ARIC Manuscript Proposal #2626

PC Reviewed: 9/8/15	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Evaluation of Selection Effect and Participants' Survival Advantage in Large Cohort Studies

b. Abbreviated Title (Length 26 characters): Selection effect & Survival Advantages

2. Writing Group:
Writing group members:

Zihe (Emma) Zheng
Kunihiro Matsushita
Judith Hoffman-Bolton
Lisa Wruck
Elizabeth Selvin
Casey Rebholz
Josef Coresh

Others are welcomed.

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

First author: **Zihe (Emma) Zheng**
Address: 2024 E. Monument Street B-314, Baltimore, MD 21205

Phone: 443-562-8918 Fax:
E-mail: zzheng11@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: **Josef Coresh**
Address: 2024 E. Monument, Suite 2-630, Baltimore, MD 21287

Phone: 410-955-0495 Fax: 410-955-0476
E-mail: coresh@jhu.edu

3. Timeline:

August 1st 2015-March 1st 2016

4. Rationale:

In epidemiological studies, study population is somewhat conditioned on selected participants or respondents and thus may not completely represent target population. This natural selection process can bias our estimate of interest. The consequence of selection bias can vary according to different study designs. For prevalence estimates, selection would be more related to external validity (referred to as "selection effect"). For association measures, selection would impact internal validity (referred as "selection bias" [1]). Usually, epidemiological studies don't specifically differentiate these two terms and use "selection bias" to refer both. However, we think it is important to distinguish these two levels of selection impact with different terminologies. We are going to mainly focused on the impact of selection effect and its implication of external validity in long-lasting cohort studies.

Selection literally can happen at all stages of a cohort study, including design stage (defining selection criteria), recruiting stage, examination stage, and follow-up stage. Selection is usually regarded as a fixed component in longitudinal studies and beyond the capability of statistical adjustment. However, we suspect that the effect of selection changes over follow-up time in a cohort towards the less biased direction, which leads to a better generalizability of conclusions at later times compared to the beginning of the cohort.

The most common selection biases in cohort studies are healthy volunteer bias, survival bias, non-response bias, and loss to follow-up bias [2]. To our knowledge, the change of these biases due to selection effect over follow-up time hasn't been well studied. It is important to understand the dynamics of study population and be able to evaluate how much confidence we have for generalizing the conclusion at different times. Given the quantitative knowledge of corresponding selection effect we would get from this study, we will be able to assess the change of external validity over follow-up time.

Comparing difference of baseline and demographic characteristics and risk factor profiles that result from non-participation or non-response have been done [3-4, 8]. The ARIC and CLUE cohort profiles of over- or under-representation of specific types of population had been published in previous study[3], but the quantitative survival profiles of participants and non-participants over more than two decades of follow-up time haven't been reported.

Among multiple cohort studies with long-term follow-up, a single measure of survival advantage measured by all cause mortality has been observed among participants compared to non-participants, which is partially attributed to selection effect [8–10]. We found there is higher total mortality among cohort respondents compared to non-respondents reported in one study[11].

In general, selection and selection bias have not been well studied due to the lack of data on non-participants, non-respondents or people who are lost to follow-up. However, our study has the information of baseline population demographics and vital statistic for both participants and non-participants with the unique linkage among Washington County private census data, Atherosclerosis Risk in Communities (ARIC) study data and Campaign Against Cancer and Heart Disease (CLUE I & II) studies data. We will be the first one to quantify the selection effect in multiple cohorts, and track its change and dynamic impacts on these cohorts' external validity.

5. Main Hypothesis/Study Questions:

Main hypotheses:

1) Volunteering for a cohort study is associated with a survival advantage. Participants have lower mortality than the general population in the community.
2) This quantifiable advantage will have same direction across two different cohorts.
3) The participation survival advantage diminishes over time due to natural dilution of cohort selection effect, but less rapidly for those who attended follow-up visits or returned questionnaires. The study participants will form a less biased cohort compared to what they were at the beginning.

The aim of our study is to quantify the cohort selection effect through the following measures:

1) Comparing the prevalence estimate of demographic factors and health risk factors between cohort participants and census population and its change over time

2) Identifying determinants of different subtypes of participation (initial, complete, partial and across-cohort participation; visit(s) examination, follow-up questionnaires and annual telephone interview participation) with census data

3) Quantifying potential survival advantage due to selection effect and its change over follow-up time

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 methodological limitations or challenges if present).

Data source

This study will use data from three cohorts (ARIC; CLUE I; and CLUE II), participants of which were recruited in Washington County, MD. Washington County Private Census data in 1975 is used as the reference/general population in our study for analyzing selection effect in large cohort studies. The linkage between cohort and census data constructs the basic framework of identifying participants, non-participants and general population (Figure.1).

For ARIC cohort in Washington County, participants aged 45-64 were recruited for visit examination after enumeration and at-home interviews between 1987-1989 (first visit). Among the 6177 potential eligible individuals, 4023 completed both the home interview and visit examination, which gave a response rate of 65% at baseline visit [3]. Three visits of re-examination happened every three years after first visit in 1990-1992, 1993-1995 and 1996-1998 and the fifth visit occurred recently from 2011-2013. To maintain contact and gather health information, annual follow-up of the cohort had been conducted since 1987 through telephone interviews, which changed to semi-annual calls starting in 2012. More details about study design and implementation have been published elsewhere [12].

In the CLUE I study conducted in 1974, 23,950 Washington County residents, aged 12-97, were recruited with mobile office trailers went through the study area. Brief health history questionnaire and blood pressures were taken and 15ml of blood was drawn at the time of enrollment. Almost a third of the adult population of the county participated in CLUE I. There was no follow-up for this population.

In the CLUE II study conducted in 1989, 25,076 Washington County residents, aged 2-98, were recruited with the same approach used in CLUE I study. Participants were asked to complete a brief health history questionnaire, and donate 20ml blood sample. A food frequency questionnaire was given to participants to complete at home with a request for including a toenail clipping when returning the questionnaire. Five follow-ups with mailed questionnaires were conducted after baseline visit: 1996 (Response Rate = 70%), 1998 (64%), 2000 (64%), 2003 (62%) and 2007 (56%). The last questionnaire was only sent to participants who responded to at least 1 previous questionnaire. Approximately 30 percent of the adult residents participated in CLUE II, with 8400 Washington County residents participating in both studies.

To identify participants and non-participants populations, study participants in ARIC and CLUE have been linked to 1975 Washington County Private Census data by first, last and maiden name and birthdate. (The number of linked participants for ARIC is 3008, for CLUE I is 19,932 and for CLUE II is 16,528). Linkage of death and causes of death were obtained from Washington County Health Department, State of Maryland Vital Statistics and National Death Index from 1975 to April 2015. Obituaries are checked daily from the local newspaper. (The data of CLUE participants is also linked to Maryland Cancer Registry (MCR). Six past linkages were done in 1996, 1998, 2005, 2007, 2011 and 2013.)

Analysis Plan

Inclusions/Exclusions:

Linkage will be established by identifying cohort participants (ARIC and CLUE) in the whole census population. People are removed from analysis if they have missing data in any one of the linkage variables, last name, first name, or date of birth. Participants of ARIC or CLUE cohort that were not included in census were mainly non-Washington County residents thus are excluded from the analysis.

All the comparison between participants, non-participants and census (general) population will be matched on age and other main cohort characteristics, considering the different eligibility criteria for each cohort.

Variables of Interest

The prevalence estimates of 1975 census demographic factors and health risk factors are compared between dynamic cohort participants, non-participants and census population alive at each time point of recruitment and follow-up visits for ARIC and CLUE studies respectively. These census variables are sex, race, education, marital status, years of residence in the county, smoking history, degree of disability, employment status, history of cancer, possession of a driver's license, and enumeration district. At each time point of interest, participants are defined as people who have date of visit data in the dataset; non-participants are defined as census population who don't have date of visit data and are still alive at that point of time; general population will be the sum of the two.

Death ascertainment will be based on census data and its linkage with multiple sources of vital statistics. Comparison between census and ARIC/CLUE death ascertainment will be conducted to evaluate the validation and potential bias of census death ascertainment.

For continuous variables, mean value and standard deviation will be calculated. Unpaired student t-test will be used for the comparison between participants and non-participants. For categorical variables, difference between participants and non-participants will be tested using chi-square. Relative volunteer bias in descriptive statistics will be calculated for each census variable as the difference in mean for continuous variables and proportion for categorical variables between participants and non participants divided by the relevant value of general population multiplied by 100 [3].

Outcome of Interest: Participation determinants

Logistic regression will be used to identify the major determinants of cohort participation at recruitment. In addition, participation is defined as four subtypes: 1) Initial participation: participants attended the baseline visit; 2) Complete participation: participants completed all the follow-up visits (ARIC: participants have visit date data for all visit 1 to visit 5 and records of annual telephone interviews; CLUE II: participants attended the baseline screen and returned all five follow-up questionnaires); 3) Partial participation: the difference in participants between initial participation and complete participation; 4) cross-cohort participation: participants attended both the ARIC and CLUE studies.

Within ARIC cohort, degree of participation will be further classified based on the participation of either annual telephone follow-up interview or five clinical examinations or both.

Outcome of Interest: Survival advantages

Comparing the all-cause mortality and disease specific mortality (CVD and cancer) between participants, non-participants and general population for each study. Cox regression model will be used to calculate the hazard of death. Adjusted Kaplan-Maier

will be used for graphical survival comparison [13]. The analysis will first stratify by study cohorts then by participation subtypes to compare the difference of survival among people with different level of participation and the change of selection effect over follow-up time. Interaction of participation type/degree with follow-up time (the length of follow-up time at each point of visit examinations, questionnaires or telephone interviews) will be used to examine the change of survival advantage over follow-up time. The main predictor variables will be drawn from the census since it has the most data on non-participants.

Limitations and Challenges

1) Linkage establishment: we adapted a relative stringent criteria for linkage establishment between two cohort studies and census data, which could lead to a lose of participants when we try to identify them in census population if any one of their 3 linkage variables (first name, last name, or date of birth) is missing in the census data.
2) Death ascertainment difference: either ARIC or CLUE cohort has better death ascertainment quality than the census vital statistics, which could lead to potential bias of the survival advantage estimation. This death ascertainment difference is inevitable due to our study design, but we will exam and quantify the magnitude of this difference by comparing vital statistics for those who have more than one data source.

7.a. Will the data be used for non-CVD analysis in this manuscript? __✔__ 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? __✔__ 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 __✔__ 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
___✔___ 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)?

ARIC Manuscript Proposal #2041: The effect of selection bias on the relationship between cardiovascular risk factors and mortality

Lisa Wruck, the senior author of the related manuscript proposal, is also a co-author of this manuscript. Their work focused on quantifying and correcting the influence of selection bias on association estimate, which is quite different from this work. Also we have more data source from multiple cohorts and census data instead of solely using ARIC data.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? ____ Yes ____ 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)* _____ _____ _____)

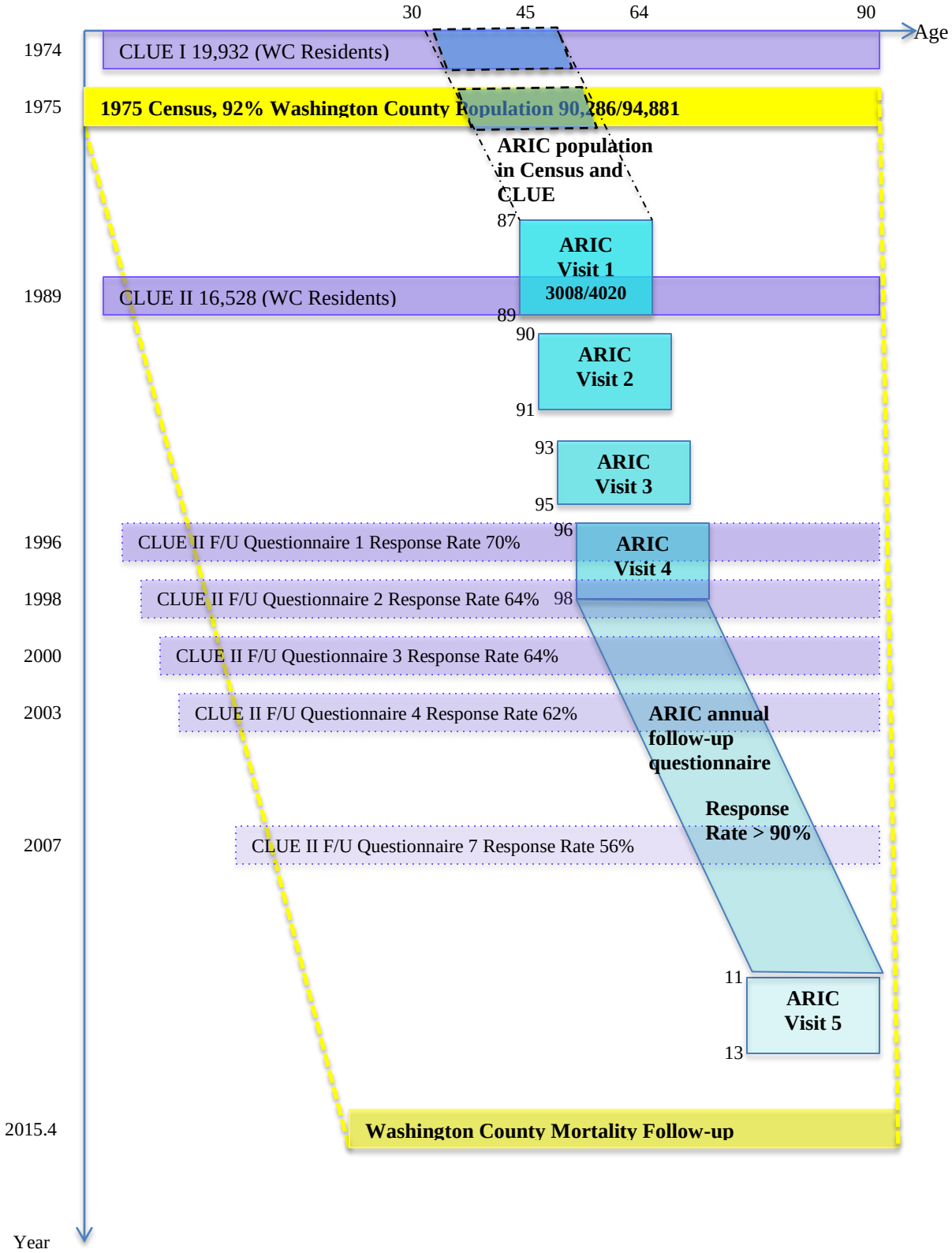
*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

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.

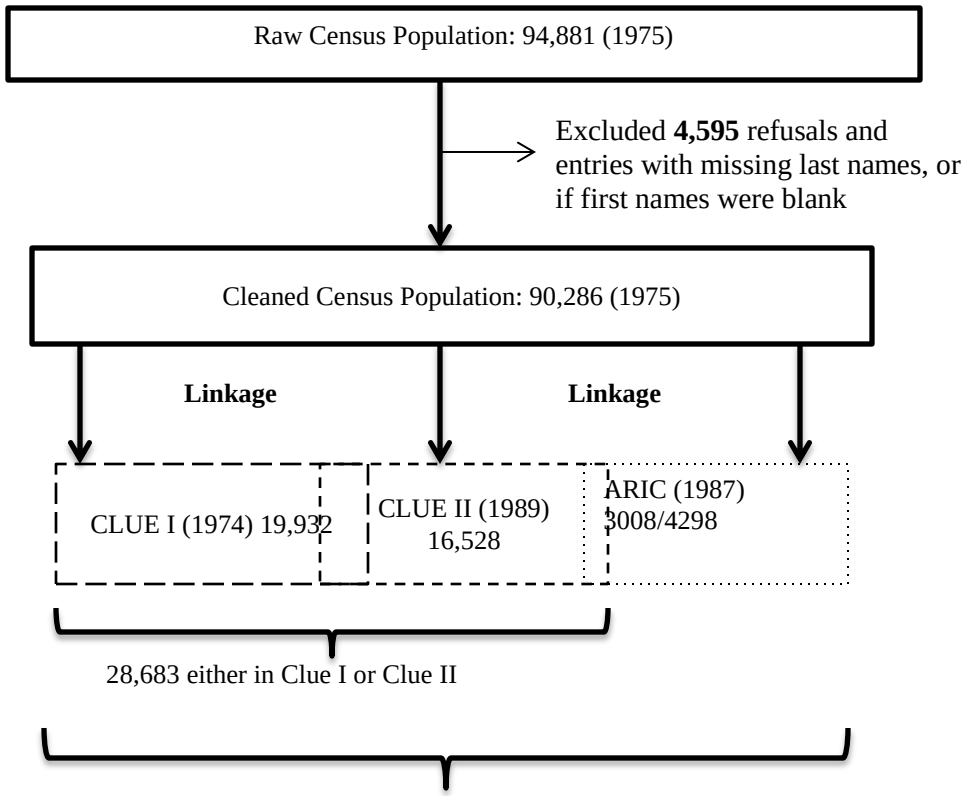
12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PUBMED Central** whenever the journal does not and be in compliance with this policy. Four files 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.

13. Per Data Use Agreement Addendum for the Use of Linked ARIC CMS Data, approved manuscripts using linked ARIC 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 ____ No.

Figure 1. Flow Chart of Cohort Structure



Inclusion criteria



Reference

1. Hernán M a, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615-625. doi:10.1097/01.ede.0000135174.63482.43.

2. Massad E, Ortega NRS, de Barros LC, Struchiner CJ. Modern epidemiology. *Stud Fuzziness Soft Comput*. 2008;232:41-57. doi:10.1007/978-3-540-69094-8_3.

3. Jackson R, Chambless LE, Yang K, et al. Differences between respondents and nonrespondents in a multicenter community-based study vary by gender and ethnicity. *J Clin Epidemiol*. 1996;49(12):1441-1446. doi:10.1016/0895-4356(95)00047-X.

4. Shahar E, Folsom AR, Jackson R. The effect of nonresponse on prevalence estimates for a referent population: Insights from a population-based cohort study. *Ann Epidemiol*. 1996;6(6):498-506. doi:10.1016/S1047-2797(96)00104-4.

5. Pizzi C, De Stavola B, Merletti F, et al. Sample selection and validity of exposure-disease association estimates in cohort studies. *J Epidemiol Community Health*. 2011;65(5):407-411. doi:10.1136/jech.2009.107185.

6. Pizzi C, De Stavola BL, Pearce N, et al. Selection bias and patterns of confounding in cohort studies: the case of the NINFEA web-based birth cohort. *J Epidemiol Community Heal*. 2012;66(11):976-981. doi:10.1136/jech-2011-200065.

7. Howe LD, Tilling K, Galobardes B, Lawlor D a. Loss to Follow-up in Cohort Studies. *Epidemiology*. 2013;24(1):1-9. doi:10.1097/EDE.0b013e31827623b1.

8. Van Loon a. JM, Tijhuis M, Picavet HSJ, Surtees PG, Ormel J. Survey non-response in the Netherlands: Effects on prevalence estimates and associations. *Ann Epidemiol*. 2003;13(2):105-110. doi:10.1016/S1047-2797(02)00257-0.

9. Heilbrun LK, Nomura a, Stemmermann GN. The effects of nonresponse in a prospective study of cancer. *Am J Epidemiol*. 1982;116(2):353-363.

10. Walker M, Shaper a G, Cook DG. Non-participation and mortality in a prospective study of cardiovascular disease. *Epidemiology*. 1987;(June 1986):295-299.

11. Barchielli A, Balzi D. Nine-year follow-up of a survey on smoking habits in Florence (Italy): higher mortality among non-responders. *Int J Epidemiol*. 2002;31(5):1038-1042. doi:10.1093/ije/31.5.1038.

12. Hill C, Gerardo D, James F, et al. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol*. 1989;129(4):687-702. http://www.ncbi.nlm.nih.gov/pubmed/2646917.

13. Nieto FJ, Coresh J. Adjusting survival curves for confounders: a review and a new method. *Am J Epidemiol*. 1996;143(10):1059-1068. doi:10.1093/oxfordjournals.aje.a008670.