Guidelines for HCHS/SOL Manuscript Verification
Version 1.0 (August 14, 2012)

Introduction

Commonly accepted best practices in the biostatistics' field include the independent verification of the data management and statistical analysis for manuscripts to be submitted to peer-reviewed journals. HCHS/SOL Steering and Publication Committees are supportive of this practice. These committees have agreed that initially at least one manuscript from each analyst working at the field center would submit their work for review by the biostatisticians at the HCHS/SOL Coordinating Center (CC). Each analyst would have an initial review of their first work to be submitted for publication. Subsequent work would be randomly selected for review over time by the Coordinating Center.

Goals

Manuscript verification entails the detailed review and evaluation of all elements that comprise the statistical analysis of a HCHS/SOL publication. These include the following:

- Datasets used in the analysis
- Inclusion and exclusion criteria employed
- Employment of the multi-stage sample design in survey analysis techniques
- Comparison of statistical approach to Publications and Presentations Subcommittee approved analysis plan
- Numerical accuracy in tables, figures, and text citations

Procedures

The lead author working with the local analyst should submit a written request to the Coordinating Center PI and/or or Project Director for manuscript verification when submitting the manuscript to the Publications Committee. In turn, the CC will assign a biostatistician to review both the analytic work performed and how those results are abstracted and used in the related publication.

Supporting documentation to be submitted to the CC for the manuscript verification process requires:

- Statistical computing request (written analysis plan) used to describe analyses performed (see below computing request for MS13 "Sleep disorders" as an example)
- Final version of the manuscript after working group review [annotated with source of statistics]
- Tables and Figures intended for publication [annotated with source of statistics]
- Software *code* (SAS, SUDAAN, R, STATA), *run time* logs, statistical software *output*

Since manuscript verification is an audit process, the only way to trace back the origins of statistical results are to know precisely how those numbers were generated by the analyst. Because all software packages offer the option of saving the log of code execution and the related output, it is routine practice for auditors to be able to review that code for warnings and exceptions to program logic. The reviewer will use the annotated manuscript, tables & figures as a guide in auditing the statistical results. Summary results from the manuscript audit will be provided to the lead author, Coordinating Center PI, and chair of the Publications and Presentations Subcommittee.

Example of a STATISTICAL COMPUTING REQUEST

1. Manuscript Number: MS13
2. Lead Author: Susan Redline
3. Statistical Analyst: Daniela Sotres
4. Manuscript or Project Short Title (Length 26): Sleep Apnea
5. Date Submitted: dd/mm/yyyy

PROGRAM SPECIFICATIONS

1. Describe data files to be used:

Use data files from HCHS/SOL Investigator Release 3.1.
PART_DERV: Participant Derived File
ANTA: Anthropometry
LABA: Lab data
SLEA: Sleep Questionnaire
SLPA: Data from Sleep Reading Center
PRBA: Pulmonary Function post-Bronchodilator

2. Exclusions/Inclusions:
Starting with 16,415 exclude (flags are defined in section 3.1):

- No sleep study (flag_sleepstudy)
- Missing background
- Very short sleep study (flag_vshort)
- Missing SLPA54 (MAIN outcome)

3. Detailed description of analysis to be performed:

3.1. Create permanent dataset for analysis for MS # 13

Keep the following variables:

PART_DERV (Participant derived variables): SUBJID weight_final_norm CONSENT BKGRD1 BKGRD1_C7 marital_status CENTER CENTERNUM FULL_AFU_ELIGIBLE age gender GENDERNUM bmi BMIGRP_C4 anta10a income income_c5 education_c2 education_c3 diabetes2 hypertension asthma_ever_md asthma_curr_md asthma_c4_md fev1_fvc_ratio cigarette_use valid_spirometry ESS ESS_GE10 AHI_GE15 dyslipidemia wave strat psu_id STRAT_CNT PSU_CNT LIST_CNT alcohol_use
ANTA (Anthropometry): anta10a
ECEA (Economic): ecea2
LABA (Lab data): laba66 laba67 laba68
SLEA (Sleep Questionnaire): slea12a slea12b slea12c slea12d slea12e slea12f slea12g slea12h slea13 slea14
SLPA (Data from Sleep Reading Center): slpa12 slpa15 slpa30 slpa36 slpa39 slpa54 slpa63 slpa66 slpa97 slpa121

Derive the following variables:

Flags (For non-missing observations: 1 if parenthesis is true and 0 otherwise)**:**

- **Sleep study or not (FLAG_SLEEPSTUDY)**
  If in_SLPA then FLAG_sleepstudy = 1; else FLAG_sleepstudy = 0;

- **Very short study or not (FLAG_VSHORT)**
  If FLAG_SLEEPSTUDY = 0 then FLAG_VSHORT = missing;
  ELSE
  IF slpa30 < 0.5 then FLAG_VSHORT = 1;
  IF slpa30 >= 0.5 then FLAG_VSHORT = 0;

  Label: "Very short study: recording time <= 0.5hr"

- **Short study or not (FLAG_SHORT)**
  SIMILAR to FLAG_VSHORT but using 4 hr as cutpoint
  Label: "Short study: recording time < 4 hr"
  Note: Secondary analyses will be restricted to participants with studies of greater than 4 hours SLPA30 >4

- **In manuscript 13 or not (KEEP_MS13)**
  KEEP_MS13 = (FLAG_sleepstudy = 1 and FLAG_VSHORT = 0 and BKGRD1_C7 ne . and SLPA54 ne .);
  Note: Variable for subpopulation analyses

Indicator variables (For non-missing observations: 1 if parenthesis is true and 0 otherwise)

- **AHI3p_GE5:** (slpa54 >= 5)
- **AHI3p_GE10:** (slpa54 >= 10)
- **AHI3p_GE15:** (slpa54 >= 15) (RENAME AHI_GE15 to AHI3p_GE15)
- **AHI3p_GE30:** (slpa54 >= 30)

- **INCOME_LT_30K:** (ECEA2=1) because it has less missing (~460) than INCOME_C5 from part_derv

Recode variables:

- **Snoring_c3 (3-level nominal variable)**
  0 "Don't know" if slea13 = 9
  1 "3-7 nights a week" if slea13 in (3, 4)
  2 "0-2 nights a week" if slea13 in (1, 2)
  Source variable(s): SLEA13

- **Knows_snores (Yes/No)**
  0 "Do not know if snores" if slea13 = 9
  1 "Do know whether snores or not" if slea13 in (1, 2, 3, 4)
  Source variable(s): SLEA13

- **Stopbreathing_c3 (3-level nominal variable)**
  Recode it similarly to snoring_c3 for SLEA13
  Source variable(s): SLEA14

Due to small cell counts for models in tables 4 and 5

- **AGEGRP2_C5** (combine 60-70 and 70-74)
  if 18<=age<30 then AGEGRP2_C5=1;
  else if 30<=age<40 then AGEGRP2_C5=2;
  else if 40<=age<50 then AGEGRP2_C5=3;
  else if 50<=age<60 then AGEGRP2_C5=4;
  else if 60<=age then AGEGRP2_C5=5;

- **BMIGRP_C3** (combine underweight and normal)
  if BMIGRP_C4 in (1,2) then BMIGRP_C3=2;
  else BMIGRP_C3= BMIGRP_C4;

3.2 Intermediate output, tables and analysis

- **Table of exclusion/inclusions**. Starting with 16,415 exclude:
  - No sleep study (flag_sleepstudy)
  - Missing background
  - Very short sleep study (flag_vshort)
  - Missing SLPA54 (MAIN outcome)

- PROC CONTENTS for permanent analysis dataset
- PROC MEANS for all variables in the permanent analysis dataset
- PROC FREQ for all categorical variables in the permanent analysis dataset

- DESCRIPTIVE STATISTICS FOR MANUSCRIPT (regular SAS output)
  By site:
  - % missing sleep studies : FREQ flag_sleepstudy
  - % very short studies (<30 min): FREQ flag_vshort
  - % short studies (< 4 hrs): FREQ flag_short
  - Recorded time (SLPA30)
    Statistics: min, Q1, median, Q3, max, mean and SD: overall and by site, gender, agegroup_c6_nhanes, and background

For Tables 1, 2 and 3 the columns are the Hispanic/Latino background plus an overall column. The rows are specified below. In all the analyses, use the SUBPOPN or DOMAIN statements with KEEP_MS13=1 to include only those with valid recording time SLEA30 > 30min and non-missing SLPA54**.**

Table 1. Age-adjusted demographic, anthropometric and health characteristics by Hispanic/Latino background
Table 1F. Same title as Table 1, Females
Table 1M. Same title as Table 1, Males

ESTIMATES SHOULD BE AGE-ADJUSTED (Include a footnote stating it). For continuous variables include mean and SE and for categorical variables include % and SE. Assess % missing and if overall > 5% then please create a missing category (EXCEPT for variables where missing is due to a SKIP PATTERNS like CURRENT OCCUPATION)

Age (yr)
Female (%)
Marital status
Education - EDUCATION_C3
Income (use < $30K and >= $30k levels)
Employment status - EMPLOYED
Current occupation - OCCUPATION_CURR
Current cigarette smoker (%) – (Cigarette_Use=3)
Former cigarette smoker (%) – (Cigarette_Use=2)
Never cigarette smoker (%) – (Cigarette_Use=1)
Current alcohol use (%) – (Alcohol _Use=3)
Former alcohol use (%) – (Alcohol _Use=2)
Never alcohol use (%) – (Alcohol _Use=1)
BMI
BMIGRP_C4
Waist Circumference (cm)
Diabetes (%) - DIABETES2 (3-level ADA)
Dyslipidemia from part_derv
Hypertension (%) - HYPERTENSION (BP≥140/90 and scanned medication use)
Asthma Diagnosed by MD (%) - Asthma_C4_MD (4-level by MD diagnosis)
FEV_1 % Predicted - PRBA29
FEV1 / FVC - FEV1_FVC_RATIO
Total cholesterol - LABA66
HDL cholesterol - LABA68
Triglycerides - LABA67

NOTE: For derived variables in part_derv for which there are several definitions (e.g. diabetes) make sure you specify which definition you are using. Also, make sure you describe it in the methods section.

Table 2. Age and BMI adjusted indices of sleep disordered breathing by Hispanic/Latino background
Table 2F. Same title as Table 2, Females
Table 2M. Same title as Table 2, Males

ESTIMATES SHOULD BE AGE AND BMI ADJUSTED (Include a footnote stating it)

1. AHI >= 5, DERIVED VARIABLE
2. AHI >= 10, DERIVED VARIABLE
3. AHI > = 15, AHI3p_GE15
4. AHI > = 30, DERIVED VARIABLE
5. Sleep duration
    a. Weekday
    b. Weekend
    c. Average
6. Snoring_c3 – DERIVED VARIABLE
7. StopBreathing_c3 – DERIVED VARIABLE
8. Epworth Sleepiness Scale (ESS) - ESS
9. Excessive Sleepiness (ESS > =10)- ESS_GE10
10. SDB5_3p (AHI3p>=5 and ESS>=10) DERIVED VARIABLE
11. SDB15_3p (AHI3p>=15 and ESS>=10) DERIVED VARIABLE

Table 4. Adjusted OR (95% CI) between Risk Factors and Sleep Apnea (AHI>=15).

Fit the following logistic regression models and provide OR and 95% CI in table for FINAL MODEL; table shell is provided at the end of the request.

Outcome: AHI3p_GE15 (i.e. Sleep Apnea; AHI (3% desat) > = 15)
Main effects (assess each one separately): snoring_c3, stopbreathing_c3, ESS_GE10, BMIGRP_C3, diabetes, hypertension

Effect modifier: male or BKGRD1_C7
Adjust for: male, BKGRD1_C7, AGEGRP2_C5, EDUCATION_C2, marital status and site

Logit{AHI3p_GE15} = Main effect + covariates

/* SNORING (Models 1 to 3) */
Model 4_1. snoring_c3 + covariates (FINAL MODEL)
Model 4_2. snoring_c3 + snoring_c3*male + covariates
Model 4_3. snoring_c3 + snoring_c3 *BKGRD1_C7 + covariates

/* STOP BREATHING (Models 4 to 6) */
Model 4_4. stopbreathing _c3 + covariates (FINAL MODEL)
Model 4_5. stopbreathing _c3 + snoring_c3*male + covariates
Model 4_6. stopbreathing _c3 + snoring_c3 *BKGRD1_C7 + covariates

/* SLEEPINESS (ESS) (Models 7 to 9) */
Model 4_7. ESS_GE10 + covariates
Model 4_8. ESS_GE10 + ESS_GE10*male + covariates
Model 4_9. ESS_GE10 + ESS_GE10*BKGRD1_C7 + covariates (FINAL MODEL)

/* BMIGRP_C3 (Models 10 to 12) */
Model 4_10. BMIGRP_C3 + covariates (FINAL MODEL)
Model 4_11. BMIGRP_C3 + BMIGRP_C3 *male + covariates
Model 4_12. BMIGRP_C3 + BMIGRP_C3 *BKGRD1_C7 + covariates

/* DIABETES2 (Models 13 and 14) */

Model 4_13.
Logit{AHI3p_GE15} = background + site + AGEGRP2_C5+ male + EDUCATION_C2 + marital status + diabetes2

Model 4_14.
Model 4_13 + BMIGRP_C3 + Waist circumference

/* HYPERTENSION (Models 15 and 16) */

Model 4_15.
Logit{AHI3p_GE15} = background + site + AGEGRP2_C5+ male + EDUCATION_C2 + marital status + hypertension

Model 4_16.
Model 4_14 + BMIGRP_C3 + Waist circumference

Table 4. Adjusted Odds Ratio (95% CI) between Risk Factors and Sleep Apnea (AHI>=15).

	Crude OR	95% CI	AOR	95% CI
Habitual Snoring				
0-2 nights a week	1		1	
3-7 nights a week	Crude OR only include risk factor as covariate			(LL, UL) From Model 4_1
Don't know				From Model 4_1
Stop breathing				
0-2 nights a week	1		1	
3-7 nights a week				From Model 4_4
Don't know				From Model 4_4
ESS_GE10& by Hispanic/Latino background				
Dominican	OR between SA and ESS_GE10 by background		AOR between SA and ESS_GE10 by background	From Model 4_9
Central American				From Model 4_9
Cuban				From Model 4_9
Mexican				From Model 4_9
Puerto Rican				From Model 4_9
South American				From Model 4_9
More than one/ Other				From Model 4_9
BMI Group				
Underweight & normal	1		1	
Overweight				From model 4_10
Obese (BMI >=30)				From model 4_10
Hypertension#				
No	1		1	
Yes				From model 4_16
Diabetes#				
Normal glucose regulation	1		1	
Impaired glucose tolerance				From model 4_14
Diabetes				From model 4_14
Gender$				
Female	1		1	
Male				From Model 4_1
Age group$				
18-29	1		1	
30-39				From Model 4_1
40-49				From Model 4_1
50-59				From Model 4_1
60-74				From Model 4_1

Adjusted Odds Ratio (AOR) from a logistic regression model for AHI_GE15 and risk symptom adjusted for age, gender, Hispanic/Latino background, education, marital status, and site.

& Interaction between ESS_GE10 and Hispanic/Latino background was significant

Adjusted Odds Ratio (AOR) from a logistic regression model for AHI_GE15 and risk symptom adjusted for age, gender, Hispanic/Latino background, education, marital status, BMI group, wait circumference, and site.

$ From model including snoring as a risk factor