ARIC Manuscript Proposal #2362

PC Reviewed: 5/13/14  Status: A  Priority: 2
SC Reviewed: _________  Status: _____  Priority: ____

1.a. Full Title:

b. Abbreviated Title (Length 26 characters):
Prediction of CHD risk

2. Writing Group:

Jie Kate Hu, University of Washington, Seattle
Gary KC, Chan, University of Washington, Seattle
Norman Breslow, University of Washington, Seattle
David Couper, University of North Carolina, Chapel Hill

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

First author: Jie Kate Hu
Address: Department of Biostatistics
University of Washington
Seattle, WA 98195-7232

Phone: 206-910-3140  Fax: 206-616-2724
E-mail: jhu@uw.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 E Franklin St, Suite 203
Campus Box 8030
Chapel Hill, NC 27599-8030

Phone: 919-962-3229  Fax: 919-962-3265
E-mail: david_couper@unc.edu
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 cohort studies to make systematic use of stratified sampling (the Cohort Random Sample) to select participants for biomarker studies of the risk of CHD and stroke. Current biostatistical research is concerned with efficient ways of combining the marker data available for disease cases and the selected sample with information on standard risk factors that is available for a much larger number of subjects in the main cohort. Using ARIC data to help develop and illustrate the new statistical methods serves the dual purpose of demonstrating their utility, or lack thereof, when applied to important studies, and of providing new insights about disease risk.

Two statistical studies using data from Ballantyne et al. (2004) have been published already. Breslow et al. (2009) studied methods of fitting Cox proportional hazards models to evaluate the effect of lipoprotein-associated phospholipase A2 (Lp-PLA\textsubscript{2}) levels in combination with standard risk factors on CHD outcomes. Calibration of the “design” weights using information available for the entire cohort increased the efficiency of most regression coefficients, though not those for Lp-PLA\textsubscript{2} itself. Kang et al. (2013) demonstrated that increased levels of high-sensitivity C-reactive protein (hs-CRP) acted additively (added to), rather than proportionally, on the baseline risks of both CHD and stroke. They mentioned in the conclusion of their paper that efficiency gains might be possible using techniques similar to those developed by Breslow et al.

The current proposal is to develop a method for risk prediction of CHD based on an additive hazards model incorporating the hs-CRP data from the Ballantyne study. This will involve joint estimation of the coefficients of excess risk associated with each risk factor and of the baseline CHD risk curve, as a function of time on study. A method that predicts individual CHD risk at 5 and 10 years in the future, for example, could provide physicians a practical tool to explain clinical results to patients and to plan for prevention. The intent is to use not only data on hs-CRP available for the limited number of CHD cases and members of the Cohort Random Sample (CRS), but also to incorporate the information on standard risk factors available for a much larger number of subjects remaining in the main cohort. Breslow and Lumley (2013) demonstrated that such risk prediction based on the Cox proportional hazards model was feasible, using 10,000 simulated stratified samples from a cohort of participants in the National Wilms Tumor Study. Calibration of the weights from information available for all subjects in the main cohort reduced the mean squared error (MSE) comparing risks estimated using the simulated samples to risks estimated by fitting the same model once to data on the same risk factors for the entire cohort. No such risk prediction methodology has yet been developed for the additive hazards model with stratified sampling. Whether or not calibration would improve risk prediction using the additive hazards model with a more realistic study such as ARIC is unknown.
The inflammation marker hs-CRP has been shown in multiple epidemiological studies to predict cardiovascular disease. A statement from the Centers for Disease Control and Prevention and the American Heart Association concluded that hs-CRP was an independent predictor of increased coronary risk. (Pearson et al., 2003) It recommended the optional use of hs-CRP to identify asymptomatic people, already known to be at intermediate risk on the basis of standard risk factors, who might be at even higher risk. Development of new statistical tools to improve the accuracy of risk prediction in this context would be an important contribution.

5. Main Hypothesis/Study Questions:

- Development of statistical methodology for joint estimation of excess risk coefficients and baseline risk functions using biomarker data from stratified samples consisting of disease cases and a Cohort Random Sample.

- Combination of the estimated coefficients and baseline risk function to yield a risk prediction equation based on risk factors measured for the stratified sample and the time since their measurement. This will include calculation of standard errors of risk estimates.

- Calibration of the design (inverse probability of sampling) weights used in the above methods using information on the association between disease outcomes and standard risk factors known for most subjects in the cohort. The goal is to include more cohort data and thereby improve the accuracy of the risk prediction.

- Illustration and evaluation of the methodology using data from the Ballantyne study by construction of two sets of CHD predictors based on hs-CRP and standard risk factors, one with and one without calibration of the weights. One hypothesis is that the standard errors for major risk factors, and for interaction terms involving hs-CRP and main risk factors (e.g., LDL-C), will be reduced by calibration. A second hypothesis is that the standard errors of the predicted risks will also be reduced by calibration.

- Using data on standard risk factors and CHD occurrence for the main ARIC cohort, conduct computer simulations of Cohort Random Sampling designs to evaluate how the regression coefficients and risk predictors based on the samples compare with those obtained by using the same procedure with the entire cohort. Again, the hypothesis is that use of calibration will result in coefficients and predicted risks that are closer, on average, to those estimated for the entire cohort.

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 plan to use exactly the same dataset that was prepared for the reports by Breslow et al. (2009) and Kang, Cai and Chambless (2013). The corresponding ARIC manuscript numbers are MS#1184 and MS#1621. This includes all information on traditional risk factors available for 12,363 participants in the full cohort together with information on Lp-PLA₂ and hs-CRP obtained for CHD cases and the non-cases in the CRS.
Data inclusion/exclusion

The study population of 12,363 subjects is a subset of the “potential full cohort” of 12,819 subjects mentioned by Ballantyne et al. (2004). This consisted of subjects who had Lp-PLA2 and hs-CRP measurements at visit 2, except for those excluded because of CHD or missing CHD information before visit 2, because they had a prior transient ischemic attack or stroke, or because they belonged to an underrepresented minority group. Further exclusions were applied to the data used for the statistical studies due to changes in consent status or missing information.

Study Design, outcome and other variables of interest

The cohort was followed up for the subsequent (after visit 2) development of a CHD event, including CHD-related death. Subjects alive and event-free at the end of 1998, or who had been lost to follow-up, had their records censored at that time. A CRS was selected using a stratified random sampling design based on 8 strata defined by sex, race (black versus white) and age at baseline (>=55 versus <55). Hs-Crp and Lp-PLA2 was assessed only for the CRS members and the subsequently identified CHD cases using plasma stored from visit 2.

Summary of Data analysis

We will first fit additive hazard models to the CRS/CHD sample data using inverse probability weighted estimating equations (Kang et al, 2013). Explanatory variables in the model will be selected from the stratification variables sex, race, and age; the traditional risk factors smoking status, diabetes, systolic blood pressure, LDL-C, and HDL-C; the nontraditional risk factor hs-CRP; and possibly the interaction term between hs-CRP and LDL-C. The excess hazard associated with increased hs-CRP will be reported with two standard errors, one assuming that the additive hazards model is correct, with respect both to the selected covariates and the additivity of their effect on the baseline hazard, and one that is still valid when the model is wrong. The second standard error is considered “robust” in the face of model misspecification. Previous studies of additive hazard models only reported the first, model based standard error (Lin and Ying, 1994; Kulich and Lin, 2000; Kang et al, 2013). The baseline cumulative hazards at various time points will also be calculated with both regular and robust standard errors.

Risk prediction

Next, based on our fitted additive hazards model, we will estimate the cumulative hazard of CHD for individuals having specific covariate values at times 2, 4 and 8 years after visit 2. We will plot estimated cumulative CHD hazard curves for subjects having fixed levels of traditional risk factors but with different hs-CRP levels. These curves may provide additional information for physicians when the traditional risk factors cannot identify the patient at a high risk of CHD.

Calibration of the weights

A second set of analyses will incorporate additional information available for the full cohort by calibration of the weights. The main idea is to adjust the weights used in the inverse probability weighted method by multipliers such that the estimates based on the sampled data agree with
known totals for auxiliary variables in the full cohort. We expect the confidence intervals obtained for regression coefficients using this method will be narrower than those obtained using only the sampled data. We will also investigate which auxiliary variables yield the most efficiency gain.

Simulation studies:

We will also conduct simulation studies based on data known for all subjects in the ARIC cohort to evaluate the properties of our proposed methods. Briefly, we will first simulate 1000 cohort random samples from the full cohort by independent stratified sampling. The sampling probability for each stratum will be the same as in the original study. Each cohort random sample will then be combined with all the CHD cases to form a simulated dataset.

A standard risk factor such as LDL-C or the LDL/HDL ratio will be selected to play the role of the biomarker in the simulation studies. We will pretend that levels of this factor are known only for the subjects selected for each simulated dataset. However, we will first fit the additive hazards model using data for the full cohort to estimate regression coefficients and cumulative hazards. The estimated coefficients and risk predictions from the model fitted to the full cohort will serve as the “gold standard” for comparison with the coefficients and risk predictions estimated using methods described above and data from each of the 1000 simulated CHD/CRS datasets. Comparisons will be made of the estimates obtained both with and without calibration of the weights. The mean squared error (MSE) of estimates obtained with the full cohort vs the simulated samples will be used to evaluate whether calibration brings any meaningful improvement to risk predictions made using the sampled data. We will compare these MSE with the differences between the squared standard errors of estimates based on the simulated samples, and the squared standard errors obtained from the full cohort analysis, to evaluate whether our methods of calculating standard errors for the weighted analyses, both with and without calibration, are accurate.

All the data analysis will be conducted in R statistical language. The R programs developed for this project will be made available to ARIC investigation 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 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  __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 list under the Study Members Area of the web site at: http://www.csc.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 most related manuscripts in ARIC are


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.csc.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.csc.unc.edu/ARIC/index.php, under
Publications, Policies & Forms. [http://publicaccess.nih.gov/submit_process_journals.htm](http://publicaccess.nih.gov/submit_process_journals.htm) shows you which journals automatically upload articles to Pubmed central.

References:


