Guidelines for HCHS/SOL Manuscript Verification
Version 2.0 (September 4, 2020)

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 and HCHS/SOL Coordinating Center (CC) has been verifying HCHS/SOL manuscripts since 2014. Originally the CC verified at least one manuscript using only baseline (2008-2011) data from each analyst working at the field center. Starting 2020, the CC will only verify manuscripts involving visit 2 (2014-2017) data. Each analyst will have an initial review of their first work to be submitted for publication. Subsequent work will be randomly selected for review over time by the Coordinating Center. Note that manuscripts using Ancillary Studies data and manuscripts using only visit 1 data are not verified by the CC. Lead authors should work with their local statistical analysis resources and their writing group members for critique and review of the results presented in their papers and/or presentations. The most common errors detected in papers selected for verification were either use of an obsolete dataset, or failure to correctly specify the analysis domain group of interest when employing survey statistical methods described in HCHS study documentation, *"HCHS Visit 2 Analysis Methods Ver 2.0 July 2020"*. See the study website Statistics and Data Analysis page for this manual and other references.

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 Committee 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 Project Director (cc CC coauthors) 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. It takes about two weeks for data verification after the Publications Committee's approval of the manuscript.

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

- Roadmap of documentation including statistical analysis plan
- Final version of the manuscript after writing 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 Committee.

Example

We will illustrate documentation needed using published MS552 "Incidence of Hypertension Among US Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos, 2008 to 2017" by Dr. Elfassy.

Elfassy T, Adina Zeki Al Hazzouri, Cai J, Baldoni PL, Llabre MM, Rundek T, Raij L, Lash JP, Talavera GA, Wassertheil-Smoller S, Daviglus M, Booth 3rd JN, Castaneda SF, Garcia G, Schneiderman N. Incidence of Hypertension Among US Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos, 2008 to 2017. *J Am Heart Assoc.* 2020 Jun 16;9(12):e015031. PMID: 32476602 PMCID: PMC7429033

MS552 Manuscript Verification Request

Manuscript Number: MS52
Lead Author: Tali Elfassy
Statistical Analyst: Tali Elfassy
Manuscript or Project Short Title (Length 26): hypertension Incidence
Date Submitted: January 29th 2019
Files submitted:

- SOL MS 552_For Data Verification and Journal Submission.docx
- MS 552 Data Verification_01_29_2019.sas
- MS 552 Data Verification_01_29_2019.log
- MS 552 Data Verification_01_29_2019.lst

1. Describe data files to be used:

Baseline data (INV4)
PART_DERV: Participant Derived Variables
ANTA: Anthropometry
LABA: Lab Data
PIEA: Questionnaire Data

Visit 2 data (INV3)
PART_DERV_V2: Participant Derived Variables

2. Exclusions/Inclusions:

Starting with 16,415 exclude:

- Anyone who did not participate in visit 2 (n=4,792)
- Anyone missing baseline information on SBP, DBP, self-reported HTN meds (n=66)
- Anyone with baseline hypertension (n=5,300)
- Anyone missing visit 2 information on SBP, DBP, self-reported HTN meds (n=76)

Final Sample size N=6,181

3. Statistical Analyses Methods

All analyses accounted for the complex survey design including cluster sampling and the use of stratification in sample selection. The sampling weights were calculated based on the baseline sampling weights and accounted for the non-responses for examination 2. The sampling weights are trimmed, calibrated to the age, sex, and Hispanic/Latino background distributions from the 2010 US Census for the 4 study field centers based on participants' age at the first examination, and normalized.

Among the population free from hypertension, we first described baseline demographic, socioeconomic, clinical, and behavioral characteristics. We calculated mean follow-up time beginning at the date of each participant's first study visit (baseline examination) and ending at the date of the second study visit (follow-up examination). Using Poisson regression through generalized linear regression models, we estimated incidence rates (IRs) of hypertension, per 1000 person-years (PY), as well as incidence rate ratios (IRRs) overall and according to baseline demographic, socioeconomic, clinical, and behavioral factors of interest. Except for the estimates reported across age groups, all IRs and IRRs were age adjusted to the overall HCHS/SOL baseline age distribution. We also estimated IRRs fully adjusted for: age, Hispanic/Latino background, education, income, nativity/years in the United States, health insurance, BMI, elevated SBP, smoking, physical activity, and AHEI 2010.

4. Description of TABLES and FIGURES for Manuscript:

Table 1: Age Adjusted Characteristics of the sample by sex. The columns are sex specific. The rows are specified below. In all the analyses we used the subpopulation statement with keep=1 to include only those in the analytical sample. The age weights we have used are based on the overall SOL population in the analytical sample---see the overall estimates in table 1. Use the variable we created above called "age4" to standardize

```
18-34:  53.5640
35-49:  31.9284
50-64:  12.3113
65+:    2.1963;
```

Table 2: Incidence of hypertension among US Hispanic/Latino men in the HCHS/SOL target population by baseline characteristics.

This table has the same rows as Table 1 above. However, it is restricted to men and provides the incidence rate of hypertension per 1,000 person years, the age adjusted incidence rate ratio of hypertension (with respective reference categories specified) and fully adjusted incidence rate ratios. Fully adjusted models include all of the variables in Table 1.

This table includes the subpopulation statements of keep=1 and female=0

Table 3: Incidence of hypertension among US Hispanic/Latino women in the HCHS/SOL target population by baseline characteristics.

This table is the same as Table 2 above but provides the estimates for women.
This table includes the subpopulation statements of keep=1 and female=1

Table 4: Prevalence of US Hispanics/Latinos recommended for hypertension treatment, treated, and controlled in the HCHS/SOL target population by sex.

This table provides the prevalence of: hypertension treatment recommendation (among all and among all with hypertension), hypertension treatment and hypertension control (among individuals recommended for treatment of hypertension). The table is stratified by sex and also provides individuals estimates for each Hispanic/Latino background group. 95% confidence intervals are provided as well as comparisons using Mexican Americans as the reference category.

FIGURE 1: Six year cumulative incidence of hypertension among US Hispanic/Latino men and women, by Hispanic/Latino background group, HCHS/SOL 2008-2017.

This graph represents the cumulative incidence of hypertension between visit 1 and visit 2 stratified by sex and by Hispanic/Latino background groups. Each cumulative incidence rate is compared to Mexican Americans as the reference categories. The error bars on the graph represent the 95% confidence interval band. These estimates are age adjusted to the distribution provided above.

Supplemental Table 1: Incidence of JNC7 defined hypertension among US Hispanic/Latino men and women in the HCHS/SOL target population.

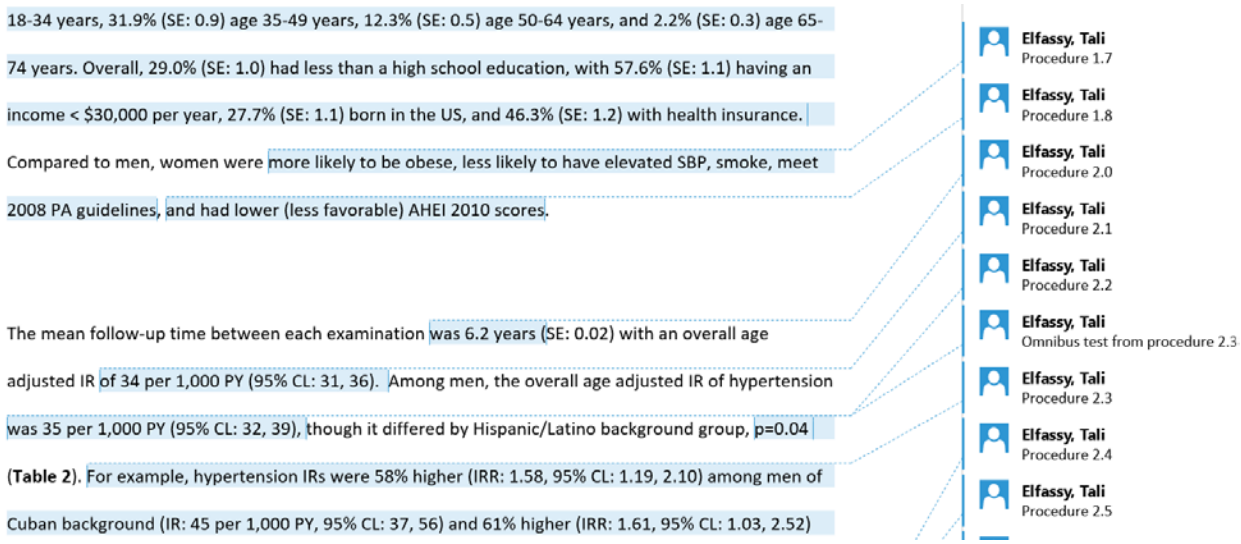
This table provides the age adjusted cumulative incidence, incidence rate, and incidence rate ratio of hypertension (defined by the previous hypertension definition). The table is stratified by sex and estimates are provided for each Hispanic/Latino background group with Mexican Americans as the reference category. The cumulative incidence estimates are age adjusted to the distribution provided above.

Rows of Table 1: ALL VARIABLES ARE % EXCEPT FOR AHEI2010 SCORE (MEAN)

	Overall			Men			Women		
	Unweighted			Unweighted			Unweighted		
	N	%	SE	N	%	SE	N	%	SE
Hispanic/Latino background									
Mexican									
Central American									
Cuban									
Dominican									
Puerto Rican									
South American									
Mixed/Other									
Women									
Age group									
18-34									
35-49									
50-64									
65-74									
Less than HS education									
Income < $30,000									
Nativity/years in the US									
For. born < 10 yrs. in US									
For. born, 10+ yrs. in US									
US born									
Health insured									
Body mass index									
Normal									
Overweight									
Obese									
Elevated SBP									
Smoking									
Never									
Current									
Former									
Meets 2008 PA guidelines									
AHEI 2010, mean									

ANNOTATED MANUSCRIPT

Excerpt from manuscript:



Corresponding excerpt from SAS code:

```
****PROCEDURE 1.7****;
proc descript data = dat filetype = sas design = WR nomarg;
subpopn keep=1;
      nest  strat psu;
      weight WEIGHT_NORM_OVERALL_V2;
      .
      .
      .
 output / FILENAME=table3 FILETYPE=SAS TABLECELL=DEFAULT REPLACE;
run;

****PROCEDURE 1.8****;
proc descript data = dat filetype = sas design = WR nomarg;
      .
      .
      .
run;
```