ARIC Manuscript Proposal #4073

PC Reviewed: 7/12/22	**Status: _____**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Proteomics and Ischemic Stroke Subtype: The ARIC Study

b. Abbreviated Title (Length 26 characters): Stroke Subtype and Proteomics

2. Writing Group:
Writing group members: Michelle C. Johansen (first author), Wendy Wang, Lin Yee Chen, Rizwan Kalani, James Floyd, Keenan A. Walker, Silvia Koton, Rebecca F. Gottesman, Myriam Fornage, Joe Coresh (senior author), with others welcome.

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. ___MCJ__

First author: **Michelle C. Johansen, MD, PhD**
Address: Phipps 446; 600 North Wolfe Street; Baltimore, MD 21287
Phone: 410-614-2381 Fax: 410-955-0672
E-mail: mjohans3@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: **Joe Coresh, MD, PhD**
Address: 2024 E. Monument Street, Suite 2-600, Baltimore, MD 21287
E-mail: coresh@jhu.edu

3. Timeline: Analysis will begin as soon as the proposal is approved. Planned abstract submission winter 2022, with manuscript submission fall 2023.

4. Rationale:

Ischemic stroke is a devastating disease and is the second leading cause of death worldwide. The treatment and anticipated longitudinal outcomes of patients differ based on the type of stroke they have had, with cardioembolic stroke representing one particularly devastating subtype.[1,2] We have recently shown in the ARIC study that patients with embolic stroke had a higher odds of a worse NIH stroke score at the time of admission, as well as a lower cumulative survival compared to patients with thrombotic stroke, even after accounting for traditional vascular risk factors (results to be published).

While vascular risk factors can impact the type of ischemic stroke a patient develops, there may be intrinsic factors that are yet to be identified that make a particular stroke subtype more likely to occur. While the overall incidence of ischemic stroke is declining, the incidence of embolic, or cardioembolic stroke is anticipated to increase, likely due to the aging population in the United States.[3] Understanding the plasma proteome prior to the development of incident stroke is something that is uniquely provided in ARIC, rather than just changes in protein modifications after stroke, which is where the burden of literature currently lies.

Improvements in the ability to explore biologic pathways and improve risk prediction for ischemic stroke provides a unique opportunity to begin to answer some of these questions.
In this study, we propose to use SOMAScan data from the ARIC study, with over 5,000 proteins measure over time to determine the proteomic signatures associated with specific stroke subtype, that is 1) incident embolic ischemic stroke and 2) incident thrombotic ischemic stroke specifically. Given that strokes associated with different mechanisms are composed of different thrombogenic material, one could easily hypothesize that the protein pathways would be different. For example, thrombotic strokes tend to be "white clots" in that they are composed of mainly platelets, with small amounts of fibrin and respond to antiplatelet drugs. Embolic strokes, particularly in the setting of atrial fibrillation, usually are fibrin enriched clots, or "red clots" and occur in the setting of stasis of flow, and respond to anticoagulants.

We propose to use ARIC visits 2 to 5 for our protein discovery analysis, which will be conducted separately for both incident embolic and thrombotic stroke, and then will perform a replication analysis from ARIC visit 5 through the end of the available data. We will also test for an interaction between the protein "hits" and stroke subtype (embolic vs. thrombotic). We will additionally test for interactions by race and sex as it may be that the impact of proteins on ischemic stroke subtype is different by race, or sex. As a final step, should we find significant associations, we will consider using genomics data as a next step to this analysis.

5. Main Hypothesis/Study Questions:

Aim 1: To determine the proteomic signatures of incident ischemic stroke classified as embolic ischemic stroke
Aim 2: To determine the proteomic signatures of incident ischemic stroke classified as thrombotic ischemic stroke
Aim 3: To determine if there are differences in the proteomic signatures, or significant proteomic hits, across stroke subtypes (embolic vs thrombotic)
Aim 4: To determine if there is a difference in the proteomic signature of incident ischemic stroke subtypes across race groups.
Aim 5: To determine if there is a difference in the proteomic signature of incident ischemic stroke subtypes by sex.

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).

Study design: This will be a prospective study examining data for eligible participants beginning at visit 2 and followed until the time of the latest follow-up visit for incident ischemic stroke.

The data will be divided into a discovery (visit 2 up to visit 5), and a validation sample (visit 5 through end of available data) Of note, we are not dividing each visit into discovery and validation since power may be limited once ischemic stroke is subdivided into the different subtypes. Discovery will be performed separately for embolic and thrombotic stroke.

Exclusion criteria: Participants will be excluded if they are missing SOMAscan data at visit 2, or if the sample failed quality standards. They will be excluded if they are missing relevant covariate data, or if they have prevalent stroke at visit 2.

Outcome: The outcome will be incident ischemic stroke after visit 2, specifically either an ischemic stroke adjudicated as an embolic ischemic stroke, or a thrombotic ischemic stroke. For our third and fourth aim, we will test for "interactions" between identified proteins and stroke subtype, and between identified proteins and race, or sex.

It should be noted that since stroke subtype reflects data at the occurrence of stroke and not at baseline, the model will not be a regular interaction model (which works for baseline variables interacting with protein levels). Instead, we will run separate regression models with each stroke subtype as the outcome and test for differences between the beta coefficients for each protein between the regression model for embolic and thrombotic stroke using seemingly unrelated regression (which uses the covariance between different models to compare coefficients in different regression models). In each stroke subtype outcome regression, we will censor individuals with the other stroke subtype at the date of stroke incidence but they will be included at study baseline.

Exposure: SOMAScan proteins (~5,000 proteins measured using SOMAScan's aptamer-based proteomics platform). We will be guided by the QC document following the current recommended data cleaning, focus on log protein levels, and use flag 2 to exclude unreliable data (flag2=0 for N = 4877 aptamers excluding flag 1 cat1-3: Fc mouse/Contaminants, or var<0.01 or soma qc cvba>50% at v3 or v5) – this is the suggested list of aptamers to analyze).

Other variables: We will include race-center, sex, and age information from ARIC baseline (race, center, sex) and visit 2 (age). In addition, estimated GFR, hypertension, diabetes, LDL cholesterol, body mass index, smoking and prevalent atrial fibrillation will be determined.

Data analysis: We will perform a cox regression analysis to determine the association between log-transformed protein levels (we will use log base 2 for ease of interpretation and divide each protein logRFU by its standard deviation to facilitate fair comparisons across proteins) and incident embolic ischemic stroke (visit 2-visit 5), and incident thrombotic ischemic stroke, in separate models with step-wise adjustment for covariates as follows:

0) Model 1: Age, sex, center-race, estimated GFR
1) Model 2: Model 1 plus baseline hypertension (defined by systolic blood pressure and hypertension medication use), diabetes, LDL cholesterol, body mass index and tobacco use.
2) Model 3: Model 2 plus atrial fibrillation. (Given that atrial fibrillation may be on the causal pathway, we will include it in a separate model).

3) *Model 4: We will consider an additional statistical model which considers the above covariates in a time dependent fashion

For Aim 3, we will use seemingly unrelated regression to determine if there is a difference in the association between the protein level of interest and incident stroke subtype, between embolic stroke and thrombotic stroke as previously described. For aims 4 and 5, will also add in a separate model, an interaction term for race, and then for sex, to determine if there is a difference in the consistency of proteomic associations with incident ischemic stroke. It may be that there are differences between the association by race as there are differences by race in the incidence of specific ischemic stroke subtypes, and some risk factors, such as atrial fibrillation, are more associated with one race (White individuals) for example.[4,5] Stroke risk also varies across sex, and therefore suggests biologic plausibility of modification by sex.[6]

We will consider applying the Bonferroni correction in the discovery phase in order to have the most conservative analysis. Again, we will use visit 2 to visit 5 for discovery. A test for interaction with baseline age older or younger than 65 years will be used as a guide as to whether replication using visit 5 data when participants are older is reasonable, which is our a priori belief.

For the replication phase, we will examine if the same proteomic associations with ischemic stroke subtypes, each, determined to be significant in the discovery phase, are also significant in the validation sample (visit 5-visit 6). We will do this at a p-value <0.05 divided by the number of significant hits. We will again examine embolic and thrombotic stroke separately for both the discovery and replication phase and will specifically test interaction of validated hits in their association across ischemic stroke subtype.

Limitations: The cause of ischemic stroke is not always known. As a result, we will exclude the small number of participants with either a possible stroke, or a stroke of unknown etiology during study follow-up, but not at study baseline, to ensure that the case definition for embolic and thrombotic stroke is robust. These represent events and definitions that have been independently adjudicated by expert reviewers. Other work is ongoing in ARIC, and CHS, looking at incident ischemic stroke of all types and proteomic signatures and these authors are co-authors on this manuscript proposal.

A limitation of the current work is the lack of a separate cohort for external validation however there are ongoing plans to perform external validation. After the planned initial papers, a meta-analysis with CHS and other cohorts with similar data can occur (project proposal has been submitted for ARIC to validate CHS and we expect the arrangement to be symmetric).

We also understand that our power may be limited to detect a difference with interaction analysis, but this is an exploratory analysis and we will conduct discovery separately for both thrombotic stroke and embolic stroke in order to increase power, and then test for differences between the two and interactions with race, or sex.

7.a. Will the data be used for non-CVD analysis in this manuscript? ____ Yes __x__ 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? __x__ Yes ____ 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"? __x__ 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)?

MP #3840: Proteomics and Risk of Incident Ischemic Stroke: The Atherosclerosis Risk in Communities Study (ARIC); Wang et al.
MP #3876: Matrix Metalloproteinases (MMPs) and risk of incident hemorrhagic and ischemic stroke: The Atherosclerosis Risk in Communities Study; Ji et al.

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* _AS2017.27_**)**
___ 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.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript ____ Yes __x__ No.

References:

1. Arboix A, Alio J. Acute cardioembolic stroke: An update. *Expert Rev Cardiovasc Ther*. 2011;9(3):367-379.

2. Arboix A, Alio J. Cardioembolic stroke: Clinical features, specific cardiac disorders and prognosis. *Curr Cardiol Rev*. 2010;6(3):150-161.

3. Kamel H, Healey JS. Cardioembolic stroke. *Circ Res*. 2017;120(3):514-526.

4. Di Tullio MR, Sacco RL, Sciacca RR, Homma S. Left atrial size and the risk of ischemic stroke in an ethnically mixed population. *Stroke*. 1999;30(10):2019-2024.

5. Wang W, Norby FL, Zhang MJ, et al. Differences in left atrial size and function and supraventricular ectopy between black and white participants in the ARIC study. *Journal of the American Heart Association*. 2021;10(21):e021723. https://www.ncbi.nlm.nih.gov/pubmed/34713724. doi: 10.1161/JAHA.121.021723.

6. Reeves MJ, Bushnell CD, Howard G, et al. Sex differences in stroke: Epidemiology, clinical presentation, medical care, and outcomes. *Lancet Neurol*. 2008;7(10):915-926.