ARIC Manuscript Proposal # 3355

PC Reviewed: 2/12/19	**Status: _____**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Leveraging partial information across data sets with privacy restrictions: proof of concept for cognitive aging research

b. Abbreviated Title (Length 26 characters): **Pooling with restrictions**

2. Writing Group:

Writing group members: Teresa Filshtein (first), Melinda C. Power (senior), M. Maria Glymour, Sarah Ackley, Xiang Li

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

F**irst author**:	**Teresa Filshtein**
Address:	550 16th Street
	Mission Hall, 2nd Fl
	San Francisco, CA 94158
Phone:	860-803-4577
E-mail:	teresa.filshtein@ucsf.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).

Senior author:	**Melinda Power**
Address:	950 New Hampshire Avenue NW, 5th floor, Washington DC 20052
Phone:	202.994.7778
E-mail:	power@gwu.edu

3. Timeline:

Completion 1 year after approval.

4. Rationale:

Despite promising observational evidence used to motivate randomized control trials (RCTs) for Alzheimer's disease (AD) prevention, trial results continue to be null. Several possible explanations for this exist, e.g. publication bias (i.e. observational literature is

biased), suboptimal design in RCT (i.e. too small/short to detect effect, eligibility criteria too restrictive, etc.), bias in observational studies (i.e. residual confounding or selection bias). Data pooling and analysis of all available observational studies could bring us a step closer to bridging the gap between observational studies and RCTs. Unfortunately, privacy restrictions preclude full data pooling and many data sets do not include identical variables. This causes difficulties by restricting sample size (power) and the available covariate set. We propose a method to leverage partial information across data sets, combined with a pre-specified causal structure, that can make use of publicly reportable summary statistics to allow construction of simulated data that can be used in lieu of formal data pooling.

We use the Health and Retirement Study (HRS, N= 12,186, "unrestricted") and ARIC ("restricted"), to illustrate this approach for the question of the effect of Hemoglobin A1c (HbA1c) on memory loss in aging. We specify a presumed causal structure in both data sets, represented as a DAG. The structural causal model is then used alongside the unrestricted data set, as a set of simulation rules to generate synthetic data sets of our unrestricted and restricted data sets. Synthetic data sets are pooled and the causal model is estimated and compared to observed associations. Variables are intentionally left out and imputed to provide further evidence of reproducibility.

5. Main Hypothesis/Study Questions:

This is a proof of concept paper. Our goal is to prove that this method of data pooling is valid and an effective way to leverage information from partial 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).

Inclusion/Exclusion:

HRS: Individuals 50+ at biomarker (HbA1c) collection comprised the initial sample. Of this initial sample, individuals without memory assessment at biomarker collection were excluded as well as those missing key covariates.

ARIC: Individuals 50+ at a biomarker (HbA1c) collection comprised the initial sample. Of this initial sample, individuals without memory assessment at biomarker collection were excluded as well as those missing key covariates.

Independent variables:

- Glycosylated hemoglobin (HbA1c) measured in either 2006 or 2008 in HRS, and at Visit 2 in ARIC.
- Age at observation

Hypothesized confounders

- Birth Cohort
- Childhood socioeconomic status (continuous composite score based on measures of human and financial capital, HRS only)
- Education (less than high school, high school, college)
- Gender
- Race (white, black, other)
- Diet (Western and prudent diet z-scores, ARIC only)

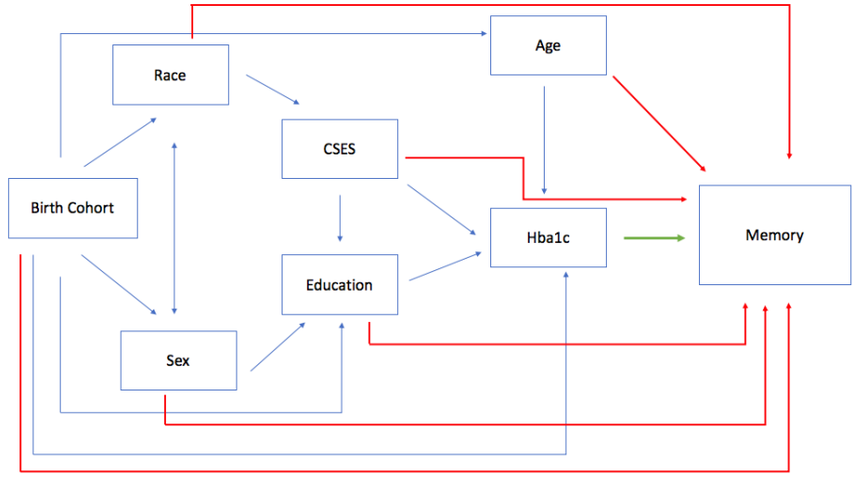
Dependent variables:

- Memory Function:
    - HRS: previously developed composite memory score combining proxy and direct memory assessments for longitudinal analyses.
    - ARIC: Z-score DWRT measures

Statistical Analyses:

We propose a method to leverage partial information across data sets, combined with a pre-specified causal structure, that can make use of publicly reportable summary statistics to reconstruct data not available. We use the Health and Retirement Study and ARIC to illustrate this approach for the question of the effect of Hemoglobin A1c (HbA1c) on memory loss in aging. We specify a presumed causal structure in both data sets, represented as a DAG (Figure 1).

Figure 1. Presumed Causal Structure - Effect of Hba1c on Memory. Red arrows represent the direct effect of each variable on the outcome, green arrow represents the main association of interest.



The structural causal model is then used alongside the unrestricted data set, as a set of simulation rules to generate synthetic data sets of our unrestricted and restricted data sets. Synthetic data sets are pooled and the causal model is estimated overall and by synthetic cohort and compared to observed associations in each original cohort. Variables not included in both datasets are intentionally part of the DAG and imputed to provide further evidence of reproducibility.

Other considerations:

All individual-level HRS analyses and construction of the simulated data from summary statistics will be conducted at UCSF. All individual-level ARIC analyses will be conducted at George Washington University, which has a DMDA allowing access to ARIC data.

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

#2630 - Hypoglycemia and cognitive function in older adults with diabetes

#1871 - Type 2 diabetes and cognitive decline over 14 years, accounting for mortality

#2160 - Diabetes and cognitive change over 20 years: the Atherosclerosis Risk in Communities Study

#2094 - Cardiovascular Risk Factor Control in Diabetes Pooling Project

#1904 - Cardiovascular Disease Risk Prediction in Combined Cohort Studies

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 : 2017.01)

___ 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.