ARIC Manuscript Proposal # 1184

PC Reviewed: _08__/_15__/06	**Status: __A___**	**Priority: __2__**
SC Reviewed: _08__/_17__/06	**Status: __A__**	**Priority: __2__**

1.a. Full Title: Reanalysis and simulation of Lp-PLA_2/hs-CRP/CHD case-cohort study data (Ballantyne et al., Circulation 109:837-42, 2004) using all analysis data available for the potential full cohort

b. Abbreviated Title (Length 26 characters): Lp-PLA2/CRP/CHD reanalysis

2. Writing Group:
Writing group members:

Christie Ballantyne, Baylor College of Medicine
Norman Breslow, University of Washington, Seattle
Lloyd Chambless, University of North Carolina
Michal Kulich, Charles University, Prague
Thomas Lumley, University of Washington

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

First author: Norman Breslow
Address: Department of Biostatistics
Mailstop 357232
University of Washington
Seattle, WA 98195-7232
Phone: 206-543-2035 Fax: 206-616-2724
E-mail:

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):
Address:

Phone: Fax:
E-mail:

3. Timeline: One year from approval of proposal and receipt of analysis files for completion of work and submission of manuscript.

4. Rationale: ARIC is one of the few major studies to have made systematic use of the case-cohort design for investigation of risk factors for CHD and stroke. The typical study involves stratified random sampling of a cohort random sample (CRS), also known as a sub-cohort, to which cases of CHD or stroke not already sampled are added for detailed covariate assessment such as genotyping or other bioassay. The resulting data on cases and sub-cohort members are

then analyzed by Cox regression analysis, with weighting of observations to account for the stratified sampling. This conventional approach to the analysis ignores data available on a large number of control subjects not included in the CRS. Modern methods for the design and analysis of two-phase stratified case-cohort studies offer the potential to incorporate these additional data into the analysis and thereby increase the precision of the results. The main purpose of the proposed study is to advertise the availability of these methods, as implemented in the freely available R statistical package, to the general community of epidemiologists. By demonstrating the efficiency gains possible through illustrative analysis of data from a large and well known cohort study, others will be encouraged to avail themselves of these new and important statistical tools.

5. Main Hypothesis/Study Questions: The principal hypothesis is that the precision of hazard ratios (HRs) estimated for Lp-PLA_2 and hs-CRP, as reflected in the width of the corresponding confidence intervals, will be narrower in the reanalysis than they were in the original publication. By incorporating data on traditional risk factors available from visit two for the 12,819 subjects in the potential full cohort, results of the reanalysis should be closer to the results that would have been obtained had Lp-PLA_2 and hs-CRP measurements been made for all of them. Because of the known correlations among the traditional factors, Lp-PLA_2 and hs-CRP, this increase in precision should be found both for HRs that are adjusted for the traditional factors and for those that are adjusted only for the stratification variables (age, sex, race). This point will be demonstrated by conducting a parallel simulation study of case-cohort sampling that involves only the traditional risk factors available for everyone, treating one of the traditional risk factors (*e.g.*, triglycerides) as the additional covariate measured only in the case-cohort sample. Here the hypothesis will be that the new methods of case-cohort analysis have smaller mean squared errors of estimation of the CHD HR for high vs. low tertiles of triglycerides, in comparison with results based on the full cohort data, than do the conventional methods of analysis.

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).

This study will use analysis files prepared for the report by Ballantyne et al. (2004). This includes all information on traditional risk factors available for the 12,819 participants in the potential full cohort together with information on Lp-PLA_2 and hs-CRP measured from stored plasma for the CHD cases and non-cases in the CRS. Thus the main cohort will consist of apparently healthy middle-aged men and women in the ARIC study who had plasma samples stored from visit two. Subjects in the main cohort who have data missing on the traditional risk factors used in the analysis will be excluded, as will those in the case-cohort samples who have data missing either on traditional factors or on Lp-PLA_2 or hs-CRP. A single analysis file will be created in which missing value indicators will be inserted for Lp-PLA_2 and hs-CRP levels for subjects not samples as cases or in the CRS.

A simulation study will be conducted in which up to 1000 CRS will be obtained by independent stratified random sampling from the main cohort, such that the same numbers of subjects are sampled in each of the 8 sex/age/race strata as in the original study. One of the traditional risk factors, possibly plasma concentration of triglycerides, will be selected to play the role of the ``missing" covariate available only for the cases and the CRS. Each of these sampled datasets will be analyzed using the same statistical methods as will be employed on the single set of actual study data where the missing covariates are Lp-PLA_2 and hs-CRP.

A series of statistical analyses will be conducted for the original and each of the simulated datasets. The first analysis will be a conventional stratified case-cohort analysis (Borgan et al., 2000), using the available R software, in which only the 8 strata defined by sex, age and race are taken into account. This uses the standard Cox partial likelihood analysis, but weights the control observations in each stratum with their inverse sampling frequencies, i.e., by the ratio of the number of main cohort controls to CRS controls in that stratum. The point estimate is identical to that proposed by Binder (2000) and Lin (2000). However the variance estimate accounts for sampling variability in the main cohort, considered to be a random sample from a super-population, as well as between sub-cohort (CRS) and main cohort (Lin, 2000). This initial analysis is expected to yield HR estimates and confidence intervals extremely close if not identical to those reported by Ballantyne et al. (2004).

A second analysis will be undertaken using these same methods for stratified case-cohort studies, but incorporating finer "post-hoc" stratification in which the strata are defined not only by age, sex and race but also by low, medium and high risk levels for as many of the major traditional risk factors as can be incorporated without running into numerical problems with very small control stratum frequencies.

Once the limit has been reached in terms of the number and size of strata, we plan to use logistic regression modeling to incorporate additional information on the traditional risk factors to "predict" which participants were included in the CRS. Specifically, as outlined in Section 7.3 of the text by Therneau and Grambsch (2000), logistic regression models for the binary outcome "sampled for CRS" will be fitted to the main cohort controls using the 8 strata and traditional risk factors as covariates. The inverse predicted values from this logistic regression will be used as sampling weights in the Cox analysis to obtain HR estimates. Their variances will be obtained from a least squares linear regression of the Cox partial likelihood score contributions on the likelihood scores from the logistic model as described by Therneau and Grambsh (2000) based on earlier work by Robins, Rotnitzky and Zhao (1994). This is expected to increase precision of the HR estimates to the extent that the covariates used in the logistic regression are correlated with the "missing" observations on Lp-PLA_2 and hs-CRP. The log HRs for Lp-PLA_2 and their standard errors will be recorded for each such analysis and for each of the models of interest: with and without adjustment and with and without restriction to those with LDL-C below the median – see Table 4 of Ballantyne et al. (2004). A similar procedure will be followed for the simulated datasets.

All of the analyses will be implemented in the R statistical language. The R programs developed for this project will be made available to ARIC investigators upon conclusion of the project.

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 ICTDER02 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 ICTDER02 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 ICTDER02 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 ___x____ 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 only closely related ARIC proposals are ms889 (Ballantyne CM, Hoogeveen RC, Bang H, Coresh J, Folsom AR, Heiss G, Sharrett AR. Lipoprotein-associated phospholipase A2, high-sensitivity C-reactive protein and risk for incident coronary heart dise ase in middle-aged men and women in the Atherosclerosis Risk in Communities. Circulation 2004;109:837-42) and ms934 (Folsom AR, Chambless LE, Ballantyne CM, Coresh J, Heiss G, Wu K, Boerwinkle E, Mosley TH, Sorlie P, Diao G, Sharrett AR. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers. Arch Intern Med 2006;166:1368-73). There is no real conflict with these published papers since we plan to re-analyze the same data, supplemented by full cohort data, using different methods, to produce more precise estimators.

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. primiarly 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/

12. 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.

Reference List

1. Ballantyne CM, Hoogeveen RC, Bang HJ, et al. Lipoprotein-associated phospholipase A(2), high-sensitivity C-reactive protein, and risk for ischemic stroke in middle-aged men and women in the atherosclerosis risk in communities study. Circulation 2004;110:641.

2. Binder DA. Fitting Cox's proportional hazards model from survey data. Biometrika 1992;79:139-147.

3. Borgan O, Langholz B, Samuelsen SO, et al. Exposure stratified case-cohort designs. Lifetime Data Anal 2000;6:39-58.

4. Lin DY. On fitting Cox's proportional hazards models to survey data. Biometrika 2000;87:37-47.

5. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994;89:846-866.

6. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer-Verlag; 2000.