ARIC Manuscript Proposal # 1753

PC Reviewed: _2_/_8_/11	**Status:** _A___	**Priority:** _2_
SC Reviewed: _________	**Status:** _____	**Priority:** ____

1.a. Full Title: **Genome-Wide Association Study of Longitudinal Change in Pulmonary Function: Meta-Analysis in the CHARGE and SpiroMeta Consortia**

b. Abbreviated Title (Length 26 characters): **GWAS of Change in PFTs**

2. Writing Group:

Writing group members: Bonnie Joubert, Nora Fransceschini, Laura Loehr, Kari North, Alanna Morrison, David Couper, Aaron Folsom, Stephanie London and other interested ARIC investigators with time to contribute*.

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

First ARIC author*: **Bonnie Joubert, Postdoctoral Fellow NIEHS**
Lead ARIC author*: **Stephanie London, Senior Investigator, NIEHS**
Address: NIEHS, PO Box 12233, MD A3-05, RTP, NC 27709

Phone: 919-541-5772 Fax: 919-541-5772
E-mail: london2@niehs.nih.gov

*Note – this manuscript is being lead by Dr. Pat Cassano and her collaborators from Health-ABC within the CHARGE consortium. They drafted an initial analysis plan which was then refined in several iterations of input from the CHARGE pulmonary group (see attached). The ARIC lead authors will not be the lead authors for the paper. In addition to the CHARGE cohorts, the paper will include cohorts from the European SpiroMeta consortium and hopefully additional European cohorts who have been participating in some recent meta-analysis with CHARGE and SpiroMeta. It is not yet clear which cohorts will be included in starred first and starred last authorships and it is not yet clear how many individual authors will be listed per cohort.

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: **Nora Franceschini**

Address: 137 E. Franklin St., Suite 306,Campus Box 8050,Chapel Hill, NC,
27514

Phone: 966-1305 Fax: 919-966-1305

E-mail: noraf@unc.edu

3. Timeline: Begin ARIC analysis March 2011. The ARIC investigators will be responsible for running the GWAS with the ARIC data according to the analysis plan agreed upon by the CHARGE pulmonary group. Drs. Joubert and London will be responsible for the computation required for the ARIC analysis and all ARIC authors will assist with checking results and providing input. The GWAS betas and P values will be uploaded to the the CHARGE sharepoint. Pat Cassano of Health ABC (CHARGE) and her team will perform the meta-analysis once all of the participating cohorts have uploaded their GWAS results. We would hope that a manuscript would be ready to submit to the ARIC manuscript committee by February 2012 but the timeline will not be completely under our control because Dr. Cassano is leading this meta-analysis requiring the cooperation of many different analysts.

4. Rationale: Recent genome-wide association studies of cross-sectional measures of pulmonary function have identified a number of novel loci. The first gene identified was the HHIP gene (1). A later meta-analysis from the CHARGE Consortium (ARIC MS #1357) identified an additional 8 novel loci and confirmed HHIP (2). Pulmonary function at a given point in adult reflects factors that influence lung and airway growth to early adulthood and then factors that influence the inevitable age-related decline. The genes identified in GWAS of cross-sectional pulmonary function are weighted toward those involved in growth and development. There are no published data from genome wide association studies on longitudinal decline in pulmonary function. However, ARIC participated in look-up replication of top hits for the ESE consortium (MS # 1676). Although this paper is not yet complete, the results suggest that distinct genes underlie cross-sectional pulmonary function and its decline with age. We speculate that genes involved in pulmonary response to environmental agents might be related to decline in pulmonary function more strongly than genes involved in development. To date candidate gene studies have not been fruitful. As was the case with cross-sectional pulmonary function, we expect that GWAS may identify novel genetic associations with longitudinal measures.

5. Main Hypothesis/Study Questions:

We are asking whether genome wide association analysis will identify novel genetic variants related to the rate of decline in pulmonary function.

The primary parameter of interest is the FEV1 (forced expiratory volume in one second). This is the pulmonary function parameter that is typically followed with respect to decline. However, it is possible that reviewers or editors may request analyses of the other major pulmonary function parameters – FVC and the FEV1/FVC. Therefore it is possible that we might need to include analysis of these parameters in the manuscript.

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).

A detailed analysis plan (**ATTACHED**) has been developed by Drs Pat Cassano, Wenbo Tang and Marty Wells (all of Cornell University). These investigators represent the HEALTH-ABC cohort (longitudinal study of aging) in CHARGE and had previously performed a longitudinal analysis in HEALTH-ABC. They are taking the lead on this meta-analysis for the CHARGE group. The analysis plan was refined based on interactive feedback from the CHARGE Pulmonary Group on several phone conferences and by email afterwards. Dr. Cassano will seek additional cohorts outside of CHARGE to participate either in the meta-analysis or for look-up replication, depending on their level of interest. She will first look for collaborators within the SpiroMeta consortium who have recently been collaborating with the CHARGE pulmonary group. However, given that there are few SpiroMeta cohorts with longitudinal data she will also look to additional cohorts with longitudinal pulmonary function data such as the ESE consortium for whom we participate for look-up replication (MS# 1676).

7.a. Will the data be used for non-CVD analysis in this manuscript? __x__ 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? __x__ 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

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 lead author (Stephanie London) heads the CHARGE Pulmonary Group which has several manuscript proposals completed or ongoing. Please see below.

# 1676 Look-up replication in ARIC of top findings from genome wide association study (GWAS) of decline in pulmonary function in the ESE consortium. S. London

The above manuscript proposal was to do look-up replication for another group that was leading a GWAS.

# 1597 Genome-wide association study of pulmonary function: joint meta-analysis of two consortia - CHARGE and SpiroMeta. S. London

The above manuscript proposal was limited to cross-sectional analysis.

#1562. Genome Wide Association Study of interaction with smoking in relation to pulmonary function and COPD. D. Hancock working with S. London. This manuscript will be restricted to cross-sectional analysis.

#1357 Genome-Wide Association Study (GWAS) of Pulmonary Function and Chronic Obstructive Pulmonary Disease (COPD) – interaction with intake of fiber and other nutrients in ARIC. S. London. This manuscript proposal was merged with #1360 and resulted in the publication of Hancock et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Genet. 2010 Jan;42(1):45-52. Epub 2009 Dec 13.

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.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. Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, Myers RH, Borecki IB, Silverman EK, Weiss ST, O'Connor GT. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS Genet. 2009;5(3):e1000429. PMCID: 2652834.

2. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, Franceschini N, van Durme YM, Chen TH, Barr RG, Schabath MB, Couper DJ, Brusselle GG, Psaty BM, van Duijn CM, Rotter JI, Uitterlinden AG, Hofman A, Punjabi NM, Rivadeneira F, Morrison AC, Enright PL, North KE, Heckbert SR, Lumley T, Stricker BH, O'Connor GT, London SJ. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Genet. 2010;42(1):45-52. PMCID: 2832852.

Genome-Wide Association Study of Longitudinal Change in FEV_1

Wenbo Tang (wt227@cornell.edu)
Marty Wells (mtw1@cornell.edu)
Pat Cassano (pac6@cornell.edu)

January 7, 2011

This analysis plan for the genome-wide association study (GWAS) of rate of decline in FEV_1 describes two-tiered approach. To begin, all cohorts will explore the optimal mixed model parameters, and report attributes of the mixed model (see 1.2.3 below), which will be reviewed prior to finalizing the best model to use across all cohorts. Following this step, all cohorts will complete two GWAS analyses, as follows: 1) using the mixed model approach in which the repeated FEV_1 measurements are the responses and 2) using the "slope as outcome" approach where the dependent variable is the individual FEV_1 slope, calculated in a separate step by regressing FEV_1 measurements on time.

The strategy for the GWAS of the trajectory in FEV_1 is to run a full GWAS using the two models described above. In subsequent investigations, for example in sensitivity analyses, the slope as outcome approach will be used to identify the subset of the most statistically significantly SNPs; final analysis of the subset of top hits will be completed using the mixed model approach to achieve more unbiased estimates of effect and more precise estimates of standard errors.

1. Mixed model approach

1.1 Data set preparation

1.1.1 Participant and observation exclusion criteria

Spirometry: The quality control standards for spirometry data are the same as in all previous CHARGE GWAS meta-analyses. Thus, FEV_1 measurements meeting the ATS/ERS criteria for acceptability will be included.

Covariates: Cohort participants with data on all covariates are included in the analysis. The covariates include: baseline age, gender, baseline height, change in height from baseline to date of spirometry for each spirometry time point, smoking pattern over follow-up, baseline pack-years, smoking status at each spirometry time point, study site (if applicable), and principal components.

1.1.2 Data set construction

The data for a cohort will typically exist in 'wide-format' as shown below for the Health ABC. The spirometry and covariate exclusion procedures and the computation of the smoking pattern variable, and height change variable (described below) is easy to achieve in this data set format.

HABCID	Baseline age	GENDER	Baseline height	Smoking pattern	Baseline pack-years	Fev1 at Y0	Fev1 at Y4	Fev1 at Y7	Fev1 at Y9
1001	73	Female	152.2	4	0	1680	1432	1323	1387
1003	73	Male	184.65	4	0	2266	.	1667	1871
1005	77	Male	170.7	3	48	1817	.	.	.
1006	75	Male	174.25	4	0	3120	2851	2740	2732

In preparation for the mixed model analysis, transpose the 'wide-format' data set into a 'long-format' data set similar in structure to the example below (only selected variables are shown). [SAS program used in Health ABC for this data set transposition is provided in Appendix 1]

HABCID	Gender	Baseline age	Baseline pack-years	Smoking pattern	Baseline height	Height change	Time	Fev1
1001	Female	73	0	4	152.2	0	0	1680
1001	Female	73	0	4	152.2	1.35	4.2137	1432
1001	Female	73	0	4	152.2	-0.9	7.27397	1323
1001	Female	73	0	4	152.2	-1.3	9.23836	1387

1003	Male	73	0	4	184.65	0	0	2266
1003	Male	73	0	4	184.65	-2.6	7.34795	1667
1003	Male	73	0	4	184.65	-3.95	9.14247	1871
1005	Male	77	48	3	170.7	0	0	1817
1006	Male	75	0	4	174.25	0	0	3120
1006	Male	75	0	4	174.25	-0.725	4.19726	2851
1006	Male	75	0	4	174.25	-1.2	7.36164	2740
1006	Male	75	0	4	174.25	-1.7	9.13973	2732

1) The number of observations (rows) for each participant depends on the number of acceptable FEV_1 measurements. For example, in the wide-format data set above, HABCID 1003 has one missing FEV_1 at Y4. As a result, in the long-format data set, HABCID 1003 has three observations in the long format file.
2) Participants with only one acceptable FEV_1 measurement are included in the analysis because the mixed model accommodates all available data on a participant (for example HABCID1005 above). This feature does not involve the assumption of missing at random of the follow-up FEV_1 measurements.
3) A new variable, time, is created to capture the time (in years) elapsed between each pulmonary function test and the study baseline (thus, time=0 for the first FEV_1 measurement). The values for the time variable are not integers, and it is treated as a continuous variable in the mixed model.

1.1.3 Cigarette Smoking

Past GWAS adjusted for 3 smoking variables including ever smoker yes/no, current smoker yes/no, and pack-years. The longitudinal models will adjust for a smoking pattern variable (summarizing smoking pattern over the follow-up), a current smoking status variable (smoking status at each time point), and cumulative pack-years smoking at the baseline, as described below.

1) Smoking pack-years, calculated at study baseline, as per prior GWAS of pulmonary phenotypes.
2) Smoking pattern over follow-up, assessed at same time points as repeated measurements of FEV_1, as follows: (a) persistent smokers (current smoker at all time points), (b) former smokers (former smoker at all time points), (c) never smokers (never smoker at all time points), and (d) intermittent smokers (all others).

   Note: Computation of this variable requires the smoking status variable measured at study baseline and each follow-up time point (corresponding to FEV_1 measurements). Therefore, a participant without baseline smoking status data has missing data, and would be excluded from the analysis. When a participant has missing information at a follow-up time point and there are no alternative sources to acquire the needed information, we assume the smoking status at the preceding time point applies to the missing data (imputation strategy—smoking status assumed to be stable, unless otherwise reported; each cohort will need to consider the best approach in the context of missing data patterns).
3) Current smoking status y/n, at each time point corresponding to a FEV_1 measurement.

In Health ABC this set of 3 variables performed well (lowest AIC) and additional consideration of accumulating pack-years over follow-up and/or average dose smoked at each follow-up visit did not add appreciably to the model (little or no change in AIC).

1.1.3 Height

The mixed model will adjust for height using two variables: height at study baseline and the change in height at each follow-up time point (height change). Individuals without baseline height data are excluded from further consideration. Height change is a time-varying variable and should always have a value of 0 at study baseline, which corresponds to the first FEV_1 measurement. When height is missing at a follow-up time point, we assume that the individual's height did not change since the last measurement (imputation strategy: assume no changes until the next available measurement).

In Health ABC this set of 2 variables performed well (lowest AIC) and additional consideration of baseline height and an interaction of height at each time point x time did not add appreciably to the model (little or no change in AIC). Based on comparisons by using mixed models both with and without SNP terms (a subset of 20 SNPs were used), the two alternatives did equally well, and both models showed the same association between change in height and rate of decline in FEV_1. The simpler model, baseline height and change in height, was chosen.

1.2 Mixed model description

1.2.1 Model setup:

1) Preliminary mixed model without SNP terms:

$$\begin{aligned}\mathbf{FEV_{1ij}} = {} & [\beta_0 + \beta_1\mathrm{age} + \beta_2\mathrm{gender} + \beta_3\mathrm{time}_{ij} + \beta_4\mathrm{aheight} + \beta_5\mathrm{heightchg} + \beta_6\mathrm{smkcat} \\ & + \beta_7\mathrm{smkcat}*\mathrm{time}_{ij} + \beta_8\mathrm{apackyr} + \beta_9\mathrm{currstat} + \beta_{10}\mathrm{site} + \beta_{11}\mathrm{pc}_1 + \beta_{12}\mathrm{pc}_2] \\ & + [U_{0j} + U_{1j}\mathrm{time}_{ij} + r_{ij}]\end{aligned}$$

2) Full GWAS mixed model:

$$\begin{aligned}\mathbf{FEV_{1ij}} = {} & [\beta_0 + \beta_1\mathrm{age} + \beta_2\mathrm{gender} + \beta_3\mathrm{time}_{ij} + \beta_4\mathrm{aheight} + \beta_5\mathrm{heightchg} + \beta_6\mathrm{smkcat} \\ & + \beta_7\mathrm{smkcat}*\mathrm{time}_{ij} + \beta_8\mathrm{apackyr} + \beta_9\mathrm{currstat} + \beta_{10}\mathrm{SNP} + \beta_{11}\mathrm{SNP}*\mathrm{time}_{ij} + \beta_{12}\mathrm{site} + \beta_{13}\mathrm{pc}_1 + \beta_{14}\mathrm{pc}_2] \\ & + [U_{0j} + U_{1j}\mathrm{time}_{ij} + r_{ij}]\end{aligned}$$

Where:

1) FEV_{1ij} represents the i^{th} FEV_1 measurement for person j.
2) Variables are: aheight = height at baseline; heightchg = height change from baseline; smkcat = longitudinal smoking pattern; apackyr = baseline smoking pack-years; currstat = point smoking status; site = study site (adjust if applicable); pc = principal components.
3) The first set of parameters (enclosed in the first pair of brackets) are fixed effects and the second set of parameters are random effects for intercept, time (slope) and within-person residual r_{ij}.
4) The covariance structure is not reflected in the model above; see following section for further details.
5) Example SAS code for the preliminary model without SNP terms is provided in Appendix 2.
6) Example R code for the full mixed model is provided in Appendix 3.

1.2.2 Preliminary modeling

The above model is used to: 1) select the most appropriate covariance structure option and 2) identify outlying FEV_1 measurements to edit data in light of the repeated measurements.

1) Run the preliminary model with different covariance structures (e.g. VC, UN, AR, and CS), and compare model fitness using the information criteria scores (AIC, BIC, AICC). When two covariance structures yield similar fitness, choose the more parsimonious structure. [See Appendix 1 for example SAS code]
2) With the optimal covariance structure identified, use the mixed model procedure to compute residuals and examine the corresponding residual diagnostic plots. Using the conditional studentized residuals, it is typical to apply a filter of ±3 to identify outliers. Examine outliers to confirm the presence of aberrant trajectories, and edit data set accordingly. With the edited data set, run the same model again and compare the residual diagnostic plot for improved residual distribution. [Note, in Health ABC we edited 46 out of ~4000 observations, and decisions were clear in examining the set of repeated PFTs and their corresponding QC scores. As mentioned above, FEV_1 measurements not meeting the ATS/ERS criteria for acceptability (with a QC score < 1) were excluded prior to this step. The outlying FEV_1 measurements identified here generally had lower QC scores than the other measurements from the same individual.]

1.2.3 Results of preliminary mixed model

To check model performance across cohorts prior to running the GWAS, each cohort should run the above preliminary mixed model (1.2.1 Model Setup, point 1) with the appropriate covariance option and the edited data. [A results reporting table is provided in Appendix 4]

1.3 Result file format for GWAS

The key coefficients are for the SNP main effect variable and the interaction term of SNP by time, which conveys the effect of adding a coded allele on the annual change in FEV$_1$. As per past GWAS, providing results in a comma delimited (csv) files with the cohort name and the GWAS approach (mixed GWAS) in the file name would be optimal.

The following fields are required for each SNP. It would be appreciated if the field names are as follows:

- **SNP**: rs number
- **Cod_all**: coded allele (effect allele) (“A” “C” “G” “T”)
- **Noncod_all**: non-coded allele (“A” “C” “G” “T”)
- **Strand**: the strand of the baseline and the coded alleles (“+” or “-“)
- **Freq**: allele frequency for coded allele (numeric data)
- **Beta_SNP**: SNP main effect size for each copy of the coded allele (numeric data)
- **SE_SNP**: standard error of beta_SNP (numeric data)
- **P_SNP**: p-value of beta_SNP (numeric data)
- **Beta_int**: SNP*time interaction effect size (numeric data)
- **SE_int**: standard error of beta_int (numeric data)
- **P_SNP**: p-value of beta_int (numeric data)
- **Cov**: covariance between SNP and SNP*time β estimates (numeric data)
- **Type**: whether the SNP was genotyped or imputed (“gen” or “imp”)
- **Imp_info**: an imputation quality score (numeric data)

Please do not apply a genomic control correction, but please provide the lambda for analyses so that this correction can be applied by us later if required.

2. “Slope as outcome” approach

2.1 Slope outcome variable preparation

1) In cohorts with two FEV_1 measurements (baseline and follow-up) per participant, compute the slope variable using the simple equation below:
   FEV_1 decline (ml/year) = [FEV_1 (ml) at $time_2$ – FEV_1 (ml) at $time_1$]/ [$time_2$-$time_1$ (years)]

2) In cohorts with more than two FEV_1 measurements, calculate the slope for each participant using a linear regression of FEV_1 on time (regressing by “participant”). [See Appendix 5 for example SAS code]

2.2 Analytical model

<u>2.2.1 Overview:</u>

1) The raw calculated FEV_1 slopes will be used as the dependent variable instead of the transformed residuals of FEV_1 slopes.
2) Baseline FEV_1 was considered as a covariate in preliminary models in Health ABC, using a set of 20 SNPs. In the model adjusting for baseline FEV_1 effect sizes were reduced, standard errors were similar and thus p-values for the SNP term were less significant compared to the model without adjusting for baseline FEV_1. In Health ABC data, the intercept and slope were not highly correlated (r=~-.15), but each cohort should check this correlation, and report it in the preliminary model results table. A final decision on the best approach will be based on the review of the preliminary model data from each cohort. (still under discussion, decision pending; complete preliminary models)
3) Adjust for annual rate of decline in height during the follow-up, which can be computed using the same regression procedure described above for calculating FEV_1 slope. The purpose of adjusting for this variable is to account for the effect of the change in height on the rate of decline in FEV_1 in a similar way as in the mixed model approach. Based on the Health ABC data, this variable was highly significant in the preliminary model (accounting for 2.2% of the variability in FEV slope), but adjusting for the height trajectory did not appreciably change the regression coefficients for the 20 SNPs tested.

<u>2.2.2 Regression Model:</u>

1) Preliminary model without the SNP term:

$$\mathbf{FEV_1\ slope = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 smkcat + \beta_4 pack\text{-}years + \beta_5 aheight + \beta_6 height\_slp + \beta_7 pc1 + \beta_8 pc2 + \beta_9 site + e}$$

2) Full GWAS model:

$$\mathbf{FEV_1\ slope = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 smkcat + \beta_4 pack\text{-}years + \beta_5 aheight + \beta_6 height\_slp + \beta_7 SNP + \beta_8 pc1 + \beta_9 pc2 + \beta_{10} site + e}$$

Where:

- FEV_1 slope: computed annual rate of decline in FEV_1 as described in 2.1;
- Variables are:
  - smkcat = longitudinal smoking pattern variable, same as in the mixed model approach;
  - pack-years: baseline smoking pack-years;
  - aheight = height at baseline;
  - height_slp = annual rate of change of height;
  - pc1 and pc2: principal components variables; gender, site: self-explanatory

<u>2.2.3 Results of preliminary slope as outcome model:</u> To check model performance across cohorts prior to running the GWAS, each cohort should run the above preliminary model and report the analysis results. [A result reporting table is provided in Appendix 6]

2.3 Result file format—same as 1.3, above.

3. Timeline (to be discussed)

3.1 Preliminary modeling of the mixed and slope as outcome approaches.
Suggested date February 1st

3.2 GWAS analyses using both approaches.
Suggested date March 1st

Appendix 1: Example SAS code for data set transposition:

```
Data datasetname2; set datasetname1;

Id=_N_;

/*Baseline=Y1*/
Time=0;
If y1exqcfev ne 0 then fev1=.; else fev1=y1fev1; /*Y1 FEV1 sets to missing any ppts with qc=0*/
Currstat=y1smoke /*point smoking status at baseline, no missing*/;
Heightchg=0;
OUTPUT;

/*Y5*/
Time = (cv5date-cv1date)/365; /*time elapsed from baseline to Y5 in year*/
If y5exqcfev ne 0 then fev1=.; else fev1=y5fev1; /*Y5 FEV1 sets to missing any ppts with qc=0*/
If y5smoke le .z then currstat=y1smoke; else currstat=y5smoke /*point smoking status at year5*/;
Heightchg=heightchg5 /*height change from baseline to Y5, previously calculated*/;
OUTPUT;

/*Y8*/
Time = (cv8date-cv1date)/365; /*time elapsed from baseline to Y8 in year*/
If y8exqcfev ne 0 then fev1=.; else fev1=y8fev1; /*Y8 FEV1 sets to missing any ppts with qc=0*/
If (y8smoke le .z) and (y5smoke le .z) then currstat=y1smoke;
        Else if y8smoke le .z then currstat=y5smoke;
        Else currstat=y8smoke /*point smoking status at year8*/;
Heightchg=heightchg8 /*height change from baseline to Y8, previously calculated*/;
OUTPUT;

/*Y10*/
Time = (cv10date-cv1date)/365; /*time elapsed from baseline to Y10 in year*/
If y10exqcfev ne 0 then fev1=.; else fev1=y10fev1; /*Y10 FEV1 sets to missing any ppts with qc=0*/
If (y10smoke le .z) and (y8smoke le .z) and (y5smoke le .z) then currstat=y1smoke;
        Else if (y10smoke le .z) and (y8smoke le .z) then currstat=y5smoke;
        Else if y10moke le .z then currstat=y8smoke; else currstat=y10smoke /*point smoking status at year10*/;
Heightchg=heightchg10 /*height change from baseline to Y10, previously calculated*/;
OUTPUT;
Run;

/*Use the statement below to delete entries with missing fev1*/;
Data datasetname2; set datasetname2;
If fev1 < .z then delete;
Run;
```

Where:

- *Datasetname1* = wide-format data set; *datasetname2* = long-format data set;
- Time is calculated based on the baseline clinical visit date variable and the corresponding date variable of a follow-up clinical visit, and then converted to years;
- Y#exqcfev is a previously created categorical variable with three values: 1) *missing* indicates no spirometry test was performed, 2) *0* indicates the y#fev1 test has an acceptable QC score, 3) *1* indicates the y#fev1 test has an unacceptable QC score and should be excluded;
- Currstat is created based on the corresponding point smoking status variable (y#smoke) at each time point. When missing y#smoke is encountered, always assume the same smoking status was maintained from the last time point;
- Heightchg is created from the three heightchg# variables that are previously calculated based on the repeated height measures.

Appendix 2: Example SAS statements to run preliminary models prior to GWAS:

SAS model statement (without SNP variables):

```
        Proc mixed data=datasetname noclprint;
        Class ID smkcat;
        Model fev1 = age gender site aheight heightchg smkcat|time pack-years currstat pc1 pc2/ddfm=kr;
        Random intercept time/ subject=ID type=VC;
        Run; quit;
```

Where:

- ID = participant id number;
- Aheight = height at baseline;
- Heightchg = change of height from baseline, time-varying;
- Smkcat = longitudinal smoking pattern variable, fixed and same value on each record;
- Smkcat|time = interaction of longitudinal smoking pattern and time (allows rate of decline to vary by smoking status);
- Pack-years = baseline smoking pack-years;
- Currstat = time-varying variable on each record for current smoker y/n;
- Pc1 and pc2 = principal components variables;
- The random statement specifies both intercept and time as random effects. The covariance structure in the example is variance components (VC). This option may differ by cohort, thus each cohort will explore the best choice and model appropriately.

Appendix 3: Example R code of the mixed model with variance components (VC) covariance matrix (assuming no covariance between intercept and time) to run the GWAS:

```
Mod <- lmer (fev1 ~ age + gender + site + aheight + heightchg + smkcat*time + pack-years + currstat + pc1 +
    pc2 + SNP*time + (1 | id) + (0 + time | id), data=mdat, na.action=na.omit)
```

More details on the model itself, and implementation in R.

See the following links for more reading on the implementation of the mixed model in R:

General information:
http://www.rensenieuwenhuis.nl/r-sessions-16-multilevel-model-specification-lme4/

Chapter 1: Introduction to Mixed Models
http://lme4.r-forge.r-project.org/book/Ch1.pdf

Chapter 2: Models with Multiple Random Effects Terms
http://lme4.r-forge.r-project.org/book/Ch2.pdf

Chapter 4: Models for Longitudinal Models
http://lme4.r-forge.r-project.org/book/Ch4.pdf

Bates, D. Linear Mixed Model implementation in lme4
http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf

Appendix 4: Results reporting table for preliminary data from the mixed model:

		Cohort (N_{ppt}=?, N_{obs}=?)		
		β	SE	P-value
Intercept				
Time (year)				
Age (year)				
Gender (1=male, 0=female)				
Smkcat	1: persistent			
	2: intermittent			
	3: former			
	4: never			
Smkcat *time	1: persistent			
	2: intermittent			
	3: former			
	4: never			
Pack-years				
Currstat (1=yes, 0=no)				
Aheight (cm)				
Heightchg (cm)				
Pc1				
Pc2				

Covariance structure used:

Appendix 5: Computing FEV1 slope using a simple linear regression:

```
Proc reg data=datasetname;
ODS output ParameterEstimates=fev1reg;
By ID;
Model fev1 = time;
```

Where:

- The data set used for this procedure is in the long-format.
- ID = participant id number;
- Fev1 = repeated FEV_1 measurements;
- Time = time elapsed (in year) between baseline FEV_1 and each subsequent FEV_1;

Appendix 6: Results reporting table for preliminary slope as outcome model:

		Cohort (N=?)		
		β	SE	P-value
Intercept*				
Age (year)				
Gender (1=male, 0=female)				
Smkcat	1: persistent			
	2: intermittent			
	3: former			
	4: never			
Pack-years				
Aheight (cm)				
Height_slp (cm/year)				
Pc1				
Pc2				

*Correlation between baseline FEV_1 and FEV_1 slope = x.xx