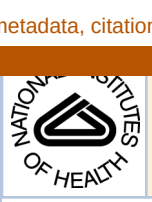


Published in final edited form as:

Ann Epidemiol. 2010 August ; 20(8): 642–649. doi:10.1016/j.annepidem.2010.05.006.

Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos

Lisa M. LaVange, Ph.D.[1], **William Kalsbeek, Ph.D.**[2], **Paul D. Sorlie, Ph.D.**[3], **Larissa M. Avilés-Santa, M.D. M.P.H.**[3], **Robert C. Kaplan, Ph.D.**[4], **Janice Barnhart, M.D., M.S.**[4], **Kiang Liu, Ph.D.**[5], **Aida Giachello, Ph.D.**[6], **David J. Lee, Ph.D.**[7], **John Ryan, Dr.P.H.**[8], **Michael H. Criqui, M.D., M.P.H.**[9], and **John P. Elder, Ph.D., M.P.H.**[10]

[1]Collaborative Studies Coordinating Center, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

[2]Survey Research Unit, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

[3]Division of Cardiovascular Sciences, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD.

[4]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY.

[5]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL.

[6]Midwest Latino Health Research, Training & Policy Center, Jane Addams College of Social Work, University of Illinois at Chicago (UIC), Chicago, IL.

[7]Department of Epidemiology & Public Health, Sylvester Comprehensive Center, University of Miami, Miami, FL.

[8]Department of Family Medicine and Community Health, University of Miami, Miami, FL.

[9]Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, CA

[10]Graduate School of Public Health, San Diego State University, San Diego, CA.

Abstract

PURPOSE—The Hispanic Community Health Study (HCHS)/Study of Latinos (SOL) is a multi-center, community based cohort study of Hispanic/Latino adults in the United States. A diverse participant sample is required that is both representative of the target population and likely to remain engaged throughout follow-up. The choice of sample design, its rationale, and benefits and challenges of design decisions are described in this paper.

METHODS—The study design calls for recruitment and follow-up of a cohort of 16,000 Hispanics/Latinos aged 18-74 years, with 62.5% (10,000) over 44 years of age and adequate

Correspondence: Lisa M. LaVange, Ph.D. CSCC, Department of Biostatistics, UNC-CH 137 E. Franklin St., Suite 203 Chapel Hill, NC 27514 .

subgroup sample sizes to support inference by Hispanic/Latino background. Participants are recruited in community areas surrounding four field centers in the Bronx, Chicago, Miami, and San Diego. A two-stage area probability sample of households is selected with stratification and over-sampling incorporated at each stage to provide a broadly diverse sample, offer efficiencies in field operations, and ensure that the target age distribution is obtained.

CONCLUSIONS—Embedding probability sampling within this traditional, multi-site cohort study design enables competing research objectives to be met. However, the use of probability sampling requires developing solutions to some unique challenges in both sample selection and recruitment, as described here.

Keywords

Probability sampling; Sampling diverse populations; Hispanic/Latino health

INTRODUCTION

The Hispanic Community Health Study (HCHS), a Study of Latinos (SOL), is a multi-center community based cohort study of Hispanics/Latinos in the United States (US). The study objectives are to provide information on the health status and disease burden of US Hispanics/Latinos and to investigate relationships between baseline risk factors and disease incidence during follow-up. A cohort of 16,000 Hispanics/Latinos aged 18-64 years will be enrolled, and upon completion of a comprehensive baseline examination, followed annually to determine the incidence of clinical events, including cardiovascular events and pulmonary exacerbations. The study is funded by the National Heart, Lung, and Blood Institute and six other Institutes, Centers, or Offices within the National Institutes of Health. Details of the study design and its various components are described elsewhere (Sorlie, et al., NHLBI, unpublished manuscript submitted jointly with this article). This paper describes the sample design used to identify and select households and persons for study participation.

Two distinct analytical objectives motivated the approach to sample selection. First, the study sample must support estimates of prevalence of baseline risk factors, both overall and by Hispanic/Latino background and other demographic subgroups. Second, the sample must support evaluation of the relationships between the various risk factors and disease outcomes measured during follow-up. To accomplish both objectives, a hybrid approach to cohort identification and selection is used that combines deliberate selection of community areas and random selection of households within those areas. The rationale for the use of probability sampling, details of the sample design, and the impact of the sampling strategy on the recruitment process are provided in the following sections.

MATERIALS AND METHODS

The four communities included in HCHS/SOL are located in the Bronx, Chicago, Miami, and San Diego. The sampled area in each community was defined by a group of neighboring census tracts to provide geographical balance and diversity with respect to Hispanic/Latino background. Each community's field center purposively selected its targeted tracts based on their proximity to the clinic, tract-level demographic distributions available from the 2000 decennial census; and local information about neighborhoods. The target population in HCHS/SOL corresponds to all non-institutionalized Hispanic/Latino adults aged 18-74 years residing in the four sampled areas. Probability sampling within these areas is employed to assure a broad representation of the target population and to minimize the various sources of bias that may otherwise enter into the cohort selection and recruitment process.

The Need for Probability Sampling

The design of a population-based sample must accommodate the specific informational needs of the study. For HCHS/SOL, the selected sample should be broadly representative of the target population in that the sample mirrors the full range of possible values for key outcome variables while also providing adequate representation of important combinations of predictor and outcome variables (1). Probability sampling provides a means for achieving such balanced representation. Probability sampling also provides a basis for making unbiased inference to target population characteristics of interest. These advantages, however, come at a cost. Probability sample requires the exclusive use of random selection so that the statistical probability of choosing each sample member can be calculated (2). Random selection requires enumeration of members of the target population, or well-defined subsets thereof, and can be costly to implement. As a result, study designs incorporating more convenient methods of selection are often utilized for population based cohort studies.

Any sample design that utilizes non-random selection (e.g., a convenience sample) produces a non-probability sample. Quasi-probability samples that combine random and non-random methods of selection (e.g., allowing interviewers to subjectively select a quota sample of households within a random sample of neighborhoods) are also non-probability samples (3,4). Accompanying the simplicity and lower cost associated with non-probability sampling are two problems. First, there is no direct theoretical basis for making estimates of population characteristics from the sample (5). Instead, one must either defend a model to explain the generation of the sample data from some underlying distribution or assume that the variability of sample based estimates is similar to that associated with simple random sampling. Both assumptions are difficult to verify. Second, non-probability samples,typically offer a skewed reflection of the sampled population due to diminished participation by population sectors (1). Self-selected samples exclude those more reluctant to volunteer and who are less accessible; allowing interviewers to decide who is selected can also exclude those not meeting personal preferences, leading to potentially biased study results. The magnitude of this bias is directly related to the extent of under-representation in the sample and the degree to which key study measurements on those included differ from those not included. While it is true that sources of error unrelated to sample selection (e.g., non-response) can bias the analysis of data from probability samples, non-probability samples are subject to these same non-sampling errors, producing estimates with bias due to both non-random selection and non-sampling sources (6).

To illustrate the potential for bias in non-probability samples, we compared health outcomes estimated from a national probability sample to outcomes estimated from a simulated clinic-users sample using data from the 2005 Medical Expenditure Panel Survey (MEPS). MEPS utilizes a national probability sample of all civilian, non-institutionalized U.S. residents (7). The subset of MEPS respondents reporting one or more physician visits in the past year (“clinic users”) mimics a convenience sample selected through physician practices alone. Estimates of the number of chronic conditions, average cost of physician visits, obesity prevalence, and number of work days missed due to illness/injury are provided in Figure 1 for the full sample and the clinic-users sample, both overall and by race/ethnicity. Sampling weights are incorporated in the analysis to account for disproportionate sampling of population subgroups in the MEPS study design. Estimates of the clinic-users sample standardized by age, race/ethnicity, and gender to the MEPS target population are also provided in Figure 1, in an attempt to adjust for skewness in the convenience sample. The findings reveal that estimates from the clinic-users sample are consistently higher than those from the probability sample. This selection bias is not unexpected due to the association between the criteria for selecting the clinic-users sample and the outcome measures (i.e., clinic users are more likely to have health problems than the population as a whole). The fact that standardizing estimates from the clinic users sample does not consistently

compensate for the skewed representation due to non-probability sampling, however, is unexpected. Standardization appears to offset the effect of selection bias for one outcome (chronic conditions), partially compensate for the effect in another (health care costs), and exacerbate the effect in the remaining two measures (obesity and days lost). For subgroup comparisons, the white-nonwhite difference in obesity prevalence for the clinic users sample overstates the actual difference, and standardization exacerbates this overstatement. While this example highlights the pitfalls of just one form of non-random selection, similar results would be expected for other forms of convenience sampling.

Rationale for Key Sample Design Features

A probability-based sampling strategy was chosen for HCHS/SOL, with specific features dictated by the goals and overall design of the study. First, the decision to identify Hispanics/Latinos from the general residential population made controlling the cost of face-to-face recruitment a priority. The mode of recruitment and data collection is an important cost factor in population-based studies, and while mail and web-based methods are inexpensive, nonsampling errors due to incomplete frame coverage and non-response can occur. Telephone screening is also relatively inexpensive but its exclusive use was impractical for HCHS/SOL due to the declining use of telephone land-lines and the fact that an extensive clinic visit is a key component of data collection. Face-to-face sample recruitment was seen as the only real option for HCHS/SOL; consequently, steps to control the associated higher costs were needed. One obvious cost-saving measure was to sample geographic clusters of households (i.e., census block groups) at the first stage of a multi-stage sample in order to reduce the cost of return visits to neighboring households. More substantial cost savings were realized through over-sampling of both clusters and households within clusters most likely to be Hispanic/Latino, thereby reducing the number of sampled households that must be screened to achieve the study's sample size goals. Geographic clusters were stratified by the proportion of the population found to be Hispanic/Latino in the 2000 decennial census, and clusters in the 'high concentration' stratum were selected at a higher rate than clusters in the "low concentration" stratum at the first-stage of sample selection. An optimal delineation point between high and low concentration was determined for each field center using Cochran's cumulative $\sqrt{f}$ rule (8). Similarly, household addresses within clusters were divided into two strata, those associated with Hispanic/Latino surnames versus all others. Hispanic/Latino surname addresses were selected at a higher rate than other addresses at the second stage. Over-sampling in multiple stages of the selection process in this way provides efficiencies in sample identification while still retaining the advantages of random selection.

Meeting the HCHS/SOL objectives requires adequate representation of the socio-economic status (SES) distribution of residents of the defined community areas. Although SES is an individual- or household- level characteristic, it is rarely possible to stratify a sample of households by a direct measure of SES. The next best option is to use census measures such as educational attainment or household income as a practical proxy indicator (9). To this end, geographic clusters were stratified by the proportion of residents aged 25 years or older with at least a high school education based on the 2000 census. The high and low SES delineation point was defined as the median value of the distribution across clusters, and the first-stage sample was allocated proportionately across strata to ensure broad SES representation. To meet the HCHS/SOL objective of identifying predictors of disease outcomes including cardiovascular events, a target sample size of 10,000 persons aged 45-74 years (62.5% of the full cohort) was set. Over-representation of this age group required sub-sampling households or persons within households according to the household's age distribution. Such a procedure is best applied during screening, with the intention of retaining a higher portion of discovered older Hispanics/Latinos than would

occur if persons were chosen at random. Sub-sampling according to age was accomplished in one of two ways. Method 1 was designed to keep all households intact, with no sub-sampling at the person level, and was adopted at study start. With this method, households in which the Hispanic/Latino adults are all aged 45-74 years are selected with certainty (probability of selection = 1) within the first-stage cluster, and all other households are sub-sampled with probability < 1. Method 2 involves dividing each household into two sub-clusters, Hispanics/Latinos aged 45-74 years and Hispanics/Latinos aged 18-44 years. The 45-74 year sub-clusters are selected with certainty (probability = 1), while the 18-44 year sub-clusters are selected with probability < 1. This method involves sub-sampling persons within a household rather than keeping households intact, but can result in fewer households needing to be screened and was adopted after study start for efficiency.

The final design consideration was the need to compare health characteristics by Hispanic/Latino background among the four field centers. Valid comparisons require comparability across sites in cohort recruitment, but not necessarily identical probability sample designs. Indeed, the same sample design structure was used (i.e., two-stage stratified sampling of households with the same sampling units and stratification variables in each stage), with some allowance for how the strata were defined and the sample allocated among the centers.

Sample Selection

A stratified two-stage area probability sample of household addresses was selected in each of the four HCHS/SOL field centers. A summary of the center-specific designs is presented in Table 1. At the first stage, a stratified simple random sample of census block groups (BGs), which served as primary sampling units (PSUs), was selected in each field center. PSU sampling strata were defined by the cross-classification of (i) high and low Hispanic/Latino concentration and (ii) high and low SES, defined above. The distribution of BGs across strata and the over-sampling ratios for high and low Hispanic concentration strata are presented in Table 2. Special strata were created as needed to target specific neighborhoods. In the Bronx, a fifth stratum was defined as a portion of a high-rise housing complex (named Co-op City) in order to provide additional income diversity, and two additional strata were appended after study start to increase coverage. In Miami, a fifth stratum was defined with high expected concentrations of Central and South Americans, and a sixth stratum corresponding to an area with a high concentration of Cuban residents was appended after study start. All BGs within these special strata were selected. Overall, 632 (73%) of the 871 BGs in the target areas were selected for the PSU sample.

Separate stratified second stage samples of household addresses were selected within each sample PSU. Address listings came from the Delivery Sequence File (DSF) available from the US Postal Service and obtained through MSG-Genesys of Ft. Washington, PA. The DSF addresses within each sample BG were cross-referenced with telephone and commercial mailing lists, and surname and telephone number were appended where available. Table 2 provides the second-stage over-sampling ratios for the Hispanic/Latino surname strata used to achieve the final sample of 123,213 addresses.

The sample addresses in each field center were randomly sub-sampled to form three waves corresponding to the three years of recruitment. Thus, the yearly sample for each field center was representative of the target community area, thereby minimizing bias due to temporal trends.

Design Modifications

A key feature of the HCHS/SOL sample design is the ability to modify components in order to adapt to recruitment experiences. The modifications made to date include the designation

of a sixth stratum in the Miami field center to append certain block groups in the Hialeah neighborhood for increased coverage of the Cuban population and designation of a sixth and seventh strata in the Bronx to capture a neighborhood adjoining the original target area, thereby increasing coverage of the Bronx Hispanic/Latino community.

Approximately six months into recruitment, a decision was made to apply Method 2 for over-sampling adults aged 45-74 years in lieu of Method 1, based on the need to accept a higher proportion of households into the sample and reduce recruitment time. The selection probabilities for both methods of over-sampling 45-74 year-olds during household screening were initially based on 2005 American Community Survey data for the geographic region of each field center. The sample age distribution is monitored continually as data on HCHS/SOL households accumulates, and the selection probabilities are adjusted as needed. Table 2 provides the sub-sampling rates for each method applied to each field center.

Sample Size and Data Analysis

Each field center will enroll 4,000 Hispanics/Latinos with the prescribed age distribution, namely 2,500 aged 45-74 years and 1,500 aged 18-44 years. In terms of Hispanic/Latino background, the Bronx field center sample is predominantly Puerto Rican and Dominican, while the majority of participants in the San Diego site are Mexican in origin. Study participants in the Miami field center are Cuban and Central/South American, and participants in the Chicago field center are Mexican, Puerto-Rican, and Central/South American. A minimum of 2,000 participants in each of the pre-specified Hispanic/Latino groups (Mexican, Puerto Rican, Cuban, and Central/South American) is required to support the analysis objectives, and sample sizes are monitored continuously to determine if adjustments to the sampling strategy are needed.

The HCHS/SOL sample size will support a broad range of analyses planned for the study. As an example, consider the possible association of an exposure variable with incident disease. The range of hazard ratios able to be detected with approximately 90 percent power are provided in Table 3 by event rate and the relative sample sizes of low to high risk groups. The estimates incorporate a design effect to account for clustering in the sample of 1.25, based on an average cluster size (persons per block group) of 24 and intra-class correlation for incident disease of 0.01. Based on the entire study cohort of 16,000, a hazard ratio of 1.6 would be able to be detected for an event occurring at the rate of 4 per 1,000 person years of follow-up and equally sized low and high risk groups, e.g., for a continuous exposure variable dichotomized at the median value. With a population subgroup of size 4,000 (e.g., a single site or Hispanic/Latino subgroup), the hazard ratio able to be detected in the same circumstances is 2.25. For higher levels of intra-class correlation, power for both comparisons would decrease.

The use of multi-stage or clustered sampling creates complexity in data analyses due to correlations among sample units at the various stages of selection, here, correlations among households within the same block group and correlations among individuals in the same household. Similarly, over-sampling through the use of differential probabilities of selection requires the use of sampling weights for unbiased estimation of population characteristics. While clustering and unequal probabilities of selection tend to increase the variability of population estimates and reduce the power available for testing associations, stratification at one or more stages of sample selection has the reverse effect. To ensure accurate estimation of variances and valid statistical tests of hypotheses therefore requires appropriately accounting for the HCHS/SOL sample design during data analysis. Initial sampling weights will correspond to the inverse probability of selection for each participant. Non-response adjustments and calibration to known population totals (from the 2010 decennial Census, when available) will be applied. Final sampling weights, stratification variables, and cluster

identifiers will be available for design specification during data analysis. A variety of statistical methods that account for multi-stage sampling are available (see, e.g., 10,11), and most standard statistical software packages are able to accommodate probability sample designs (e.g., SAS, STATA). Special purpose software (e.g., SUDAAN) for complex sample designs is also available.

Sample Recruitment

Successful implementation of probability sampling requires a systematic approach to recruitment in order to realize the benefits of the sample design. If subjective factors such as interviewer preference enter into the recruitment process, then the objectivity associated with random selection will not be achieved. The goals of HCHS/SOL recruitment are to optimize the ability to establish contact with, determine eligibility of, and actively engage households at every sample address, regardless of the neighborhood or living conditions encountered in the field. Recruitment teams inform potential participants of the study objectives and associated benefits of their participation. The research nature of the study is emphasized, including the information it is designed to provide and the impact the study results may have on policy making and health care for future generations of US Hispanic/Latinos. Extensive community engagement efforts provide the context for this information exchange, including collaborations with community based organizations and targeted media campaigns.

The recruitment protocol consists of three steps: (i) initial mailings to sample addresses describing the study; (ii) optional telephone contacts for households with telephone numbers available; and (iii) in-person contacts. Once contact is established, a brief household screener is administered via a digital hand-held device to determine eligibility and implement the age sub-sampling procedure (12). Upon obtaining agreement to participate, a roster of household members is created, and individual eligibility confirmed. Persons on active duty military service, not currently living at home, planning to move from the area in the next six months, or are physically unable to attend the clinic examination are considered ineligible.

Household- and individual-level screening and eligibility rates and clinic participation rates are monitored continuously, and adjustments to selection parameters (for age) or fielding of sample addresses (for SES and background) are made as needed. At the conclusion of HCHS/SOL recruitment, final household and individual level participation rates will be computed among those eligible for the study. A goal of 60% participation was set at the onset of recruitment.

DISCUSSION

Study design decisions are typically made to accommodate competing priorities; the National Children's Study provides a recent example (13,14). If the HCHS/SOL research objectives were limited to baseline prevalence estimates and comparisons thereof, then a probability sample representing a broad cross-section of US Hispanics/Latinos would be the choice. Had the sole objective been to support valid inference of relationships among baseline risk factors and disease incidence during follow-up, then enrolling a cohort most likely to remain active in the study for years to come would be optimal. A hybrid design was chosen to simultaneously meet both objectives, with communities defined based on proximity to clinical centers and Hispanic/Latino diversity, and probability sampling nested within.

Two aspects of the HCHS/SOL design represent a novel approach for epidemiology studies with similar objectives. First, formal probability sampling methods are embedded into the

study design at each site, thereby allowing the advantages of probability sampling to be available within a traditional, multi-site study model. Secondly, methods to efficiently sample Hispanics/Latinos in already enriched community areas are employed, where efficiency is defined in both a statistical and operational sense. Techniques are applied at each stage of sample selection such that field operations are optimized without unnecessarily sacrificing precision of estimates.

Several trade-offs occur as a result of incorporating probability sampling in the HCHS/SOL design. First, the lack of recent census data upon which to base key design parameters produced some inefficiencies early in recruitment. Second, areas with low Hispanic/Latino concentration are included in the sample for diversity, although with lower representation due to undersampling, and their coverage can substantially increase field costs. Finally, the complexity of data analyses is increased by the need to account for the sample design.

In summary, the HCHS/SOL sampling strategy was chosen to provide broad representation of the US Hispanic/Latino population living in the communities surrounding the four field centers with sufficient diversity to support the research objectives. A rigorous recruitment protocol is required to realize the benefits of probability sampling; however, the design is flexible in that modifications can be incorporated with minimal disruption of ongoing recruitment activities. The hybrid design employed for HCHS/SOL can serve as a model for the design of future studies with similar objectives.

Acknowledgments

Funding Support The Hispanic Community Health Study/Study of Latinos is funded by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Center on Minority Health and Health Disparities, the National Institute of Deafness and Other Communications Disorders, the National Institute of Dental and Craniofacial Research, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Neurological Disorders and Stroke, and the Office of Dietary Supplements.

ABBREVIATIONS

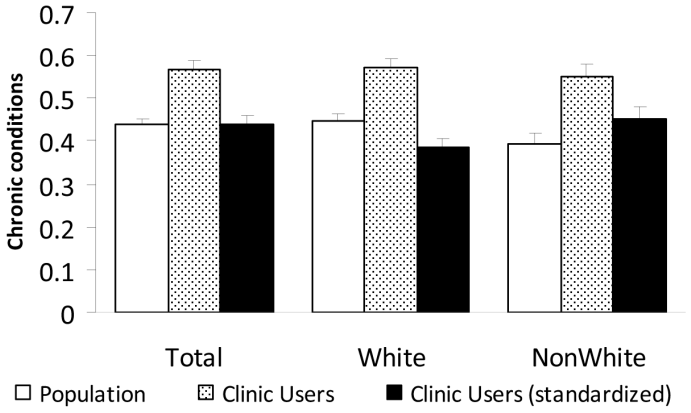
HCHS	Hispanic Community Health Study
SOL	Study of Latinos
NHLBI	National Heart, Lung, and Blood Institute
MEPS	Medical Expenditure Panel Survey
SES	Socio-economic status
BG	block group
PSU	primary sampling unit
DSF	Delivery Sequence File

REFERENCES

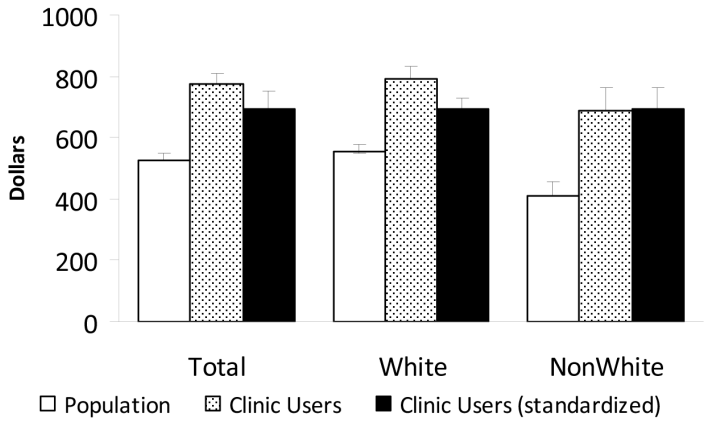
1. Kruskal W, Mosteler F. Representative Sampling, III: The Current Statistical Literature. International Statistical Review 1979;47:245–265.
2. Sarndal, CE.; Swensson, B.; Wretman, J. Model Assisted Survey Sampling. Springer-Verlag; New York: 1992.
3. Kish, L. Second Printing. John Wiley and Sons, Inc.; New York, NY: 1965. Survey Sampling.
4. Kish, L. Statistical Design for Research. John Wiley & Sons, Inc.; New York, NY: 1987.

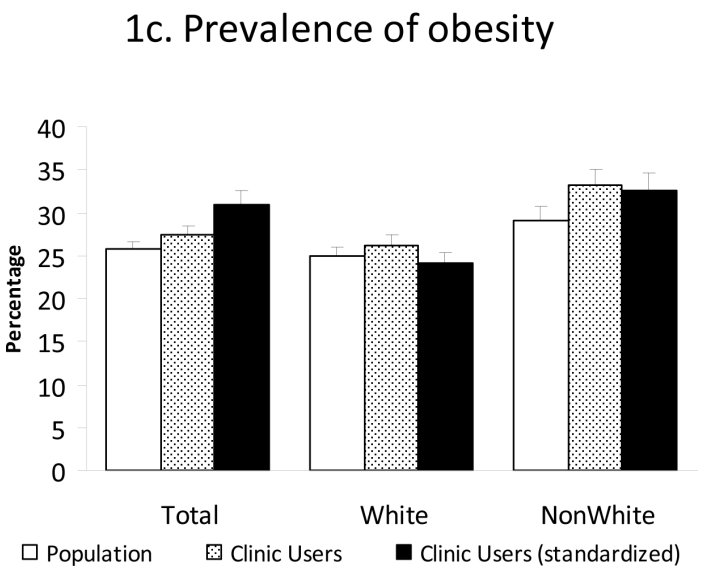
5. Groves, RM.; Fowler, FJ.; Couper, MP.; Lepkowski, JM.; Singer, E.; Tourangeau, R. Survey Methodology. Wiley and Sons; New York: 2004.
6. Lessler, JT.; Kalsbeek, WD. Nonsampling Errors in Surveys. Wiley and Sons; New York: 1992.
7. Ezzati-Rice, TM.; Rohde, F.; Greenblatt, J. Sample Design of the Medical Expenditure Panel Survey. Agency for Healthcare Research and Quality; Rockville, MD: 2008. Methodology Report 22
8. Cochran, WG. Sampling Techniques. 3rd Edition. John Wiley & Sons, Inc.; New York, NY: 1977.
9. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. American Journal of Public Health 1992;82:816–820. [PubMed: 1585961]
10. Korn, EL.; Graubard, BI. Analysis of Health Surveys. Wiley; New York, NY: 1999.
11. Little RA. To model or not to model? Competing modes of inference for finite population sampling. Journal of the American Statistical Association 2004;99(466):546–556.
12. Bryan, H.; Mehlman, T.; Gildner, P. A Study Recruitment System Using Ultra-Mobile Computers with Handwriting Recognition. Poster presentation at the Society for Clinical Trials 30th Annual Meeting; Atlanta, GA. 2009;
13. National Children's Study. Final Report from the NCS Sampling Design Workshop. National Children's Study; 2004.
14. Michael RT, O'Muircheartaigh CA. Design priorities and disciplinary perspectives: The case of the US National Children's Study. Journal of the Royal Statistical Society: Series A 2008;171(2):465–480.

1a. Average number of chronic conditions per year



1b. Average cost of physician office visits per year





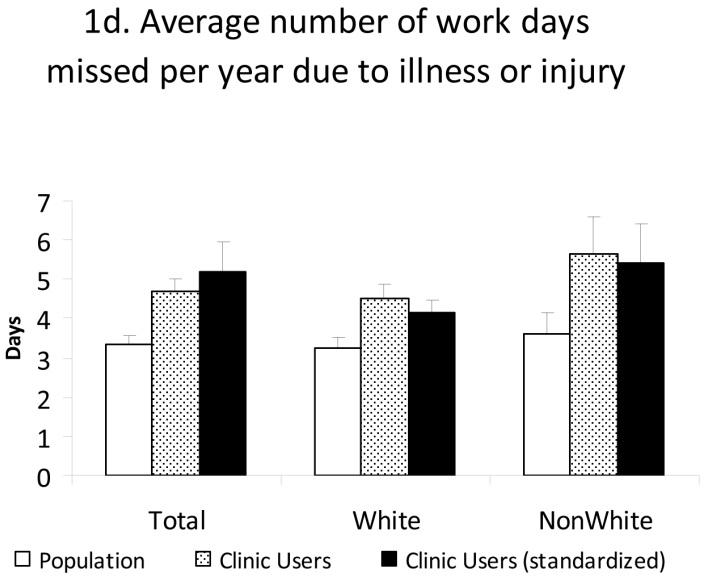


Figure 1.
Results of a Simulated Comparison of Probability and Non-Probability Samples Using Data from the 2005 Medical Expenditure Panel Survey (MEPS)
Legend. 'Population' corresponds to estimates based on the MEPS sample and therefore representing the 2005 US resident, non-institutionalized population. 'Clinic Users (Raw)' corresponds to estimates based on the subset of the MEPS sample reporting one or more physician visits in the past year. 'Clinic Users (Standardized)' corresponds to estimates based on the MEPS clinic-users sample, standardized to the 2005 US resident, non-institutionalized population distributions for age, race/ethnicity, and gender.

Table 1

Summary of HCHS/SOL Sample Design Features

Sampling Stage	Sampling Unit and Frame Source(s)	Stratification and Stratum Allocation	Random Selection Method in Each Stratum
1	**Sampling unit:** Block group (BG) as defined for the 2000 Census **Frame:** List of BGs created from the designated community area at the site, defined in each stratum as a set of contiguous BGs.	• Explicit strata formed by crossclassification of: (i) "high"/"low" categories according to % Hispanic among total population in 2000 and (ii) "high"/"low" socio-economic status as measured by % of the population with at least a high school education. One or more special strata were defined in a subset of sites to target specific population subgroups. • Disproportionately higher BG sampling rates for strata in the "high" % Hispanic category; proportionate allocation to "high" and "low" SES categories within each % Hispanic category	• Simple random sampling
2	**Sampling Unit:** Household address **Frame source:** US Postal Service listing of addresses available through MSG-Genesys	• Explicit strata formed by whether/not the occupant has an Hispanic surname • Disproportionately higher address sampling rate for Hispanic surname addresses; uniform stratum-specific address sampling rates among sample BGs	• Simple random sampling

Ann Epidemiol. Author manuscript; available in PMC 2011 August 1.

Table 2

Design Characteristics of the HCHS/SOL Sample

Design Characteristic	Bronx	Chicago	Miami	San Diego	Total Count: All Field Centers Combined
STAGE 1 (Sampling Block Groups):					
Total number of BGs on Census frame	376	170	158	221	925
Delineation point: “High” vs. “Low” % Hispanic	44.69%	45.79%	64.71%	44.67%	---
Delineation point: “High” vs. “Low” SES (% >= high school education)	46.80%	44.11%	48.50%	53.32%	---
Number of selected BGs:	238	125	147	160	670
By stratum (% Hispanic / SES)					
Low Concentration / High SES	14	19	9	51	93
Low Concentration / Low SES	5	5	0	11	21
High Concentration / High SES	78	46	46	35	205
High Concentration / Low SES	100	55	34	63	252
Special stratum#1	3	---	11	---	14
Special stratum #2	16	---	47	---	63
Special stratum #3	22				22
BG Over-sampling ratio for % Hispanic strata: High/Low 1	2.49	2.48	1.57	1.98	---
STAGE 2 (Sampling Addresses within Block Groups):					
Total number of addresses on USPS frame	188,932	83,950	98,072	126,769	497,723
Total number of addresses in selected BGs	117,319	59,666	90,298	92,061	359,344
Total number of selected addresses	30,718	31,143	22,929	42,423	127,213
Address over-sampling ratio:					
By stratum (% Hispanic - SES)					
Low-Low	10.0	3.0	2.0	3.0	---
Low-High	6.0	3.5	---	3.0	---
High-Low	4.0	2.3	1.0	3.1	---
High-High	4.0	3.0	1.0	3.6	---

Design Characteristic	Bronx	Chicago	Miami	San Diego	Total Count: All Field Centers Combined
Special stratum #1	15.0	---	1.0	---	---
Special stratum #2	6.0	---	1.0	---	---
Special stratum #3	4.0				
FINAL SAMPLE (Hispanics/Latinos aged 18-74 years):					
Age over-sampling during household screening:					
Method 1: Selection probability for mixed-age households	0.21	0.14	0.38	0.12	---
Method 2: Selection probability for younger adults within a household (aged 18-44 years)	0.63	0.30	0.45	0.50	---
Targeted participant sample size	4,000	4,000	4,000	4,000	16,000
Target distribution by background (%):					
Central/South American	8%	11%	35%	2%	2,240
Cuban	1%	1%	60%	0%	2,480
Mexican	7%	61%	1%	97%	6,640
Puerto Rican/Dominican Republic	84%	27%	4%	1%	4,640

[1]The over-sampling ratio is calculated by dividing the sampling rate in a stratum isolating those to be oversampled by the sampling rate for the corresponding stratum isolating those to be under-sampled.

Table 3

Hazard Ratios Detected with Approximately 90% Power by Event Rate and Lowto-High Risk Group Ratio

3.a Analyses Based on the Total HCHS/SOL Sample (n = 16,000)			
	Ratio of low-risk to high-risk		
Rate in low-risk group	**1:1**	**3:1**	**15:1**
2/1000 person-years	1.85	1.90	2.50
4/1000 person-years	1.60	1.65	2.10
6/1000 person-years	1.45	1.50	1.90
8/1000 person-years	1.40	1.45	1.75

3.b Field Center- or Subgroup-Specific Analyses (n = 4,000)			
	Ratio of low-risk to high-risk		
Rate in low-risk group	**1:1**	**3:1**	**15:1**
2/1000 person-years	2.95	2.95	4.15
4/1000 person-years	2.25	2.35	3.20
6/1000 person-years	2.00	2.05	2.80
8/1000 person-years	1.85	1.90	2.55
10/1000 person-years	1.75	1.80	2.40

Assuming 3-year accrual and 2-year follow-up periods and a design effect due to clustering of 1.25.

Assuming 3-year accrual and 2-year follow-up periods and a design effect due to clustering of 1.25.