ARIC Manuscript Proposal #2862

PC Reviewed: 10/11/11 **Status: _____** **Priority: 2**
SC Reviewed: _________ **Status: _____** **Priority: ____**

1.a. Full Title: Harmonization of respiratory outcomes data from nine US population-based cohorts: the NHLBI Pooled Cohorts Study

b. Abbreviated Title (Length 26 characters): Harmonization of Pooled Cohorts

2. Writing Group:

First author: Elizabeth C Oelsner, Columbia University, New York, NY, USA, eco7@cumc.columbia.edu
Senior author: R Graham Barr, Columbia University, New York, NY, USA, rgb9@cumc.columbia.edu
Co-authors:
Pallavi Balte, Columbia University, New York, NY, USA, ppb2119@cumc.columbia.edu
Pat Cassano, Cornell University, Ithaca, NY, USA, pac6@cornell.edu
Paul Enright, University of Arizona, Tuscon, AZ, USA, lungguy@gmail.com
Aaron Folsom, University of Minnesota, Minneapolis, MN, USA, folso001@umn.edu
John Hankinson, Hankinson Consulting, Inc., Athens, GA, USA, john@occspiro.com
David Jacobs, University of Minnesota, Minneapolis, MN, USA, jacob004@umn.edu
Ravi Kalhan, Northwestern University, Chicago, IL, USA, RKalhan@nm.org
Robert Kaplan, Albert Einstein College of Medicine, New York, NY, USA, Robert.kaplan@einstein.yu.edu
Richard Kronmal, University of Washington, Seattle, WA, USA, kronmal@u.washington.edu
Leslie Lange, University of Colorado, Denver, CO, USA, leslie.lange@ucdenver.edu
Laura Loehr, University of North Carolina, Chapel Hill, NC, USA, lloehr@email.unc.edu
Stephanie London, NIH/NIEHS, Research Triangle park, NC, USA, london2@niehs.nih.gov
Ana Navas Acien, Columbia University, New York, NY, USA, an2737@cumc.columbia.edu
Anne Newman, University of Pittsburgh, Pittsburgh, PA, USA, NewmanA@edc.pitt.edu
George O'Connor, Boston University, Boston, MA, USA, goconnor@bu.edu
Joseph Schwartz, Columbia University, New York, NY, 10032, jes2226@cumc.columbia.edu
Lewis Smith, Northwestern University, Chicago, IL, USA, ljsmith@northwestern.edu
Wendy White, Tougaloo College, Tougaloo, MS, USA, wendywhite2001@yahoo.com
Sachin Yende, University of Pittsburgh, Pittsburgh, PA, USA, Yendes@upmc.edu

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

First author: **Elizabeth C Oelsner**
Address: 622 West 168th Street
Division of General Medicine, Columbia University Medical Center
622 West 168th Street, PH9E-105
New York, NY, 10032

Phone: 917-880-7099 Fax: 212-342-0560
E-mail: eco7@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: **David Couper**
Address: 137 East Franklin Street
Suite 203
CB #8030
Chapel Hill, NC 27599

Phone: (919) 962-3229 Fax:
E-mail: david_couper@unc.edu

3. Timeline: Data from all cohorts with the exception of HCHS/SOL, JHS, and SHS have already been obtained, harmonized, and pooled under the auspices of approved paper/ancillary study proposals relating to specific biological hypotheses that are to be tested in this data. We hope to submit an abstract on our methods for harmonization and pooling for the 2017 American Thoracic Society Conference (abstract deadline November 2, 2016) and to prepare the relevant manuscript in spring 2017.

4. Rationale:

Chronic obstructive pulmonary disease (COPD) is the third leading cause of death worldwide (1). COPD is currently defined functionally by airflow limitation on spirometry that does not fully reverse with bronchodilators, together with respiratory symptoms (2). Smoking, occupational and environmental exposures, gene variants, and early life factors have been identified as risk factors for COPD, yet considerable variability in COPD incidence and prognosis remains unexplained (3). An accelerated rate of decline in lung function, which may be caused by smoking, may lead to the development of COPD (4); however, there is increasing recognition that low lung function in early adulthood may be another trajectory leading to COPD (5-7). This has encouraged reconsideration of classical paradigms in order to identify novel risk factors as well as strategies for primary prevention – which, beyond smoking cessation and avoidance, are currently lacking (8).

To advance research on the determinants and prognostic implications of lung function and lung function trajectories in adulthood, well-characterized population-based epidemiologic cohorts are needed. Unfortunately, the utility of the National Health and Nutrition Examination Survey (NHANES) and administrative datasets for this purpose is limited by lack of reliable measurements of major risk factors (e.g., pack-years, lung function) in these data (9, 10). Although national registries in Europe and Asia provide attractive resources (11, 12), the generalizability of findings from these cohorts to the multiethnic, increasingly non-smoking, and increasingly overweight US population is questionable, especially in light of the racial and ethnic disparities in COPD and asthma morbidity (13-15).

The NHLBI Pooled Cohorts Study (NIH/NHLBI R21-HL121457, R21-HL129924, K23-HL130627) intends to harmonize data on respiratory outcomes from NIH-funded cohorts conducted over the past forty years. While the importance of data harmonization is drawing increasing attention from the research community (16-19) – driven, at least in part, by the growing availability of heterogeneous "big data" – a current search of pubmed.gov for articles on "harmonization AND spirometry" yields zero records. This is despite the fact that standardization of spirometry measures, which are highly effort-dependent, has been the subject of considerable attention from the clinical community, resulting in a series of evolving guidelines in recent decades (20-23).

We therefore propose a "methods paper" to describe our approach in the NHLBI Pooled Cohorts Study to applying contemporary spirometry standards across nine US epidemiologic cohorts so as to ensure valid harmonization of these data, which were acquired using differing technologies and according to varying standards. We will characterize the impact of applying these standards on sample size, inter- and intra-subject variability in lung function, inter- and intra-cohort differences in age-adjusted lung function, and quantity of empirically and clinically defined "outliers." We will furthermore describe our approach to defining clinical endpoints including respiratory symptoms and severe obstructive lung disease events (SOLE).

5. Main Hypothesis/Study Questions:

1. Contemporary spirometry standards can be applied consistently to spirometry data obtained from NIH-funded epidemiologic cohorts, and exclusion of spirometry observations that do not meet contemporary spirometry standards will have the following impact on the NHLBI Pooled Cohorts sample:
    a. Decreased sample size, although the total number of participants and observations will remain the largest currently available US population-based sample with spirometry as well as reliable measures of smoking

history, anthropometrics, socio-demographics, biomarkers, and respiratory events follow-up.
    b. Decreased inter- and intra-subject variability in lung function.
    c. Decreased inter- and intra-cohort variability in age-adjusted lung function, controlling for the known effects of sex, race/ethnicity, and smoking status.
    d. Fewer "outliers," defined by statistical criteria (> 2 standard deviations from the mean) and also *a priori* as lung function measures that are substantially (> 15%) lower than one or more follow up measures, therefore diverging from the classical paradigm of gradual age-related lung function decline, and potentially reflecting falsely low lung function attributable to poor effort or other factors.
2. Measurements of respiratory symptoms ascertained via questionnaires can be harmonized across cohorts, and missing data may be imputed using statistical methods.
3. Severe obstructive lung events (SOLE) – defined as hospitalizations/deaths with COPD, chronic bronchitis, emphysema, or asthma as the primary discharge diagnosis code or underlying cause of death (24) – can be identified in those cohorts with events follow-up that includes ascertainment of discharge diagnosis and death certificate data.

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 harmonize and pool pre-bronchodilator spirometry measurements performed over the past four decades, as well as identifying and harmonizing data on respiratory symptoms and SOLE, in nine NIH-funded United States epidemiologic cohort studies:

1. Atherosclerosis Risk in Communities (ARIC) Study
2. Coronary Artery Risk Development in Young Adults (CARDIA) Study
3. Cardiovascular Health Study (CHS)
4. Framingham Heart Study (FHS)
5. Health Aging and Body Composition (Health ABC) Study
6. Hispanic Community Health Study/Study of Latinos (HCHS/SOL)
7. Jackson Heart Study (JHS)
8. Multiethnic Study of Atherosclerosis (MESA)
9. Strong Heart Study (SHS)

We will use the following data (as available):

- All spirometry measures, including QC variables, from all available exams

- Symptoms: self-reported respiratory symptoms including dyspnea, wheeze, cough
- Events: occurrence and time-to-event from study baseline for hospitalizations/deaths with international classification of disease (ICD) codes for asthma, COPD, chronic bronchitis, emphysema
- Socio-demographics: age, sex, race/ethnicity, insurance status, socioeconomic status
- Anthropometric: height, weight, BMI, waist-to-hip ratio
- Smoking: smoking status, cigarettes per day, pack-years, pipe use
- Medical history: history of COPD, asthma, relevant medication use

Of note, data from all cohorts with the exception of HCHS/SOL, JHS, and SHS have already been obtained, harmonized, and pooled under the auspices of approved paper/ancillary study proposals relating to specific biological hypotheses that are to be tested in this data.

Analytic Plan

We will apply the following grading criteria, which incorporate acceptability according to American Thoracic Society (ATS)/European Respiratory Society (ERS) standards (23) as well as reproducibility (25):

- A: 3 or more acceptable curves, with the largest two values within 100mL
- B: 2 or more acceptable curves, with the largest two values within 150mL
- C: at least two acceptable curves, with the largest two values within 200mL
- D: at least two acceptable curves, with the largest two values within 250mL OR only one acceptable curve
- F: No acceptable curves

These criteria will be applied to pre-bronchodilator measures of:

- Forced expiratory volume in one second (FEV1)
- Forced vital capacity (FVC)

“Valid” spirometry will be defined as measures graded A, B, or C; “invalid” spirometry will be defined as measures graded D or F. The impact of defining grade D as “valid” will be tested in sensitivity analyses.

For exams/cohorts without the necessary data to determine grade according to the abovementioned criteria, an alternative grading rubric will be developed and described.

To determine the suitability of harmonizing across different grading schemes, differences in the distribution and variance of the valid versus invalid spirometry data using the alternative grading scheme will be compared to results using the standardized grading scheme in other exams/cohorts.

We will compare the sample excluding invalid spirometry ("valid spirometry") to entire sample ("any spirometry") on the basis of:

- Sample size; we will also construct "CONSORT" flow charts to visualize those included versus excluded based on spirometry validity versus loss-to-follow-up or other factors, and compare the socio-demographics of those with "valid" versus "invalid" spirometry
- Inter- and intra-subject variability in lung function. We will test the within and between subject variances and their ratio, the intra-class correlation, in mixed models including adjustment for age, sex, height, and race/ethnicity.
- Inter- and intra-cohort variability in lung function. We will test the within and between cohort variances and their ratio, the intra-class correlation, in mixed models including adjustment for age, sex, height, and race/ethnicity.
- Outliers
  - We will determine the number, proportion, and distribution of observations > 2 standard deviations from the mean.
  - We will determine the number, proportion, and distribution of observations that are substantially (> 15%) lower than one or more follow-up measures in the same participants.

Questionnaires regarding respiratory symptoms will be compared and evaluated qualitatively, and responses to identical or very similar questions will be cautiously combined. Prevalence and incidence of respiratory symptoms will be compared by cohort, and between-cohort differences in symptom profiles will be evaluated with respect to objective measurements of lung function in order to evaluate potential intra-cohort differences attributable to questionnaires, demographics, and/or temporal trends and birth cohort effects. The efficacy and validity of available options to account for missing data (e.g., inverse probability weighting, multiple imputation) will be explored.

In cohorts with the necessary data, SOLE, or hospitalizations and deaths attributable to chronic lower respiratory diseases, will be identified via ICD codes (ICD-9 490-493, 496, 506.4; ICD-10 J40-J45). Events will be sub-classified by code position (primary diagnosis code or underlying cause of death versus any code position) and sub-type (e.g., COPD versus bronchitis). Adjudicated exacerbations and deaths due to CLRD (available in HABC, MESA, and HCHS/SOL) will be evaluated as an alternate definition of SOLE. Of note, based on prior work in MESA and HCHS/SOL, in which adjudication compared favorably with ICD-based classifications (24), it is anticipated that ICD-based and adjudicated SOLE will be suitable for harmonization.

Statistical analyses will be performed in R or SAS, Version 9.3.

7.a. Will the data be used for non-CVD analysis in this manuscript? __x__ Yes ____ 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? __x__ Yes ____ 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)?

The authorship group for this proposal has several approved proposals that test non-overlapping, specific biological hypotheses in the harmonized and pooled data (AS 2013.04, 2014.41, 2016.09).

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? ____ Yes __x__ No

11.b. If yes, is the proposal
___ A. primarily the result of an ancillary study (list number* _________)
___ 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.