CHIS california health interview survey

June 9, 2014

CHIS 2011-2012 Methodology Report Series

Report 3

Data Processing Procedures

CALIFORNIA HEALTH INTERVIEW SURVEY

CHIS 2011-2012 METHODOLOGY SERIES

REPORT 3

DATA PROCESSING PROCEDURES

VERSION DATE: JUNE 9, 2014

This report was prepared for the California Health Interview Survey by Howard King and Sherman Edwards of Westat.

CHIS california health interview survey

www.chis.ucla.edu

This report describes the data processing and editing procedures for CHIS 2011-2012 that were performed by Westat. It discusses standard data editing procedures and addresses the steps taken for ensuring data quality. It also presents discussions on special procedures of editing and coding of geography and race and ethnicity survey items.

Suggested citation:

California Health Interview Survey. *CHIS 2011-2012 Methodology Series: Report 3 – Data Processing Procedures*. Los Angeles, CA: UCLA Center for Health Policy Research, 2014.

The California Health Interview Survey is a collaborative project of the UCLA Center for Health Policy Research, the California Department of Public Health, and the Department of Health Care Services. Funding for CHIS 2011-2012 came from multiple sources: the Blue Shield of California Foundation, the California Department of Health Care Services, the California Department of Mental Health, the California Department of Public Health, the California Endowment, First 5 California, the California Office of the Patient Advocate, Kaiser Permanente, the NIH National Cancer Institute, NIH Office of Research on Women's Health, NIH Office of Behavioral and Social Sciences Research, San Diego County Health and Human Services Agency, United American Indian Involvement, Inc./CA Indian Health Service, Archstone Foundation, Susan B. Komen for the Cure, Centers for Disease Control and Prevention, NCI American Cancer Society, NIH RAND, and the California Wellness Foundation.

PREFACE

Data Processing Procedures is the third report in a series of methodological reports describing the 2011 California Health Interview Survey (CHIS 2011-2012). The other reports are listed below. This report describes the data processing procedures that took place at Westat. It does not include the additional processing procedures performed later by UCLA. Please check the CHIS website (www.chis.ucla.edu) for availability of reports on the data processing procedures at UCLA.

CHIS is a collaborative project of the University of California, Los Angeles (UCLA) Center for Health Policy Research, the California Department of Public Health, the Department of Health Care Services, and the Public Health Institute. Westat was responsible for data collection and the preparation of five methodological reports from the 2011 survey. The survey examines public health and health care access issues in California. The telephone survey is the largest state health survey ever undertaken in the United States. The plan is to monitor these issues and examine changes over time by conducting surveys in the future.

Methodological Reports

The first five methodological reports for CHIS 2011-2012 are as follows:

- Report 1: Sample Design;
- Report 2: Data Collection Methods;
- Report 3: Data Processing Procedures;
- Report 4: Response Rates; and
- Report 5: Weighting and Variance Estimation.

The reports are interrelated and contain many references to each other. For ease of presentation, the references are simply labeled by the report numbers given above.

This report describes the data processing and editing procedures for CHIS 2011-2012. One chapter details the data editing procedures and addresses the steps taken for ensuring data quality. Delivery of the final data sets is also discussed. Another chapter presents information about geographic coding. The next chapter describes how the race and ethnicity survey items were coded for CHIS.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

List of Tables

1. CHIS 2011-2012 Sample Design and Methodology Summary

1.1 Overview

The California Health Interview Survey (CHIS) is a population-based telephone survey of California conducted every other year since 2001 and continually beginning in 2011. CHIS is the largest state health survey conducted and one of the largest health surveys in the nation. CHIS is conducted by the UCLA Center for Health Policy Research (UCLA-CHPR) in collaboration with the California Department of Public Health, the Department of Health Care Services, First 5 California, The California Endowment, the National Cancer Institute, and Kaiser Permanente. CHIS collects extensive information for all age groups on health status, health conditions, health-related behaviors, health insurance coverage, access to health care services, and other health and health related issues.

The sample is designed to meet and optimize two objectives:

1) Provide estimates for large- and medium-sized counties in the state, and for groups of the smallest counties (based on population size), and

2) Provide statewide estimates for California's overall population, its major racial and ethnic groups, as well as several Asian and Latino ethnic subgroups.

The CHIS sample is representative of California's non-institutionalized population living in households. CHIS data and results are used extensively by federal and State agencies, local public health agencies and organizations, advocacy and community organizations, other local agencies, hospitals, community clinics, health plans, foundations, and researchers. These data are used for analyses and publications to assess public health and health care needs, to develop and advocate policies to meet those needs, and to plan and budget health care coverage and services. Many researchers throughout California and the nation use CHIS data files to further their understanding of a wide range of health-related issues (visit the CHIS Research Clearinghouse at http://healthpolicy.ucla.edu/chis/research/Pages/default.aspx for many examples of these studies).

This series of reports describes the methods used in collecting data for CHIS 2011-2012, the sixth CHIS data collection cycle, which was conducted between June 2011 and January 2013. The previous CHIS cycles (2001, 2003, 2005, 2007, and 2009) are described in similar series, available at http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx.

1.2 Switch to a Continuous Survey

From the first CHIS cycle in 2001 through 2009, CHIS data collection was biennial, with data collected during a 7-9 month period every other year. Beginning in 2011, CHIS data are collected continually over each 2-year cycle. This change was driven by several factors including the ability to track and release information about health in California on a more frequent and timely basis and to eliminate potential seasonality in the biennial data.

The CHIS 2011-2012 data included in these files were collected between June 2011 and January 2013. Approximately half of the interviews were conducted during the 2011 calendar year and half during the 2012 calendar year. As in previous CHIS cycles, weights are included with the data files and are based on the State of California's Department of Finance population estimates and projections, adjusted to remove the population living in group quarters (such as nursing homes, prisons, etc. and not eligible to participate in CHIS). When the weights are applied to the data, the results represent California's residential population during that one year period for the age group corresponding to the data file in use (adult, adolescent, or child).

See what else is new in the 2011-2012 CHIS sampling and data collection here:
http://healthpolicy.ucla.edu/chis/design/Documents/whats-new-chis-2011-2012.pdf

In order to provide CHIS data users with more complete and up-to-date information to facilitate analyses of CHIS data, additional information on how to use the CHIS sampling weights, including sample code, is available at: http://healthpolicy.ucla.edu/chis/analyze/Pages/sample-code.aspx

Additional documentation on constructing the CHIS sampling weights is available in CHIS 2011-2012 Methods Report #5—Weighting and Variance Estimation, available at:
http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx. Other helpful information for understanding the CHIS sample design and data collection processing can be found in the four other methodology reports for each CHIS cycle year, described in the Preface above.

1.3 Sample Design Objectives

The CHIS 2011-2012 sample was designed to meet two sampling objectives discussed above: (1) provide estimates for adults in most counties and groups of counties with small populations; and (2) provide estimates for California's overall population, major racial and ethnic groups, and for several smaller ethnic subgroups.

To achieve these objectives, CHIS employed a dual-frame, multi-stage sample design. The random-digit-dial (RDD) sample included telephone numbers assigned to both landline and cellular service. The random-digit-dial (RDD) sample was approximately 80% landline and 20% cellular phone numbers. For the landline RDD sample, the 58 counties in the state were grouped into 44 geographic sampling strata, and 14 sub-strata were created within two of the largest metropolitan areas in the state (Los Angeles and San Diego). The Los Angeles County stratum included 8 sub-strata for Service Planning Areas, and the San Diego County stratum included 6 sub-strata for Health Service Regions. Most of the strata (39 of 44) are made up of a single county with no sub-strata (counties 3-41 in Table 1-1), with three multi-county strata comprised of the 17 remaining counties (see Table 1-1). A sufficient number of adult interviews were allocated to each stratum and sub-stratum to support the first sample design objective—to provide health estimates for adults at the local level. The same geographic stratification of the state has been used since CHIS 2005. In the first two CHIS cycles (2001 and 2003) there were 47 total sampling strata, including 33 individual counties and one county with sub-strata (Los Angeles).

Within each geographic stratum, residential telephone numbers were selected, and within each household, one adult respondent (age 18 and over) was randomly selected. In those households with adolescents (ages 12-17) and/or children (under age 12), one adolescent and one child were randomly selected; the adolescent was interviewed directly, and the adult most knowledgeable about the child's health completed the child interview.

The RDD CHIS sample is of sufficient size to accomplish the second objective (produce estimates for the state's major racial/ethnic groups, as well as many ethnic subgroups). To increase the precision of estimates for Koreans and Vietnamese, areas with relatively high concentrations of these groups were sampled at higher rates. These geographically targeted oversamples were supplemented by telephone numbers associated with group-specific surnames drawn from listed telephone directories to further increase the sample size for Koreans and Vietnamese.

Table 1-1. California county and county group strata used in the CHIS 2011-2012 sample design

1. Los Angeles
   1.1 Antelope Valley
   1.2 San Fernando Valley
   1.3 San Gabriel Valley
   1.4 Metro
   1.5 West
   1.6 South
   1.7 East
   1.8 South Bay
2. San Diego
   2.1 N. Coastal
   2.2 N. Central
   2.3 Central
   2.4 South
   2.5 East
   2.6 N. Inland
3. Orange
4. Santa Clara
5. San Bernardino
6. Riverside
7. Alameda
8. Sacramento
9. Contra Costa
10. Fresno
11. San Francisco
12. Ventura
13. San Mateo
14. Kern
15. San Joaquin
16. Sonoma
17. Stanislaus
18. Santa Barbara
19. Solano
20. Tulare
21. Santa Cruz
22. Marin
23. San Luis Obispo
24. Placer
25. Merced
26. Butte
27. Shasta
28. Yolo
29. El Dorado
30. Imperial
31. Napa
32. Kings
33. Madera
34. Monterey
35. Humboldt
36. Nevada
37. Mendocino
38. Sutter
39. Yuba
40. Lake
41. San Benito
42. Colusa, Glen, Tehama
43. Plumas, Sierra, Siskiyou, Lassen, Modoc, Trinity, Del Norte
44. Mariposa, Mono, Tuolumne, Alpine, Amador, Calaveras, Inyo

Source: UCLA Center for Health Policy Research, 2011-2012 California Health Interview Survey.

To help compensate for the increasing number of households without landline telephone service, a separate RDD sample was drawn of telephone numbers assigned to cellular service. In CHIS 2011-2012, the goal was to complete approximately 8,000 interviews (20% of all RDD interviews statewide) with adults from the cell phone sample. Telephone numbers assigned to cellular service cannot be geographically stratified at the county level with sufficient precision, so the cell RDD sample was geographically stratified into 28 strata using 7 CHIS regions and telephone area codes. If a sampled cell number was shared by two or more adult members of a household, one household member was selected for the adult interview. Otherwise, the adult owner of the sampled number was selected. Cell numbers used exclusively by children under 18 were considered ineligible. About 550 teen interviews and 1,500 child interviews were completed from the cell phone sample in CHIS 2011-2012.

The CHIS 2011-2012 and 2009 cell phone sampling method differed from that used in CHIS 2007 in two significant ways. First, in CHIS 2011-2012, all cell phone sample numbers used for non-business purposes by adults living in California were eligible for the extended interview, while in 2007 only cell numbers belonging to adults in cell-only households were eligible. Thus, adults in households with landlines who had their own cell phones or shared one with another adult household member could have been selected through either the cell or landline sample. The second change to the cell phone sample was the inclusion of child and adolescent extended interviews.

Unlike both CHIS 2007 and CHIS 2009, where the cell phone sample quotas were treated separately from the landline sample, the CHIS 2011-2012 cell sample respondents were included in the overall and county specific target sample sizes. Twenty-eight cell phone sampling strata were created using CHIS 2007 and 2009 cell phone respondents' data and their pre-assigned FIPS county code, supplied by the sampling vendor. The statewide target of 8,000 adult cell phone interviews was also supplemented with an oversample to yield approximately 1,150 adult cell phone interviews. The oversample focused on six counties; Los Angeles, Orange, Santa Clara, Alameda, San Francisco, and San Mateo.

Finally, the CHIS 2011-2012 sample included an American Indian/Alaska Native (AIAN) oversample. This oversample was sponsored by Urban American Indian Involvement, Inc., and California Indian Health Services. The purpose of this oversample was to increase the number of AIAN participants and improve the statistical stability and precision of estimates for this group. The oversample was conducted using a list provided by Indian Health Services.

1.4 Data Collection

To capture the rich diversity of the California population, interviews were conducted in five languages: English, Spanish, Chinese (Mandarin and Cantonese dialects), Vietnamese, and Korean. These languages were chosen based on analysis of 2000 Census data to identify the languages that would cover the largest number of Californians in the CHIS sample that either did not speak English or did not speak English well enough to otherwise participate.

Westat, a private firm that specializes in statistical research and large-scale sample surveys, conducted CHIS 2011-2012 data collection under contract with the UCLA Center for Health Policy Research. For all samples, Westat staff interviewed one randomly selected adult in each sampled household, and sampled one adolescent and one child if they were present in the household and the sampled adult was the parent or legal guardian. Thus, up to three interviews could have been completed in each household. In landline sample households with children where the sampled adult was not the screener respondent, children and adolescents could be sampled as part of the screening interview, and the extended child (and adolescent) interviews could be completed before the adult interview. This "child-first” procedure was new for CHIS 2005 and has been continued in subsequent CHIS cycles; this procedure substantially increases the yield of child interviews. While numerous subsequent attempts were made to complete the adult interview for child-first cases, there are completed child and/or adolescent interviews in households for which an adult interview was not completed. Table 1-2 shows the number of completed adult, child, and adolescent interviews in CHIS 2011-2012 by the type of sample (landline RDD, surname list, cell RDD, and American Indian/Alaska Native list).

Table 1-2. Number of completed CHIS 2011-2012 interviews by type of sample and instrument

Type of sample	Adult	Child	Adolescent
Total all samples	42,935[1]	7,334	2,799
Landline RDD	32,692	5,600	2,164
Surname list	825	161	57
Cell RDD	9,151	1,523	557
American Indian/Alaska Native list	267	50	21

Source: UCLA Center for Health Policy Research, 2011-2012 California Health Interview Survey.

Interviews in all languages were administered using Westat’s computer-assisted telephone interviewing (CATI) system. The average adult interview took about 35 minutes to complete. The average child and adolescent interviews took about 15 minutes and 23 minutes, respectively. For “child-first” interviews, additional household information asked as part of the child interview averaged about 9 minutes. Interviews in non-English languages generally took longer to complete. More than 14 percent of the adult interviews were completed in a language other than English, as were about 27 percent of all child (parent proxy) interviews and 7 percent of all adolescent interviews.

Table 1-3 shows the major topic areas for each of the three survey instruments (adult, child, and adolescent).

[1] Numbers in this table represent the data publically released and available through our Data Access Center. Total sample sizes may differ for specific calculations within the five methodology reports, or for specific analyses based on CHIS data.

Table 1-3. CHIS 2011-2012 survey topic areas by instrument

Health status	**Adult**	**Teen**	**Child**
General health status	✓	✓	✓
Days missed from school due to health problems	✓	✓	✓
Health-related quality of life (HRQOL)		✓	
Health conditions	**Adult**	**Teen**	**Child**
Asthma	✓	✓	✓
Diabetes, gestational diabetes, pre- /borderline diabetes	✓		
Heart disease, high blood pressure, stroke	✓		
Arthritis, physical disability	✓		
Epilepsy		✓	
Physical, behavioral, and/or mental conditions			✓
Mental health	**Adult**	**Teen**	**Child**
Mental health status	✓	✓	
Perceived need, access and utilization of mental health services	✓	✓	
Functional impairment, stigma	✓		
Suicide ideation and attempts	✓		
Health behaviors	**Adult**	**Teen**	**Child**
Dietary intake, fast food	✓	✓	✓
Physical activity and exercise, commute from school to home		✓	✓
Walking for transportation and leisure	✓		
Doctor discussed nutrition/physical activity		✓	✓
Flu Shot	✓		✓
Alcohol and cigarette use	✓	✓	
Illegal drug use		✓	
Sexual behavior	✓	✓	
HIV/STI testing		✓	
Elderly falls	✓		
Women's health	**Adult**	**Teen**	**Child**
Mammography screening	✓		
Pregnancy	✓	✓	
Dental health	**Adult**	**Teen**	**Child**
Last dental visit, main reason haven't visited dentist		✓	✓
Neighborhood and housing	**Adult**	**Teen**	**Child**
Safety, social cohesion	✓	✓	✓
Homeownership, length of time at current residence	✓		
Park use		✓	✓
Civic engagement	✓	✓	

Table 1-3. CHIS 2011-2012 survey topic areas by instrument (continued)

Access to and use of health care	**Adult**	**Teen**	**Child**
Usual source of care, visits to medical doctor	✓	✓	✓
Emergency room visits	✓	✓	✓
Delays in getting care (prescriptions and medical care)	✓	✓	✓
Medical home, timely appointments, hospitalizations	✓	✓	✓
Communication problems with doctor	✓		✓
Internet use for health information	✓		✓

Food environment	**Adult**	**Teen**	**Child**
Access to fresh and affordable foods	✓		
Where teen/child eats breakfast/lunch, fast food at school		✓	✓
Availability of food in household over past 12 months	✓		

Health insurance	**Adult**	**Teen**	**Child**
Current insurance coverage, spouse's coverage, who pays for coverage	✓	✓	✓
Health plan enrollment, characteristics and plan assessment	✓	✓	✓
Whether employer offers coverage, respondent/spouse eligibility	✓		
Coverage over past 12 months, reasons for lack of insurance	✓	✓	✓
Difficulty finding private health insurance	✓		
High deductible health plans	✓	✓	✓
Partial scope Medi-Cal	✓		

Public program eligibility	**Adult**	**Teen**	**Child**
Household poverty level	✓		
Program participation (CalWORKs, Food Stamps, SSI, SSDI, WIC, TANF)	✓	✓	✓
Assets, alimony/child support, social security/pension	✓		
Medi-Cal and Healthy Families eligibility	✓	✓	✓
Reason for Medi-Cal non-participation among potential beneficiaries	✓	✓	✓

Bullying and interpersonal violence	**Adult**	**Teen**	**Child**
Bullying, personal safety, interpersonal violence		✓	

Parental involvement/adult supervision	**Adult**	**Teen**	**Child**
Adult presence after school, role models, resiliency		✓	
Parental involvement			✓

Child care and school attendance	**Adult**	**Teen**	**Child**
Current child care arrangements			✓
Paid child care	✓		
First 5 California: Kit for New Parents			✓
Preschool/school attendance, name of school		✓	✓
Preschool quality			✓
School instability		✓	

Table 1-3. CHIS 2011-2012 survey topic areas by instrument (continued)

Employment	**Adult**	**Teen**	**Child**
Employment status, spouse’s employment status	✓		
Hours worked at all jobs	✓		
Income	**Adult**	**Teen**	**Child**
Respondent’s and spouse’s earnings last month before taxes	✓		
Household income , number of persons supported by household income	✓		
Respondent characteristics	**Adult**	**Teen**	**Child**
Race and ethnicity, age, gender, height, weight	✓	✓	✓
Veteran status	✓		
Marital status, registered domestic partner status (same-sex couples)	✓		
Sexual orientation	✓		
Language spoken with peers, language of TV, radio, newspaper used	✓		
Education, English language proficiency	✓		
Citizenship, immigration status, country of birth, length of time in U.S., languages spoken at home	✓	✓	✓

Source: UCLA Center for Health Policy Research, 2011-2012 California Health Interview Survey.

1.5 Response Rates

The overall response rate for CHIS 2011-2012 is a composite of the screener response rate (i.e., success in introducing the survey to a household and randomly selecting an adult to be interviewed) and the extended interview response rate (i.e., success in getting one or more selected persons to complete the extended interview). To maximize the response rate, especially at the screener stage, an advance letter in five languages was mailed to all landline sampled telephone numbers for which an address could be obtained from reverse directory services. An advance letter was mailed for 48.3 percent of the landline RDD sample telephone numbers not identified by the sample vendor as business or nonworking numbers, 81.1 percent of surname list sample numbers, and 94.3 percent of the AIAN list with landline numbers after removing nonworking and business numbers. Addresses were not available for the cell sample. As in all CHIS cycles since CHIS 2005, a $2 bill was included with the CHIS 2011-2012 advance letter to encourage cooperation.

The CHIS 2011-2012 screener response rate for the landline sample was 31.6 percent, and was higher for households that were sent the advance letter. For the cell phone sample, the screener response rate was 33.0 percent in all households. The extended interview response rate for the landline sample varied across the adult (47.4 percent), child (73.2 percent) and adolescent (42.7 percent) interviews. The adolescent rate includes getting permission from a parent or guardian. The adult interview response rate for the cell sample was 53.8 percent, the child rate was 73.4 percent, and the adolescent rate 42.6 percent. Multiplying the screener and extended rates gives an overall response rate for each type of interview. The percentage of households completing one or more of the extended interviews (adult, child, and/or adolescent) is a useful summary of the overall performance of the landline sample. For CHIS 2011-2012, the landline/list sample household response rate was 17.0 percent (the product of the screener response rate and the extended interview response rate at the household level of 53.9 percent). The cell sample household response rate was 18.3 percent, incorporating a household-level extended interview response rate of 55.5 percent. All of the household and person level response rates vary by sampling stratum. For more information about the CHIS 2011-2012 response rates please see *CHIS 2011-2012 Methodology Series: Report 4 – Response Rates*.

Historically, the CHIS response rates are comparable to response rates of other scientific telephone surveys in California, such as the California Behavioral Risk Factor Surveillance System (BRFSS) Survey. However, comparing the CHIS and BRFSS response rates requires recomputing the CHIS response rates so they match the BRFSS response rate calculation methods. The 2011 California BRFSS landline response rate is 37.4 percent, the cell phone response rate is 20.4 percent, and the combined landline and cell phone rate is 35.4 percent.[2] In contrast, the CHIS 2011-2012 landline response rate is 39.5, cell phone response rate is 32.1 percent, and the combined landline and cell phone response rate is 35.1 percent, all these computed using the BRFSS methodology. California as a whole and the state's urban areas in particular are among the most difficult parts of the nation in which to conduct telephone interviews. The 2011 BRFSS, for example, shows the refusal rate for California (31.4%) is the highest in the nation and twice the national median (16.0%). Survey response rates tend to be lower in California than nationally, and over the past decade response rates have been declining both nationally and in California.

Further information about CHIS data quality and nonresponse bias is available at http://healthpolicy.ucla.edu/chis/design/Pages/data-quality.aspx.

[2] As reported in the Behavioral Risk Factor Surveillance System 2011 Summary Data Quality Report (Version #5--Revised: 2/04/2013 , available online at http://www.cdc.gov/brfss/pdf/2011_Summary_Data_Quality_Report.pdf.

After all follow-up attempts to complete the full questionnaire were exhausted, adults who completed at least approximately 80 percent of the questionnaire (i.e., through Section K which covers employment, income, poverty status, and food security), were counted as “complete.” At least some responses in the employment and income series, or public program eligibility and food insecurity series were missing from those cases that did not complete the entire interview. They were imputed to enhance the analytic utility of the data.

Proxy interviews were conducted for frail and ill persons over the age of 65 who were unable to complete the extended adult interview in order to avoid biases for health estimates of elderly persons that might otherwise result. Eligible selected persons were re-contacted and offered a proxy option. For 283 elderly adults, a proxy interview was completed by either a spouse/partner or adult child. A reduced questionnaire, with questions identified as appropriate for a proxy respondent, was administered.

1.6 Weighting the Sample

To produce population estimates from CHIS data, weights are applied to the sample data to compensate for the probability of selection and a variety of other factors, some directly resulting from the design and administration of the survey. The sample is weighted to represent the non-institutionalized population for each sampling stratum and statewide. The weighting procedures used for CHIS 2011-2012 accomplish the following objectives:

- Compensate for differential probabilities of selection for households and persons;
- Reduce biases occurring because non-respondents may have different characteristics than respondents;
- Adjust, to the extent possible, for under-coverage in the sampling frames and in the conduct of the survey; and
- Reduce the variance of the estimates by using auxiliary information.

As part of the weighting process, a household weight was created for all households that completed the screener interview. This household weight is the product of the “base weight” (the inverse of the probability of selection of the telephone number) and a variety of adjustment factors. The household weight is used to compute a person-level weight, which includes adjustments for the within-household sampling of persons and nonresponse. The final step is to adjust the person-level weight using an iterative proportional fitting method or raking, as it is commonly called, so that the CHIS estimates are consistent with the marginal population control totals. This iterative procedure forces the CHIS weights to sum to known population control totals from an independent data source (see below). The procedure requires iteration to make sure all the control totals, or raking dimensions, are simultaneously satisfied within a pre-specified tolerance.

Population control totals of the number of persons by age, race, and sex at the stratum level for CHIS 2011-2012 were created primarily from the California Department of Finance's (DOF) 2012 Population Estimates and 2012 Population Projections. The raking procedure used 12 raking dimensions, which are combinations of demographic variables (age, sex, race, and ethnicity), geographic variables (county, Service Planning Area in Los Angeles County, and Health Region in San Diego County), household composition (presence of children and adolescents in the household), and socio-economic variables (home ownership and education). The socio-economic variables are included to reduce biases associated with excluding households without landline telephones from the sample frame. One limitation of using Department of Finance (DOF) data is that it includes about 2.4 percent of the population of California who live in "group quarters" (i.e., persons living with nine or more unrelated persons and includes, for example nursing homes, prisons, dormitories, etc.). These persons were excluded from the CHIS target population and, as a result, the number of persons living in group quarters was estimated and removed from the Department of Finance control totals prior to raking.

DOF control totals used to create the CHIS 2011-2012 weights are based on 2010 Census counts, while those in previous CHIS cycles were based on Census 2000 counts (with adjustments made by the Department of Finance). Please pay close attention when comparing estimates using CHIS 2011-2012 data with estimates using data from previous CHIS cycles. The most accurate California population figures are available when the US population count is conducted (every 10 years). Population-based surveys like CHIS must use estimates and projections based on the decennial population count data between Censuses. For example, population control totals for CHIS 2009 were based on DOF estimates and projections, which were based on Census 2000 counts with adjustments for demographic changes within the state between 2000 and 2009. These estimates become less accurate and more dependent on the models underlying the adjustments over time. Using the most recent Census population count information to create control totals for weighting produces the most statistically accurate population estimates for the current cycle, but it may produce unexpected increases or decreases in some survey estimates when comparing survey cycles that use 2000 Census-based information and 2010 Census-based information. See *CHIS 2011-2012 Methodology Series: Report 5 – Weighting and Variance Estimation* for more information on the weighting process.

1.7 Imputation Methods

Missing values in the CHIS data files were replaced through imputation for nearly every variable. This was a massive task designed to enhance the analytic utility of the files. Westat imputed missing values for those variables used in the weighting process and UCLA-CHPR staff imputed values for nearly all other variables.

Two different imputation procedures were used by Westat to fill in missing responses for items essential for weighting the data. The first imputation technique was a completely random selection from the observed distribution of respondents. This method was used only for a few variables when the percentage of the items missing was very small. The second technique was hot deck imputation without replacement. The hot deck approach is one of the most commonly used method for assigning values for missing responses. With a hot deck, a value reported by a respondent for a particular item is assigned or donated to a "similar" person who did not respond to that item. The characteristics defining "similar" vary for different variables. To carry out hot deck imputation, the respondents who answer a survey item form a pool of donors, while the item non-respondents are a group of recipients. A recipient is matched to the subset pool of donors based on household and individual characteristics. A value for the recipient is then randomly imputed from one of the donors in the pool. Once a donor is used, it is removed from the pool of donors for that variable. Hot deck imputation was used to impute the same items in CHIS 2003, CHIS 2005, CHIS 2007, CHIS 2009, and CHIS 2011-2012 (i.e., race, ethnicity, home ownership, and education).

UCLA-CHPR imputed missing values for nearly every variable in the data files other than those imputed by Westat and some sensitive variables in which nonresponse had its own meaning. Overall, item nonresponse rates in CHIS 2011-2012 were low, with most variables missing valid responses for less than 2% of the sample. However, there were a few exceptions where item nonresponse rate was greater than 20%, such as household income.

The imputation process conducted by UCLA-CHPR started with data editing, sometimes referred to as logical or relational imputation: for any missing value, a valid replacement value was sought based on known values of other variables of the same respondent or other sample(s) from the same household. For the remaining missing values, model-based hot-deck imputation with donor replacement was used. This method replaces a missing value for one respondent using a valid response from another respondent with similar characteristics as defined by a generalized linear model with a set of control variables (predictors). The link function of the model corresponds to the nature of the variable being imputed (e.g., linear regression for continuous variables, logistic regression for binary variables, etc.). Donors and recipients are grouped based on their predicted values from the model.

Control variables (predictors) used in the model to form donor pools for hot-decking always included the following: gender, age group, race/ethnicity, poverty level (based on household income), educational attainment, and region. Other control variables were also used depending on the nature of the imputed variable. Among the control variables, gender, age, race/ethnicity and regions were imputed by Westat. UCLA-CHPR then imputed household income and educational attainment in order to impute other variables. Household income, for example, was imputed using the hot-deck method within ranges from a set of auxiliary variables such as income range and/or poverty level.

The imputation order of the other variables followed the questionnaire. After all imputation procedures were complete, every step in the data quality control process is performed once again to ensure consistency between the imputed and non-imputed values on a case-by-case basis.

1.8 Methodology Report Series

A series of five methodology reports is available with more detail about the methods used in CHIS 2011-12:

- Report 1 – Sample Design;
- Report 2 – Data Collection Methods;
- Report 3 – Data Processing Procedures;
- Report 4 – Response Rates; and
- Report 5 – Weighting and Variance Estimation.

For further information on CHIS data and the methods used in the survey, visit the California Health Interview Survey Web site at http://www.chis.ucla.edu or contact CHIS at CHIS@ucla.edu.

2. DATA EDITING PROCEDURES

Survey data for all CHIS 2011-2012 samples – landline RDD, surname list, and cellular RDD – were collected using the same computer-assisted telephone interview (CATI) system. While the screening interview varied somewhat by sample, the same editing procedures were followed for all CHIS 2011-2012 cases.

In a CATI environment, the data collection and interview process is controlled using a series of computer programs to ensure consistency and quality. (*CHIS 2011-2012 Methodology Series: Report 2 - Data Collection Methods* provides a thorough discussion of the interview process and a description of how the survey data were collected.) The CATI system programming determines which questions are asked based on household composition, respondent characteristics or preceding answers, and the order in which the questions are presented to interviewers. The system also presents the response options that are available for recording answers.

CATI range and logic edits help ensure the integrity of the data during collection. Editing at the time of the interview greatly reduces the need for post-interview editing, and allows most questionable entries to be reviewed in real time with the respondent as part of the collection process. Although the CATI system virtually eliminates out-of-range responses and many other anomalies, some consistency and edit issues may arise. For example, interviewers may note concerns or problems that must be handled by data preparation staff after the interview is complete. Updating activities require that both manual and machine editing procedures be developed to correct interviewer, respondent, and CATI program errors and to check that updates made by data preparation staff were input correctly. Because data editing resulted in changes to the survey data, specific quality control procedures were implemented. CHIS 2011-2012 survey data were carefully examined and edited before Westat delivered final data files to UCLA. Quality control procedures involved limiting the number of staff who made updates, using the CATI specifications to resolve issues in complex questionnaire sections, carefully checking updates, and performing computer runs to identify inconsistencies or illogical patterns in the data.

The data editing procedures for CHIS 2011-2012 consisted of four main tasks: (1) managing and resolving problem cases, (2) reviewing interviewer comments to determine if data updates to the data in CATI were needed, (3) coding question responses that were recorded as text strings (i.e., “up-coding” responses captured in “other specify” fields), and (4) verifying data editing updates. The final step was to convert the edited data from the CATI system to the SAS data delivery files. The sections below describe each of these processes in turn.

2.1 Resolving Problem Cases

One of the important tasks for ensuring high-quality data was managing and resolving problem cases. The data preparation staff, as well as project staff and staff from the Telephone Research Center (TRC), worked collectively to resolve problem cases. The method interviewers used to communicate problems is described in this section, along with the system used by data editing and preparation staff to update or modify the data.

An interviewer who experienced a problem while working a case during data collection could alert the project team in one of two ways. One method was to fill out an electronic problem sheet for the case. All problem sheets were transmitted to a single staff member who distributed them to the appropriate department or project staff person. Data preparation staff often used these problem sheets as a guide to review cases and to make certain that any required updates were made accurately.

The second method of communicating problems was to assign a specific result code to cases within the CATI system,. The problem result code category had three sub-categories for special queues to which these problem cases could be assigned for review. These sub-categories were used to indicate the Westat staff person or group responsible for investigating the case further—1) TRC staff who work directly with the interviewers on a daily basis, 2) project staff who oversee design and implementation of the project, or 3) data processing staff who handle data cleaning and processing. Problem cases were reviewed electronically by a TRC supervisor and either re-fielded to the interviewing staff or distributed to the appropriate TRC, data processing, or project staff.

CATI database updates were not done if a problem could be resolved by simply releasing the case for general interviewing and including a message telling the interviewer what to do. If, for example, an adult extended interview was stopped during the middle of Section E, the interviewer would enter a detailed comment explaining why the case could not proceed (e.g., "Respondent wanted to change several answers. I was unable to back up properly"). The solution for these types of cases was to re-field the interview with a message stating, "Case will restart at the beginning of in Section E, in AD13[3]," and so the entire series of questions could be asked again. Most restart cases were made available to the general interviewing staff. For unusual or complex problems, the case could be assigned to a specific interviewer with experience in handling these types of problems.

[3] Note that questions from earlier CHIS cycles that were also asked in CHIS 2011 retained their original CATI screen names, in addition to having a sequential number appropriate to the 2011 interview. In this example, the first question in Section E for CHIS 2011 has screen name AD13.

Some examples of cases reviewed by Westat project staff were those in which an error was made in enumerating a household member or when a change in the person named as most knowledgeable about the sampled child was needed. Other types of problems required special interviewer handling, even after changes were made to the CATI database.

One specific category of problems—enumeration errors where some household members were either incorrectly identified or their characteristics were entered in error—was somewhat more challenging than other types of errors to resolve. If enough information was not available to complete the screener accurately the data manager could reload the case by using a utility created for CHIS and allow the next interviewer to enter data anew.

2.2 Interviewer Comments

Another important data editing task is reviewing all comments that interviewers type in a special entry window accessed by a “hot key” in the CATI system. Comments are used to record answers and statements that don’t fit into programmed response options that interviews see on their screens. Some comments merely elaborate on previously-recorded response, express an opinion, or are otherwise not directly related to the survey. These kinds of comments usually do not require modifying or updating survey responses. In other situations, substantive comments indicate that a data update is needed. For example, if the weight that a respondent reports is outside the pre-determined acceptable range programmed in CATI, the interviewer would first ask the respondent to confirm the response, then would enter “Don’t Know” as the answer in CATI, and then would add a comment with the respondent’s actual weight. In this case, the data preparation staff reviewing the comment later would enter the correct weight value CATI data file.

At the beginning of the each CHIS cycle or when new questions are added mid-cycle, comments are also used to identify problems such as misunderstood questions or logic in a series of questions, or issues with the response options for a question. In previous CHIS cycles, response option sets for some question items were amended or updated in the CATI system during the survey field period. Other such changes have occurred in preparation for the next CHIS cycle. These changes have helped reduce the number of interviewer comments and lessen the amount of data preparation work. For CHIS 2011-2012, the only changes to the response options were made after data collection had been completed. <u>However</u>, for CHIS 2011-12, the only changes to the response options were made after data collection had been completed. New codes were created after a number of similar of responses were found during the review of “other specify” text. The decision to create a new response options was made if the total number of entries that could be grouped under a new category was larger than the number of entries for any of the existing codes.

Several items yielded substantial numbers of responses outside the standard response set. An example is AK25 from the adult extended interview, “Do you own or rent your home?” Interviewers recorded responses in the comment field for this item such as “I own my home but rent the space it occupies.” Table 2-1 provides examples of items and responses that interviewers initially had difficulty coding. These examples are unchanged from CHIS 2009, as these items have continued to be among the most difficult to code.

Westat data preparation and project staff held weekly meetings during data collection to cover data-related issues, review comments, and developed case-specific solutions for pending or interim problem cases. Comments and cases under review included both complete and incomplete (interim status) interviews.

Table 2-1. Examples of difficult responses to code in CHIS 2011-2012

CATI Screen ID	Question and Response Options	Respondents’ Answers:
AK25	Do you own or rent your home? 1. OWN 2. RENT 3. OTHER ARANGEMENT -7. REFUSED -8. DON'T KNOW	“Own the home, but rent the space it occupies.”
AK1	Which of the following were you doing last week? 1. Working at a job or business, 2. With a job or business but not at work, 3. Looking for work, or 4. Not working at a job or business? -7. REFUSED -8. DON'T KNOW	“Working as a volunteer.”
AL9	Are you legally blind? 1. YES 2. NO -7. REFUSED	“I am blind in one eye.”

2.3 Coding with Text Strings

Most items in the CHIS 2011-2012 had only closed-ended response options, so coding of open-ended responses was not needed. However, the survey had a number of other-specify questions, that required coding of narrative text strings recorded by interviewers. Other-specify questions had specific response categories but also allowed for text or values to be typed into an “other” category. For example, question AA5 in the adult extended interview asked respondents “And what is your Latino or Hispanic ancestry or origin? Such as Mexican, Salvadoran, Cuban, Honduran -- and if you have more than one, tell me all of them.” An “other” category was available for responses that fell outside the list of categories that were read as a part of the question. Additional questions with an “other” category from the CHIS 2011-2012 adult extended interview included:

- Racial/ethnic ancestry (AA5, AA5A, AA5E, AA5E1);
- Tribal names (AA5B, AA5D);
- Sexual orientation (AD46);
- Country of birth (AH33, AH34, AH35);
- Languages spoken at home (AH36);
- Place visited for health care (AH3);
- Place visited for flu vaccine (AB57);
- How first found out about breast cancer (AB60);
- Health insurance coverage items (AI15, KAI15, AI15A, KAI15A, AI17A, KAI17A, AI45, KAI45, AI45A, KAI45A, AI36, KAI36, AI24, KAI24, AL19);
- Child/adolescent health insurance coverage items (CF7, KCF7, CF18, KCF18, IA18, KIA18, CF29, KCF29, IA29, KIA29, CF1A, CF2A, KCF2A, IA2A, IA7, KIA7).
- Adult/child/adolescent Insurance plan names (AH50, AI22A, MA2, MA7, KAH50, KAI22A, KMA2, KMA7);
- Reason no longer receiving behavioral health treatment (AF80);
- Country of birth (AI56, AI56C, AI56T);
- Languages used by doctor to speak to respondent (AJ50);

Questions with an “other (specify)” category in the child and adolescent interviews included:

- Child condition or disability (CA10A);
- Adolescent race and ethnicity (TI1A, TI2, TI2A,TI2C,TI2D,TI2D1);
- Child race and ethnicity (CH2, CH3, CH4, CH6, CH7, CH7A);
- Child/teen languages spoken at home (CH17, TI7);
- Child/mother/father place of birth (CH8, CH11, CH14);
- Adolescent country of birth (TI3);
- Child/adolescent school name/type of school (CB22, TA4B);

- Child/adolescent usual source of health care (CD3, TF2);
- Place child got last flu vaccine (CD42);
- Language used by child’s doctor to talk to parent (CD31)
- Type of STD adolescent tested for (TH32);
- Reason for adolescent not visiting dentist in past year (TM1);
- Reason for adolescent to have changed school (TA7);
- Place where adolescent usually eats breakfast (TD20);
- Place where adolescent usually eats lunch (TD21); and
- Person teen admires (TH23).

Westat data preparation staff reviewed these responses and up-coded them to the existing categories whenever possible. Additional response codes were added to a limited number of survey items to accommodate answers recorded in the other-specify category. The updated response codes for these items are given in Table 2-2. These items are the same as those presented in the 2009 report; the codes were added to the CATI database in CHIS 2009, but not to the 2011 CATI instruments.

CATI edit specifications were initially prepared by Westat staff and then forwarded to UCLA for review, comment, and approval. The specifications were then implemented to improve data quality by informing interviewers when an out-of-acceptable-range or seemingly improbable response was recorded. Edit specifications enabled interviewers to identify and correct potential errors with the respondent during the interview and eliminated the need for a call back.

Soft-range edits were activated during the interview when the respondent gave an unlikely response (a value outside the specified range). The CATI system responded by placing a message on the screen and required the interviewer to re-enter the response. This system feature gives the interviewer an opportunity to verify that the response is recorded accurately or re-ask the question to be certain the respondent understood what was being asked as needed. Hard-range edits prevented recording unacceptable values. For example, for a question on how many glasses of juice the adolescent respondent had the previous day, the soft range is 0-9, the hard range 0-20. During data collection, one hard-range edit specification (variable AE7, number of servings of vegetables eaten in the past month, from 120 to 300) was revised to accept the actual range of responses being collected. Also, moving from 2011 to 2012 during data collection, all items incorporating a specific year were updated appropriately.

In circumstances when the respondent insisted on giving a response that violated the soft- or hard-edit specifications, interviewers recorded the respondent’s answer in the comment field and data preparation staff reviewed and updated the case as needed.

Table 2-2. Response codes added to CHIS 2011-2012

Variable Name	Question Name	Question Text	New Code	Response Description
Adult Interview Questions:				
AH3	QA11_H2	{What kind of place do you go to most often—a medical/Is your doctor in a private} doctor's office, a clinic or hospital clinic, an emergency room, or some other place?	4 5 6	Complementary and alternative medicine Family/friend is health provider Internet/library
AI24	QA11_H72	What is the ONE MAIN reason why you do not have any health insurance?	9 10	Feels no need/healthy No reason/has not thought about insurance
AI36	QA11_H70	What is the ONE MAIN reason why you did not have any health insurance during those months?	9 10	Feels no need/healthy No reason/has not thought about insurance
CF18	QA11_I26	What is the ONE MAIN reason {CHILD} does not have any health insurance?	9 10	Feels no need/healthy No reason/has not thought about insurance
CF29	QA11_I36	What is the ONE MAIN reason {CHILD} did not have any health insurance during the time {he/she/he or she} wasn’t covered?	9 10	Feels no need/healthy No reason/has not thought about insurance
IA18	QA11_I62	What is the ONE MAIN reason why {ADOLESCENT} does not have any health insurance?	9 10	Feels no need/healthy No reason/has not thought about insurance
Child Interview Questions:				
CD3	QC11_D2	{What kind of place do you take {him/her} to most often—a medical/Is {his/her} doctor in a private} doctor's office, a clinic or hospital clinic, an emergency room, or some other place?	4	Complementary and alternative medicine
Adolescent Interview Questions:				
TH23	QT11_L3	Is this person a family member, an athlete, an entertainer, a teacher, a friend your own age, or someone else?	9	Writer/Author

Source: UCLA Center for Health Policy Research, 2011 California Health Interview Survey.

2.4 Verifying Data Updates

Updates to the original interview data were required in a variety of circumstances as described above. A series of techniques verified that the data were updated accurately. The CATI case identification number was also recorded to ensure that updates were associated with the appropriate case. A printout was created and checked for accuracy, logical effects on any other questions, or skip patterns in the questionnaire. Next, the updates were entered into the database and verified again by matching the resulting information against the printout. For more complicated circumstances, the data preparation staff carefully reviewed interviewer comments, messages, and problem descriptions to verify data updates.

An entry in an electronic transaction journal was created automatically for each data update. Transaction journal entries maintained information such as the CATI case identification number, the initial data value(s), the updated value(s), and the date that the update was made. The editing and verification process started as soon as completed interviews became available and continued during the entire the data collection period. Approximately 103,500 database values were updated and verified for CHIS 2011-2012.

Cases with similar problems were reviewed together and then updated at one time in manageable batches. This process ensured consistency in the handling of discrete data problems. Following the series of updates, a program checked for the full set of errors that had been identified to date to ensure that data editing had not created any new errors. Frequency distributions and cross-tabulations of survey variables were used extensively by data preparation staff to verify data updates.

Structural edits assessed the integrity of the CATI database (e.g., verifying that all database records that should exist actually existed, and those that should not exist did not), and, as necessary, edits that evaluated complex skip patterns were run periodically during data collection. When discrepancies were discovered, problem cases were identified and reviewed and updates were made as necessary. If data were incorrectly keyed in the database, the audit trail for the interview (a keystroke-by-keystroke record of all responses entered during the CATI interview) could be retrieved to determine the appropriate response. The interview audit trail was especially useful for reconstructing interviews interrupted unexpectedly by a power failure or system crash. A report was created every morning to find any instances of crashes during the previous day of interviewing. The number of interviews restored in CHIS 2011-2012 averaged roughly five per week. Most of these were re-fielded after the update and completed in the usual manner.

3. Geographic Coding

For CHIS 2011-2012, Westat was responsible for delivering geo-coded survey data for items from the adult extended interview, or the child interview in “child-first” cases, related to geographic location of the respondent’s residence. The self-reported county was used to assign cases to landline sample strata as described in *CHIS 2011-2012 Methodology Series: Report 1: Sample Design.* Westat also prepared and delivered more specific geocodes based on the respondent-reported address and other information. The geographic coding process for CHIS 2011-2012 used the 2011 NAVTEQ database of roads and corresponding NAVTEQ Census Block boundary definitions.

3.1 County of Residence

The CHIS 2011-2012 adult extended interview asked all respondents the name of the county where they lived: “To be sure we are covering the entire state, what county do you live in?” (AH42). In addition, for cases in which an address had been matched to the sampled telephone number[4], interviewers verified the street address and ZIP code with the adult respondent (AO1) and then collected the name of a nearby cross-street (AM9). These same questions were asked of adults who completed the child interview under the “child first” protocol. The child-first protocol allowed completion of the child interview before the adult extended interview was conducted. See *CHIS 2011-2012 Methodology Series: Report 2 – Data Collection Method* for details regarding the child-first protocol.

If there was no matched address for a given case, respondents were asked to provide their ZIP code (AM7), their street address (AO2) and then the name of the nearest cross-street (AM9). Adult respondents who refused to provide a complete street address with house number were asked just for the name of the street they lived on (AM8) and the nearest cross street.

Because telephone numbers were assigned to sampling strata based on the telephone area code and exchange (see *CHIS 2011-2012 Methodology Series: Report 1 - Sample Design*), and some exchanges serve more than one county or city, the actual stratum where the respondent resides may differ from the sampling stratum. Both to monitor the sample yield during data collection and to ensure that the analysis file reflects the sampled person’s actual residence, it was important to assign each adult who completed the extended interview to the correct stratum that the adult self-reported as the residence.

[4] The verification was not done if the telephone number was unlisted or if the sample vendor indicated that the number was on the “do not call” list.

The following two questions are asked toward the end of the adult extended interview and were used to make the self-reported stratum assignment that is used for data collection targets:

AH42. "To be sure we are covering the entire state, what county do you live in?"

and

AM7. "What is your ZIP code?"

Table 3-2 is a list of ZIP codes within each landline sampling stratum[5] for CHIS 2011-2012. The final self-reported stratum that was included in the final data file was determined by applying the geocodes developed by UCLA and CHIS staff. See *CHIS 2011-2012 Methodology Series: Report 5 - Weighting and Variance Estimation,* Section 8.5, for a fuller discussion of this process.

The final distribution of completed landline sample adult extended interview cases by self-reported and original sampling stratum is presented in Table 3-2 at the end of this chapter. Generally, the frequency counts show that there is good correspondence between the original sampling stratum and the self-reported stratum. The self-reported stratum may differ from the original sampling stratum, however, because the sampling stratum may have been incorrect or the respondent may have incorrectly reported the county of residence.

3.2 Geocoding Process

Two methods of geocoding using NAVTEQ software were employed for CHIS 2011-2012. The first option was to have the software automatically match (batch match) the input addresses to a spatial database of roads, which returned the address's latitude/longitude, state FIPS and county FIPS. If the software was unable to match to the street address, it automatically matched to the geographic ZIP centroid as a fallback. In such cases, the latitude/longitude, state FIPS code and county FIPS code of the ZIP code centroid was provided.

The second method performed the same batch process as described in option 1 above, but did not automatically default to a ZIP centroid match. If a batch match was not obtained, Westat staff interactively examined the unmatched records (excluding PO boxes and rural routes) to try and determine the reason why the software could not automatically match the address. Sometimes this was due to misspelled street names, city names, etc., or to missing house numbers. Westat corrected the address to

[5] The cell sample used larger geographic areas as strata. See *CHIS 2011-2012 Methodology Series: Report 1 – Sample Design.*

match the street database, or matched to the segment's nearest intersection. If the street address or nearest intersection could not be identified, Westat would then match to geographic ZIP centroid. The frequencies of assigned geocodes by rule and sample type are shown in table 3-1.

Table 3-1. Number of geocodes assigned by rule and by sample type

Rule Number		1=RDD	2=CELL	3=KRVT	4=AIAN	TOTAL
(1)	address provided or confirmed by respondent in section n	30,596	7,734	799	220	39,349
(3)	address provided by respondent for mailing incentive	8	256	0	1	265
(4)	address provided by respondent to get a copy of the pre-notification letter	749	2	26	8	785
(5)	cross street (am8 and am9) and zip (am7)	1,019	10	9	10	1,048
(6)	matched to street centroid within zip	366	122	4	16	508
(7)	am7 - zip only (zip centroid)	987	260	26	22	1295
(8)	matched to city centroid	1	2	0	0	3
(9)	matched to zip centroid using best zip when multiple zips available	798	46	12	1	857
(10)	matched to the population centroid of respondent-reported county	0	258	0	0	258
(11)	matched to the population centroid of sampled county or stratum	0	51	0	0	51
(12)	original geocode outside ca recoded to population centroid of respondent-reported county	0	8	0	0	8
(13)	original geocode outside ca recoded to population centroid of sampled county or stratum	0	1	0	0	1
(14)	respondent-reported or sampled Los Angeles or San Diego geocoded through reverse directory look-up	2	129	0	0	131
Total		34,526	8,879	876	278	44,559

Table 3-3. Final distribution of adult extended completed cases by self-reported and original sampling stratum, landline/list sample for CHIS 2011-2012

Stratum Name	Sampling Stratum Count	Removed	Added	Final Self-reported Stratum Count
1.1 - LA SPA 1	609	5	9	613
1.2 - LA SPA 2	1,137	27	37	1,147
1.3 - LA SPA 3	1,255	30	37	1,262
1.4 - LA SPA 4	1,396	123	81	1,354
1.5 - LA SPA 5	440	13	46	473
1.6 - LA SPA 6	566	128	92	530
1.7 - LA SPA 7	772	114	74	732
1.8 - LA SPA 8	953	29	88	1,012
2 - SAN DIEGO	4,168	9	10	4,169
3 - ORANGE	1,925	48	26	1,903
4 - SANTA CLARA	1,162	9	46	1,199
5 - SAN BERNARDINO	1,056	16	31	1,071
6 - RIVERSIDE	1,350	13	17	1,354
7 - ALAMEDA	1,049	50	13	1,012
8 - SACRAMENTO	995	9	14	1,000
9 - CONTRA COSTA	657	5	52	704
10 - FRESNO	442	9	7	440
11 - SAN FRANCISCO	621	19	7	609
12 - VENTURA	465	3	19	481
13 - SAN MATEO	505	36	18	487
14 - KERN	473	6	3	470
15 - SAN JOAQUIN	355	2	3	356
16 - SONOMA	356	6	12	362
17 - STANISLAUS	408	15	1	394
18 - SANTA BARBARA	424	5	4	423
19 - SOLANO	400	13	10	397
20 - TULARE	379	3	6	382
21 - SANTA CRUZ	406	15	4	395
22 - MARIN	444	6	2	440
23 - SAN LUIS OBISPO	411	4	5	412
24 - PLACER	388	14	24	398
25 - MERCED	419	5	17	431
26 - BUTTE	373	3	13	383
27 - SHASTA	407	20	4	391
28 - YOLO	370	19	8	359
29 - EL DORADO	378	4	9	383
30 - IMPERIAL	460	3	1	458
31 - NAPA	462	7	15	470
32 - KINGS	445	1	3	447
33 - MADERA	470	7	1	464

Table 3-3. Final distribution of adult extended completed cases by self-reported and original sampling stratum, landline/list sample for CHIS 2011-2012 (continued)

Stratum Name	Sampling Stratum Count	Removed	Added	Final Self-reported Stratum Count
34 - MONTEREY	298	3	11	306
35 - HUMBOLDT	325	6	3	322
36 - NEVADA	442	14	7	435
37 - MENDOCINO	436	5	-	431
38 - SUTTER	417	12	36	441
39 - YUBA	482	55	5	432
40 - LAKE	462	6	1	457
41 - SAN BENITO	465	1	2	466
42 - TEHAMA, ETC.	334	1	18	351
43 - DEL NORTE, ETC.	299	4	5	300
44 - TUOLUMNE, ETC.	345	4	7	348
Total	34,056	964	964	34,056

Source: UCLA Center for Health Policy Research, 2011 California Health Interview Survey.

3.3 School Name and Geographic Coding

In CHIS 2011-2012, the child and adolescent interviews included an item that collected the name of the school attended by the selected child or adolescent (CB22 and TA4B, respectively). The sampled adult or the most knowledgeable adult (MKA) reported the child's school name, and the sampled adolescent answered for him- or herself. Interviewers recorded the respondent's answers as a verbatim text entry in the CATI system. School latitude and longitude were then assigned to each school-aged child and adolescent case for which a school name was reported.

A review of the child interview data showed a number of spelling problems associated with item CB22 ("What is the name of the school {CHILD NAME /AGE/SEX} goes to or last attended"?). In many problem cases, the English-speaking adult respondent was reporting a Spanish school name (and was speaking to an English speaking interviewer). Asian and some Latino respondents, whose first language is not English, had similar difficulties in accurately reporting or spelling the school name.

Westat data preparation staff used the California School Directory, https://maps.google.com, and www.publicschooolreview.com in conjunction with the respondent's ZIP code as resources to improve the quality of school names and their location before release to UCLA for geocoding. SAS statistical programming was used to merge in open text from CB22 and TA4B as well as county of residence with relevant data fields in the school list database. Full matches were assigned a successful matching code. For cases that could not be automatically matched using statistical programming (e.g. spelling errors, county mismatch), additional CHIS variables were used to accurately identify and manually assign the name of the school. These included age of respondent, ZIP code, city, and county of home residence. Additional information in the state school database was used to verify the child or adolescent’s school, including school district, school county, school city, school ZIP code, and school grade range. Web-based searches were also used to assign geographic school information not found in the California School Directory.

For all matched public schools, latitude and longitude were provided in the state-issued school database of California. Geocoding for private schools was performed by UCLA. Cases for which the child or adolescent attended a home school or non-traditional program or where a school could not be identified were assigned a value indicating “undetermined.” Children under the age of 5 years were assigned an inapplicable value.

4. Race and Ethnicity Coding

This section describes how we handled situations when the respondent reported a race or ethnicity that was not classified into one of the pre-existing categories. These responses were recorded in the “other specify” category as a text string. The procedures for coding these “other specify” responses into existing codes (up-coding) or leaving them in the other category are presented here.

The first question in the series of items related to race and ethnicity (question AA4 in the adult extended interview) asked if the respondent was Latino or Hispanic. If the response to this item was “yes,” then a question (AA5) was asked about the specific origin (Mexican, etc.) and this includes an “other” category with responses entered by interviewers as text in item AA5OS. Question AA5A from CHIS 2007 asked respondents for their race: “Please tell me which one or more of the following you would use to describe yourself. Would you describe yourself as Native Hawaiian, Other Pacific Islander, American Indian, Alaska Native, Asian, Black, African American, or White?” The race question allowed the respondent to indicate that they belonged to any or all of the coded races (Native Hawaiian, Other Pacific Islander, American Indian or Alaska Native, Asian, African American, or White) and also to say “other” race. The “other specify” race was recorded in text (AA5AOS). Another item followed if the respondent indicated they identified with more than one race or ethnicity. That item asked which race or ethnicity the respondent most identified with (AA5F). This item did not allow interviewers to collect an “other-specify,” but responses to this item could be used in the coding decisions for other items.

4.1 Coding Procedures

The procedures for the race and ethnicity coding Westat performed were designed specifically to support the data needs for weighting the CHIS sample. If codes could not be assigned for race or ethnicity they were left as missing and were later imputed. The imputation procedures are described in *CHIS 2011-2012 Methodology Series: Report 5 - Weighting and Variance Estimation*.

The procedures used were consistent with those used to code the 2010 Census data and with those used in prior CHIS iterations. The methods used in the 2010 Census are documented in Census 2010 Redistricting Data (Public Law 94-171) Summary File – Technical Documentation (U.S. Census Bureau, 2011) available at http://www.census.gov/prod/cen2010/doc/pl94-171.pdf. The specific sections of interest are in Appendix B, pages B-2 and B-3. When we refer to the Census procedures, we mean our interpretation of the information in this document.

An initial review of cases showed that the largest group of cases with “other race” categories were ones in which the respondent identified as being Hispanic or Latino and did not identify with any pre-coded race categories. The typical response to the “other race” was “Hispanic.” Following the Census procedures, the person was left in the “other race” category and the “other specify” text remained as it was.

The specific procedures and guidelines we used are detailed below and are unchanged from those used in the past administrations of the survey. Responses captured in the other specify text field were retained and included in the final data set delivery to UCLA to accommodate other research and analytic needs.

- If the “other specify” text clearly should have been included in an existing code (following the Census procedures), then it was up-coded and removed from the “other” category. For example, if the respondent was coded only as other race and the “other specify” was “Irish,” then the code for “white” was up-coded to “yes,” other race was revised to “no” and the other specify text eliminated.

- If the “other specify” text did not fit into an existing code (following the Census procedures), then it was left in the “other” category with the existing text in the “other specify.” For example, if the “other specify” text for race was “American” and no other race category was identified, then no changes were made in the responses.

- If the respondent was coded as being Hispanic or Latino, we never revised this code based upon information in the other specify comments of the other variables. For example, if the person was coded as “Hispanic” and the specific Hispanic origin item was only coded as “other” with the text “Jewish,” then the Hispanic code was not altered.

- If the respondent was coded as not being Hispanic or Latino but the text in the “other specify” field for race indicated they were Hispanic or Latino, then the Hispanic or Latino coding was revised to “yes.” In addition, the specific Hispanic origin code was made consistent with text in the “other specify” text from the race variable, if it was possible to do so. In the case where this was not possible, the “other” Hispanic origin category was coded and the text copied from the race variable to the “other specify” for Hispanic origin. (This procedure is an elaboration of the ones above to deal with the cross-variable coding.) For example, if the race “other specify” code was “Mexican,” then the Hispanic or Latino category was revised to be “yes” and the Hispanic origin code was coded as “yes” for Mexican.

- If the “other race” text was similar to “none of above,” we left the response as it was.

- If the “other race” text was similar to “human race,” we coded this as a refusal. The race was then imputed along with other cases that were more direct refusals.

The Census procedures clearly state that persons who say they have European, Middle Eastern, or North African origin are to be classified as “White” race. This rule has many implications. For example, if a person says they are not Hispanic and only identify the “other race” as being “Spanish”, we would up-code Hispanic origin to “yes” (to be consistent with the Census procedures for Hispanic origin) and then up-code “race” to “White” (since the person is of European origin).