October 2019

CHIS 2017-2018 Methodology Report Series

Report 3

Data Processing Procedures

CALIFORNIA HEALTH INTERVIEW SURVEY

CHIS 2017-2018 METHODOLOGY SERIES

REPORT 3

DATA PROCESSING PROCEDURES

OCTOBER 2019

This report was prepared for the California Health Interview Survey by David Dutwin, Susan Sherr, Arina Goyle, Kathy Langdale, and Jonathan Best of SSRS.

www.chis.ucla.edu

This report describes the data processing and editing procedures for CHIS 2017-2018 performed by SSRS. This report discusses standard data editing procedures and addresses the steps taken for ensuring data quality. It also present discussions on special procedures of editing and coding of geography as well as race and ethnicity survey items.

Suggested citation:

California Health Interview Survey. *CHIS 2017-2018 Methodology Series: Report 3 - Data Processing Procedures*. Los Angeles, CA: UCLA Center for Health Policy Research, 2019.

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 2017-2018 came from multiple sources: the California Department of Health Care Services, the California Department of Health Care Services (Mental Health Services Division), the California Department of Public Health, The California Endowment, the California Health Benefit Exchange, the California Health Care Foundation, the California Wellness Foundation, First 5 California, Kaiser Permanente, San Diego County Health and Human Services Agency, Imperial County Public Health Department, UC San Francisco, the Stupski Foundation, California Rural Indian Health Board, and San Francisco Department of Public Health.

PREFACE

Data Processing Procedures is the third in a series of methodological reports describing the 2017-2018 California Health Interview Survey (CHIS 2017-2018). The other reports are listed below.

CHIS is a collaborative project of the University of California, Los Angeles (UCLA) Center for Health Policy Research, the California Department of Public Health, and the Department of Health Care Services. SSRS was responsible for data collection and the preparation of five methodological reports from the 2017-2018 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.

Methodological Report Series for CHIS 2017-2018

The methodological reports for CHIS 2017-2018 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. After the Preface, each report includes an "Overview" (Chapter 1) that is nearly identical across reports, followed by detailed technical documentation on the specific topic of the report.

Report 3: Data Processing Procedures (this report) describes the data processing and editing procedures for CHIS 2017-2018. 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.

For further methodological details not covered in this report, refer to the other methodological reports in the series at http://www.chis.ucla.edu/chis/design/Pages/methodology.aspx. General information on CHIS data can be found on the California Health Interview Survey Web site at http://www.chis.ucla.edu or by contacting CHIS at CHIS@ucla.edu.

Table of Contents

List of Tables

1. CHIS 2017-2018 SAMPLE DESIGN AND METHODOLOGY SUMMARY

1.1 Overview

A series of five methodology reports are available with more detail about the methods used in CHIS 2017-2018.

- 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. For methodology reports from previous CHIS cycles, go to http://www.chis.ucla.edu/chis/design/Pages/methodology.aspx

The CHIS is a population-based telephone survey of California's residential, noninstitutionalized population conducted every other year since 2001 and continually beginning in 2011. CHIS is the nation's largest state-level health survey and one of the largest health surveys in the nation. The UCLA Center for Health Policy Research (UCLA-CHPR) conducts CHIS in collaboration with the California Department of Public Health and the California Department of Health Care Services. 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 and optimized to meet 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 racial and 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 UCLA-CHPR's publication page at http://healthpolicy.ucla.edu/publications/Pages/default.aspx for examples of CHIS studies).

1.2 Switch to a Continuous Survey

From the first CHIS cycle in 2001 through 2009, CHIS data were collected during a 7 to 9 month period every other year. Beginning in 2011, CHIS data have been collected continually over a 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.

CHIS 2017-2018 data were collected between June 2017 and January 2019. 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 thus not eligible to participate in CHIS. When the weights are applied to the data, the results represent California's residential population during the two year period for the age group corresponding to the data file in use (adult, adolescent, or child). In CHIS 2017-2018, data users will be able to produce single-year estimates using the weights provided (referred to as CHIS 2017 and CHIS 2018, respectively).

See what's new in the 2017-2018 CHIS sampling and data collection here:
http://www.chis.ucla.edu/chis/design/Documents/whats-new-chis-2017-2018.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 statistical code, is available at http://www.chis.ucla.edu/chis/analyze/Pages/sample-code.aspx.

Additional documentation on constructing the CHIS sampling weights is available in the *CHIS 2017-2018 Methodology Series: Report 5—Weighting and Variance Estimation* posted at http://www.chis.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.

1.3 Sample Design Objectives

The CHIS 2017-2018 sample was designed to meet the two sampling objectives discussed above: (1) provide estimates for adults in most counties and in groups of counties with small populations; and (2) provide estimates for California's overall population, major racial and ethnic groups, and for several smaller racial and 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 RDD sample was designed to achieve the required number of completed adult interviews by using approximately 50% landline and 50% cellular phone numbers. For the RDD sample, the 58 counties in the state were grouped into 44 geographic sampling strata, and 14 sub-strata were created within the two most populous counties in the state (Los Angeles and San Diego). The same geographic stratification of the state has been used since CHIS 2005. The Los Angeles County stratum included eight sub-strata for Service Planning Areas, and the San Diego County stratum included six sub-strata for Health Service Districts. Most of the strata (39 of 44) consisted of a single county with no sub-strata (see counties 3-41 in Table 1-1). Three multi-county strata comprised the 17 remaining counties (see counties 42-44 in Table 1-1). A sufficient number of adult interviews were allocated to each stratum and sub-stratum to support the first sample design objective for the two-year period—to provide health estimates for adults at the local level. Asian surname sample list frames added 127 Korean, and 214 Vietnamese adult interviews based on self-identified ethnicity for the 2017-2018 survey year.[1] Additional samples from both the landline and cell phone frames produced 1,375 interviews in 2017-2018 within San Diego County. In 2018, an oversample of American Indian and Alaska Native residents of California added 317 completed interviews, and specific gender and ethnic oversamples in San Francisco provided an additional 498 interviews. Furthermore, an address-based sample from the USPS Delivery Sequence File produced 339 landline or cell phone interviews in 2017 within the northern part of Imperial County.

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

[1] For the 2017-2018, RDD landline and cell sample frames produced totals of 290 Korean, and 235 Vietnamese adult interviews.

Table 1-1. California county and county group strata used in the CHIS 2017-2018 sample design

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

Source: UCLA Center for Health Policy Research, 2017-2018 California Health Interview Survey.

The CHIS RDD 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). However, given the smaller sample sizes of one-year data files, two or more pooled cycles years of CHIS data are generally required to produce statistically stable estimates for small population groups such as racial/ethnic subgroups, children, teens, etc. 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 increase the sample size further for Koreans and Vietnamese.

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 2017-

2018, the goal was to complete approximately 50% of all RDD interviews statewide with adults contacted via cell phone. Because the geographic information available for cell phone numbers is limited and not as precise as that for landlines, cell phone numbers were assigned to the same 44 geographic strata (i.e., 41 strata defined by a single county and 3 strata created by multiple counties) using a classification associated with the rate center linked to the account activation. The cell phone stratification closely resembles that of the landline sample and has the same stratum names, though the cell phone strata represent slightly different geographic areas than the landline strata. The adult owner of the sampled cell phone number was automatically selected for CHIS. Cell numbers used exclusively by children under 18 were considered ineligible. A total of 880 teen interviews and 3,186 child interviews were completed in CHIS 2017-2018 with approximately 48% of teen interviews and 65% of child interviews coming from the cell phone sample.

The cell phone sampling method used in CHIS has evolved significantly since its first implementation in 2007 when only cell numbers belonging to adults in cell-only households were eligible for sampling adults. These changes reflect the rapidly changing nature of cell phone ownership and use in the US.[2] There have been three significant changes to the cell phone sample since 2009. First, all cell phone sample numbers used for non-business purposes by adults living in California were eligible for the extended interview. 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 was the inclusion of child and adolescent extended interviews. The third, enacted in CHIS 2015-2016 was to increase the fraction of the sample comprised of cell phones from 20% to 50% of completed interviews. In 2017-2018, we additionally sampled out-of-area cell phone numbers. These are cell phone numbers with exchanges outside of California that can be matched to an address that is within California, indicating that the owner of the cell phone resides in California but purchased a cell phone in another state.

The cell phone sample design and targets by stratum of the cell phone sample have also changed throughout the cycles of the survey. In CHIS 2007, a non-overlapping dual-frame design was implemented where cell phone only users were screened and interviewed in the cell phone sample. Beginning in 2009, an overlapping dual-frame design has been implemented. In this design, dual phone users (e.g., those with both cell and landline service) can be selected and interviewed from either the landline or cellphone samples.

[2] https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201906.pdf

The number of strata has also evolved as more information about cell numbers has become available. In CHIS 2007, the cell phone frame was stratified into seven geographic sampling strata created using telephone area codes. In CHIS 2009 and 2011-2012, the number of cell phone strata was increased to 28. These strata were created using both area codes and the geographic information assigned to the number. Beginning in CHIS 2011, with the availability of more detailed geographic information, the number of strata was increased to 44 geographic areas that correspond to single and grouped counties similar to the landline strata. The use of 44 geographic strata continued in CHIS 2017-2018.

1.4 Data Collection

To capture the rich diversity of the California population, interviews were conducted in six languages: English, Spanish, Chinese (Mandarin and Cantonese dialects), Vietnamese, Korean, and Tagalog. Tagalog interviews were conducted for part of the CHIS 2013-2014 cycle, but 2015-2016 were the first cycle years that Tagalog interviewers were conducted from the beginning of data collection. These languages were chosen based on analysis of 2010 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.

SSRS designed the methodology and collected data for CHIS 2017-2018, under contract with the UCLA Center for Health Policy Research. SSRS is an independent research firm that specializes in innovative methodologies, optimized sample designs, and reaching low-incidence populations. For all sampled households, SSRS 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 their parent or legal guardian. Thus, up to three interviews could have been completed in each household. Children and adolescents were generally sampled at the end of the adult interview. If the screener respondent was someone other than the sampled adult, 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 first used in CHIS 2005 and has been continued in subsequent CHIS cycles because it substantially increases the yield of child interviews. While numerous subsequent attempts were made to complete the adult interview for child-first cases, the final data contain completed child and 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 2017-2018 by the type of sample (landline RDD, surname list, cell RDD, and ABS). Note that these figures were accurate as of data collection completion for 2017-2018 and may differ slightly from numbers in the data files due to data cleaning and edits. Sample sizes to compare against data files you are using are found online at http://www.chis.ucla.edu/chis/design/Pages/sample.aspx.

Table 1-2. Number of completed CHIS 2017-2018 interviews by type of sample and instrument

Type of sample[1]	Adult[2]	Child	Adolescent
Total all samples	42,330	3,186	880
Landline RDD[3]	18,896	1,049	434
Cell RDD	21,554	1,996	409
Vietnamese surname list landline	188	10	5
Vietnamese surname list cell phone	80	10	3
Korean surname list landline	354	16	3
Korean surname list cell phone	56	5	1
Both Korean and Vietnamese landline	48	1	1
Imperial County ABS Oversample	339	42	15
AIAN Oversample landline	130	10	-
AIAN Oversample cell phone	187	20	3
San Francisco Oversample landline	148	4	1
San Francisco Oversample cell phone	350	23	5

Source: UCLA Center for Health Policy Research, 2017-2018 California Health Interview Survey.

[1] Completed interviews listed for each sample type refer to the sampling frame from which the phone number was drawn. Interviews could be conducted using numbers sampled from a frame with individuals who did not meet the target criteria for the frame but were otherwise eligible residents of California. For example, only 157 of the 190 adult interviews completed from the Vietnamese surname list involved respondents who indicated being having Vietnamese ethnicity.

[2] Includes interviews meeting the criteria as partially complete.

[3] Breakdown of completes by frame deviates slightly from original sample numbers due to numbers changing frames following post-sampling database processing.

Interviews in all languages were administered using SSRS's computer-assisted telephone interviewing (CATI) system. The average adult interview took about 42 minutes to complete. The average child and adolescent interviews took about 19 minutes and 24 minutes, respectively. For "child-first" interviews, additional household information asked as part of the child interview averaged about 14 minutes. Interviews in non-English languages typically took longer to complete with an average length of about 50 minutes for the adult interview, 29 minutes for the teen, and 23 minutes for the child. More than eight percent of the adult interviews were completed in a language other than English, as were about 13 percent of all child (parent proxy) interviews and six percent of all adolescent interviews.

Table 1-3 shows the major topic areas for each of the three survey instruments (adult, child, and adolescent). If questions were asked in only one year of survey implementation, the specific year is indicated in the table.

Table 1-3. CHIS 2017-2018 survey topic areas by instrument

Health status	**Adult**	**Teen**	**Child**
General health status	✓	✓	✓
Days missed from work or school due to health problems		✓	✓
Health conditions	**Adult**	**Teen**	**Child**
Asthma	✓	✓	✓
Diabetes, gestational diabetes, pre-diabetes/borderline diabetes	✓		
Heart disease, high blood pressure	✓		
Physical disability	✓		
Physical, behavioral, and/or mental conditions			✓
Developmental assessment, referral to a specialist by a doctor			✓
Mental health	**Adult**	**Teen**	**Child**
Mental health status	✓	✓	
Perceived need, access and utilization of mental health services	✓	✓	
Functional impairment, stigma, three-item loneliness scale (2017)	✓		
Suicide ideation and attempts	✓	✓	
Health behaviors	**Adult**	**Teen**	**Child**
Dietary and water intake, breastfeeding (younger than 3 years)	✓	✓	✓
Physical activity and exercise		✓	✓
Commute from school to home		✓	✓
Walking for transportation and leisure (2017)	✓		
Alcohol, cigarette, and E-cigarette use	✓	✓	
Marijuana use	✓	✓	
Opioid use	✓		
Chewing tobacco, tobacco flavors (2018)	✓	✓	
Exposure to second-hand smoke (2018)	✓		
Sexual behaviors	✓	✓	
HIV testing, HIV prevention medication	✓	✓	
Sleep and technology		✓	
Sedentary time		✓	✓
Contraceptive use	✓	✓	

(continued)

Table 1-3. CHIS 2017-2018 survey topic areas by instrument (continued)

Women's health	**Adult**	**Teen**	**Child**
Pregnancy status, postpartum care	✓		
Dental health	**Adult**	**Teen**	**Child**
Last dental visit, main reason haven't visited dentist	✓	✓	✓
Current dental insurance coverage	✓		✓
Condition of teeth	✓	✓	
Neighborhood and housing	**Adult**	**Teen**	**Child**
Safety, social cohesion	✓	✓	✓
Homeownership	✓		
Length of time at current residence (2017)	✓		
Park use, park and neighborhood safety		✓	✓
Civic engagement	✓	✓	
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)	✓	✓	✓
Communication problems with doctor	✓		✓
Discrimination (2017)	✓		
Timely appointment	✓	✓	✓
Access to specialist and general doctors	✓		
Tele-medical care	✓		
Care coordination (2018)	✓	✓	✓
Voter engagement	**Adult**	**Teen**	**Child**
Voter engagement	✓		
Food environment	**Adult**	**Teen**	**Child**
Access to fresh and affordable foods	✓		
Availability of food in household over past 12 months	✓		
Hunger	✓		
Health insurance	**Adult**	**Teen**	**Child**
Current insurance coverage, spouse's coverage, who pays for coverage	✓	✓	✓
Health plan enrollment, characteristics and assessment of plan	✓	✓	✓
Whether employer offers coverage, respondent/spouse eligibility	✓		
Coverage over past 12 months, reasons for lack of insurance	✓	✓	✓
High deductible health plans	✓	✓	✓
Partial scope Medi-Cal	✓		
Medical debt, hospitalizations	✓		

(continued)

Table 1-3. CHIS 2017-2018 survey topic areas by instrument (continued)

Public program eligibility	**Adult**	**Teen**	**Child**
Household poverty level	✓		
Program participation (CalWORKs, Food Stamps, SSI, SSDI, WIC, TANF)	✓	✓	✓
Assets, child support, Social security/pension	✓		
Medi-Cal eligibility, Medi-Cal renewal	✓		
Reason for Medi-Cal non-participation	✓	✓	✓
Bullying	**Adult**	**Teen**	**Child**
Bullying, school safety		✓	
Parental involvement/adult supervision	**Adult**	**Teen**	**Child**
Parental involvement			✓
Parental support, teach support		✓	
Child care and school	**Adult**	**Teen**	**Child**
Current child care arrangements			✓
Paid child care	✓		
First 5 California: Talk, Read, Sing Program / Kit for New Parents			✓
Preschool/school attendance, school name		✓	✓
Preschool quality			✓
School instability, school programs and organizational involvement		✓	
Employment	**Adult**	**Teen**	**Child**
Employment status, spouse's employment status	✓		
Hours worked at all jobs	✓		
Industry and occupation, firm size	✓		
Income	**Adult**	**Teen**	**Child**
Respondent's and spouse's earnings last month before taxes	✓		
Household income, number of persons supported by household income	✓		
Placement on quality of life ladder (2018)	✓		
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, gender identity	✓		
Gender expression		✓	
Living with parents	✓		
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, 2017-2018 California Health Interview Survey.

1.5 Response Rates

The overall response rates for CHIS 2017-2018 are composites of the screener completion rate (i.e., success in introducing the survey to a household and randomly selecting an adult to be interviewed) and the extended interview completion rate (i.e., success in getting one or more selected persons to complete the extended interview). For CHIS 2017-2018, the landline/list sample household response rate was 5.6 percent (the product of the screener response rate of 10.8 percent and the extended interview response rate at the household level of 52.0 percent). The cell sample household response rate was 3.5 percent, incorporating a screener response rate of 7.1 percent and household-level extended interview response rate of 49.0 percent. CHIS uses AAPOR response rate RR4 (see more detailed in *CHIS 2017-2018 Methodology Series: Report 4 – Response Rates*).

Within the landline and cell phone sampling frames for 2017-2018, the extended interview response rate for the landline/list sample varied across the adult (43.8 percent), child (60.0 percent) and adolescent (25.6 percent) interviews. The adolescent rate includes the process of obtaining permission from a parent or guardian.

The adult interview response rate for the cell sample was 40.9 percent, the child rate was 57.5 percent, and the adolescent rate was 18.0 percent in 2017-2018 (see Table 1-4a). Multiplying these rates by the screener response rates used in the household rates above gives an overall response rate for each type of interview for each survey year (see Table 1-4b). As in previous years, household and person level response rates vary by sampling stratum. CHIS response rates are similar to, and sometimes higher than, other comparable surveys that interview by telephone.

Table 1-4a. CHIS 2017-2018 response rates – Conditional

Type of Sample	Screener	Household	Adult (given screened)	Child (given screened & eligibility)	Adolescent (given screened & permission)
Overall	8.0%	49.9%	42.3%	58.3%	21.3%
Landline RDD/List	10.8%	52.0%	43.8%	60.0%	25.6%
Cell RDD/List	7.1%	49.0%	40.9%	57.5%	18.0%

Source: UCLA Center for Health Policy Research, 2017-2018 California Health Interview Survey.
Note. This table does not include the Imperial County, AIAN, and San Francisco oversamples.

Table 1-4b. CHIS 2017-2018 response rates – Unconditional

Type of Sample	Screener	Household	Adult (given screened)	Child (given screened & eligibility	Adolescent (given screened & permission)
Overall	8.0%	4.0%	3.4%	4.6%	1.7%
Landline RDD/List	10.8%	5.6%	4.7%	6.4%	2.8%
Cell RDD/List	7.1%	3.5%	2.9%	4.1%	1.3%

Source: UCLA Center for Health Policy Research, 2017-2018 California Health Interview Survey.
Note. This table does not include the Imperial County, AIAN, and San Francisco oversamples

To maximize the response rate, especially at the screener stage, an advance letter in six 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 39.1 percent of the landline RDD sample telephone numbers not identified by the sample vendor as business numbers or not identified by SSRS's dialer software as nonworking numbers, and for 100 percent of surname list sample numbers. Combining these two frames, advance letters were sent to 41.0 percent of all fielded landline telephone numbers. From the onset of 2017 fielding until April of 2018, cell phone sample with matched telephone numbers also received an advance letter. However, after a randomized experiment confirmed that the accuracy of the matching for cell phone sample did not warrant continuing these mailings, they were discontinued (for full experiment details, see Section 7.1 in *CHIS 2017-2018 Methodology Series: Report 4 – Response Rates*). Overall, across the two years, for cell sample, an advance letter was mailed for 27.2 percent of the RDD sample telephone numbers not identified by the sample vendor as business numbers or not identified by SSRS's dialer software as nonworking numbers, and for 100 percent of surname list sample numbers. Combining these two frames, advance letters were sent to 30.4 percent of all fielded cell telephone numbers. As in all CHIS cycles since CHIS 2005, a $2 bill was included with the CHIS 2017-2018 advance letter to encourage cooperation. Unlike previous cycles, additional incentives were not offered to cell phone and nonresponse follow up (NRFU) respondents.

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 any adult who was unable to complete the extended adult interview for themselves, in order to avoid biases for health estimates of chronically ill or handicapped people. Eligible selected persons were re-contacted and offered a proxy option. In CHIS 2017-2018, either a spouse/partner or adult child completed a proxy interview for 20 adults. A reduced questionnaire, with questions identified as appropriate for a proxy respondent, was administered.

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

1.6 Weighting the Sample

To produce population estimates from CHIS data, weights were 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 was weighted to represent the noninstitutionalized population for each sampling stratum and statewide. The weighting procedures used for CHIS 2017-2018 accomplish the following objectives:

- Compensate for differential probabilities of selection for phone numbers (households) and persons within household;
- Reduce biases occurring because non respondents may have different characteristics than respondents;
- Adjust, to the extent possible, for undercoverage 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 was used to compute a person-level weight, which includes adjustments for the within-household sampling of persons and for nonresponse. The final step was to adjust the person-level weight using weight calibration, a procedure that forced the CHIS weights to sum to estimated population control totals simultaneously from an independent data source (see below).

Population control totals of the number of persons by age, race, and sex at the stratum level for CHIS 2017-2018 were created primarily from the California Department of Finance’s (DOF) 2017 and 2018 Population Estimates, and associated population projections. The procedure used several

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), and education. 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 calibration.

The DOF control totals used to create the CHIS 2017-2018 weights are based on 2010 Census counts, as were those used for the 2015-2016 cycle. Please pay close attention when comparing estimates using CHIS 2017-2018 data with estimates using data from CHIS cycles before 2010. The most accurate California population figures are available when the U.S. Census Bureau conducts the decennial census. For periods between each census, population-based surveys like CHIS must use population projections based on the decennial count. For example, population control totals for CHIS 2009 were based on 2009 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.

1.7 Imputation Methods

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

Three different imputation procedures were used by SSRS 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. The hot deck approach is one of the most commonly used methods for assigning values for missing responses. Using a hot deck, a value reported by a respondent for a specific item was 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 answered a survey item formed a pool of donors, while the item non respondents formed a group of recipients. A recipient was matched to the subset pool of donors based on household and individual characteristics. A value for the recipient was then randomly imputed from one of the donors in the pool. SSRS used hot deck imputation to impute the same items that have been imputed in all CHIS cycles since 2003 (i.e., race, ethnicity, home ownership, and education). The last technique was external data assignment. This method was used for geocoding variables such as strata, Los Angeles SPA, San Diego HSR, and zip where the respondent provided inconsistent information. For such cases geocoding information was used for imputation.

UCLA-CHPR imputed missing values for nearly every variable in the data files other than those imputed by SSRS and some sensitive variables for which nonresponse had its own meaning. Overall, item nonresponse rates in CHIS 2017-2018 were low, with most variables missing valid responses for less than 1% of the sample. Questions that go to fewer overall respondents or that ask about more sensitive topics can have higher nonresponse.

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 without donor replacement was used. This method replaced 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 corresponded to the nature of the variable being imputed (e.g. linear regression for continues variables, logistic regression for binary variables, etc.). Donors and recipients were 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 standard measures of demographic and socioeconomic characteristics, as well as geographic region; however, the full set of control variables varies depending on which variable is being imputed. Most imputation models included additional characteristics, such as health status or access to care, which are used to improve the quality of the donor-recipient match.

Among the standard list of control variables, gender, age, race/ethnicity, educational attainment and region of California were imputed by SSRS. UCLA-CHPR began their imputation process by imputing household income so that this characteristic was available for the imputation of other variables. Sometimes CHIS collects bracketed information about the range in which the respondent's value falls

when the respondent will not or cannot report an exact amount. Household income, for example, was imputed using the hot-deck method within ranges defined by a set of auxiliary variables such as bracketed income range and/or poverty level.

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

2. DATA EDITING PROCEDURES

Survey data for all CHIS 2017-2018 samples – landline and cellular RDD, surname list, supplemental address-based sample (ABS) in Imperial County, supplemental interviews of Hispanic men, African Americans, and Chinese men in San Francisco County (SF OS), and statewide American Indian/Alaska Natives (AIAN) – 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 2017-2018 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 2017-2018 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 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 include both manual and machine editing procedures to correct interviewer, respondent, and CATI program errors and to check that updates made by data preparation staff are input correctly. Because data editing results in changes to the survey data, specific quality control procedures were implemented. CHIS 2017-2018 survey data were thoroughly examined and edited before SSRS 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 simulation computer runs to identify inconsistencies or illogical patterns in the data.

The data editing procedures for CHIS 2017-2018 consisted of three main tasks: (1) managing and resolving problem cases, (2) coding question responses that were recorded as text strings (i.e., “upcoding” responses captured in “other specify” fields), and (3) 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 important task for ensuring high-quality data was managing and resolving problem cases. The data preparation staff, as well as project staff and CATI staff, 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 could alert the project team and programmer by filling out a problem sheet for the case. Data preparation staff used these problem sheets as a guide to review cases and to make certain that any required updates were made accurately.

Not all problems required CATI database updates. Some could be resolved by simply releasing the case for general interviewing with 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 and all questions in Section E 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.

Some examples of common cases reviewed by SSRS project staff were those in which an error was made in enumerating the number of people in the household (SC5a) 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.

2.2 Coding with Text Strings

Most items in CHIS 2017-2018 had only close-ended response options, but several of them had the option of entering an ‘other-specify’ response that required coding of narrative text strings recorded by interviewers. For example, question AA5 in the adult extended interview was asked of respondents who had reported being of Hispanic or Latino ancestry or origin: “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.” The list of potential responses in AA5 included 10 different nationalities, and

interviewers could use an “other (specify:)” category for responses outside this list. Additional questions with an “other (specify:)” category from the CHIS 2017-2018 adult extended interview included:

- Racial/ethnic ancestry (AA5, AA5A, AA5E, AA5E1, AA5F);
- Tribal names (AA5B, AA5D);
- Sexual orientation (AD46B);
- Gender identity (AD67B);
- Country of birth (AH33, AH34, AH35, AI56);
- Languages spoken at home (AH36);
- Diabetes (AB51);
- Reasons for using E-cigarettes (AC83B);
- Secondhand tobacco smoke (AC145);
- Industry and Occupation (AK5, AK6);
- Health insurance coverage items (AI15, KAI15, AI15A, KAI15A, AI17A, KAI17A, AI45, KAI45, AI45A, KAI45A, AI36, KAI36, AI24, KAI24, AL19, AH104, KAH104, AH105, KAH105, AH106, KAH106, AH122, KAH122, AH101h, KAH101, AH114h, KAH114, AH121, KAH121, AI22A);
- Medicare coverage (AH124, AH125);
- Child/adolescent health insurance coverage items (CF7, KCF7, CF18, KCF18, IA18, KIA18, CF29, KCF29, IA29, KIA29, CF1A, CF2A, KCF2A, IA1A, KIA1A, IA2A, KIA2A, IA7, KIA7, AI90, KAI90, AI91, KAI91, AI92, KAI92, AI115, KAI115, AI94, KAI94, AI95, KAI95, AI96, KAI96, AI116, KAI116);
- Adult/child/adolescent insurance plan names (AH50, AI22A, MA2, MA7, KAH50, KAI22A, KMA2, KMA7);
- Marijuana use (AC125);
- Painkillers/Medicine use (AC133);
- HIV Testing (AD84);
- Reason no longer receiving behavioral health treatment (AF80);
- Immigration (AG26, AG27);
- Usual source of health care (AH3);
- Language used by doctor to speak to respondent (AJ50);
- Nature of video or telephone conversation with doctor (AJ153);
- Reason for delay in getting needed health care (AJ131, AE101, AF80);
- Main birth control method (AJ154, AJ174);

- Where received birth control method (AJ143, AJ146);
- Main reason NOT using birth control (AJ170, AJ175);
- Medical care unfair treatment (DMC6B);
- WIC (AL61, AL72, AL85);
- Medi-Cal non-participation and renewal (AL43, AL19);
- Reason for moving (AM38);
- Reason for not being registered to vote (AP71).

Questions with an "other (specify:)" category in the child and teen interviews:

- Child condition or disability (CA10A);
- Child/adolescent race and ethnicity (CH2, CH3, CH4, CH6, CH7, CH7A, TI1A, TI2, TI2A, TI2C, TI2D, TI2D1);
- Child/adolescent 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, CB22TYPE, TA4B, TA4BTYPE);
- Reason for adolescent to have changed school (TA7);
- Child/adolescent usual source of health care (CD3, TF2);
- Child/adolescent reason for delay in getting health care (CD68, TH59);
- Language used by child's doctor to talk to parent (CD31);
- Reasons for using E-cigarettes (TE68);
- Adolescent marijuana use (TE77);
- Adolescent birth control method (TG19, TG23);
- Adolescent reason not using birth control (TG20, TG24);
- Adolescent HIV testing (TL48);
- Reason for child not getting dental care (CB28);
- Child/adolescent/spouse healthcare coverage (KAH124, KAH125, KAH104, KAH105, KAH106, KAI15, KAI15A, KAI45, KAI45A, KAH122, KAI22A, KAI36, KAI24, KAI90, KAI91, KAI92, KCF7, KCF1A, KAI115, KMA2, KCF18, KCF29, KAI94, KAI95, KAI96, KIA7, KIA1A, KAI116, KMA7, KIA18, KIA29).

SSRS data preparation staff reviewed these responses and up-coded them to existing categories whenever possible. Text responses were also reviewed to remove indications to respondents' names (or initials) and to summarize long responses.

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.

When a respondent insisted on giving a response that violated the hard-edit specifications, interviewers recorded the answer and interaction in a problem sheet, and data preparation and project staff reviewed and updated the case as needed.

2.3 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, effects on any other questions, or logical skip patterns in the questionnaire. For more complicated circumstances, the data preparation staff and project staff carefully reviewed interviewer comments, messages, and problem descriptions to verify data updates.

Cases with similar problems were reviewed and updated together in manageable batches to ensure consistency in handling data problems. Following the series of updates, a program checked for all errors identified to date to ensure that editing had not created new errors. Frequency distributions and cross-tabulations 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 existed, and those that should not exist did not), and edits that evaluated complex skip patterns were run periodically during data collection. When discrepancies were discovered, problem cases were reviewed and updated as necessary.

3. GEOGRAPHIC CODING

For CHIS 2017-2018, SSRS delivered geo-coded survey data for any household where at least one interview had been completed, identifying the approximate (i.e., not "rooftop") location of the respondent's residence. The self-reported county was used to assign cases to landline sample strata as described in *CHIS 2017-2018 Methodology Series: Report 1 – Sample Design.* SSRS also prepared and delivered more specific geocodes based on the respondent-reported address and other information. The geographic coding process for CHIS 2017-2018 used Esri's ArcGIS mapping software that calls upon the TomTom streets dataset (primary source) and Census TIGERLine street dataset (secondary source) to geocode CHIS addresses. The TomTom dataset is updated twice a year and the Census TIGERLine dataset is updated once a year.

3.1 County of Residence

The CHIS 2017-2018 survey 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/SAH42). In addition, for cases in which an address had been matched to the sampled telephone number[3], 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 2017-2018 Methodology Series: Report 2 – Data Collection Methods* 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 2017-2018 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

[3] 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.

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.

The following two questions were 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?" (If the adult respondent was the same as the screener respondent, they answered this question in the screener (SAH42) and were not re-asked their county in the adult interview), and
- AM7. "What is your ZIP code?" (For address matched sample, we confirmed their zip code at AO1)

The final self-reported stratum included in the final data file was determined by applying the geocodes developed SSRS staff as described below. See *CHIS 2017-2018 Methodology Series: Report 5 - Weighting and Variance Estimation*, Section 8.2.2, 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, and the final distribution of completed cell sample adult extended interview cases by self-reported and original sampling stratum is presented in Table 3-3. Generally, the frequency counts show that there is good correspondence between the original sampling stratum and the self-reported stratum for the landline sample. 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

The geocoding for CHIS 2017-2018 was accomplished using the Esri ArcGIS mapping software package. This package calls upon the TomTom streets dataset (primary source) and Census TIGERLine street dataset (secondary source) to geocode CHIS addresses. Addresses are geocoded using an address locator (ArcGIS). The address locator attempts to match each address record to a street segment. Street segments are typically short in length with start/end points occurring at street intersections, geo-physical boundaries (i.e. rivers, lakes, etc.), and changes in 100 blocks. The following information is used when matching and address to a street segment:

- 5 digit ZIP code: This was the most critical piece of information needed for the address locator to find a match. Address records with missing or invalid ZIP code information are unable to be geocoded.
- Street name and house number: Second most critical piece of address information needed to find a match.

Once the address locator successfully matched the ZIP code, the locator filtered through all the street names associated with the ZIP code match. If a street match was made, the address was successfully geocoded. Since the USPS address data contains all required address components, we were able to geocode every address contained in both the Computerized Delivery Sequence (CDS) file, a file used to build the ABS sampling frame, and the NOSTAT file which includes a combination of newly-constructed buildings, gated communities without individual delivery, and blighted structures. Both files, the CDS and NOSTAT, were sourced from the United States Postal Service (USPS). The type of address locator used for geocoding is composite style. This allows the locator to reference and search through multiple street segment datasets. We reference two different street datasets. If an address was unable to be matched to a street segment from the TomTom dataset, the locator automatically searched for a match in the Census TIGERLine street dataset. If a match could not be found in either streets dataset, the address locator attempted to geocode the address to the next best level of precision. The order of precision matching was as follows:

- Street address (TomTom, TIGERLine)
- 5 digit ZIP code centroid

Addresses were geocoded using a method known as linear interpolation. The geocoded location of each address was estimated based on the range of numeric values between the starting and ending nodes for each street segment. Generally, every node is assigned two (2) values – one odd and one even. The values of each node correspond to the known starting and ending addresses found on both sides of the street.

SSRS staff reviewed the geocoded stratum to ensure it accurately reflected respondent provided data. If a batch match was not obtained, SSRS 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. SSRS corrected the address to match the street database, or matched to the segment's nearest intersection. If the street address or nearest intersection could not be identified, SSRS

would then match to the geographic ZIP centroid. If no zip code or address information was provided, a zip code was imputed using hot-deck imputation with area code and county as imputation classes. The frequencies of assigned geocodes by rule and sample type are shown in Table 3-1. Final distributions of adult completes by stratum corrected from the original sample are included in Table 3-2 for landline and list sample and Table 3-3 for cell phone sample.

Table 3-1. Number of Geocodes Assigned by Rule and by Sample Type

Rule	Cell	LL	Korean	Vietnamese	SF OS	AIAN	Total
1. Address assigned by matching to TomTom dataset	11,849	9,153	335	224	0	0	21,561
2. Address assigned by matching to Census TIGERLine dataset	208	117	1	4	4	11	345
3. Matched to ZIP centroid	9,197	10,935	188	182	496	306	21,304
Total	21,254	20,205	524	410	500	317	43,210

Source: UCLA Center for Health Policy Research, 2017-2018 California Health Interview Survey.

Table 3-2. Final distribution of adult extended completed cases by self-reported and original sampling stratum, landline/list sample for CHIS 2017-2018

Stratum name	Sampling stratum count	Removed	Added	Final self-reported stratum count
1 - LOS ANGELES	3,584	35	58	3,607
2 - SAN DIEGO	2,209	19	13	2,203
3 - ORANGE	1,381	36	12	1,357
4 - SANTA CLARA	656	10	32	678
5 - SAN BERNARDINO	629	10	20	639
6 - RIVERSIDE	887	13	21	895
7 - ALAMEDA	555	17	10	548
8 - SACRAMENTO	532	9	10	533
9 - CONTRA COSTA	382	5	24	401
10 - FRESNO	345	5	4	344
11 - SAN FRANCISCO	492	15	6	483
12 - VENTURA	303	8	19	314
13 - SAN MATEO	280	18	13	275
14 - KERN	289	4	4	289
15 - SAN JOAQUIN	222	1	4	225
16 - SONOMA	214	2	7	219
17 - STANISLAUS	223	10	1	214
18 - SANTA BARBARA	205	5	2	202
19 - SOLANO	197	4	4	197

(continued)

Table 3-2. Final distribution of adult extended completed cases by self-reported and original sampling stratum, landline/list sample for CHIS 2017-2018 (continued)

Stratum name	Sampling stratum count	Removed	Added	Final self-reported stratum count
20 - TULARE	232	3	3	232
21 - SANTA CRUZ	214	8	7	213
22 - MARIN	254	5	7	256
23 - SAN LUIS OBISPO	258	7	4	255
24 - PLACER	234	7	15	242
25 - MERCED	219	3	7	223
26 - BUTTE	290	2	9	297
27 - SHASTA	307	2	5	310
28 - YOLO	239	8	5	236
29 - EL DORADO	232	11	11	232
30 - IMPERIAL	564	10	0	554
31 - NAPA	231	4	4	231
32 - KINGS	233	5	1	229
33 - MADERA	272	10	2	264
34 - MONTEREY	183	4	6	185
35 - HUMBOLDT	340	2	2	340
36 - NEVADA	257	11	2	248
37 - MENDOCINO	224	5	0	219
38 - SUTTER	252	26	12	238
39 - YUBA	228	25	23	226
40 - LAKE	228	4	1	225
41 - SAN BENITO	150	1	2	151
42 - COLUSA, ETC	224	9	4	219
43 - DEL NORTE, ETC	207	7	8	208
44 - AMADOR, ETC	210	6	7	211

Source: UCLA Center for Health Policy Research, 2017-2018 California Health Interview Survey.

Table 3-3. Final distribution of adult extended completed cases by self-reported and original sampling stratum, cell phone sample for CHIS 2017-2018

Stratum name	Sampling stratum count	Removed	Added	Final self-reported stratum count
1 - LOS ANGELES	3,825	509	484	3,800
2 - SAN DIEGO	2,303	268	246	2,281
3 - ORANGE	987	200	209	996
4 - SANTA CLARA	805	210	178	773
5 - SAN BERNARDINO	673	154	254	773
6 - RIVERSIDE	967	232	268	1,003
7 - ALAMEDA	711	230	217	698
8 - SACRAMENTO	586	137	271	720
9 - CONTRA COSTA	468	112	217	573
10 - FRESNO	323	51	153	425
11 - SAN FRANCISCO	799	184	155	770
12 - VENTURA	309	74	97	332
13 - SAN MATEO	325	109	108	324
14 - KERN	325	62	82	345
15 - SAN JOAQUIN	250	72	74	252
16 - SONOMA	179	46	163	296
17 - STANISLAUS	279	60	58	277
18 - SANTA BARBARA	267	63	75	279
19 - SOLANO	333	114	93	312
20 - TULARE	286	62	56	280
21 - SANTA CRUZ	274	69	67	272
22 - MARIN	294	134	68	228
23 - SAN LUIS OBISPO	235	62	72	245
24 - PLACER	257	109	121	269
25 - MERCED	278	71	56	263
26 - BUTTE	209	57	121	273
27 - SHASTA	303	83	36	256
28 - YOLO	298	110	91	279
29 - EL DORADO	295	90	67	272
30 - IMPERIAL	298	69	29	258
31 - NAPA	351	126	50	275
32 - KINGS	328	90	13	251
33 - MADERA	300	82	30	248
34 - MONTEREY	252	61	103	294
35 - HUMBOLDT	291	50	54	295
36 - NEVADA	307	103	54	258
37 - MENDOCINO	313	66	33	280

(continued)

Table 3-3. Final distribution of adult extended completed cases by self-reported and original sampling stratum, cell phone sample for CHIS 2017-2018 (continued)

Stratum name	Sampling stratum count	Removed	Added	Final self-reported stratum count
38 - SUTTER	461	239	57	279
39 - YUBA	227	94	135	268
40 - LAKE	266	51	62	277
41 - SAN BENITO	486	171	32	347
42 - COLUSA, ETC	139	34	77	182
43 - DEL NORTE, ETC	189	33	70	226
44 - AMADOR, ETC	174	33	80	221

Source: UCLA Center for Health Policy Research, 2017-2018 California Health Interview Survey.

3.3 School Names

CHIS 2017-2018 child and adolescent interviews collected the names of schools attended by selected children or adolescents (CB22 and TA4B, respectively). A sufficiently knowledgeable adult (SKA) 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.

A review of the child interview data showed several 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). Respondents whose first language was not English had similar difficulties in accurately reporting or spelling school names. SSRS performed spell-check and abbreviation corrections to the school names list and merged in school names as well as county of residence with relevant data fields in the California school list database to identify automatic matches.

For cases that could not be automatically matched using statistical programming due to reasons such as spelling issues, abbreviations, and county mismatch, additional CHIS variables were used to accurately identify and manually assign the name of the school. These variables 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 should be used to facilitate spell-check and abbreviation corrections to the school names.

4. INDUSTRY AND OCCUPATION CODING

This section describes the CHIS 2017-2018 Industry and Occupation (I&O) open-ended response coding process. The open-ended industry question was AK5 while occupation was AK6. The first step involved translating Spanish language open-ended responses into English, correcting any spelling errors, reviewing abbreviations, and reducing text to accommodate the requirements of the National Institutes for Occupational Safety and Health's (NIOSH) NIOSH Industry and Occupation Computerized Coding System (NIOCCS)[1].

After these steps were completed, any records with an open-ended response to either AK5 or AK6 were submitted to the National Institutes for Occupational Safety and Health's (NIOSH) NIOSH Industry and Occupation Computerized Coding System (NIOCCS) V3.0. The NIOSH Industry and Occupation Computerized Coding System (NIOCCS) was upgraded to V3.0 in March 2018. Depending upon the quality of data input, the new version of the computerized system improved autocoding rates by 10-25%. The option for High and Medium confidence level coding was removed and V3.0 added a 'Suggest Review' flag on complex autocoded records. The new version also included additional variables such as Industry and Occupation scores. This coding system was developed to translate English language text entries to standardized I&O codes. As stated in the online documentation, the I&O codes are "based on the Census Industry and Occupation Classification system supplemented with special codes developed by CDC/NIOSH for non-paid workers, non-workers, and the military."[2] This means that the codes are in the same four-digit format that the Census coding system utilizes. For this process, we used Census 2010 as the classification scheme.

For CHIS 2017-2018, 72.0% industry responses matched. For occupation text, 69.9% matched. Although 76.2% of records had either their industry or occupation response match using the NIOCCS system only 65.6% matched both their industry and occupation responses. The new version of NIOCCS used for the CHIS 2017-2018 coding removed the previous option to code with high and medium confidence levels and added a suggest review flag on complex auto-coded records.

All remaining records that did not match both their industry and occupation responses using the NIOCCS system were sent to the Census National Processing Center (NPC) for coding using the Demographic Survey's Division (DSD) computer-assisted I&O coding system. Census coded industry using census codes based on the 2012 North American Industry Classification System. The occupation

[1] https://www.cdc.gov/niosh/topics/coding/overview.html
[2] https://www.cdc.gov/niosh/topics/coding/how.html

fields used census codes based on the 2010 Standard Occupational Classification Manual. First the fields are coded and then verified. There was a 10% verification used. With any discrepancies, the verifier made a determination. There was no third-party adjudication. Census NPC provided output files containing I&O codes for all remaining records. The Census I&O codes were combined with the NIOCCS system codes and appended to the adult data as the translated I&O coding responses for each record. In situations where both Census and NIOCCS codes existed for a record the Census code was retained.

5. RACE AND ETHNICITY CODING

This chapter describes handling of race and ethnicity responses outside of the pre-existing categories. These "other (specify:)" responses were recorded as text strings, and were either "up-coded" into existing codes or left in the "other (specify:)" category.

The first question in the race and ethnicity series (question AA4 in the adult interview) asked if the respondent was Latino or Hispanic. If the response to this item was "yes," the next question (AA5) asked about the specific origin (Mexican, Cuban, etc.) and allowed an "other (specify:)" response entered as text in item AA5OS. Question AA5A then 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?" This item allowed multiple responses and included an "other race" category. The "other (specify:)" text was recorded in item AA5AOS. Respondents who identified as American Indian, Asian, or Pacific Islanders were asked one or two follow-up questions about their tribal or national origin (AA5B, AA5D, AA5E, AA5E1). Each of these items also included an option for "other (specify:)'. Respondents indicating more than one race or ethnicity were asked which they most identified with (AA5F). This item listed the response already entered under "other (specify:)," if any, but did not allow interviewers to collect a new "other (specify:)" response.

5.1 Coding Procedures

The procedures for race and ethnicity coding employed by SSRS supported 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 2017-2018 Methodology Series: Report 5 - Weighting and Variance Estimation*.

The coding procedures were consistent with those from the 2010 Census data and with those used in prior CHIS cycles. Census methods are documented in the 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 indicative of Hispanic ethnicity such as "Hispanic" or "Latino." Following the Census procedures, the person was left in the "other race" category and the "other (specify:)" text was standardized to "HISPANIC-LATINO."

The specific procedures and guidelines we used are detailed below. Responses captured in the "other (specify:)" text field were retained and included in the final data set delivery 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 (specify:)" category. For example, if the respondent was coded only as other race and the "other (specify:)" was "Irish," then the code for "white" was upcoded 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 (specify:)" 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 "other (specify:)" text indicated multiple races with no specific races mentioned (such as "mixed"), then the code for "other (specify:)" race was changed to "yes" for both the first and second mention.
- If the respondent was coded as being Hispanic or Latino, this could be revised based upon information in the "other (specify:)" comments of other variables which clearly indicated a non-Hispanic identity.
- 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 (specify:)" 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.
- Similarly, if any case was upcoded to Asian, American Indian, or Other Pacific Islander, then the follow-up questions about specific origin (AA5B, AA5D, AA5E, AA5E1) were also upcoded to be consistent with the "other (specify:)" text from AA5A if it was possible to do so. In cases where this was not possible, the "other (specify:)" origin category was coded and the text copied from the race variable to the "other (specify:)" for the follow-up question. For example, if the race "other (specify:)" code was "Filipino," then code for "Asian" was upcoded to "yes," "other (specify:)" race was revised to "no" and the "other (specify:)" text eliminated.

After doing that, the code for AA5E for “Filipino” was revised to ‘yes.” In some cases, we also looked to the answers from AH33, AH34, AH35, and AH36 to find the correct code for AA5E. This happened most often when the other (specify) text for AA5A simply said “Indian.” The aforementioned questions helped us determine if this meant Asian Indian or Native American.

- If the “other race” text was similar to “none of above,” and the respondent was coded as being Hispanic or Latino, the “other (specify:)” text was standardized to “HISPANIC-LATINO.” If the respondent was not coded as Hispanic or Latino we left the response as it was.
- Hispanic or Latino respondents who reported American Indian or Alaska Native (AIAN) as their race, but did not report a tribal affiliation, are coded as having AIAN racial identity in the data. In prior cycles Hispanic or Latino respondents with unknown AIAN tribal identities were generally reclassified as non-AIAN.

After upcoding the “other (specify:)” specify responses for the race question (AA5A), SSRS also reviewed all “other (specify:)” responses to the follow-up origin questions (AA5B, AA5D, AA5E, and AA5E1). These were upcoded when possible to the existing codes using a similar procedure. 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 upcode Hispanic origin to “yes” (to be consistent with the Census procedures for Hispanic origin) and then upcode “race” to “White” (since the person is of European origin).

6. IMPERIAL COUNTY ADDRESS-BASED SAMPLE (ABS)

Data collection for the Imperial County ABS was conducted from August through mid-December 2017. Sampled addresses with a matched telephone numbers received an introductory letter. Sampled addressed without a matched telephone number received the introductory letter and a short screening questionnaire, the primary purpose of which was to obtain a telephone number. All sample received visits from field staff, who encouraged respondents to call in and complete the survey right there and then. If respondents were reluctant to call in and complete the survey at the moment, field staff attempted to administer the screening questionnaire and collect up-to-date phone numbers.

Returned questionnaires were receipted and scanned daily. A total of 1,605 questionnaires with telephone numbers were returned. Captured data were reviewed, and telephone information was updated on all records where a screening questionnaire provided a phone number. Once a sampled address was associated with a telephone number, whether through the vendor match or from the screener, the Imperial sample was fielded and processed the same way as the RDD and list sample cases. Tables detailing the Imperial ABS can be found in the other CHIS 2017-2018 methodology reports.