ARIC Manuscript Proposal #3879

PC Reviewed: 6/8/21	**Status: _____**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Identifying Signature Inflammatory Mediators in Gingival Crevicular Fluid in Periodontal Profile Class-Stages

b. Abbreviated Title (Length 26 characters): **Perio PPC mediator network**

2. Writing Group:

Writing group members:
Shaoping Zhang, Kevin Moss, Jim Beck, Carissa L. Comnick, Miyuraj Hikkaduwa, Erliang Zeng, Xianjin Xie

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

First author:
Address: Shaoping Zhang
N401 Dental Science Building
801 Newton Rd.
University of Iowa College of Dentistry
Iowa City, IA, 52242

Phone: 31-335-7386
E-mail: shaoping-zhang@uiowa.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: **James Beck**
Address: Room 3501-D Koury Oral Health Science
Department of Dental Ecology
Adams School of Dentistry
University of North Carolina at Chapel Hill
385 S. Columbia St,
Chapel Hill, NC 27599 CB#7450
Phone: Fax:
E-mail: James_Beck@unc.edu

We invite ARIC investigator(s) to participate in this manuscript

3. Timeline: Nine months to one year for manuscript draft.

4. Rationale: Periodontitis, which is an inflammatory disease, irreversibly damages tooth-supported structures that eventually leads to tooth loss. This advanced form of periodontal disease affects more than 47% of the adult Americans (1). This quite common oral disease is also associated with systemic inflammatory diseases including type 2 diabetes (2). Periodontitis patients who exhibit a similar clinical disease phenotype, such as attachment level, may have different risks for disease progression. However, the most frequently used classification systems **fail** to identify the heterogeneous pattern of periodontitis. The traditional expert opinion-based classification approach is **unable** to categorize patients into relatively homogenous subclasses, creates barriers for risk assessment, prevention and effective treatment of periodontitis.

Recently, by applying a latent class analysis (LCA), which is an unbiased data clustering statistical method, to the clinical data collected from 6,793 individuals in the atherosclerosis risk in community (ARIC) cohort, we created our periodontal profile class (PPC)-staging model for periodontal disease classification(3, 4). This data-driven PPC staging tool includes seven mutually exclusive disease categories based on the analysis of seven tooth-level indices. This PPC-staging system **outperformed** the traditional data classification systems regarding the strength of associations with the prevalence of systemic conditions and in risk prediction of incident diabetes and stroke. For example, PPC-Stage IV ("severe periodontitis") and VII ("severe tooth loss") categories cross-sectionally associated with the prevalence of diabetes and longitudinally associated with the incident diabetes in the ARIC cohort. However, the four-level 1999 AAP/CDC Periodontal Disease Classification did not identify significant associations with diabetes.

The inflammatory trait that is unique to each PPC stages remains under-investigated. The improved clinical utility of this PPC staging system justifies the further assessment of the inflammatory biomarker network in the gingival crevicular fluid (GCF) that reflects the inflammatory state of local gingival tissue in periodontitis patients. The objective of this manuscript proposal is to identify signature GCF inflammatory mediators specific for each PPC-stage. Within the scope of this proposal, we seek to quantify the expression level of 22 inflammatory mediators determined by Luminex multiplex immunoassays in 385 archived GCF samples that were collected from 55 individuals different ARIC participants within each of the seven PPC stages.

This manuscript proposal is a follow-up study of PPC-Stages in the same ARIC population. We seek to identify the biological underpinning through a proteomics approach in GCF for a novel periodontal disease classification system that improves the homogeneity of clinical disease phenotypes.

5. Main Hypothesis/Study Questions: GCF inflammatory mediators form different patterns or "signature networks" that are associated with each or several PPC-stages. Each PPC-Stage or key PPC-Stages such as PPC-Stage IV and VII harbor unique signature networks that serve the biological underpinning of this staging tool that was derived from clinical data. By measuring the level of those network mediators from patients, we aim to 1) assess the inflammatory traits for PPC-Stages and mechanisms of how those disease-associating network mediators interact; 2)

predict the PPC-Stages using the GCF mediator measurements to improve diagnostic precision of periodontal disease.

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).

We propose to leverage the data and the gingival crevicular fluid samples obtained from the dental examination at Visit 4 (1996-1998) of ARIC cohort study.

We will first categorize ARIC participants whose GCF samples are available into the seven PPC-Stages using the current PPC-Stages algorithm to dichotomized tooth-level clinical measurements and indices that we reported previously(3). We will then randomly select 55 subjects from each of the seven PPC-Stages and perform the Lumnex multiplex immunoassays (Bio-Rad) to determine the levels of 18 mediators in the 385 GCF samples. Before GCF sample selection, we will first exclude 1) participants who were smokers at the time of enrollment; and 2) participants who were diabetic at Visit 4, since both conditions are risk factors for periodontitis. The diabetic condition is defined by self-reported physician diagnosis, self-reported diabetes medication use, fasting glucose of 126 mg/dL or higher, non-fasting glucose of 200 mg/dL or higher, or 2-hour glucose of 200 mg/dL or higher following the oral glucose tolerance test. The multiplex immunoassay, which is more efficient than ELISA, allows simultaneous detection of up to approximately 40 analytes in the same sample. Those 22 selected mediators play key roles in inflammatory (innate) and adaptive immunity pathways encompassing a diverse range of immune cell activities involved in periodontal disease. This GCF biomarker panel includes: 1) chemokines CXCL1 and CXCL5 for recruiting neutrophils, CCL2 for monocytes and dendritic cells, CCL4 for monocytes, and CCL5 for T lymphocytes; 2) pro-inflammatory cytokinesIL-1α, IL-1β, IL-6, IL-17, IFN-β, and TNF-α; 3) anti-inflammatory cytokines: IL-10 and IL-1RA; 3) lympokines IFN-γ, IL-4, granulocyte-colony stimulating factor (CSF3), granulocyte-macrophage-stimulating factor (CSF2) and TGF-β; 4) growth factor vascular endothelial growth factor (VEGF); 5) inflammatory metabolite: prostaglandin E2; 6) bone resorption maker RANKL and bone formation marker osteoactivin(GPNMB).

We propose to use **two approaches** to identify signature mediators that are specific for each PPC Stage. First, we will apply general linear model (GLM) to the mediator data to identify significant mediators ($p<0.05$ adjusting for multiple comparisons) that are unique to each or certain PPC-stages. This is a statistical approach to identify mediators significantly associated with PPC-stages especially key stages of PPC-IV and PPC-VII.
Sample size estimation: Based on our preliminary data analysis of IL-6 and CCL2/MCP-1 on GCF samples, we determined that to observe 0.12 log-transformed difference with a variance of 0.04 between groups, 55 GCF samples are required to achieve a power of 80% ($\beta=0.80$) with a 95% confidence level ($\alpha=0.05$).

Statistical analysis: the quantified multiplex-measured mediator levels will be first log10 transformed to improve normality distribution. This data transformation has been routinely performed in similar studies. We will apply GLM to the mediator data adjusting for age, gender, field center/race, and batch effect for multiplex assays. The statistical significance threshold

(p=0.05) will be adjusted for multiplex comparisons using the Benjamini and Hochberg method to avoid false positive (Type I) error.

Second, we will apply feature selection methods to the GCF mediator data to identify informative mediators that can represent or predict each or certain PPC-stages. This is a machine learning approach to predict mediators that can be used as a surrogate for each PPC-Stage.

Data analysis: Feature (GCF mediator) selection techniques are widely used in gene expression data analysis to choose a subset of molecules unique to each phenotype, which are PPC-stages in this proposal. Using a computational framework, we will leverage different feature selection methods for each Classifier (prediction model) to first identify ranked candidate discriminative mediators (features), and then evaluate the performance of these mediators for predicting PPC-stages by calculating the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). Specifically, five base feature selection methods including Information Gain, Information Gain Ratio, Relief, Symmetrical Uncertainty, and Correlation, will be used for each classifier or classification model to select candidate informative mediators as feature sets. We will use a total of eight classifiers or classification models for this proposed study including Naïve Bayes, Support Vector Machine (SVM) linear, SVM Radial Basis Function (SVM-RBF), AdaBoost, Nearest Neighbor, Gaussian Process, Logistic Regression, and Random Forest. After selecting five feature sets, each by a single base feature selection method, different ensemble feature sets will be integrated by aggregating five base individual feature sets into a series of ensemble feature sets, including 1) union of all five base individual feature sets (Union or At Least 1), 2) feature set containing features that are shared by at least two of five individual base feature sets (At Least 2), 3) feature set containing features that are shared by at least three of five individual base feature sets (At Least 3), 4) feature set containing features that are shared by at least four of five individual base feature sets (At Least 4), and 5) feature set containing features that are shared by all five individual base feature sets (At Least 5). Finally, the ensemble feature sets will be evaluated using the performance of ensemble classification methods for predicting PPC stages measured by the area under the curve (AUC). The AUC will be calculated from receiver operating characteristic (ROC) curve, which is a plot of the true positive rate (TPR) against false positive rate (FPR). The larger the AUC (maximal value=1), the better the prediction. This proposed feature selection pipeline will be performed on Python and Scikit-learn (0.21.3) platform (13).

7.a. Will the data be used for non-CVD analysis in this manuscript? __X __ Yes __ __ No
***we realize diabetes is a major risk factor for CVD and some may consider it a CVD analysis.**

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 __X__ 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? ____ Yes _X___ 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"? ____ 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)?
There are many manuscript proposals that use dental variables as an exposure including but not limited to #2890, 2914, 2891,2874, 3194,3382.

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* __ 1996.01_)
___ 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.