1. Missing data in COPD status

In HCHS/SOL baseline data, there are missing data in both first and second spirometry measures. Among the 16,415 participants, 947 (5.8%) had missing first spirometry (806 not done + 141 grade F). Of those 15,468 participants who had first spirometry, 1,403 were abnormal (FEV1/FVC < 0.7 or FEV1 < LLN) and therefore were expected to have a second spirometry after bronchodilation. However, 330 of those participants had missing second spirometry, representing 23.5% of those expected to have the second spirometry. Since the diagnosis of COPD depends on both spirometry measures, overall there are 947 + 330 = 1,277 participants whose COPD status are missing, representing 7.8% of the total sample. Although the overall missing percentage is not particularly high, its potential impact on the estimates of the COPD prevalence could be large because the prevalence itself is fairly low. Therefore, adjustment for missing data is needed to obtain unbiased estimates for the COPD prevalence.

The HCHS/SOL Coordinating Center (CC) recommends using a combination of inverse probability weighting (IPW) and multiple imputation (MI) to adjust for the missing COPD data. This document describes this missing data adjustment method in detail and provides sample SAS and SUDAAN code, as well as suggested text for statistical methods sections of manuscripts.

As of September 2015 this method is currently in use for manuscripts 10a “Pulmonary Disease and Age of Immigration among Hispanic: Results from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)” (R Graham Barr, *et al*).

2. COPD missing data adjustment method

As an overview, the missing data adjustment method uses IPW to adjust for the missingness in the first spirometry and MI to impute the missing COPD status due to the missing second spirometry. The IPW method is a convenient choice of missing data adjustment in a survey sample framework as the IPW can be easily incorporated into the analysis by multiplying it to the sampling weight. The rationale for using MI for the missing COPD status in the second stage is that the first spirometry measure is believed to be strongly associated with the COPD status, and therefore more accurate imputation can be obtained if the first measure is used to impute the COPD status. The next few paragraphs briefly introduce the IPW and MI methods.

Inverse probability weighting (IPW) is one missing data adjustment method under MAR assumption, which allows correcting for the bias of the estimates obtained by complete-case analyses and can be implemented for complex survey designs. The complete cases (i.e., having first spirometry measure) are weighted by the inverse of their probability of being a complete case. For further background, Seaman and White (2013) provide a review of the implementation and advantages and disadvantages of using IPW to handle missing data in epidemiological research. The method has been implemented for physical activity (Actical) data in HCHS/SOL due to the ease with which it can be applied to complex survey data.

The probability of being a complete case is calculated by fitting a logistic regression model where the outcome is the binary variable which takes value 1 if the participant has first spirometry measure and 0 otherwise. The model is fitted without weighting by the HCHS/SOL sampling weight because what we want to predict is the conditional probability of being a complete case, given inclusion into the HCHS/SOL sample. We then use the predicted probability of being a complete case to compute an inverse probability weight for each participant. This additional weight is combined with HCHS/SOL sampling weight into a single weight for analyses involving the first spirometry data.

Multiple imputation (MI) is a commonly used approach to handle missing data under MAR assumption. It is a three step approach:

1. Generate m (typically 5 to 10) possible values for each missing observation that reflect uncertainty about the missing value.

2. Analyze each of the m datasets using complete data methods.

3. Combine the results of m separate analyses using Rubin's rule (Rubin 1987), accounting for uncertainty in the imputation.

There are various methods to impute missing data in the first step depending on the type and pattern of missing data. Commonly used methods include chained equations, conditional Gaussian approach, predictive mean matching, *etc*. Readers are referred to section 2.4 of Horton and Kleinman (JASA 2007) for more details. For the COPD study, the chained equation method (implemented in SAS PROC MI as Fully Conditional Specification method) is used to impute missing data because the variables being imputed are a mixture of continuous and categorical variables and the missing pattern is not monotone. Specifically, missing values in each variable are imputed based on all other variables. This process is done iteratively over all variables until convergence. Imputation models for specific types of covariates are:

o Logistic regression was used for binary or ordinal variables;

o Discriminant method was used for nominal variables;

o Linear regression was used for continuous variables.

In order to compute the IPW for the missingness in the first spirometry measure, a logistic regression is fitted with a set of covariates. However, there are missing data in the covariates as well. There are 690 (4.2%) participants missing at least one of the covariates included in the logistic regression model. **Table 1** summarizes the missing percentages of these covariates. The missing covariates need to be imputed before carrying out the logistic regression. The four steps of the missing data adjustment method for COPD study are described in details below.

Table 1. Missing percentages for covariates used to compute IPW

Covariate	Description	Scale	Missing data	
			N	%
GENDER	Gender	Binary	0	0
AGE	Age	Continuous	0	0
INCOME_C3	Missing were combined with refused creating 3-level nominal variable: <$30K, ≥$30K and unknown	Nominal	0	0
BKGRD1_IMP_C7	7-level Imputed Re-classification of Hispanic/Latino Background	Nominal	0	0
WEIGHT_FINAL_NORM_OVERALL	Sampling weight	Continuous	0	0
STRAT	Stratification	Nominal	0	0
BMI	Body mass index (kg/m2)	Continuous	71	0.4
AGE_US_C3	Age immigrated to US: US born, 0 to 15 years, >15 years	Nominal	73	0.4
PRECHD_ANGINA	Prevalent Cardiovascular Heart Disease, including angina	Binary	76	0.5
MHEA22	Other lung disease	Binary	89	0.5
EDUCATION_C3	3-level education level	Ordinal	91	0.6
YRSUS_C2	2-level grouped years lived in US	Binary	120	0.7
GPAQ_LEVEL	3-level physical activity level	Ordinal	140	0.9
CIGARETTE_PACK_YEARS	Cigarette Pack Years	Continuous	163	1.0
N_HC	Health Insurance Coverage - Current	Binary	323	2.0
FAMILY_HX_ASTHMA	Family history of asthma	Binary	326	2.0
FAMILY_HX_COPD	Family history of COPD	Binary	335	2.0
COPD_EVER	Self-report of ever had COPD/Emph or CB	Binary	339	2.1
CESD10	CESD 10-item total summary score	Continuous	351	2.1
ASTHMA_EVER_MD	Ever had asthma with MD diagnosis	Binary	384	2.3

STEP 1: MI on missing covariates

In the full dataset (n = 16,415), use MI to impute covariates with missing data that are to be used to compute IPW. The MI is done using the fully conditional specification (FCS method in SAS PROC MI, also known as chained equation method in the literature). Five imputations are created.

Since the order of covariates in the VAR statement of PROC MI affects the model fitting and results, we use the descending order of percent of missing values. It has been shown in the analysis of PA data that the descending order facilitates the model convergence. We also set boundaries for continuous variables to avoid nonsensical imputed values. We included STRAT as a covariate in the MI to account for the stratified sampling design. We did not conduct MI for each level of STRAT separately (i.e. using BY STRAT statement in PROC MI) because the model failed to converge, likely due to some small stratum sizes. We did not test interactions between STRAT and other covariates because this step is to impute covariates rather than outcomes, and it would be impractical to test interactions in each model for each covariate. The

SAS code for MI on covariates is listed below. Note that STRAT is not used to impute PRECHD_ANGINA due to convergence failure.

```
proc mi data=home.HC0745_fulldata seed=1 nimpute=5 out=home.S1_covar_MI
maximum=100 . 30 . . . . . . . . . . . . . . 30.2 . . .
minimum=0 . 0 . . . . 0 . . . . . . 0 . . -27.8 0 . .;
class strat gender GPAQ_level EDUCATION_C3 BKGRD1_IMP_C7 INCOME_C4 YRSUS_C2
PRECHD_ANGINA MHEA22 N_HC age_us_c3 COPD_EVER ASTHMA_EVER_MD family_hx_asthma
family_hx_copd;
var FEV1_FVC_RATIO ASTHMA_EVER_MD CESD10 COPD_EVER family_hx_COPD
family_hx_asthma N_HC CIGARETTE_PACK_YEARS GPAQ_LEVEL YRSUS_C2 EDUCATION_C3
MHEA22 PRECHD_ANGINA age_us_c3 BMI BKGRD1_IMP_C7 INCOME_C4 agec
weight_final_N_O gender strat;
fcs logistic(PRECHD_ANGINA=FEV1_FVC_RATIO ASTHMA_EVER_MD CESD10 COPD_EVER
family_hx_COPD family_hx_asthma N_HC CIGARETTE_PACK_YEARS GPAQ_LEVEL YRSUS_C2
EDUCATION_C3 MHEA22 age_us_c3 BMI BKGRD1_IMP_C7 INCOME_C4 agec
weight_final_N_O gender); * doesn't converge with strat;
fcs logistic reg;
run;
```

STEP 2: Compute IPW for the missingness of the first spirometry

For each of the 5 imputed datasets from Step 1, fit a logistic regression (n = 16,415) on the missing status of the first spirometry to compute IPW. The main effects of all covariates from Step 1 are included. A model with all pairwise interactions among them does not converge. So we only include interactions between age, age2 (both centered at sample mean) and covariates with significant main effect, as well as interactions between gender and covariates with significant main effect. Note that the logistic regression does not adjust for the sampling weight (WEIGHT_FINAL_NORM_OERALL), but it uses it as a covariate.

The IPW for the first spirometry is computed for the 15,468 participants with observed first spirometry. For each of the 5 logistic regression results the IPW is computed using the formula:

$$\widehat{IPW}^{(m)} = \frac{1}{1-\hat{p}^{(m)}} = 1 + \exp\left(X^{(m)}\hat{\beta}^{(m)}\right), \quad \text{m=1 to 5}$$

The 5 sets of IPWs are then averaged to obtain the final IPW for the 15,468 participants with observed first spirometry for subsequent analyses.

The SAS code for the logistic regression and IPW calculation is listed below.

```
proc logistic data=home.S1_covar_MI desc;
by _imputation_;
class strat gender GPAQ_level EDUCATION_C3 BKGRD1_IMP_C7 INCOME_C4 YRSUS_C2
PRECHD_ANGINA MHEA22 N_HC age_us_c3 COPD_EVER ASTHMA_EVER_MD family_hx_asthma
family_hx_copd;
model missing_S1 = FEV1_FVC_RATIO ASTHMA_EVER_MD CESD10 COPD_EVER
family_hx_COPD family_hx_asthma N_HC CIGARETTE_PACK_YEARS GPAQ_LEVEL YRSUS_C2
EDUCATION_C3 MHEA22 PRECHD_ANGINA age_us_c3 BMI agec agec2 gender
BKGRD1_IMP_C7 weight_final_N_O INCOME_C4 strat agec*gender agec2*gender
BKGRD1_IMP_C7*gender cesd10*gender age_us_C3*gender weight_final_N_O*gender
```

```
income_C4*gender PreCHD_Angina*agec PreCHD_Angina*agec2 MHEA22*agec
MHEA22*agec2 N_HC*agec N_HC*agec2 BKGRD1_IMP_C7*agec BKGRD1_IMP_C7*agec2
cesd10*agec cesd10*agec2 age_us_c3*agec age_us_C3*agec2 weight_final_N_O*agec
weight_final_N_O*agec2 income_C4*agec income_C4*agec2/covb;
output out = home.S1_logit xbeta= xb;
ods output ParameterEstimates=parms CovB=covb;
run;

data S1_logit_ipw; set home.S1_logit; prob=exp(xb)/(1+exp(xb));
ipw=1/(1-prob); keep _imputation_ subjid ipw; run;
proc sort data=S1_logit_ipw; by subjid; run;
proc transpose data=S1_logit_ipw out=S1_logit_ipw_wide prefix=ipw;
by subjid;
id _imputation_;
var ipw;
run;
data S1_ipw; set S1_logit_ipw_wide;
drop _name_ _label_;
IPW_inv_fulint=mean(of ipw1-ipw5);
run;
```

STEP 3: MI on missing COPD status due to missing data in the second spirometry

Perform MI with fully conditional specification (FCS) method on the 15,468 participants with observed first spirometry for COPD status and a number of covariates as listed in the SAS code below. Five imputations are done. Note that the MI does not adjust for the sampling weight or the IPW as in the survey data analysis, but it uses both of them as covariates.

As in Step1, the covariates are ordered in the descending order of percent of missing values in the VAR statement of PROC MI. We again set boundaries for continuous variables to avoid nonsensical imputed values. The MI also includes STRAT as a covariate to account for the stratified sampling design. No interaction between STRAT and other covariates is included in the MI because some stratum sizes are too small for any models with interaction with STRAT to converge. The interaction terms included in the logistic imputation model for COPD status are chosen based on scientific sense. The SAS code for MI on COPD status is listed below.

```
proc mi data=home.IPW_obsS1_Table6 seed=1 nimpute=5
out=home.IPW_MI_Table6_obsS1
maximum=. . 30 . . . . . . . 70.4 . . . . . . . . . . . . . . . . .
minimum=. . 0  . . . . . . . 13.8 . . . . . . . . . . . . . . . . .;
class strat COPD_BY_BD2 gender GPAQ_level EDUCATION_C3 BKGRD1_IMP_C7
INCOME_C4 PRECHD_ANGINA N_HC age_US_C3 COPD_EVER ASTHMA_EVER_MD
family_hx_asthma family_hx_copd prba3 prba4 YRSUS_C2 CIGARETTE_USE
ASTHMA_ONSET_BIN;
var COPD_BY_BD2 ASTHMA_EVER_MD CESD10 N_HC COPD_EVER family_hx_COPD
family_hx_asthma CIGARETTE_PACK_YEARS GPAQ_LEVEL YRSUS_C2 BMI EDUCATION_C3
CIGARETTE_USE PRECHD_ANGINA agec agec2 BKGRD1_IMP_C7 INCOME_C4 age_US_C3
prba3 prba4 PRBA5 gender weight_final_N_O IPW_inv_fulint FEV1_FVC_RATIO STRAT
ASTHMA_ONSET_BIN;
fcs logistic(COPD_BY_BD2=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4 GPAQ_level
EDUCATION_C3 N_HC CIGARETTE_PACK_YEARS PRECHD_ANGINA age_US_C3 COPD_EVER
ASTHMA_EVER_MD family_hx_asthma family_hx_copd prba3 prba4 prba5 BMI
WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO strat agec*N_HC
agec2*N_HC agec*gender agec2*gender agec*BKGRD1_IMP_C7 agec2*BKGRD1_IMP_C7
agec*INCOME_C4 agec2*INCOME_C4 agec*PRECHD_ANGINA agec2*PRECHD_ANGINA
agec*age_US_C3 agec2*age_US_C3 gender*N_HC gender*BKGRD1_IMP_C7
gender*INCOME_C4 gender*PRECHD_ANGINA gender*age_US_C3);
fcs logistic(GPAQ_level=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4 COPD_BY_BD2
EDUCATION_C3  N_HC CIGARETTE_PACK_YEARS PRECHD_ANGINA age_US_C3 COPD_EVER
ASTHMA_EVER_MD family_hx_asthma family_hx_copd prba3 prba4 prba5 BMI
WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO strat);
fcs logistic(EDUCATION_C3=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4
COPD_BY_BD2 GPAQ_level  N_HC CIGARETTE_PACK_YEARS PRECHD_ANGINA age_US_C3
COPD_EVER ASTHMA_EVER_MD family_hx_asthma family_hx_copd prba3 prba4 prba5
BMI WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO strat);
fcs logistic(PRECHD_ANGINA=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4
COPD_BY_BD2 GPAQ_level EDUCATION_C3 N_HC CIGARETTE_PACK_YEARS age_US_C3
COPD_EVER ASTHMA_EVER_MD family_hx_asthma family_hx_copd prba3 prba4 prba5
BMI WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO);
fcs logistic(N_HC=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4 COPD_BY_BD2
GPAQ_level EDUCATION_C3 CIGARETTE_PACK_YEARS PRECHD_ANGINA age_US_C3
COPD_EVER ASTHMA_EVER_MD family_hx_asthma family_hx_copd prba3 prba4 prba5
BMI WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO);
fcs logistic(COPD_EVER=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4 COPD_BY_BD2
GPAQ_level EDUCATION_C3 N_HC CIGARETTE_PACK_YEARS PRECHD_ANGINA age_US_C3
ASTHMA_EVER_MD family_hx_asthma family_hx_copd prba3 prba4 prba5 BMI
WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO);
fcs logistic(ASTHMA_EVER_MD=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4
COPD_BY_BD2 GPAQ_level EDUCATION_C3 N_HC CIGARETTE_PACK_YEARS PRECHD_ANGINA
age_US_C3 COPD_EVER family_hx_asthma family_hx_copd prba3 prba4 prba5 BMI
WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO strat);
fcs logistic(family_hx_asthma=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4
COPD_BY_BD2 GPAQ_level EDUCATION_C3 N_HC CIGARETTE_PACK_YEARS PRECHD_ANGINA
age_US_C3 COPD_EVER ASTHMA_EVER_MD family_hx_copd prba3 prba4 prba5 BMI
WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO strat);
fcs logistic(family_hx_copd=agec agec2 gender BKGRD1_IMP_C7 INCOME_C4
COPD_BY_BD2 GPAQ_level EDUCATION_C3 N_HC CIGARETTE_PACK_YEARS PRECHD_ANGINA
age_US_C3 COPD_EVER ASTHMA_EVER_MD family_hx_asthma prba3 prba4 prba5 BMI
WEIGHT_FINAL_N_O IPW_inv_fulint cesd10 FEV1_FVC_RATIO strat);
fcs logistic reg;
run;
```

STEP 4: Estimate the prevalence of COPD from the imputed datasets

In the imputed datasets, compute a new weight that combines the original sampling weight and the IPW:

$$\text{WEIGHT\_COPD\_IPW\_OVERALL} = \text{WEIGHT\_FINAL\_NORM\_OVERALL} * \text{IPW}$$

Then use SUDAAN PROC RLOGIST on each imputed dataset to compute the background-specific COPD prevalences and their standard errors adjusting for survey sampling design with the new sampling weight WEIGHT_COPD_IPW_OVERALL and controlling for various covariates. A sample SUDAAN code for COPD prevalence estimation in the imputed datasets is listed below.

```
proc rlogist data=analys1  filetype=sas design=wr MI_count=5;
nest strat PSU_ID / NOSORTCK;
subpopn GLOBAL_SUBPOP2_age=1;
class  gendernum BKGRD1_IMP_C7  Education_c3 yrsus_C2 AGE_US_C3 CIGARETTE_USE
ASTHMA_ONSET_BIN;
reflevel GENDERNUM=1  BKGRD1_IMP_C7=3 Education_c3=1 yrsus_C2=1 ;
weight WEIGHT_COPD_IPW_OVERALL;
model COPD_BY_BD2 = AGE gendernum BKGRD1_IMP_C7 EDUCATION_C3 YRSUS_C2
AGE_US_C3 CIGARETTE_USE CIGARETTE_PACK_YEARS ASTHMA_ONSET_BIN;
PREDMARG BKGRD1_IMP_C7;
run;
```

The “MI_count=5” option in the “proc rlogist” statement of the above code tells SUDAAN to perform a logistic regression on each of the 5 imputed datasets and then combine the results using Rubin’s rule. Note that the names of the 5 imputed datasets should end with consecutive number starting with 1, and only the first dataset name needs to be listed in the “data=” option in the “proc rlogist” statement. SUDAAN will search for the other four datasets automatically by name. The background-specific COPD prevalence is computed by the statement “predmarg BKGRD1_IMP_C7”. This statement computes the predicted marginal for each ethnic background level by averaging the predicted risk of COPD based on all observations in the dataset assuming the variable BKGRD1_IMP_C7 takes on the value for the corresponding ethnic background level for all observations. For details please see section 10.3 (page 232) of Sudaan 11 manual.

The above predicted marginal is not readily available in SAS PROC SURVEYLOGISTIC. The usual “lsmeans” statement only computes the linear predictor Lβ for a certain covariate matrix L, but not the marginal predicted risk of COPD as described in the previous paragraph. It is possible to manually compute the predicted marginal through programming in SAS, but we recommend using Sudaan to obtain the result directly.

3. Suggested wording for statistical methods section

Missing assessment of COPD was accounted for using a three-step process. Missing data for the first spirometry were accounted for by inverse probability weighting (Seaman and White, 2011). Weights were developed using a logistic regression model from a set of baseline covariates. Four percent of participants had sporadic missing data for one or more covariates;

these missing values were first imputed by multiple imputation. COPD status for participants with missing post-bronchodilator spirometry was then imputed with multiple imputation. The background-specific covariate-adjusted COPD prevalences were estimated by logistic regression on each of the imputed dataset and combined by Rubin's rule. The weight used in the logistic regression was the product of the IPW weight and the HCHS/SOL sampling weight. All analyses were performed using SAS version 9.4 (SAS Institute) and SUDAAN release 11.0.0 (RTI).

REFERENCES

Horton & Kleinman (2007). "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models." The American Statistician 61(1):79-90.

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

Seaman, S. R., & White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research, 22(3), 278-295.