ARIC Manuscript Proposal #3388

1.a. Full Title:
Proteomic Profiling of Gout Risk in ARIC

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

2. Writing Group:
Writing group members:
Adrienne Tin, Peggy Sekula, Bing Yu, Megan Grove, Jan Bressler, Christine Ladd-Acosta, Morgan Grams, Allan Gelber, Eric Boerwinkle, Joe Coresh and Anna Kottgen. Others are welcome.

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

First author: Adrienne Tin, PhD
Address: Department of Epidemiology
        Johns Hopkins Bloomberg School of Public Health
        615 N. Wolfe Street, W6021
        Baltimore, MD 21205
        Phone: 201-281-9577
        atin1@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: Anna Kottgen, MD, MPH
Address: Johns Hopkins Bloomberg School of Public Health and
        Medical Center - University of Freiburg
        Institute of Genetic Epidemiology
        Hugstetter Str. 49
        79106 Freiburg im Breisgau
        Germany

3. Timeline:
Analysis will begin immediately using visit 5 data and focusing on gout prediction in older adults. The manuscript is expected to be completed in 3-6 months.
When the proteomic data from visits 2 and 3 are available, we plan to amend this manuscript proposal to include visits 2 and 3 data.

4. Rationale:

Gout is the most common form of inflammatory arthritis and affects ~8 million U.S. adults.\(^1,2\) Emergency department visits for gout have been increasing and the annual gout hospitalization rate has doubled to 8.8 per 100,000 over the last two decades.\(^3,4\) Discovering novel risk factors and improving our understanding of mechanisms underlying gout pathogenesis may contribute to its prevention, treatment, and management.\(^5\)

Gout attacks result from an inflammatory response to the deposition of monosodium urate crystals in articular tissues, including the synovial lining, causing rapidly intensifying painful and swollen joint(s). Thus, hyperuricemia has long been recognized as a major risk factor for gout.\(^6\) However, most individuals with hyperuricemia do not develop gout.\(^1\) As such, what renders some hyperuricemic individuals susceptible to developing gout is a major unanswered question. The immune response is known to be involved in gout flares, including activation of the inflammasome.\(^7,8\) The circulating proteome may contain markers related to gout susceptibility, such as markers of the immune system, or markers that relate to protective factors against the development of gout. The ARIC study has ~5,000 plasma proteins measured using SOMAscan at visit 5. These protein measures are valuable resources for discovering novel risk factors for gout, improving our understanding of gout pathogenesis, and the construction of gout risk prediction models in older adults.

5. Main Hypothesis/Study Questions:

- Proteins in the circulating proteome will be associated with incident gout flare
- In particular, proteins related to the immune response will be associated with incident gout flare and meaningfully improve the prediction of gout.

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:** Longitudinal prospective study

**Inclusion criteria:** Participants with SOMAscan, outcome, and covariate data.

**Exclusion criteria:** Participants who only consented for CVD research.

**Outcome:**

**Visit 5 as baseline:** incident gout flare will be defined by ICD codes (ICD 9: 274.x, ICD 10: M10.x) obtained from ARIC cohort surveillance data. The primary analysis will include participants with prevalent gout. Given suboptimal treatment due to patient compliance, inadequate drug dosing, or lack of the practice of treat to target urate levels, initial incident and recurrent gout episodes likely share common factors.\(^9,10\)
Exposure:
Human plasma protein concentrations measured by SOMAscan assay (Somalogic Inc., Boulder, CO). We will investigate an appropriate exclusion threshold for proteins that were FLAGGED in at least one plate, and the appropriate transformation of the relative florescence units (RFU) of protein levels, such as log2 and/or inverse normal transformation.

Covariates:
Known risk or protective factors of gout that are available in ARIC: age, sex, race-center, serum/plasma levels of uric acid, CRP, triglycerides, body mass index (BMI), current drinking, systolic blood pressure (SBP), diabetes, diuretic use, cholesterol lowering medication, alcohol consumption, estimated glomerular filtration rate (eGFR) from both cystatin C and serum creatinine.

Statistical data analysis:
The development of the prediction models will consist of the following steps:

1) Use cross validation to screen for model predictors and estimate out-of-sample prediction error. The participants will be divided into 10 folds. Nine folds, the training set, will be used for variable screening, and the left-over fold, the test set, will be used for prediction. For each of the 10 training set-test set combinations, we will conduct the following analyses:
   A) Select significant covariates in the training set using Cox regression, e.g. using backward selection based on AIC with p< 0.05.
   B) Assess protein independent association in the training set. The independent association of each protein with incident gout will be assessed using Cox regression controlling for significant covariates in step 1A.
   C) Assess protein association stability in the training set. To select a subset of proteins that are likely robust predictors of gout, we will use the significant covariates in step 1A and the top 50 proteins with the most significant independent association in step 1B as input into elastic net with Cox regression incorporating an inner cross validation and the one standard error rule.
   D) Assessment of prediction accuracy in the test set. Predictors with non-zero coefficient from step 1C will be used to predict gout in the test set. The average of the prediction accuracy in all test sets is an estimate of the out-of-sample prediction accuracy.

2) Model development. With the predictors most often selected in cross validation in step 1C, we will construct prediction model(s) and assess discrimination and calibration. Several considerations will be investigated at this stage:
   A) Survival analysis methods. We will start with Cox regression. If the proportional hazards assumption is found to be violated by testing of time interaction terms or the model is not well calibrated, we will investigate the use of other survival analysis methods, such as Weibull and generalized gamma regression.
   B) Statistical significance threshold for predictors. Elastic net with the one standard error rule uses a penalized method to shrink the coefficients of some variables to 0. While the variables with non-zero coefficients may add value to prediction in the ARIC
population, some may have association p-values >0.05, the traditional significance threshold, in a survival model, and these predictors may not be generalizable. We will determine the significance threshold for predictors based on model discrimination and calibration statistics.

**Model performance metrics.** We will use measures of calibration and discrimination to evaluate all risk prediction models. We will evaluate calibration by comparing predicted levels of risk against observed events using a modified Hosmer-Lemeshow $\chi^2$ statistic. We will assess discrimination by computing Harrell’s C statistic, which allows for censoring.\textsuperscript{14} We will calculate the differences, standard errors, and $z$-statistics.\textsuperscript{15} We will further calculate the Brier score as a measure of overall model performance,\textsuperscript{16} the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) for model comparison.\textsuperscript{14, 17, 18} If the calibration is suboptimal, we will investigate whether the baseline hazard estimate is the source of miscalibration.\textsuperscript{19} When discrimination is suboptimal, we will evaluate the impact of each predictor by varying its coefficient.

**External validation of gout prediction models.** We will contact studies with SOMAscan data, e.g. HUNT3 and MESA, on the possibility of validating the prediction accuracy and calibration of the gout prediction models.

**Validation of protein identity.** We will take advantage of the protein identity validation offered by SomaLogic to verify the binding specificity of the aptamers of the proteins used in the final gout prediction model(s).

**Secondary analyses.**
A) To increase our understanding of the protein association with gout, we assess independent association of each protein in the whole cohort, i.e. conduct step 1A and 1B in the whole cohort instead of in 9 folds of the data and use these results to conduct pathway analysis.
B) To inform the causal relation between associated proteins and gout, we will use genetic variants associated with the proteins to perform Mendelian randomization analysis. The genetics of the proteome is covered in another manuscript proposal.
C) We will evaluate model performance excluding participants with prevalent gout and among participants with high urate levels (>6.8 or 7.0 mg/dL) only.\textsuperscript{20}

7.a. Will the data be used for non-CVD analysis in this manuscript? \textbf{X} Yes \hspace{1cm} \textbf{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? \textbf{X} Yes \hspace{1cm} \textbf{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, this proposal does not include genetic analysis. If we decide to include proteomic genetic findings in the manuscript, we will submit an amendment.

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/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)?

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 https://www2.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:


