ARIC Manuscript Proposal # 1585

PC Reviewed: 12/8/09	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Detection of susceptibility loci with estimated glomerular filtration rate (eGFR) using multiple measures in a linear mixed model.

b. Abbreviated Title (Length 26 characters): GFR LME

2. Writing Group:
Writing group members: Adrienne Tin, Joe Coresh, Anna Kottgen, Brad Astor, Eric Boerwinkle, Nora Franceschini, Linda Kao. Others welcome.

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

F**irst author**: Adrienne Tin (may change depending on the phenotype being examined)
Address: 615 N. Wolfe St., Room W6014
Baltimore, MD 21205

Phone: 201-281-9577 Fax: 410-955-0863
E-mail: atin@jhsph.edu

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

Linda Kao
Address: 615 N. Wolfe St., Room W6513
Baltimore, MD 21205

Phone: 410-614-0945 Fax: 410-955-0863
E-mail: wkao@jhsph.edu

3. Timeline:
Starting Analyses: December 2009
First Draft: August 2010
Submission for Publication: October 2010

4. Rationale:

Genome-wide association studies (GWAS) of complex diseases tend to require large sample sizes to detect genetic variants with modest effects due to locus and allelic heterogeneity and errors in the measurement of the variables.(1,2) Recently the CHARGE consortium published the results of a genome-wide association study of renal function with a total of 40,000 participants. (3) While the use of repeated (or multiple) measures of an outcome can reduce random measure error and increase statistical power (4), there has been relatively little empirical data demonstrating this. Therefore, we propose to conduct a genome-wide association analysis using multiple glomerular filtration rate (GFR) estimates in a linear mixed model. GFR can be estimated in the Atherosclerosis Risk in Communities (ARIC) study from serum creatinine at Visits 1, 2, and 4, and from serum cystatin C at Visit 4. We will compare the results from the mixed model approach and simple association analysis and determine whether the associations are consistent between the two approaches and, if so, whether the associations are stronger in the mixed model approach. This analysis can help investigators to decide whether they can use existing data more efficiently by applying a mixed model approach. We have two slightly different reasons why this analysis will work better: 1) repeated eGFR will reduce variance, or 2) cystatin-C provides another correlated biomarker that should give the outcome greater sensitivity and specificity (i.e. true renal genes and not creatinine genes).

GFR is an important measure of kidney function and used in the diagnosis of chronic kidney disease. (5) The gold standard for estimating GFR is to measuring the clearance of iothalamate or other exogenous substances. (6) These procedures are not practical in epidemiological studies, including large scale genetic association studies. In these studies, it is common to use the Modification in Diet and Renal disease (MDRD) study equation to estimate GFR based on calibrated serum creatinine level (eGFRcreat). (3,7) Serum cystatin C and other biomarkers have also been proposed for calculating eGFR. (8,9). Regardless of the biomarkers, estimated GFR is determined by three components: the true GFR, measurement bias, and random error. Measurement bias is consistent and generated by the measurement method and it can bias the results of an association study in either direction depending on the direction of the bias. (10) On the other hand, random error fluctuates and may be due to day-to-day changes in the physiological condition of the person and laboratory condition during the assay of the biomarker. Random fluctuations widen the precision of the measurement and bias the results of an association study toward the null. (10) However, random error can be reduced with repeated measures.

With respect to measurement bias, comparing eGFRcreat with GFR based on the clearance of iothalamate (mGFR), eGFRcreat has little bias when mGFR is below 60 ml/min per $1.73m^2$ and has a downward bias of around 8% when mGFR is above 60 ml/min per $1.73m^2$. Regarding precision, the overall percentage of estimates within 30% of mGFR (P_{30}) is 83%. There is little difference in P_{30} between the high and low strata of mGFR. (11) While some have advocated that an eGFR based on serum cystatin C (eGFRcys) might be a more valid estimate of true GFR, this issue remains equivocal. (6)

A study using the data from the Third National Health and Nutrition Examination Survey shows that eGFRcys may underestimate GFR when body mass index (BMI) is high because adipocytes may secrete cystatin C (12). Another study that includes mostly patients with chronic kidney disease concludes that both eGFRcreat and eGFRcys have minimal bias and similar precision when mGFR is less than 90 ml/min per $1.73m^2$. This study proposes an equation that uses both serum creatinine and cystatin because it performs better than equations that use one biomarker alone. (13) Both measurement bias and random error contribute into the value of $P_{30.}$

Recently a new equation, the CKD-EPI equation, is proposed for estimating GFR using serum creatinine. (14) The new GFR estimates (eGFRckdepi) has less bias and higher precision than the MDRD study equation. Using the CKD-EPI equation, the overall median difference between eGFRckdepi and mGFR is 2.5 versus 5.5 mL/min/$1.73m^2$ using the MDRD equation. The overall P_{30} improved from 80.6% to 84.1%. Since the CKD-EPI equation results in little bias, the P_{30} value is mostly influence by random error. A P_{30} of 84.1% is still quite substantial.

The ARIC study has serum creatinine from visits 1, 2, 4 and cystatin C from visit 4 for estimating GFR. These GFR estimates can be used together to reduce random errors and increase the power of association analysis. The biases from using serum creatinine and cystatin C may stem from different sources. For example, serum creatinine is related to muscle mass, while cystatin C may related to adiposity. (12) Using both biomarkers may shrink the bias. However, there is very little research on the bias direction of eGFRcys in non-clinical population.

One way to use all these measures in the detection of genetic association is to model these measures as correlated outcomes in a linear mixed model. At the simplest level, a mixed model for repeated measures using a genetic variant as a predictor has the following form:

Outcome = individual random intercept + fixed effect intercept + covariates + genetic variant + error

The covariates include age and sex. The random intercept estimates individual deviation from the population average. The fixed effect intercept is the population average at the referenced allele. The coefficient and p-value of the genetic variant term assess the association of the genetic variant. A key aspect of the mixed model is that the correlation in errors across outcomes is modeled. (15)

A limitation of this analysis is that serum creatinine was measured from visits which were approximately three years apart. The sources of within-subject variation include both random error and changes in clinical conditions that had affected GFR. We can only detect the genetic effects that persist during the study period. This assumption is similar to the assumption in simple association analyses that do not control for clinical confounders because this kind of analysis can only detect the genetic effects that are independent of comorbidities in the study population.

To calculate eGFR using serum creatinine, we will use the CKD-EPI equation (14) based on calibrated serum creatinine values. (16) For eGFRcys, we will use the equation in Stevens et al. (13) We will conduct one analysis using the three eGFRckdepi values and another analysis with the three eGFRckdepi values plus eGFRcys. The comparison between the two results can quantify the improvement gained from combining different biomarkers.

5. Main Hypothesis/Study Questions:

We hypothesize that using multiple measures of estimated GFR based on creatinine and cystatin C will increase our power to detect quantitative trait loci (QTL) for eGFR.

6. Data (variables, time window, source, inclusions/exclusions):

Inclusion:

All white ARIC participants

1) giving consent for use of DNA
2) have successful GWAS data

Outcome: eGFRcrkdepi at Visits 1, 2, 4 and eGFRcys at visit 4. We are also interested in other GFR biomarkers, such as beta-trace protein and beta(2)-microglobulin when they are available (currently being measured in all participants at Visit 4). In addition, we will expore the impact of including the data on serum creatinine and cystatin C from the ARIC Carotid MRI study (n~2,000 during 2004-2006).

Exposure: Affy 6.0 imputed and unimputed SNPs. The imputed platform provides more extensive coverage, while the unimputed platform includes more variants with smaller minor allele frequencies.

Covariates include, but are not limited to age, sex, and principal components that are associated with the outcomes.

Analysis Plan

1. Generate sex specific unstandardized residuals adjusting for age, center, and significant principal components. Sex specific residual will be combined. Using residuals as outcomes reduces the complexity of the mixed model.
2. Transform the combined residuals to match the multivariate-normal assumption of the linear mixed model and then standardized to have mean of 0 and standard deviation of 1, so that they are in the same units.
3. Use simple association analysis results to select SNPs for mixed model analysis. Since linear mixed model is computationally intensive, we will first perform simple association analysis on the standardized residuals of each outcome separately and use the results to set a criteria for selecting SNPs that have the potential to achieve high significance level (e.g. 10^-4) in the mixed model analysis. We will explore a couple of strategies and try to take into the account the

selection criteria in phase 1 (simple analysis of each trait) in interpreting phase 2 (mixed model analysis).
One approach is to select all SNPs with a borderline result for any of the analyses (p<10e-5) and see if the mixed model analysis results in a p-value that is smaller than all the simple linear model analysis and cross the usual threshold for GWA significance ($p<5*10^{-8}$). Another way for determining the criteria empirically is to select a small subset of SNPs with various minor allele frequencies and small p-values with <u>all outcomes</u> in simple association analyses, then check for the p-value of these SNPs in a mixed model analysis. For example, if the outcomes are not correlated and the SNPs are independent of each other, the proportion of SNPs with a p-value of 0.1 with all outcomes is approximately 0.1^4 (4 is the number of outcomes). Since the outcomes are correlated, the proportion will be higher. This analysis will select SNPs most likely to benefit from looking at a mixed model of all measures of kidney disease. Examination of these results will provide information on the criteria to select SNPs for mixed model analysis.

4. Assess mixed model fit. We will compare mixed models with and without a time factor in the fixed and random effects to determine model fit. Nested models of fixed effects can be compared by likelihood ratio test of maximum likelihood. Nested models of random effects can be compared by likelihood ratio test of restricted maximum likelihood. (17) Different correlation structures will also be compared with the unstructured specification to determine the optimal structure.
5. Use PLINK for simple association analysis, and SAS PROC MIXED for mixed model analysis and assume an additive genetic model.
6. Compare results between the linear model and the mixed model. Compare the list of SNPs with p-value < 10^{-4} in mixed model and simple association analyses in terms the number of SNPs exceeding the threshold and the effect size. Whether the mixed model is more sensitive in detecting genetic association depends on the correlation among the outcome measures and the amount of missing data at follow-up visits. To generalize the results beyond the current analysis will require a simulation study that considers a range of correlations and the amount of missing data during follow-ups.
7. Pending results, we will collaborate with other CHARGE cohorts in our replication effort.

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?
__X__ 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? __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 ICTDER02 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

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)?

Lead authors of related proposals are in the writing group.

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 (GWAS data collection, 2004.10 cystatin-C, Astor)

__ 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/

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.

References

1. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. Size matters: Just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 2009 Feb;38(1):263-73.

2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature 2009 Oct 8;461(7265):747-53.

3. Kottgen A, Glazer NL, Dehghan A, Hwang SJ, Katz R, Li M, Yang Q, Gudnason V, Launer LJ, Harris TB, Smith AV, Arking DE, Astor BC, Boerwinkle E, Ehret GB, Ruczinski I, Scharpf RB, Ida Chen YD, de Boer IH, Haritunians T, Lumley T, Sarnak M, Siscovick D, Benjamin EJ, Levy D, Upadhyay A, Aulchenko YS, Hofman A, Rivadeneira F, Uitterlinden AG, van Duijn CM, Chasman DI, Pare G, Ridker PM, Kao WH, Witteman JC, Coresh J, Shlipak MG, Fox CS. Multiple loci associated with indices of renal function and chronic kidney disease. Nat Genet 2009 May 10.

4. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Hoboken, NJ: John Wiley & Sons, Inc.; 2004.

5. Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes MW, Hogg RJ, Perrone RD, Lau J, Eknoyan G, National Kidney Foundation. National kidney foundation practice guidelines for chronic kidney disease: Evaluation, classification, and stratification. Ann Intern Med 2003 Jul 15;139(2):137-47.

6. Laterza OF, Price CP, Scott MG. Cystatin C: An improved estimator of glomerular filtration rate? Clin Chem 2002 May;48(5):699-707.

7. Bash LD, Coresh J, Kottgen A, Parekh RS, Fulop T, Wang Y, Astor BC. Defining incident chronic kidney disease in the research setting: The ARIC study. Am J Epidemiol 2009 Aug 15;170(4):414-24.

8. Dharnidharka VR, Kwon C, Stevens G. Serum cystatin C is superior to serum creatinine as a marker of kidney function: A meta-analysis. Am J Kidney Dis 2002 Aug;40(2):221-6.

9. Filler G, Priem F, Lepage N, Sinha P, Vollmer I, Clark H, Keely E, Matzinger M, Akbari A, Althaus H, Jung K. Beta-trace protein, cystatin C, beta(2)-microglobulin, and creatinine compared for detecting impaired glomerular filtration rates in children. Clin Chem 2002 May;48(5):729-36.

10. Rothman KJ, Greenland S, Lash TL. Modern epidemiologyRothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd edition ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

11. Stevens LA, Manzi J, Levey AS, Chen J, Deysher AE, Greene T, Poggio ED, Schmid CH, Steffes MW, Zhang YL, Van Lente F, Coresh J. Impact of creatinine calibration on performance of GFR estimating equations in a pooled individual patient database. Am J Kidney Dis 2007 Jul;50(1):21-35.

12. Vupputuri S, Fox CS, Coresh J, Woodward M, Muntner P. Differential estimation of CKD using creatinine- versus cystatin C-based estimating equations by category of body mass index. Am J Kidney Dis 2009 Jun;53(6):993-1001.

13. Stevens LA, Coresh J, Schmid CH, Feldman HI, Froissart M, Kusek J, Rossert J, Van Lente F, Bruce RD,3rd, Zhang YL, Greene T, Levey AS. Estimating GFR using serum cystatin C alone and in combination with serum creatinine: A pooled analysis of 3,418 individuals with CKD. Am J Kidney Dis 2008 Mar;51(3):395-406.

14. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF,3rd, Feldman HI, Kusek JW, Eggers P, Van Lente F, Greene T, Coresh J, CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 5;150(9):604-12.

15. McCullagh P, Nelder JA. Generalized linear models. 2nd edition ed. Boca Raton: Chapman & Hall/CRC; 1989.

16. Coresh J, Astor BC, McQuillan G, Kusek J, Greene T, Van Lente F, Levey AS. Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate. Am J Kidney Dis 2002 May;39(5):920-9.

17. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer Science+Business Media, Inc.; 2000.