ARIC Manuscript Proposal #3836

PC Reviewed: 5/11/21 **Status: _____** **Priority: 2**
SC Reviewed: _____ **Status: _____** **Priority: ____**

1.a. Full Title: Establishing a DNA methylation-based smoking index and its predictive value for lung function decline, incident airflow limitation, and all-cause mortality

b. Abbreviated Title (Length 26 characters): DNA methylation smoking index

2. Writing Group:
Writing group members:

Christina Eckhardt, cme2113@cumc.columbia.edu
Haotian Wu, hw2694@cumc.columbia.edu
Diddier Prada, dgp2114@cumc.columbia.edu
Myriam Fornage, Myriam.Fornage@uth.tmc.edu
Jan Bressler, jan.bressler@uth.tmc.edu
James Pankow, panko001@umn.edu
Weihua Guan, wguan@umn.edu
Stephanie J. London, london2@niehs.nih.gov
Andrea Baccarelli, ab4303@cumc.columbia.edu

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

First author: Christina Eckhardt
Address: 622 West 168th Street, PH 8-101
New York, NY, 10032

Phone: 212-305-2972 Fax: 212-342-3144
E-mail: cme2113@cumc.columbia.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: Eric A. Whitsel, MD, MPH
Address: Departments of Epidemiology and Medicine
University of North Carolina
123 W. Franklin St., Suite 410, Room 4226
Chapel Hill, NC, 27516-8050

Phone: 919-966-3168 Fax: 919-966-9800
E-mail: eric_whitsel@unc.edu

3. Timeline

4. Rationale:

Tobacco smoking is the single leading cause of preventable death and disease in the United States,[1] and causes more than 480,000 premature deaths in the US per year. Even decades after smoking cessation, former smokers carry an increased risk of lung disease, lung cancer, and stroke[2-4]. Accurate monitoring and detection of tobacco use is therefore critically important for identifying at-risk populations who may benefit from cessation assistance and targeted medical screening. However, self-reported tobacco smoke exposure and second-hand smoke exposure have been shown to underestimate true smoke exposure[5,6], and current biomarkers including urine cotinine are unable to quantify long-term tobacco smoke exposure[7]. Developing a robust index of cumulative tobacco smoke exposure may enable health care practitioners to identify individuals who would benefit from early interventions, including targeted tobacco cessation assistance and smoking-related medical screening, which may help minimize smoking-related morbidity and mortality.

Smoking-induced epigenetic alterations may offer a unique biologic window into the mechanisms by which smoking exerts nocuous effects on the human body and impacts clinical outcomes. Tobacco smoke alters DNA methylation at thousands of cytosine-phosphate-guanine (CpG) sites in nucleated blood cells[8-10], some of which localize to genes associated with lung disease[11]. Epigenetic changes regulate tissue-specific gene expression[12], and may provide a link between tobacco smoke exposure and smoking-related lung diseases including chronic obstructive pulmonary disease[13]. In this context, constructing a smoking score that indexes smoking-related changes in DNA methylation could produce a reliable biomarker of cumulative tobacco smoke exposure, which may in turn be useful in predicting incident lung disease and prospective mortality. While several prior studies generated DNA methylation-based smoking scores[14-16], the scores either had limited generalizability[14,15] or frequently misclassified former smokers due to limited sensitivity[16]. We therefore propose to construct an innovative DNA methylation-based smoking index (DNAm-smoke) using a mixed effects elastic net regression model and Bayesian kernel machine regressions in order to classify cumulative tobacco smoke exposure. We will then analyze associations of DNAm-smoke with longitudinal lung function, incident airflow limitation, and all-cause mortality in a large sample of adults.

5. Main Hypothesis/Study Questions:

Aim 1: Determine whether a blood-based biomarker can classify prior tobacco smoke exposure. *Hypothesis:* A mixed effects elastic net regression model and Bayesian kernel machine regressions can be implemented in a whole blood dataset to generate a DNA methylation-based smoking index (DNAm-smoke) that classifies cumulative tobacco smoke exposure **(Figure 1)**.

Aim 2: Determine whether DNAm-smoke is associated with longitudinal lung function. *Hypothesis 2a:* A higher DNAm-smoke score will be associated with lower forced expiratory volume in 1 second (FEV1) and lower FEV1/Forced Vital Capacity (FVC) on spirometry measurements obtained after the baseline visit. *Hypothesis 2b:* A higher DNAm-smoke score will be associated with increased risk of incident airflow limitation (FEV1/FVC < 0.7).

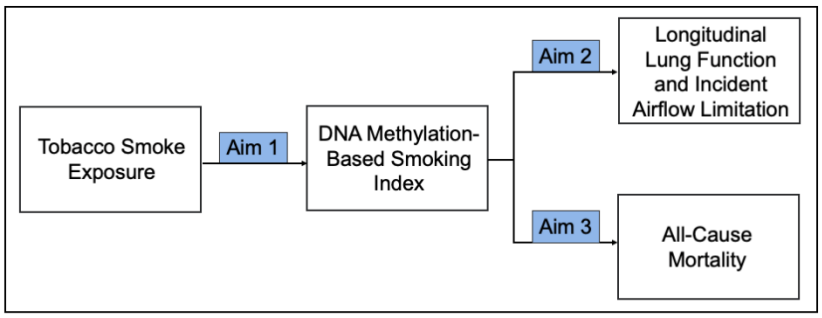

Figure 1. Conceptual model of the project

Aim 3: Determine whether DNAm-smoke is associated with all-cause mortality. *Hypothesis*: A higher DNAm-smoke score will be associated with increased all-cause mortality.

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).

Sample

- We propose to use data from four cohorts:
  - Atherosclerosis Risk in Communities (ARIC, n = 2,905)
  - Coronary Artery Risk Development in Young Adults (CARDIA, n = 1,042)
  - Normative Aging Study (NAS, n = 794)
  - Women's Health Initiative (WHI, n = 6,769)
- Aim 1 Sample: In accordance with prior studies that successfully pooled multiple DNA methylation datasets to create a training dataset,[17,18] we will pool all participants who submitted whole blood for DNA methylation analysis (N = 11,510).
- Aims 2 and 3 Samples: Our primary analysis strategy for Aims 2 and 3 will be to analyze each cohort individually using the available variables in each cohort. Secondarily, we will meta-analyze the results to derive a single estimate for each aim. When evaluating associations with incident airflow limitation, we will exclude participants with airflow limitation present on baseline spirometry.

Exposures:

- Smoking (total pack-years, duration of smoking, years since quitting smoking)
- DNA methylation levels extracted from peripheral blood using the HM450 chip during ARIC Visits 2 (1990 - 1992) and 3 (1993 - 1995).

Endpoints: **Table 1** outlines which endpoints are available in each cohort.

- FEV1, FVC, FEV1/FVC measured during ARIC Visits 2 (1990 - 1992) and 5 (2011 - 2013)
- Incident airflow limitation
- All-cause mortality

Table 1. Endpoints available in each cohort

Aim	Endpoints	Cohorts	N
1	DNAm-smoke	ARIC, CARDIA, NAS, WHI	11,510
2	Spirometry	ARIC, CARDIA, NAS	4,741
3	Mortality*	ARIC, NAS, WHI	10,468

*Due to younger age at enrollment, we will not include CARDIA participants

Covariates:

- Sociodemographic: age, sex, race/ethnicity, education, clinical site
- Anthropometric: height, weight, BMI
- Physical Activity
    - In ARIC, we will adjust for self-reported physical activity as reported on the modified Baecke Questionnaire during Visits 1, 3 and 5[19].
- Medical comorbidities: hypertension, diabetes, coronary heart disease
- DNA methylation: Microarray batch, sample, plate, chip, blood cell type composition

Statistical Analysis

General Strategies:

Cohort Differences and Cohort Specific Estimates: As the four cohorts differ in their design, population characteristics, data collection methods, and variable structures, our primary analysis strategy for Aims 2 and 3 will be to analyze each cohort individually using the extant variables for each cohort. This approach will help us to assess the robustness of the newly-derived DNAm-smoke index to different cohorts while still allowing us to meta-analyze the resulting effect estimates to derive a single effect estimate. It also bypasses the need to harmonize data that may not be naturally compatible. Lastly, it allows us to leverage other cohorts to conduct sensitivity analyses that would otherwise not be possible in certain cohorts. For example, race effects are difficult to estimate in ARIC (where it is strongly confounded by center) and NAS (strong homogeneity in race). For this variable, we can leverage CARDIA and WHI to assess the impact of race on the relationship between DNAm-smoke index and health outcomes. It is otherwise important to note that Aim 1 does not require variables beyond smoking pack-years, DNA methylation (which were measured using the same platform) and sex. All analyses will be performed at Columbia University Irving Medical Center.

Selective Loss to Follow-Up (i.e. Selection Bias). Prior to analyses of Aims 2 and 3, we will calculate and apply inverse probability weights of censoring to the appropriate models.[20] In addition, as a sensitivity analysis, we can apply quantitative bias analysis to estimate the potential association under both plausible and extreme scenarios.

Aim 1: If possible, we will obtain unprocessed DNA methylation data and perform uniform data processing across cohorts. We will perform batch correction for each individual cohort before proceeding with the analyses. To derive the smoking index, we will emulate two prior studies that successfully combined multiple DNA methylation datasets to perform elastic net regressions of gestational age[18] and methylation age,[17] and will pool DNA methylation data from the four cohorts. Pooling across these four distinct cohorts will enhance our demographic coverage, which will in turn increase the generalizability of the smoking index. We will randomly split participants into discovery (n = 9,208) and validation samples (n = 2,302). We will use a mixed effects elastic net regression model constructed in the *R* package *caret* to regress a smoking index in the training dataset, which will allow us to account for cohort- and site-specific effects. We will input the normalized methylation β-values and the participants' smoking pack-years and will include sex as a covariate. The elastic net regression model will select an array of smoking-related probes. Because elastic net regression treats methylation scores as linear and additive, which are assumptions that may not hold (pack-years may vary non-linearly with some β-values and some may be interactive in predicting DNAm-smoke), we will subsequently fit Bayesian kernel machine regressions (BKMR) to the selected probes to derive the final probes that will comprise the smoking index. We will calculate the root-mean-square error to assess the fit of the regression model.

Aim 2: We will test associations of DNAm-smoke with longitudinal FEV1, FVC, and FEV1/FVC using linear mixed models. Models will be adjusted for age, sex, race/ethnicity, clinical site, education, height, weight, and physical activity. We calculated power conservatively without considering the efficiency gains from longitudinal repeated measurements design. With a sample size of 4,741, we will have 79% power to detect a minimum effect size of 0.04 standard deviation (SD) change in FEV1, FVC, and FEV1/FVC per SD increase in DNAm-smoke. We will also test associations of DNAm-smoke with incident airflow limitation using Cox proportional hazards models. Models will be adjusted for age, sex, race/ethnicity, clinical site, education, height, weight, and physical activity. Assuming a 10% rate of incident airflow limitation during follow-up[21], we will have 81% power to detect a minimum hazard ratio (HR) of 1.14 per SD increase in DNAm-smoke **(Figure 2A).**

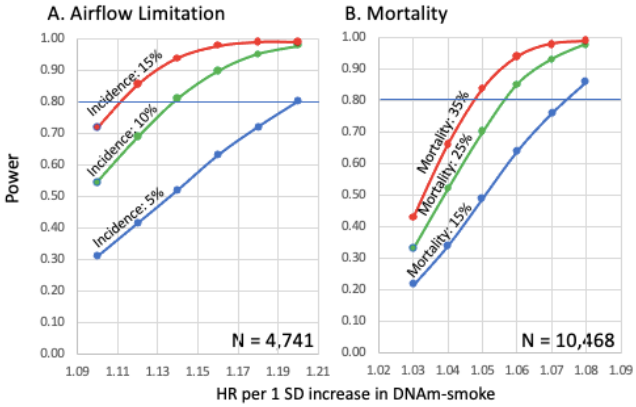

Figure 2. Achieved power with our sample sizes for Cox proportional hazards models under varying assumptions

Aim 3: We will model associations of DNAm-smoke with all-cause mortality using Kaplan-Meier curves and Cox proportional hazards models. Cox proportional hazards models will be adjusted for age, sex, race/ethnicity, clinical site, study, education, height, weight, physical activity, hypertension, diabetes and coronary heart disease. Assuming a mortality rate of 25% during follow-up[22-24], we will have 85% power to detect a minimum HR of 1.06 per SD increase in DNAm-smoke **(Figure 2B**).

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? __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" ? N/A
(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? __Yes__

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__

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
__Yes__

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)?

The present manuscript proposal differs from the following proposals in that we plan to derive a novel DNA methylation-based smoking score that reflects cumulative tobacco smoke exposure and can be applied in studies without cumulative tobacco smoke measurements. We will then evaluate the utility of that smoking score as a predictor of lung function decline and all-cause mortality in order to determine its clinical utility and relevance.

2342: Epigenome-wide association of DNA methylation with smoking in the Atherosclerosis Risk in Communities Study

2345: A prospective study of the association of DNA methylation age with lung function and type 2 diabetes in the Atherosclerosis Risk in Communities Study

3120: Meta-analysis of the relation of DNA methylation patterns, lung function, and chronic obstructive pulmonary disease

3414: Assessment of smoking-related cancer risk using DNA methylation as a measure of adult smoking history

DNA methylation-based risk score and prediction of all-cause mortality in the Atherosclerosis Risk in Communities Study

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __Yes__

11.b. If yes, is the proposal

_X_ A. primarily the result of an ancillary study (list number* 2012.10, 2014.20)

___ 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. The Health Consequences of Smoking-50 Years of Progress: A Report of the Surgeon General. Atlanta (GA)2014.
2. Pan B, Jin X, Jun L, Qiu S, Zheng Q, Pan M. The relationship between smoking and stroke: A meta-analysis. Medicine (Baltimore) 2019;98:e14872.

3. Oelsner EC, Balte PP, Bhatt SP, et al. Lung function decline in former smokers and low-intensity current smokers: a secondary data analysis of the NHLBI Pooled Cohorts Study. Lancet Respir Med 2019.
4. Tindle HA, Stevenson Duncan M, Greevy RA, et al. Lifetime Smoking History and Risk of Lung Cancer: Results From the Framingham Heart Study. J Natl Cancer Inst 2018;110:1201-7.
5. Connor Gorber S, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. Nicotine Tob Res 2009;11:12-24.
6. Elf JL, Kinikar A, Khadse S, et al. Secondhand Smoke Exposure and Validity of Self-Report in Low-Income Women and Children in India. Pediatrics 2018;141:S118-S29.
7. Jarvis MJ, Russell MA, Benowitz NL, Feyerabend C. Elimination of cotinine from body fluids: implications for noninvasive measurement of tobacco smoke exposure. Am J Public Health 1988;78:696-8.
8. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics 2015;7:113.
9. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. Circ Cardiovasc Genet 2016;9:436-47.
10. Zeilinger S, Kuhnel B, Klopp N, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One 2013;8:e63812.
11. Wauters E, Janssens W, Vansteenkiste J, et al. DNA methylation profiling of non-small cell lung cancer reveals a COPD-driven immune-related signature. Thorax 2015;70:1113-22.
12. Orr BA, Haffner MC, Nelson WG, Yegnasubramanian S, Eberhart CG. Decreased 5-hydroxymethylcytosine is associated with neural progenitor phenotype in normal brain and shorter survival in malignant glioma. PLoS One 2012;7:e41036.
13. Hogg JC. Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. Lancet 2004;364:709-21.
14. Elliott HR, Tillin T, McArdle WL, et al. Differences in smoking associated DNA methylation patterns in South Asians and Europeans. Clin Epigenetics 2014;6:4.
15. Zhang Y, Florath I, Saum KU, Brenner H. Self-reported smoking, serum cotinine, and blood DNA methylation. Environ Res 2016;146:395-403.
16. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. Epigenomics 2019.
17. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol 2013;14:R115.
18. Knight AK, Craig JM, Theda C, et al. An epigenetic clock for gestational age at birth based on blood methylation data. Genome Biol 2016;17:206.
19. Richardson MT, Ainsworth BE, Wu HC, Jacobs DR, Jr., Leon AS. Ability of the Atherosclerosis Risk in Communities (ARIC)/Baecke Questionnaire to assess leisure-time physical activity. Int J Epidemiol 1995;24:685-93.
20. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ, Jr. Selection Bias Due to Loss to Follow Up in Cohort Studies. Epidemiology 2016;27:91-7.
21. Luoto JA, Elmstahl S, Wollmer P, Pihlsgard M. Incidence of airflow limitation in subjects 65-100 years of age. Eur Respir J 2016;47:461-72.
22. Manson JE, Aragaki AK, Rossouw JE, et al. Menopausal Hormone Therapy and Long-term All-Cause and Cause-Specific Mortality: The Women's Health Initiative Randomized Trials. JAMA 2017;318:927-38.
23. Seidelmann SB, Folsom AR, Rimm EB, Willett WC, Solomon SD. Dietary carbohydrate intake and mortality: reflections and reactions - Authors' reply. Lancet Public Health 2018;3:e521.
24. Turiano NA, Hill PL, Roberts BW, Spiro A, 3rd, Mroczek DK. Smoking mediates the effect of conscientiousness on mortality: The Veterans Affairs Normative Aging Study. J Res Pers 2012;46:719-24.