prodromal phase of dementia. While the former objective may lead to the development of a minimally-invasive method for dementia risk stratification for clinical trials, the latter objective would shed light on the complex systemic changes that occur outside the central nervous system in the years before dementia onset and thereby highlight novel pathways for therapeutic intervention.

The current study will use recently developed SOMAscan Multiplexed Proteomic technology to examine the relationship between the level of ~5,200 plasma proteins and dementia risk within the Atherosclerosis Risks in Communities (ARIC) Cohort, a large, biracial community-based sample of older adults. Candidate proteins identified in the discovery cohort will then be validated within a separate subset of the ARIC cohort. We will conduct a series of secondary analyses to determine whether there are sex-, race-, age-, and \(APOE\ v4\)-specific associations between plasma protein levels and risk of incident dementia. In the event that a number of candidate proteins are identified, we will (1) examine the predictive value of specific proteins and empirically-derived protein risk scores, (2) apply systems-level analyses to determine whether specific biologically informed protein pathways are overrepresented, and (3) determine the degree to which dysregulation of identified protein networks may be independently associated with dementia risk. We will also use machine learning based data-driven approaches, such as feedforward multilayer neural networks, to identify candidate proteins and develop dementia prediction models. These deep learning networks are ideal for...
modeling complex associations among a large number of variables for the purposes of outcome prediction. Results from these analyses will be used as preliminary data to inform the hypothesis and scientific approach for the omics aim of the ARIC-Neurocognitive Study (NCS) renewal application.

5. Main Hypothesis/Study Questions:

**H1.** Multiple blood-based proteins (measured at Visit 5) will be abnormally elevated or reduced among non-demented participants (cognitively normal and with mild cognitive impairment) who progress to dementia after ARIC Visit 5.

i. We will test several proteins that have been previously associated with Alzheimer’s disease in multiple cohorts. This will test the utility of a number of Alzheimer’s disease biomarkers (e.g., Alpha-1-antitrypsin, Alpha-2-macroglobulin, Apolipoprotein E, Complement C3, Complement factor H, Serum amyloid p-component) as measured in the SomaScan (recognizing predictive utility may vary across platforms which recognize different protein features).

ii. We will identify a number of novel proteins and proteomic features that are associated with incident dementia after correction for multiple comparisons.

iii. Identified proteins will be used to develop a protein-driven risk score, which will be validated within the ARIC Cohort (i.e., the validation sample) and within external cohorts.

**H2.** Among participants who meet criteria for mild cognitive impairment at Visit 5, we will identify a number of novel blood-based proteins associated with progression from mild cognitive impairment to dementia over approximately five years (Visit 5 to Visit 6). We will also consider examining proteins associated with conversion from normal cognition to mild cognitive impairment over five years to facilitate an understanding of which proteins may be differentially associated with early- versus late-stage cognitive decline.

i. The set of proteins associated with progression from mild cognitive impairment to dementia will be partially overlapping with the set of proteins associated with incident dementia in the full sample (H1). A unique set of candidate proteins will also be identified.

ii. The set of proteins associated with progression from normal cognition to mild cognitive impairment will be partially overlapping with the set of proteins associated with incident dementia in the full sample (H1). A unique set of candidate proteins will also be identified.

**H3.** In stratified analyses we will identify multiple blood-based proteins at Visit 5 that are differentially associated with incident dementia according to sex, race, age (dichotomized at the median), and APOE ε4 status. This will shed light on interactions and differences (modest quantitative as well as qualitative) in risk relationships and the biology in these subgroups.

**H4.** Protein pathway enrichment and protein co-expression network analyses will implicate complement and coagulation, cytokine signaling, and synaptic transmission protein networks (among others) in progression to dementia.
H5. We will build a prediction model using machine learning methods. Compared to the prediction model using pre-selected proteins and other predictors, the model using machine learning method will achieve better performance and generalizability.

Initial development work will be conducted in a 2/3 sample of the ARIC cohort with validation in the other 1/3 and using bootstrapping techniques. Subsequent external validation will be done in collaborating cohorts such as AGES or consortia which have approached ARIC.

H6. We will also consider comparing the predictive utility of several the protein-driven risk scores (identified as part of H1) to that of several previously defined genetic risk scores\(^{21}\) for dementia using the ARIC Cohort.

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).

**Inclusion criteria:** We will include all participants who (1) attended Visit 5, (2) were classified as non-demented at Visit 5 (i.e., cognitively normal or mild cognitive impairment [MCI]), and (3) have SOMAscan protein measurement available from blood collected at Visit 5.

**Exclusion Criteria:** We will exclude non-white and non-African-American participants and non-white participants in Washington Co. and Minnesota, participants missing the education level variable, and participants missing information needed to classify Visit 5 cognitive status (i.e., normal/MCI/dementia classification).

**Proteomic measurement (exposure variables):** Using plasma collected at Visit 5 (2011-13), proteins were measured using a Slow Off-rate Modified Aptamer (SOMAmer)-based capture array (SomaLogic, Inc, Boulder, Colorado). Using chemically modified nucleotides, this process transforms protein signals to a nucleotide signal quantifiable using relative florescence on microarrays. Previous work indicates a median intra- and inter-run coefficient of variation of approximately 5% and intra-class correlation coefficients of ~0.9.\(^{11,12,22,23}\)

**Primary outcome variables:**
- **Incident Dementia:** Dementia incidence between Visit 5 and Visit 6 (2015-2017) will be examined among participants who were classified as either cognitively normal or MCI at Visit 5. Dementia will be defined using both the information from the full Visit 6 examination with expert committee diagnosis and information captured in annual follow-up (AFU) interviews using the Six Item Screener (SIS) and the Ascertain Dementia 8-item Informant Questionnaire (AD8). Date of dementia onset will be captured using the SIS and AD8, and dementia diagnosis will be confirmed at Visit 6 for those who attend Visit 6. Participants who attended Visit 5, but not Visit 6, and have SIS and AD8 information available from the AFU will also be included. For participants who did not attend Visit 6, the SIS, AD8, hospital discharge codes, and death certificates will be used to define dementia diagnosis and date of onset.

**Secondary outcome:**
- **Mild cognitive impairment (MCI):** Among participants classified as cognitively normal at Visit 5, MCI will be adjudicated at Visit 6 by an expert panel of physicians and neuropsychologists based on CDR sum of boxes, FAQ, and the full ARIC cognitive battery (ARIC Visit 6 Manual.
Analytic Plan

Plasma proteins associated with incident dementia. Participants will be randomly assigned to either a discovery cohort or a validation cohort with a roughly 66%:33% split, respectively. Using the discovery cohort, we will use Cox proportional hazard models to examine the extent to which each protein (after log transformation) is individually associated with 1) incident dementia occurring between Visit 5 and Visit 6 within the full sample and 2) progression from MCI to dementia among participants with MCI at Visit 5. Analyses will be adjusted for potentially confounding variables, including age at sample acquisition, sex, education, and APOE ε4 status. We may additionally adjust models for cardiovascular risk factors (i.e., BMI, total cholesterol, HDL cholesterol, hypertension, diabetes, and coronary artery disease) and other physiological variables known to jointly affect protein levels and neurocognitive outcomes. We will repeat the above analyses after stratifying participants by sex, race (black/white), age at sample acquisition (median split ~75 years), and APOE ε4 allele number (0/≥1). Bonferroni correction will be applied to resulting p-values (e.g., .05/5,000 protein=1.00x10^-5) to identify candidate proteins. Using only the set of candidate proteins that surpass this p-value threshold, we will repeat this series of analyses using the validation cohort.

Predictive performance of protein-driven risk score. To determine the potential for clinical application of individual proteins for predicting incident dementia among non-demented persons and persons with MCI, we will select multiple predictive proteins based on one or more of the following criteria: 1) top 10 proteins based on Cox regression coefficients, 2) top 5 proteins based on Cox regression coefficients, 3) top 3 proteins based on Cox regression coefficients, and 4) all predictive proteins based on a univariate AUC >0.70 and a correlation coefficient <0.40 between each pair of proteins, indicating high predictive accuracy and relative independence. Using these criteria, we will construct a protein-driven risk score based on a linear combination of selected proteins, with coefficients for each protein in the predictive risk score weighted according to Cox regression coefficients. We will use Cox proportional hazard models to determine whether risk scores act as independent predictors of dementia after adjusting for demographic variables, APOE ε4 status, and cardiovascular risk factors. Next, we will use time-dependent AUC to evaluate the predictive utility of each protein-based risk score alone and in combination with demographic characteristics and APOE ε4 status. We will also examine a stepwise Cox regression model constructed with p-enter=0.001 and p-exit=0.001.

Build prediction model using machine-learning. In addition to selecting predictors using traditional step-wise based approaches, we will also explore prediction models using machine learning methods. These include: 1) protein selection using penalized regression; 2) prediction model using random forest; 3) prediction model using feedforward multilayer neural network. Penalized regressions, like lasso regression and elastic net, have the advantage that they will conduct variable selection and build up the prediction model simultaneously. We will explore a Cox prediction model incorporating lasso and elastic net. Random forest is a tree-based ensemble method which generally performs well with high dimensional data. It can provide variable importance scores for each predictor, which could help us identify important proteins. Neural networks are good at modeling complex associations in a prediction task. Usually, a neural
network with one or two hidden layers and dense connectivity is capable of modeling any arbitrary associations between predictors and outcome in a high dimensional space. We will experiment with several networks with different numbers of layers and nodes. For all the three methods, hyper-parameter tuning will be conducted using cross-validation. We will compare the performance of the models developed using the traditional approach and using machine learning approaches by AUC or equivalent metrics.

**Protein pathway enrichment.** In order to examine the potential role of identified proteins in the dementia disease process, we will use DAVID bioinformatics resources (the database for annotation, visualization and integrated discovery) to extract biological features/meaning associated with protein lists by identifying over-represented functional annotations assigned to each of the candidate proteins based on biological function. Specifically, we will perform a KEGG enrichment analysis based on KEGG PATHWAY, which is a collection of manually drawn pathway maps, which integrates molecular-level information from large scale datasets to represent the current knowledge of molecular interaction and reaction networks. Additionally, we may use the weighted protein co-expression network analysis to construct a co-expression network to define distinct modules of protein co-expression enriched for gene ontologies associated with discrete biological processes. The network strength of each co-expression module will be quantified and examined in relation to cognitive status (normal/MCI/dementia) at Visit 6.

**Relative importance of pathways predicting AD.** To determine the relative importance (or predictive value) of identified protein pathways as potential determinants of dementia risk, a pathway score will be defined by deriving the linear combination of the coefficients for proteins in each pathway as defined using the techniques outlined in the section above. We will examine time-dependent area under the receiver operating characteristic (ROC) curve for each protein pathway score associated with incident dementia.

7.a. Will the data be used for non-CVD analysis in this manuscript? 

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? 

8.a. Will the DNA data be used in this manuscript? 

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”? 

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/mantrack/maintain/search/dtSearch.html

___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)?

MP# 3051. The association of middle and late-life blood pressure with conversion to MCI and dementia: The ARIC Study

MP# 3058. The association of late-life glycemia status with 3-year late-life cognitive decline and incident MCI/dementia: The ARIC Study

MP# 3903. Multi-omic data integration using systems approaches for mechanistic understanding of disease in the Atherosclerosis Risk in Communities (ARIC) Study

MP#3113. Identification of novel genetic variants associated with Alzheimer’s disease in the Alzheimer’s Disease Sequencing Project (ADSP)

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* 2017.27)  “Proteomic longitudinal ARIC study: SOMAscan of multiple visits”

         ___ 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 https://www2.cscc.unc.edu/aric/approved-ancillary-studies

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.
Understood

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


