ARIC Manuscript Proposal #3895

PC Reviewed: 7/13/21	**Status: _____**	**Priority: ____**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Proteomic analysis for the risk of infectious disease: The ARIC Study

b. Abbreviated Title (Length 26 characters): **Proteomics and infectious disease**

2. Writing Group:

Writing group members: Junichi Ishigami, Jingsha Chen, Adrienne Tin, David Dowdy, Christie Ballantyne, Morgan Grams, Josef Coresh, Kunihiro Matsushita, Others welcome

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

First author: **Junichi Ishigami**

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: jishiga1@jhu.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: Kunihiro Matsushita

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: kmatsus5@jhmi.edu

3. Timeline: Once the data it obtained, data analysis and manuscript preparation will be done in the next 6-12 months.

4. Rationale:

Acute infections, such as pneumonia and bloodstream infections, are among the leading causes of hospitalization and deaths in the US, posing significant public health and economic burden.[1] The transmission of infectious diseases involves pathogen-specific mechanisms but also depends on individual's immune systems. For example, epidemiological studies have shown that older adults and individuals with chronic conditions (e.g., diabetes, chronic kidney disease, heart failure) are at high risk for infectious diseases.[2] Moreover, some protein biomarkers representing some clinical conditions such as inflammation (e.g., interleukin-6),[3] mineral bone metabolism (e.g., fibroblast growth factor 23),[4,5] and cardiac damage (e.g., troponin T)[6] have been associated with an increased risk of infection. However, few studies have explored multiple proteins representing different pathophysiological pathways for infection risk in a single study population, leaving uncertainty on complex interplay among proteins and their associated pathways behind infectious disease.

A recently developed technology to measure the level of plasma proteins in an aptamer-based platform may provide a unique opportunity to address this knowledge gap. The Atherosclerosis Risk in Communities (ARIC) Study has measured the level of >5,000 plasma proteins using the SOMAscan Multiplexed Proteomic technology.[7] Using these data, we propose to explore proteins that are associated with the risk of infectious disease. We will divide the ARIC cohort into discovery and validation cohorts, and identify candidate proteins in the former and then replicate in the latter. We will apply advanced analytical techniques, such as lasso regression analysis. Once we find proteins significantly associated with infection risk, we will extend our analysis to pathway analysis and Mendelian randomization analysis to explore the underlying mechanisms of an increased risk of infectious disease.

5. Main Hypothesis/Study Questions:

Hypothesis 1: We will identify plasma proteins that are associated with the risk of incident hospitalization with infections in the general population.

Hypothesis 2: We will explore protein pathways and protein-protein interactions that underlie the risk of infection in the general population.

Hypothesis 3: We will explore causal pathways of proteins to increase the risk of infection through applying a Mendelian randomization analysis, if plausible.

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 and exclusion criteria: We will include all ARIC participants who have SOMA scan protein measurements at visit 2, 3, or 5. We will exclude from the analysis if participants had end-stage kidney disease or prior history of hospitalization with infection at the time of SOMA

scan measurement, self-identified as non-white and non-African-American (and non-white in Washington Co. and Minnesota), and had missing covariates of interest.

Exposures: Plasma concentrations of protein analytes are quantified by standard DNA microarrays in relative fluorescence units (RFUs) using a Slow Off-rate Modified Aptamer (SOMAmer)-based capture array (SomaLogic, Inc, Boulder, Colorado). Each protein analyte measurement underwent the standardization and normalization according to the SOMAscan protocol.[7] Visit 2 will provide us the largest number of study population and longest follow-up period and thus be used for our primary analysis. Although visits 3 and 5 will be used for confirmatory analysis, the results for visit 5 will be carefully interpreted since protein profiles related to infection can be different in midlife (visits 2 and 3) and late-life (visit 5).

Outcomes: The outcome of interest will be incident hospitalization with infection. We will use the same approach as we did in our previous ARIC paper.[8] Specifically, hospital discharge records with the *International Classification of Diseases, Clinical Modification* (ICD-9-CM) codes have been captured through active surveillance for all the ARIC study participants. The definitions of infection-related ICD-9 codes are based on the analytical methods used in United States Renal Data System (USRDS),[9] including any pathogen-, organ-, or symptom-based diagnoses. In addition, the a priori determined four most common types of infection[10] will be separately analyzed, including pneumonia (ICD-9, 480-486), kidney and urinary tract infections (590, 590.0-4, 597, 598, 599.0, 601, 604, 607, and 608), bloodstream infections (038 and 790.7), and cellulitis (681 and 682).

Covariates: Age, race, sex, study center, body mass index, smoking status, alcohol status, education level and income (base on visit 1 data), estimated glomerular filtration rate, albuminuria (for visit 5 analysis), history of stroke, coronary heart disease, atrial fibrillation, heart failure, peripheral artery disease, kidney disease, hypertension, lung diseases, liver diseases, and cancer, and the use of medications (e.g., antihypertensive agents, immunosuppressive agents).

Analytic plan:
Plasma protein and risk of infection: At each visit (i.e., visit 2, 3, and 5), ARIC participants will be randomly divided into two subsets of discovery and validation cohort at a rough ratio of two to one (i.e., 66% for discovery cohort, 33% for validation cohort). Then, we will use multivariable Cox regression models to explore the association of plasma proteins with the risk of hospitalization with infection in the discovery cohort. To account for multiple comparison, we will apply Bonferroni correction, where a 2-sided p-value threshold for statistical significance will be adjusted to 0.05 divided by the number of proteins tested (e.g., 0.5/5000 if 5000 proteins are tested). We will also use the Benjamin-Hochberg procedure to optimize the false discovery rate, since Bonferroni correction may be too conservative leading to a high chance of type I error. Candidate proteins that have passed the test for statistical significance will be re-assessed in the validation cohort.

Candidate proteins and non-infection outcomes: Candidate proteins may be non-specific risk markers that reflect poor health conditions and a high risk of hospitalization in general. To examine the specificity of the association, we will test the models accounting for measures of

general health perceptions, such as Health and Life Profile (visit 2) and 12-item Short-Form Health Survey (SF-12) (visit 5). We will also explore the association of candidate proteins with hospitalization with osteoarthritis (ICD-9CM, 715) as a non-infection related hospitalization comparator. We select osteoarthritis, since it is the most common cause of hospitalization among middle to older age adults in the US,[1] but presumably its etiology is relatively distinct from that of infectious disease compared to other major causes of hospitalization (e.g., cardiovascular disease).

Lasso regression models: To determine the relative importance of candidate proteins compared to other known risk factors (e.g., kidney function), we will apply shrinkage methods, such as lasso regression models, which are a supervised learning algorithm that can continuously select relevant features by penalizing extra features in the model. Lasso regression shrinks the regression coefficients by imposing a penalty on their effect size. In addition, because lasso models can estimate the coefficients in a single model (unlike traditional regression models that can handle one independent variable at a time), we will consider pairwise variables to account for potential interactions between the variables (e.g., protein-protein, protein-kidney function), in addition to the models only accounting for individual factors.

Pathway analysis: Pathway analysis and network analysis can potentially enrich our findings by incorporating prior knowledge about protein features (e.g., pathways, complexes, functions), and thus may be helpful to understand the potential role of proteins in a broader context of their metabolic pathways and biological features. We will utilize bioinformatics resources (e.g., DAVID, PANTHER, KEGG PATHWAY) and perform an enrichment analysis (e.g., http://geneontology.org/, https://www.genome.jp/kegg/pathway.html) to relate candidate proteins to known biological process, molecular function, and cellular component.

Mendelian randomization analysis: If plausible, we will conduct a Mendelian randomization analysis for candidate proteins that show the significant association with the risk of infection. Specifically, we will link the candidate proteins to ARIC genome-wide association studies (GWAS) data, and explore the potential causal association between candidate protein and the risk of infection.

Limitations: Residual confounding is possible. Measurements based on an aptamer-based platform may lead to inaccurate protein identifications.

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? ____ Yes _X_ No

b. If Yes, is the author aware that the current derived consent file ICTDER05 must be used to exclude persons with a value RES_OTH and/or RES_DNA = "ARIC only" and/or "Not for Profit" ? ____ Yes ____ No
(The 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 current derived consent file ICTDER05 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/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)?

To date, we are not aware of any ARIC proposals relating proteomics to the risk of infectious disease.

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 https://sites.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.

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. Healthcare Cost and Utilization Project. Trends in Hospital Inpatient Stays in the United States, 2005–2014. Available: https://www.hcup-us.ahrq.gov/reports/statbriefs/sb225-Inpatient-US-Stays-Trends.pdf. Accessed 21 Jan 2020.
2. Centers for Disease Control and Prevention. People at High Risk For Flu Complications. Available : https://www.cdc.gov/flu/highrisk/index.htm. Accessed 22 May 2019.
3. Ishigami J, Taliercio J, H IF, et al. Inflammatory Markers and Incidence of Hospitalization With Infection in Chronic Kidney Disease. *American journal of epidemiology.* 2020;189(5):433-444.
4. Ishigami J, Jaar BG, Rebholz CM, et al. Biomarkers of Mineral and Bone Metabolism and 20-Year Risk of Hospitalization With Infection: The Atherosclerosis Risk in Communities Study. *The Journal of clinical endocrinology and metabolism.* 2017;102(12):4648-4657.
5. Ishigami J, Taliercio JT, Feldman HI, et al. Fibroblast Growth Factor 23 and Risk of Hospitalization with Infection in Chronic Kidney Disease: The Chronic Renal Insufficiency Cohort (CRIC) Study. *Journal of the American Society of Nephrology : JASN.* 2020;31(8):1836-1846.
6. Ishigami J, Hoogeveen RC, Ballantyne CM, et al. Associations of High-Sensitivity Cardiac Troponin and Natriuretic Peptide With Subsequent Risk of Infection in Persons Without Cardiovascular Disease: The Atherosclerosis Risk in Communities Study. *American journal of epidemiology.* 2019;188(12):2146-2155.
7. Tin A, Yu B, Ma J, et al. Reproducibility and Variability of Protein Analytes Measured Using a Multiplexed Modified Aptamer Assay. *J Appl Lab Med.* 2019;4(1):30-39.
8. Ishigami J, Grams ME, Chang AR, Carrero JJ, Coresh J, Matsushita K. CKD and Risk for Hospitalization With Infection: The Atherosclerosis Risk in Communities (ARIC) Study. *American journal of kidney diseases : the official journal of the National Kidney Foundation.* 2017;69(6):752-761.
9. United States Renal Data System. ESRD analytical methods. Available: http://www.usrds.org/2014/view/v2_00_appx.aspx. Accessed 14 Feb 2016.
10. Christensen KL, Holman RC, Steiner CA, Sejvar JJ, Stoll BJ, Schonberger LB. Infectious disease hospitalizations in the United States. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America.* 2009;49(7):1025-1035.