ARIC Manuscript Proposal #1969

PC Reviewed: 7/10/12	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: More efficient estimators for generalized case-cohort studies

b. Abbreviated Title (Length 26 characters): **Generalized case-cohort studies**

2. Writing Group:

Writing group members: Soyoung Kim, Jianwen Cai, Donglin Zeng, David Couper, others welcome

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _SK_

F**irst author**: Soyoung Kim
Address: Department of Biostatistics, CB#7420
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420

Phone: 919-357-1203 Fax: 919-966-3804
E-mail: kimso@live.unc.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: Department of Biostatistics
University of North Carolina at Chapel Hill
137 E. Franklin St., Suite 203
Chapel Hill, NC 27599-7420

Phone: 919-962-3229 Fax: 919-962-3265
E-mail: david_couper@unc.edu

3. Timeline: Work can begin as soon as approval is received. Expect to be able to submit to a journal within 1 year.

4. Rationale:

The case-cohort design is originally developed for large cohort studies, where the cost to assemble the exposure and covariates is huge. Such a design is appealing especially when the event rate is low and the exposure measurements are expensive to obtain. The design effect needs to be taken into consideration when data from case-cohort studies are analyzed. A key advantage of the case-cohort study design is its capacity to use the same subcohort for several diseases or for several subtypes of disease. In recent years, in order to preserve the raw material collected in the study, the case-cohort study design has also been used in situations when the disease is not rare. In such studies, it is not desirable to conduct a traditional case-cohort study which collects the expansive covariate information on all cases. Sampling only a fraction of the cases is more practical. Recently, Kang and Cai (2009) proposed methods for fitting failure time data from generalized case-cohort studies with multiple disease outcomes under marginal proportional hazards models. In that study, information on expensive exposure measurement was available on the subcohort as well as a subset of failures or cases. However, they did not make full use of the available information on covariates in estimating the effect of a risk factor. For example, information about the second disease was not used when constructing estimating equations for the first disease in generalized case-cohort studies. In our research, we will propose an estimating procedure and study the properties of the proposed method by using all available information. As part of the research, we will also apply our method to a real case-cohort data set to illustrate its use. This is the reason we are requesting data from the ARIC study. These data will be published just as part of a methodological paper with acknowledgement of the ARIC study.

5. Main Hypothesis/Study Questions:

We want to use the proposed method to compare the effects of the biomarkers (checkpoint kinase 2 gene, KCNQ1 locus, DUSP9, and HNF1A) on type 2 diabetes and on incident coronary heart disease while adjusting for possible confounding factors using data collected in ARIC. These biomarkers are collected on the entire cohort of ARIC. We will artificially create two generalized case-cohort studies using the same subcohort. We will apply our proposed methods to study such effects using data from the generalized case-cohort study. We will compare the results to those based on the entire ARIC 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).

The study design is a generalized case-cohort study. Two outcomes (i.e. types of cases) will be considered – incident diabetes and incident CHD. Predictors of interest are various genetic loci (in/near *CHEK2*, *KCNQ1*, *DUSP9*, *HNF1A*) that have already been investigated for associations with diabetes. The second case group will be incident CHD cases. Other covariates that will be considered include age and gender. The generalized case-cohort sample will be selected by sampling from the ARIC cohort with the required genetic data.

A new statistical model is being developed by Ms Kim as part of her Biostatistics PhD dissertation. This new model will be used to analyze the data. Part of her model development includes using simulation studies to examine the properties of the model. Then the model will be applied using the ARIC data as an example of its potential use.

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 ____ No
(This file ICTDER03 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? __X__ 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"?
__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/search.php

___X___ Yes _____ No

Since this is a methodology paper and includes no novel risk factors, it does not overlap with ongoing research.

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 published manuscripts that have already reported on associations between the genetic markers to be considered here and diabetes are:

MS#1238: DNA-Damage Pathway and Genetic Susceptibility to Type 2 Diabetes Mellitus and Insulin Resistance States

MS1470B Genome-wide association study of prevalent type 2 diabetes in the Atherosclerosis Risk in Communities (ARIC) Study

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

___ A. primarily the result of an ancillary study (list number* _____)

_X__ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* ___2006.03__ _2006.12___ _____)

*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.