HCHS/SOL Manuscript Writing Recommendations

November 18, 2016

Version 2.1

Prepared by the
HCHS/SOL Coordinating Center
Collaborative Studies Coordinating Center
UNC Department of Biostatistics
Jianwen Cai
Daniela Sotres-Alvarez
Sonia Davis
Franklyn Gonzalez II
Natalia Gouskova
William Kalsbeek
Donglin Zeng
Marston Youngblood

Recommendations for HCHS/SOL Manuscript Writing

Preface

This document was prepared by the biostatisticians at the HCHS/SOL Coordinating Center as an aid to authors and writing groups working on manuscripts that use the study data. We have organized its layout to follow the standard sections for publications so that the user can relate this material to their own work in progress. Our comments and explanations have been developed to address specific questions commonly raised by journal editors and reviewers of HCHS/SOL publications.

Title

- The title must include the full name of the study "Hispanic Community Health Study/Study of Latinos" in complete form or abbreviated as HCHS/SOL.

Abstract

- Include full name of the study "Hispanic Community Health Study / Study of Latinos"
- Provide sample size, age group, and enrollment dates
- Provide design. Important terms: population-based, cohort, probability sample

Methods

1. Sampling Design:
   The Hispanic Community Health Study (HCHS)/Study of Latinos (SOL) is a community based prospective cohort study of 16,415 self-identified Hispanic/Latino persons aged 18-74 years at screening from randomly selected households in four U.S. field centers (Chicago, IL; Miami, FL; Bronx, NY; San Diego, CA) with baseline examination (2008 to 2011) and yearly telephone follow-up assessment for at least three years. HCHS/SOL cohort includes participants who self-identified as having Hispanic/Latino background, the largest groups being Central American (n=1,732), Cuban (n=2,348), Dominican (n=1,473), Mexican (n=6,472), Puerto-Rican (n=2,728), and South American (n=1,072). The goals of the HCHS/SOL are to describe the prevalence of risk and protective factors for chronic conditions (e.g. cardiovascular disease (CVD), diabetes and pulmonary disease), and to quantify all-cause mortality, fatal and non-fatal CVD and pulmonary disease, and pulmonary disease exacerbation over time. The baseline clinical examination (Sorlie et al, 2010) included comprehensive biological (e.g., anthropometrics, blood draw, oral glucose tolerance test, ankle brachial pressure index, electrocardiogram), behavioral (e.g. dietary intake assessed with two 24-hour recalls, physical activity assessment by accelerometer and self-report, overnight sleep exam for apneic events, tobacco and alcohol assessed by self-report), and socio-demographic (e.g., socioeconomic status, migration history) assessments.

The sample design and cohort selection has been previously described (LaVange et al, 2010). Briefly, a stratified two-stage area probability sample of household addresses was selected in each of the four field centers. The first sampling stage randomly selected census block groups with stratification based on Hispanic/Latino concentration and proportion of high/low socio-economic status. The second sampling stage randomly selected households, with stratification, from US Postal Service registries that covered the randomly-selected census block groups. Both stages oversampled certain strata to increase the likelihood that a selected address yielded a Hispanic/Latino household. After households were sampled, in-person or telephone contacts were made to screen eligible households and to roster its members. Lastly, the study oversampled the 45-74 age group (n=9,714, 59.2%) to facilitate examination of target outcomes. As a result, participants included in HCHS/SOL cohort were selected with unequal probabilities of selection, and these probabilities need to be taken into account during data analysis to appropriately represent the target population. HCHS/SOL sampling weights are the product of a "base weight" (reciprocal of the probability of selection) and three adjustments: 1) non-response adjustments made relative to the sampling frame, 2) trimming to handle extreme values (to avoid a few weights with extreme values being overly influential in the analyses), and 3) calibration of weights to the 2010 U.S. Census according to age, sex and Hispanic background.

2. Shorter Version for Referring to Sampling Weight:
   All reported values (means and prevalence rates) were weighted to account for the disproportionate selection of the sample and to at least partially adjust for any bias effects due to differential nonresponse in the selected sample at the household and person levels. The adjusted weights were also trimmed to limit precision losses due to the variability of the adjusted weights, and calibrated to the 2010 Census characteristics by age, sex and Hispanic background in each field site's target population. All analyses also account for cluster sampling and the use of stratification in sample selection.

3. Target Population:
   The HCHS/SOL target population is defined as all non-institutionalized Hispanic/Latino adults aged 18-74 years and residing in the defined geographical areas (census block groups) across the four participating field centers. The choice of the census block groups was designed to provide diversity among cohort participants with regard to socioeconomic status and national origin or background. HCHS/SOL participants were selected using a probability sampling design within these areas to provide a representative sample of the target population.

4. Response Rate:
   Household-level response rate was 33.5%. Of 39 384 individuals who were screened and selected and who met eligibility criteria, 41.7% were enrolled, representing 16 415 persons from 9872 households. This study was approved by the Institutional Review Boards at the data coordinating center and at each field center where all subjects gave written consent.

5. Hispanic/Latino definition used in HCHS/SOL:
   In order to recruit Spanish or English speaking individuals who self-identify with belonging to a Hispanic/Latino background, participants were asked the following question during screening: “Do you consider yourself to be Hispanic/Latino?” (See the Eligibility form QxQ: “The populations of interest for HCHS/SOL are persons or descendants of persons from Cuba, Mexico, Puerto Rico, and Spanish speaking countries in the Caribbean and Central and South America. A complete list of countries of interest is provided in Appendix I. This list is provided as a reference tool for recruiters and is NOT to be read during recruitment visits nor shown to respondents.”)

6. Field center and background adjustment
   Clearly describe whether field center was adjusted for and if so how. For example, the following statements describe the situation when the final analysis step was an adjustment for study center: The fact that people with specific Hispanic/Latino backgrounds tend to concentrate in specific geographic areas meant that not all backgrounds were present in each study center, creating confounding between background and center. In particular, Cubans were predominantly in Miami, Dominicans were predominantly in the Bronx, and participants from San Diego were predominantly Mexican. We therefore examined possible center effects within background by fitting additional multivariable regression models adjusting for a 17-level background-by-center interaction variable in place of the background variable, with levels corresponding to the 13 combinations of center and background that had 100 or more participants in the analysis sample, and one combined level per center for the mixed/other background category plus all other cells with count < 100.

7. Age adjustment and age standardization:
   For internal age adjustment provide mean age that is being adjusted to. For age standardization provide population and year of reference (e.g. US 2010 Census Population). When the association is not linear, using age groups is preferred because this allows for better model fit and the interpretation of age effects is easier with age groups than with non-linear terms.

8. The choice of conditional and predicted marginals:
   Conditional marginals can be helpful when the goal is to adjust estimates to a particular mean value in order to compare prevalence externally, such as to a different population or a different study. Predicted marginals adjust to the distribution of the target population, and are most helpful for internal comparisons of subgroups within the target population. In addition, the predicted marginal tends to be closer to the prevalence estimate from the linear regression. When planning and reporting prevalence estimates using logistic regression, authors should clearly specify which of the two marginal estimates will be reported, and clearly describe their interpretation.

Recommended Wordings for Linear[Logistic] Model:

For continuous [categorical] response, age-adjusted means [prevalence estimates] for the target population of Hispanic/Latinos in the 4 HCHS/SOL communities were calculated using survey linear regression weighted least squares [logistic regression: predicted <conditional> marginals], adjusting each subgroup to the age distribution of the target population <age of 60>.

Note: contents in [ ] are for logistic regression. < > indicates alternative wordings.

7. Referencing statistical software:
   All the analyses were performed using SAS 9.3 software (SAS Institute, Cary, NC) and/or SUDAAN software Release 11 (RTI International, Research Triangle Park, NC) <or Stata Statistical Software, Release 13 (StataCorp LP, College Station, TX)>.

   Note: All commercial statistical software used for analyses must be referenced, typically at the end of the Statistical Analysis section. It is a good idea to also reference any free software which was used, such as R packages, publicly available SAS macros, etc: “Predicted nutrient intake was computed using NCI macros MIXTRAN and INDIVIDINT <or R package locfit()>.”

Results

1. The Sample Size:
   For socio-demographic characteristics (typically in table 1) unweighted sample sizes can be presented. Do not present unweighted percentages in tables or text.

2. Sampling Weights:

   When presenting results, be specific and clear whether these refer to the specific sample we happen to observe or the target population. For example, the cohort (or analytic sample) consisted of 9,835 women. In the target population, 52.3% were female and the overall mean age was 41 years.

   Note that HCHS/SOL sampling weights are calibrated (age, gender and Hispanic/Latino background) to the US 2010 Census within the specific HCHS/SOL target areas whereas conducting external age standardization to the US 2010 Census refers to the United States age distribution. However, note that HCHS/SOL estimates after external standardization to the US 2010 age-distribution do not generalize to the entire US Hispanic/Latino population but rather to the Hispanic/Latino population living in the target areas had they followed the same age-distribution as from the US 2010 Census.

Discussion

1. Target Population Representativeness (Note: the first part of the following paragraph is usually included in the Methods section. The part in italics is more appropriate in the Discussion section to address the concerns of the representativeness of the HCHS/SOL sample.)

   The HCHS/SOL cohort was selected through a stratified multi-stage area probability sample of four communities (LaVange et al., 2010). The probability-based sampling allows HCHS/SOL to estimate prevalence of diseases and baseline risk factors in the target population, which includes all non-institutionalized Hispanic/Latino adults aged 18-74 years residing in the four defined community areas. The selected communities are diverse regions of the US, each with high concentrations of specific Hispanic/Latino backgrounds, allowing the study to estimate prevalence of diseases and risk factors for each background. *Although the target population is limited to the four communities rather than the entire nation, HCHS/SOL's hybrid design, which uses probability sampling within pre-selected diverse regions, is superior to the convenience samples which are typically exploited in epidemiological cohort studies.*

2. Addressing Low Response Rate

   HCHS/SOL study employed a probability sampling design. A stratified two-stage area probability sample of household addresses was selected from subjectively designated Hispanic neighborhoods defined by the set of census tracts serving as the sampling target population in each of the four field centers. After households were randomly selected, in-person visits and/or telephone contacts were made to screen eligible households and to roster its members. The household-level response rate was 33.5%. Of 39 384 individuals who were selected, screened and met eligibility criteria, 41.7% were enrolled, representing 16 415 persons from 9872 households. Even though the response rate is low, a widely accepted statistical adjustment protocol was followed to reduce the potential bias of estimates due to study non-participation. To minimize this bias effect while controlling the precision loss implications of adjustment, the sample weight of each participant was: (1) calculated based on its selection probability; (2) adjusted for differential non-response at the household and person levels, and trimmed to reduce the variability of the adjusted weights; and (3) calibrated to the 2010 U.S. Census count by age, gender, and Hispanic background in each field center's target population. This three-step approach to calculate sample weights is consistent with weighting strategies used in all major health surveys utilizing probability sampling (e.g., NHANES, NHIS, and MEPS). Thus, as with other comparable population-based sample, to address the potential bias in the respondent sample, sample weights should be used in the analysis. In other words, analysis using sample weights helps to correct for the potential bias associated with low response rate.

Cross-Sectional Results

When reporting cross-sectional results in a paper, avoid the words “outcome” and predictors”. Instead of saying “outcome” you could say “dependent variable” or “occurrence of disease,” or “occurrence of the condition". Since this is cross-sectional, you cannot infer prediction. Instead of “predictors”, you should say “variables associated with….” as for example: “Variables associated with the health condition”. Or you can refer to these as independent variables. The main idea is to avoid the implication of directionality of the relationship.

Acknowledgement

The authors thank the staff and participants of HCHS/SOL for their important contributions. A complete list of staff and investigators has been provided by Sorlie P., et al. in Ann Epidemiol. 2010 Aug;20: 642-649 and is also available on the study website http://www.cscc.unc.edu/hchs/

FUNDING

See study website posted document titled “Special Notes for Primary Authors.”

DISCLOSURE

Drs. *X and Y* had full access to the study data and take responsibility for the integrity of the data and accuracy of analyses. All authors have reviewed and approved the final manuscript. None of the authors had any financial or other conflicts of interest.

References

LaVange LM, Kalsbeek WD, Sorlie PD, Aviles-Santa LM, Kaplan RC, Barnhart J, et al. Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. Ann. Epidemiol. 2010;20(8):642-49

Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, Kaplan RC, Daviglus ML, Giachello AL, Schneiderman N, Raij L, Talavera G, Allison M, Lavange L, Chambless LE, Heiss G. Design and implementation of the Hispanic Community Health Study/Study of Latinos. Ann Epidemiol. 2010 Aug;20(8):629-41.

Recommendations for Good Practices for Statistical Analysts:

i. **Familiarize with HCHS/SOL documentation and check website for updates:**
- Analysis Methods Manual
- Overview documents: Investigator Use Database, Dietary Data, Physical Activity
- Codebooks (baseline examination, diet data, physical activity) with unadjusted descriptive statistics and frequencies of response items.
- Data Dictionaries (participant derived, actical data, audiometry) describe HCHS/SOL derived variables.
- Guidelines for HCHS/SOL Manuscript Verification
- Recommendations for manuscript writing
- MEMOS from different working groups providing findings or recommendations

ii. Checking for missing numeric values (., .L, .M, .N)
syntax differences: “EQ .” vs. “LE .Z”

iii. Check the data distribution for variables used in the analysis (obtain frequencies for categorical variables, and at least mean, SD, minimum, maximum for continuous variables). Check the number of missing observations for each variable.

iv. Use of permanent analysis file instead of run-time datasets

v. Use derived file first (PART_DERV), instead of re-creating variables

vi. Handling missing covariates (race, income, education)

Recommendations for Good Practices for Manuscript Writing:

- When describing results based on the weighted analysis, do not use the word “participant” because the inference is to the target population. One suggestion is to use “individual”.
- Do not include and describe variables that are not discussed later in the paper.
- Focus the discussion on the *final* model.
- Make clear when discussing weighted or unweighted results. Restrict the unweighted results to unweighted N only.
- Interpretation of results needs to be consistent throughout the text.
- Repeat key terms.
- Involve your coauthors at all phases of the manuscript.
- Keep coauthors informed about the timeline for publication and the submission process.

Some Useful Contents that the Readers Might Want to Know:

- Number of persons excluded and whether they were equally distributed among Hispanic/Latino groups.
- Reliability and validity of the survey tools used
- How do the results compare to the national data (e.g. NHW, AA)?
- How do the data compare to other publications on Latinos in the U.S.?
- May need to discuss potential confounding variables that were not measured or were not included in the analysis.

Some Background Information about the Study Design and Sample Weights:

1. NIH contract specifications for HCHS called for a design utilizing probability sampling, which implies sample identification based on the use of random methods applied to a well-defined target population. The study was designed based on the NIH specifications, and with basic features included to: (1) control costs (by oversampling higher Hispanic concentrations in the target area to increase screening efficiency), and (2) seriously oversample older Hispanics (through the use of Bernoulli sampling for within-household selection of individuals), The design is a probability sample of a specific target population where steps were taken to deal with the nonresponse, which occurs in all population-based studies.
2. Normalization of the sample weight does not affect the size of any reported estimates and is done primarily for the benefit of the statistical analyst (and not the reader) to avoid reporting exaggerated statistical significance levels in certain types of analysis protocols where degrees of freedom are determined by sums of respondent weights.

Decisions by Steering Committee and Statistical Methods Subcommittee:

1. Include the 9 participants who are over 75 at the time of the clinic visit. The age eligibility criterion is that they are 18-74 years old at screening. Only exclude participants based upon age (any age, not just ages 75+) if the reference measures used in the analysis of interest only apply to a specific age group (ex. Framingham CVD 10-year risk population of interest are ages 30-74 only) or when presenting age-standardized to US Census population as the last age group reference is < 75 years.