ARIC Manuscript Proposal Form

ARIC Publication Admin Use Only

Publication Committee Review Date:
_**11**_/**21**/_**23**_
ARIC Manuscript Proposal Number: # **4353**

1.a. Full Title: Protein biomarkers of ultra-processed food consumption and risks of chronic kidney disease, coronary heart disease, and all-cause mortality.

b. Abbreviated Title (Length 26 characters): Proteomics of UPF and outcomes

2. Writing Group [please provide a middle initial if available; EX: Adam L Williams]:
Writing group members: Shutong Du, MHS; Jingsha Chen, MS; Hyunju Kim, PhD; Alice H. Lichtenstein, PhD; Bing Yu, PhD; Nilanjan Chatterjee, PhD; Peter Ganz, MD; Josef Coresh, MD, PhD; Casey M. Rebholz, PhD, MS, MNSP, MPH

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

First author [please provide a middle initial; EX: Adam L Williams]:
Shutong **Du**
First name Middle initial Last name

Address: Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe St
Baltimore MD 21205

Phone: 206-601-3274
E-mail: sdu12@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 (The ARIC author should be involved enough in ARIC to be able to point the lead author to appropriate ancillary study PIs and to be able to search ARIC manuscript proposals if the lead author doesn't have the access needed to do such a search).
Name: Casey M. Rebholz
Address: 2024 E. Monument St, Suite 2-500
Baltimore MD 21287

Phone: 410-502-2359
E-mail: crebhol1@jhu.edu

3. Timeline: Data analysis and manuscript preparation is anticipated to take place within one year of approval of this proposal.

4. Rationale: Ultra-processed foods are defined as industrially manufactured, ready to eat food or drinks that contain additives and nonculinary substances, with little or no intact foods.[1] They are typically high in sugar, salt and refined carbohydrates, and low in micronutrients and fibers. In the U.S., these foods are becoming easily accessible and the consumption has been continuously increasing in both adults and children over the past two decades.[2,3] The consumption of ultra-processed food in the U.S. is estimated to constitute over half of total energy intake.[4] Previous studies have also linked higher ultra-processed food consumption with adverse health outcomes including cardiovascular diseases, diabetes mellitus and cancer.

In recent years, there have been debates surrounding the classification of ultra-processed food.[8,9] Although useful as a holistic approach to assess a unique characteristic of food consumption, ultra-processed food classification do not disentangle processing from nutritional content. High throughput, untargeted proteomic profiling provides an unprecedented opportunity to identify levels of plasma proteins that represent ultra-processed food consumption. Studying the proteomics of ultra-processed food allows for the discovery of objective biomarkers of ultra-processed food. Additionally, since proteomics analyses can provide substantial information, including protein-protein interactions, food metabolisms, and protein involvement in biological pathways, assessing prospective associations between ultra-processed foods-related proteins and health outcomes can provide further insights regarding the biological activity that links ultra-processed food consumption to disease.

5. Main Hypothesis/Study Aims:
Specific Aim 1: To discover novel biomarkers of ultra-processed food consumption using untargeted proteomics.

Specific Aim 2: Assess the prospective associations between ultra-processed foods-related proteins and health outcomes including chronic kidney disease, coronary heart disease, and all-cause mortality.

Hypothesis 1: We hypothesize that a set of proteins will be associated with levels of ultra-processed food intake.

Hypothesis 2: We hypothesize that a subset of ultra-processed foods-related proteins will be prospective associated with health outcomes including chronic kidney disease, coronary heart disease, and all-cause mortality.

6. Design and analysis - please address the following aspects:

a) inclusion/exclusion

For specific aim 1, we will exclude participants based on following criteria:

1) Missing proteomics data
2) Missing or implausible dietary data
    a. Missing 10 or more items on the food frequency questionnaire
    b. <500 or >3500 kcal/day energy intake for women, <600 or >4500 kcal for men
3) Non-black and non-white participants, black participants from Minneapolis, Minnesota, and black participants from Washington County, Maryland
4) Missing covariates [age, sex, race, center, body mass index, education level as a proxy for socioeconomic status, smoking status, physical activity level, total energy intake, diabetes status, hypertension status, coronary heart disease, and estimated glomerular filtration rate (eGFR)]
5) Prevalent coronary heart disease cases; eGFR less than 60 mL/min/1.73m^2; missing incident chronic kidney disease or incident coronary heart disease information.

b) study design

We will investigate whether plasma proteins levels are associated with levels of ultra-processed foods intake, and if ultra-processed food-related proteins are associated with health outcomes. We will use available plasma proteomics data measured by the SomaScan® (SomaLogic, Inc., Boulder, CO, USA) platform from plasma specimens collected from study participants at study visit 3, and ultra-processed foods intake calculated based on dietary intake data collected with a food frequency questionnaire at study visit 3. Outcome ascertainments were continued from visit 3 (1993-1995) until December 31st, 2020.

c) outcome and other variables of interest with specific reference to the time of their collection

Dietary Assessment:

Dietary intake was assessed using a semi-quantitative 66-item food frequency questionnaire (FFQ) which was modified from the original Willett FFQ.[10,11] Participants reported their usual intake of food items of a specified portion size during the preceding year. The FFQ was administered at study visit 3 by trained interviewers. Nutrient intake was derived by multiplying the servings of each food item by its nutrient content which was obtained primarily from U.S. Department of Agriculture sources.

Classification of Ultra-Processed Food:

All food items on the FFQ were classified into one of four categories according to the NOVA classification system based on processing level: 1) unprocessed or minimally processed foods, including foods obtained directly from the nature, and undergo little or no alteration other than methods such as freezing, drying or pasteurization, and with no addition of culinary ingredients; 2) processed culinary ingredients, including substances extracted from natural foods or by natural process such as pressing, grinding or crushing. Examples such as salt, sugar, vegetable oils and fats; 3) processed foods, including products that have undergone specific preparation or preservation with the use of level 2 added to level 1 foods to preserve or to make them more palatable; 4) ultra-processed foods, including products that

were made entirely or mostly from industrial formulations with artificial additives and nonculinary substances.[12]

We previously published the classification of food items FFQ administered in the ARIC study according to processing level.[13, 14]

Outcome Assessment:

Incident chronic kidney disease will be defined using the following criteria:

1. A reduced kidney function, defined by eGFR less than 60 ml/min/1.73 m^2, along with a decline of at least 25% in eGFR in comparison to the initial visit.
2. Stage 3 or higher chronic kidney disease-related hospitalizations (based on International Classification of Disease [ICD] 9 and 10 code).
3. Death related to stage 3 or higher diagnosis of chronic kidney disease according to ICD code, identified through linkage to National Death Index
4. End-stage renal disease diagnosis, kidney dialysis or transplantation identified through linkage with the U.S. Renal Data System (USRDS) registry.

Incident coronary heart disease will be defined as hospitalized myocardial infarction or fatal coronary heart disease. Incident cases were identified through active surveillance on cardiovascular disease-related hospitalizations and deaths by study staff and were adjudicated by a group of experts.

All-cause mortality defined as death caused by any reason, as determined through linking to the National Death Index, conducting telephone interviews with relatives or proxies, and actively monitoring local hospital discharge and state death records.

Proteomic Profiling:

Plasma specimens were stored at -80°C since collection at visit 3 (1993-1995). In the ARIC study, participants were instructed to fast for 12 hours prior to the study visit. As such, the proteomic profiling of plasma specimens reflects the impact of usual dietary intake on the proteome rather than recent intake. The proteome could be considered the internal dose of dietary intake as it reflects not only the consumption of food but also the metabolism of food.

Modified aptamers, which are protein-specific binding reagents that have been prepared using chemically modified DNA, are used to quantify proteins by SomaLogic (Boulder, Colorado).[15,16] The modifications provide nucleotides with protein-like functional groups allowing for binding with high affinity and specificity to proteins. The current version of the SOMAscan platform quantifies approximately 5,000 proteins. Proteomic data underwent the SOMAscan standardization and normalization process. Additional data cleaning and quality control procedures were conducted by ARIC study investigators. [17]

d) summary of data analysis

For specific aim 1:

To characterize the proteomic profile of ultra-processed food, we will use visit 3 data. We will conduct analyses to examine individual proteins as well as identify a composite of proteins that are representative of ultra-processed food. Proteins will be log-transformed to improve the normality of their distribution. We will use linear regression models to investigate the association between log-transformed protein levels as the outcome variables and ultra-processed food as the

exposure variable. Multivariable regression models will adjust for age, sex, race, study center, education level, smoking status, body mass index, physical activity levels, eGFR, total energy intake, and disease status including diabetes status, hypertension status, and coronary heart disease.

We will combine multiple candidate proteins that are significantly associated with ultra-processed food since panels of multiple biomarkers provide better estimates of dietary intake than single biomarkers.[18] We will dichotomize ultra-processed food empirically (highest quartile vs. lower 3 quartiles) given the lack of a clinically important threshold, as we have previously done in a previous publication on the metabolomics of dietary acid load.[19] We will calculate the area under the curve (C statistic) for incremental ability of each individual protein (as well as the panel of proteins) to predict the highest quartile vs. lower 3 quartiles of ultra-processed food. We will calculate Pearson's correlation coefficients between the significant proteins to describe their interrelationships.

We will use the Bonferroni method to account for multiple comparisons.[20] We will randomly divide the study population into a 2/3 discovery sample and a 1/3 validation sample. We will examine which proteins that were statistically significant ($p<0.05$/number of proteins analyzed) in the discovery (2/3) sample are also statistically significant ($p<0.05$/number of proteins statistically significantly associated with ultra-processed food in the discovery sample) in the validation sample.

We will examine potential differences by sex and race in stratified analyses. We will formally test for differences by sex, race and body mass index categories using likelihood ratio tests. We will also conduct pathway analysis using a broader list of proteins, that are statistically significant at FDR<0.05 threshold.

<u>For specific aim 2:</u>

After obtaining the list of proteins that are significantly associated with ultra-processed food consumption in aim 1, we will use Cox proportional hazards models to evaluate the prospective association between ultra-processed foods-related proteins and time to incident coronary heart disease, chronic kidney disease and all-cause mortality.

We will calculate hazard ratios for each outcome with the corresponding 95% confidence intervals per doubling of proteins levels (log2-transformed). We will account for known confounders including age, sex, race, center, education level, smoking status, physical activity level, body mass index, eGFR, and total energy intake. We will use Bonferroni correction and false discovery rate methods to adjust the statistical threshold to account for multiple comparisons and reduce the likelihood of detecting false positive findings.

e) Any anticipated methodologic limitations or challenges if present

Ultra-processed foods intake was recorded using self-reported food frequency questionnaire, which relies on respondents' ability to report frequency and portion size and is prone to biases and errors. Future research is warranted to examine the proteomic profile of ultra-processed food intake in well controlled feeding studies. Exposure misclassification of food into processing level categories using the NOVA classification system is possible. Specifically, the relatively short length of the questionnaire could result in an underestimation of the absolute intake of ultra-processed food. However, we should be able to adequately rank participants within the population according to consumption of ultra-processed food. There is relatively less detailed

information available about specific foods in the ARIC study compared to dietary data collected using other assessment tools (e.g., 24-hour dietary recall). We will use the rubric used in previous manuscripts on ultra-processed food and coronary heart disease and ultra-processed food and chronic kidney disease in the ARIC study, which have passed peer review and are now published. [13, 14]

The SomaLogic platform estimates proteins in relative levels, thus, subsequent work is needed to develop methods for absolute quantification of ultra-processed foods biomarkers.

The FFQ used in this study were not specifically designed to answer food processing questions, hence some common ultra-processed food items were not included, and the dietary intake data may not accurately reflect the current American's diet.

f) Will the author need Limited data to complete the proposed manuscript? ☐ Yes, Limited data is needed (Provide a brief (2-3 sentences) justification for requesting PHI data) . ☒ No, De-identified data will be sufficient.

*Please note, Limited dataset access is strict and rarely provided. Limited data includes identifiable information such as dates (birthdays, visit dates, etc.). CMS, Genomic, Geocoded, Proteomics/Somalogic, and other -omic data all fall under the limited data category. De-identified data does not include dates. All dates are date adjusted to "Days since Visit 1".

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? (Non-ARIC analysis means that the authors are not regarded as ARIC investigators and the "ARIC author" is essentially just a facilitator rather than an integral part of the writing group.) ☐ Yes ☒ 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 is distributed to ARIC PIs annually, 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 ☒ 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"? ☐ Yes ☐ No

9. The lead author or the "sponsoring" ARIC 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 website at: https://aric.cscc.unc.edu/aric9/proposalsearch [ARIC Website□Publications□Proposal Search]

☒ 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 #3727 - Association between ultra-processed food consumption and risk of coronary heart disease and chronic kidney disease in the Atherosclerosis Risk in Communities Study

This previous proposal did not incorporate proteomic data.

MP #3755 – Protein Biomarkers of the DASH Diet and Hypertension Risk in the ARIC study

This previous proposal was focused on another dietary exposure.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or does it use current [or ongoing] ancillary study data (this includes ACHIEVE)? ☒ Yes ☐ No → Skip to question 12

11.b. If yes to 11.a., is the proposal

☒ **A. primarily the result of an ancillary study**

☐ **B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables)**

11.c. If yes to 11.a., list number* 2021.27

*ancillary studies are listed by number
https://aric.cscc.unc.edu/aric9/researchers/ancillary_studies/approved_ancillary_studies [ARIC Website□Ancillary Studies□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 https://aric.cscc.unc.edu/aric9/publications/policies_forms_and_guidelines [ARIC Website□Publications□Publication Policies, Forms, and Guidelines]. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

References:

1. Monteiro CA, Cannon G, Levy RB, et al. Ultra-processed foods: what they are and how to identify them. *Public Health Nutr.* 2019;22(5):936-941. doi:10.1017/S1368980018003762
2. Juul F, Parekh N, Martinez-Steele E, Monteiro CA, Chang VW. Ultra-processed food consumption among US adults from 2001 to 2018. *Am J Clin Nutr.* 2022;115(1):211-221. doi:10.1093/ajcn/nqab305
3. Wang L, Martínez Steele E, Du M, et al. Trends in Consumption of Ultraprocessed Foods Among US Youths Aged 2-19 Years, 1999-2018. *JAMA.* 2021;326(6):519-530. doi:10.1001/jama.2021.10238
4. Baraldi LG, Martinez Steele E, Canella DS,Monteiro CA. Consumption of ultra-processed foods and associated sociodemographic factors in the USA between 2007 and 2012: evidence from a nationally representative cross-sectional study. BMJ Open 2018;8:e020574.
5. Srour B, Fezeu LK, Kesse-Guyot E, et al. Ultra-processed food intake and risk of cardiovascular disease: prospective cohort study (NutriNet-Santé). *BMJ*. 2019;365:l1451. Published 2019 May 29. doi:10.1136/bmj.l1451
6. Srour B, Fezeu LK, Kesse-Guyot E, et al. Ultraprocessed Food Consumption and Risk of Type 2 Diabetes Among Participants of the NutriNet-Santé Prospective Cohort. *JAMA Intern Med*. 2020;180(2):283-291. doi:10.1001/jamainternmed.2019.5942
7. Fiolet T, Srour B, Sellem L, et al. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. *BMJ*. 2018;360:k322. Published 2018 Feb 14. doi:10.1136/bmj.k322
8. Poti JM, Braga B, Qin B. Ultra-processed Food Intake and Obesity: What Really Matters for Health-Processing or Nutrient Content?. *Curr Obes Rep*. 2017;6(4):420-431. doi:10.1007/s13679-017-0285-4
9. Gibney MJ, Forde CG, Mullally D, Gibney ER. Ultra-processed foods in human health: a critical appraisal [published correction appears in Am J Clin Nutr. 2018 Mar 1;107(3):482-483]. *Am J Clin Nutr*. 2017;106(3):717-724. doi:10.3945/ajcn.117.160440
10. Stevens J, Metcalf PA, Dennis BH, Tell GS, Shimakawa T, Folsom AR. Reliability of a food frequency questionnaire by ethnicity, gender, age and education. *Nutr Res.* 1996;16(5):735-745.
11. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am J Epidemiol*. 1985;122(1):51-65. doi:10.1093/oxfordjournals.aje.a114086
12. Monteiro CA, Cannon G, Moubarac JC, Levy RB, Louzada MLC, Jaime PC. The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutr*. 2018;21(1):5-17. doi:10.1017/S1368980017000234
13. Du S, Kim H, Rebholz CM. Higher Ultra-Processed Food Consumption Is Associated with Increased Risk of Incident Coronary Artery Disease in the Atherosclerosis Risk in Communities Study. *J Nutr*. 2021;151(12):3746-3754. doi:10.1093/jn/nxab285
14. Du S, Kim H, Crews DC, White K, Rebholz CM. Association Between Ultraprocessed Food Consumption and Risk of Incident CKD: A Prospective Cohort Study. Am J Kidney Dis. 2022;80(5):589-598.e1. doi:10.1053/j.ajkd.2022.03.016

15. Gold L, Ayers D, Bertino J, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One*. 2010;5(12):e15004. Published 2010 Dec 7. doi:10.1371/journal.pone.0015004
16. Brody E, Gold L, Mehan M, et al. Life's simple measures: unlocking the proteome. *J Mol Biol*. 2012;422(5):595-606. doi:10.1016/j.jmb.2012.06.021
17. 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. doi:10.1373/jalm.2018.027086
18. Garcia-Aloy M, Rabassa M, Casas-Agustench P, Hidalgo-Liberona N, Llorach R, Andres-Lacueva C. Novel strategies for improving dietary exposure assessment: Multiple-data fusion is a more accurate measure than the traditional single-biomarker approach. *Trends in Food Science & Technology*. 2017;69:220-229. doi:10.1016/j.tifs.2017.04.013
19. Rebholz CM, Surapaneni A, Levey AS, et al. The Serum Metabolome Identifies Biomarkers of Dietary Acid Load in 2 Studies of Adults with Chronic Kidney Disease. *J Nutr*. 2019;149(4):578-585. doi:10.1093/jn/nxy311
20. Curtin F, Schulz P. Multiple correlations and Bonferroni's correction. *Biol Psychiatry*. 1998;44(8):775-777. doi:10.1016/s0006-3223(98)00043-2

To view publications materials, click "Log in" at the top right of the ARIC website. Click "Forgot Password" if you are experiencing issues with logging in.