CHIS 2019-2020 Methodology Report

Report 3

Data Processing Procedures

September 2021

UCLA CENTER FOR HEALTH POLICY RESEARCH

CALIFORNIA HEALTH INTERVIEW SURVEY

CHIS 2019-2020 METHODOLOGY SERIES

REPORT 3

DATA PROCESSING PROCEDURES

SEPTEMBER 2021

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

www.chis.ucla.edu

This report describes the data processing and editing procedures for CHIS 2019-2020 performed by SSRS. This report 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 as well as race and ethnicity survey items.

Suggested citation:

California Health Interview Survey. *CHIS 2019-2020 Methodology Series: Report 3 - Data Processing Procedures*. Los Angeles, CA: UCLA Center for Health Policy Research, 2021.

The California Health Interview Survey is a collaborative project of the UCLA Center for Health Policy Research with multiple funding sources. Funding for CHIS 2019-2020 came from the following 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 Mental Health Services Authority, the California Tobacco Control Program, the California Wellness Foundation, First 5 California, Kaiser Permanente, the Long Term Services and Supports Content Development Project, and San Diego County Health and Human Services Agency.

PREFACE

Data Processing Procedures is the third in a series of methodological reports describing the 2019-2020 California Health Interview Survey (CHIS 2019-2020). The other reports are listed below.

CHIS is a collaborative project of the University of California, Los Angeles (UCLA) Center for Health Policy Research with multiple funding sources from public, private, and non-profit organizations. SSRS was responsible for data collection and the preparation of five methodological reports from the 2019-2020 survey. The survey examines public health and health care access issues in California. The survey is the largest state health survey ever undertaken in the United States.

Methodological Report Series for CHIS 2019-2020

The methodological reports for CHIS 2019-2020 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 2019-2020. 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

Table **Page**

1. CHIS 2019-2020 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 2019-2020.

- 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 multimode (web and 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 multiple funding sources from public, private, and non-profit organizations. 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 in Sampling and Data Collection Methodology

Starting in 2019-2020, the CHIS transitioned from a dual-frame landline/cellphone random digit dial (RDD) methodology to an address-based sample (ABS) methodology with multimode data collection that takes place on the web or by telephone. The CHIS research team deemed this change necessary due to decreasing response to telephone surveys, the improved geographical precision available for stratification when using the US Postal Service Delivery Sequence file of addresses as a sampling frame, and the lower cost of a study where the majority of interviews are completed online.

Prior to launching data collection in 2019, CHIS conducted two experiments in 2018 to test the effectiveness of an ABS mail push-to-web design with a telephone nonresponse follow-up. The first experiment was limited to three counties (Los Angeles, Santa Clara, and Tulare) to achieve a preliminary assessment of the efficacy of the proposed design (see Wells et al., 2018). Following the initial positive results from that test, a statewide pilot test was conducted in the late 2018 implementing a number of additional experiments and improvements based on the previous lessons learned (see Wells et al., 2019). Given that these additional improvements resulted in higher response and reductions in cost compared to maintaining the 2017-2018 design, CHIS committed to transitioning to the new design for the 2019-2020 cycle.

For CHIS 2019-2020, respondents are invited to either complete the survey online or call in to be interviewed by a member of the SSRS interviewing staff. Respondents receive an initial invitation letter with a $2.00 pre-incentive. This is followed by a reminder postcard and, in 2019, a final certified mail letter for all nonresponders[1]. In 2020, the certified mail letter was replaced with a standard letter and final postcard. Where addresses can be matched to a listed telephone number, the nonresponding households are also called six times to attempt to complete an interview before the sampled household is considered to be a resolved nonresponse.

See more about what's new in the 2019-2020 CHIS sampling and data collection here:

[1] For the last 2019 mailing, the certified letter was replaced with a standard letter.

https://chis.ucla.edu/chis/design/Documents/whats-new-chis-2019-2020.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 2019-2020 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 and year.

1.3 Sample Design Objectives

The CHIS 2019-2020 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 an address-based sample design. For the ABS 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 cycle—to provide health estimates for adults at the local level.

In addition, for CHIS 2019-2020, statistical modeling was used to determine the likelihood that specific targeted groups of interest for oversampling resided at addresses in the sample, and a hierarchy was established to determine the degree of over or undersampling among these strata. CHIS 2017-2018 data were used to build the models. All available auxiliary data from voter registration databases, consumer databases, Marketing Systems Group database information (specifically, all ranges of surnames), and Census Planning Database data were appended to the CHIS 2017-2018 data. All these

appended data served as the independent variables (features) in random forest models, while self-reported attributes (demographics, etc.) served as the dependent variables.

Models for CHIS 2019-2020 were specifically designed to predict the following household attributes:

1. Korean
2. Vietnamese
3. Other Asian
4. Hispanic or Spanish-Speaker
5. Low Educational Attainment or not a US Citizen
6. Have children (under 19)

Since these six models are run independently, households can be predicted to include more than one of the six target groups. For this reason, models were applied to the sample hierarchically with preference to the higher listed model (for example, a household predicted to be Korean was scored as Korean no matter what else they might have been predicted to be).

Utilizing these models results in two additional sample groups, or strata: 1) sample records for which none of the models predicted any attribute ("Residual" sample) and 2) sample for which no auxiliary data were found ("No Match" sample). The final step in utilizing the models was to develop relative sampling fractions by which households were selected within the modeled strata.

Within each geographic and modeled stratum combination, residential addresses 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 via CATI or Web. Most frequently the child interview was completed by the randomly selected respondent who was the parent or guardian. Less frequently and only within the CATI program, an adult sufficiently knowledgeable about the child's health could complete the child interview.

Table 1-1. California county and county group strata used in the CHIS 2019-2020 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, 2019-2020 California Health Interview Survey.

The CHIS two-year ABS sample is of sufficient size to accomplish the second objective as well, to produce statistically stable estimates for small population groups such as racial/ethnic subgroups, children, adolescents, etc.

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 dialect), Vietnamese, Korean, and Tagalog. Tagalog was administered by phone only. 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 collaborated with UCLA on the methodology and collected data for CHIS 2019-2020, 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, one randomly selected adult in each sampled household either completed an on-line survey or was interviewed by telephone by an SSRS interviewer. In addition, the study 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. The child interview was moved in 2019-2020 to take place immediately after Section A of the adult survey and the rostering of the household. The adolescent survey took place either immediately after the adult with phone interviews or in a separate session online.

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 if the interview was completed by phone. 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. Table 1-2 shows the number of completed adult, child, and adolescent interviews in CHIS 2019-2020 by mode of interview. Note that these figures were accurate as of data collection completion for 2019-2020 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 2019-2020 interviews by mode of interview and instrument

Type of sample[1]	Adult	Child	Adolescent
Total ABS	44,109[1]	6,557	2,212
Completes by Web	40,072	6,295	2,000
Completes by phone	4,037	262	212

Source: UCLA Center for Health Policy Research, 2019-2020 California Health Interview Survey.

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

Interviews in all languages were administered using SSRS's computer-assisted web interviewing and computer-assisted telephone interviewing (CAWI/CATI) system. As expected, the CATI interviews were longer in duration. The duration of the CATI interviews averaged almost 48 minutes, 26 minutes, and 23minutes for the adult, child, and adolescent interviews, respectively; the duration of the CAWI interviews averaged around 35 minutes, 13 minutes, and 17 minutes for the adult, child, and adolescent interviews, respectively. Interviews in non-English languages typically took longer to complete across

both modes: the non-English CATI interviews had an average length of about 64 minutes, 31 minutes, and 29 minutes for the adult, child, and adolescent interviews respectively; the non-English CAWI interviews had an average length of about 47 minutes, 18 minutes, and 20 minutes for the adult, child, and adolescent interviews, respectively. Just over four and half percent of the adult interviews were completed in a language other than English, as were about nine percent of all child (parent proxy) interviews and one 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 2019-2020 survey topic areas by instrument

Health status	**Adult**	**Adolescent**	**Child**
General health status	✓	✓	✓
Days missed from work or school due to health problems	✓	✓	✓
Health conditions	**Adult**	**Adolescent**	**Child**
Asthma	✓	✓	✓
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			✓
Covid-19	**Adult**	**Adolescent**	**Child**
Covid testing history and effects of pandemic	✓		
Mental health	**Adult**	**Adolescent**	**Child**
Mental health status	✓	✓	
Perceived need, access and utilization of mental health services	✓	✓	
Functional impairment, stigma, three-item loneliness scale	✓		
Suicide ideation and attempts	✓	✓	
Mental health and technology	✓	✓	
Health behaviors	**Adult**	**Adolescent**	**Child**
Dietary and nutritional intake, breastfeeding (younger than 3 years)	✓	✓	✓
Physical activity and exercise, sedentary time		✓	✓
Commute from school to home		✓	✓
Alcohol use/abuse		✓	
Cigarette and E-cigarette use	✓	✓	
Marijuana use	✓	✓	
Opioid use	✓		
Chewing tobacco, tobacco flavors	✓		
Exposure to second-hand smoke	✓		
Sexual behaviors	✓	✓	
HIV testing, HIV prevention medication (PrEP/Truvada)	✓	✓	
Contraceptive use, birth control	✓	✓	
Sexual violence	**Adult**	**Adolescent**	**Child**
Past unwanted sexual encounter	✓		

(continued)

Table 1-3. CHIS 2019-2020 survey topic areas by instrument (continued)

Women's health	**Adult**	**Adolescent**	**Child**
Pregnancy status/plans and birth control	✓	✓	
Dental health	**Adult**	**Adolescent**	**Child**
Last dental visit, main reason haven't visited dentist	✓	✓	✓
Delays in getting care			✓
Current dental insurance coverage	✓		✓
Condition of teeth	✓	✓	
Neighborhood and housing	**Adult**	**Adolescent**	**Child**
Safety, social cohesion	✓	✓	✓
Homeownership	✓		
Park use, park and neighborhood safety		✓	✓
Civic engagement, community involvement	✓	✓	
Access to and use of health care	**Adult**	**Adolescent**	**Child**
Usual source of care, visits to medical doctor	✓	✓	✓
Emergency room visits	✓	✓	✓
Delays in getting care (prescriptions and medical care)	✓	✓	✓
Communication problems with doctor	✓		✓
Timely appointment	✓	✓	✓
Access to specialist and general doctors	✓		
Tele-medical care	✓		
Care coordination	✓	✓	✓
Voter engagement	**Adult**	**Adolescent**	**Child**
Voter engagement	✓		
Food environment	**Adult**	**Adolescent**	**Child**
Access to affordable foods	✓		
Availability of food in household over past 12 months	✓		
Hunger	✓		
Health insurance	**Adult**	**Adolescent**	**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	✓	✓	✓
Medical debt, hospitalizations	✓		

(continued)

Table 1-3. CHIS 2019-2020 survey topic areas by instrument (continued)

Public program eligibility	**Adult**	**Adolescent**	**Child**
Program participation (CalWORKs, Food Stamps, SSI, SSDI, WIC, TANF)	✓	✓	✓
Assets, child support, Social security/pension, worker's compensation	✓		
Medi-Cal renewal	✓		
Reason for Medi-Cal non-participation	✓	✓	✓
Parental involvement/adult supervision	**Adult**	**Adolescent**	**Child**
Parental involvement			✓
Child care and school	**Adult**	**Adolescent**	**Child**
Current child care arrangements			✓
Paid child care	✓		
First 5 California: Talk, Read, Sing Program / Kit for New Parents			✓
Preschool/school attendance, school name		✓	✓
Caregiving	**Adult**	**Adolescent**	**Child**
Caregiving	✓		
Employment	**Adult**	**Adolescent**	**Child**
Employment status, spouse's employment status	✓		
Hours worked at all jobs	✓		
Industry and occupation, firm size	✓		
Income	**Adult**	**Adolescent**	**Child**
Respondent's and spouse's earnings last month before taxes	✓		
Household income, number of persons supported by household income	✓		
Respondent characteristics	**Adult**	**Adolescent**	**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, 2019-2020 California Health Interview Survey.

1.5 Response Rates

The overall response rates for CHIS 2019-2020 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 2019-2020, the overall household response rate was 12.2 percent (the product of the screener response rate of 16.2 percent and the extended interview response rate at the household level of 75.2 percent). CHIS uses the RR4 type response rate described in the AAPOR (The American Association for Public Opinion Research), 2016 guidelines (see more detailed in *CHIS 2019-2020 Methodology Series: Report 4 – Response Rates*).

The extended interview response rate for the ABS sample varied across the adult (72.0 percent), child (85.7 percent) and adolescent (33.2 percent) interviews. The adolescent rate includes the process of obtaining permission from a parent or guardian.

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 2019-2020 (see Table 1-4b).

Table 1-4a. CHIS response rates - Conditional

Type of Sample	Screener	Household (given screened)	Adult (given screened)	Child (given screened & eligibility)	Adolescent (given screened & permission)
Overall	16.2%	75.2%	72.0%	85.7%	33.2%

Source: UCLA Center for Health Policy Research, 2019-2020 California Health Interview Survey.

Table 1-4b. CHIS response rates - Unconditional

Type of Sample	Screener	Household (given screened)	Adult (given screened)	Child (given screened & eligibility)	Adolescent (given screened & permission)
Overall	16.2%	12.2%	11.6%	13.9%	5.4%

Source: UCLA Center for Health Policy Research, 2019-2020 California Health Interview Survey.

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 2019-2020, either a spouse/partner or adult child completed a proxy interview for eight 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 2019-2020 accomplish the following objectives:

- Compensate for differential probabilities of selection for addresses (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 frame 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 address) and several 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 2019-2020 were created primarily from the California Department of Finance's (DOF) 2019 and 2020 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 and Human Services Agency (HHSA) region in San Diego County), and education. One limitation of using 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 DOF control totals prior to calibration.

The DOF control totals used to create the CHIS 2019-2020 weights are based on 2010 Census counts, as were those used for the 2017-2018 cycle. Please pay close attention when comparing estimates using CHIS 2019-2020 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 HSSA region, 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 2019-2020 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 the CHIS 2019-2020 sample was collected using a combination of computer assisted web interviewing (CAWI) and computer assisted telephone interviewing (CATI). While the screening interview varied somewhat by whether the interview was conducted via CATI or CAWI, the same editing procedures were followed for all CHIS 2019-2020 cases.

In both, the CATI and CAWI environment, the data collection and interview process was controlled using a series of computer programs to ensure consistency and quality. The same base computer program was used for both CATI and CAWI interviews. (*CHIS 2019-2020 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 system programming determines which questions are asked based on household composition, respondent characteristics or preceding answers, and also determines the order in which the questions are presented to interviewers. The system also presents the response options available for recording answers.

The system 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 program virtually eliminates out-of-range responses and many other anomalies, some consistency and edit issues may arise. For example, for CATI interviewers, 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 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 2019-2020 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 program 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 2019-2020 consisted of four 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), (3) verifying data editing updates, and (4) assigning special codes. The final step was to convert the edited data 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 operations staff, worked collectively to resolve problem cases. The method used to communicate problems is described in this section, along with the system used by data editing and preparation staff to update or modify both the CATI and CAWI systems data.

For CATI interviews, 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.

During CAWI interviews, respondent had the option to reach out via to the project staff via a help feature in the program. In some instances, respondents wanted their responses adjusted after completing the survey. These cases were reviewed by SSRS project staff and, if deemed appropriate, the edits were made to the data stored in the system.

2.2 Coding with Text Strings

Most items in CHIS 2019-2020 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 2019-2020 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);
- Industry and Occupation (AK5, AK6);
- Health insurance coverage items (AI15, KAI15, AI15A, KAI15A, AI45, KAI45, AI45A, KAI45A, AI36, KAI36, AI24, KAI24, AL19, AH104, KAH104, AH105, KAH105, AH106, KAH106, AH122, KAH122, AH101h, AH114h, AH121h, AI22A);
- Child/adolescent health insurance coverage items (CF7, KCF7, CF18, KCF18, IA18, KIA18, CF29, KCF29, IA29, KIA29, CF1A, IA1A, KIA1A, IA7, KIA7, AI90, KAI90, AI91, KAI91, AI115, KAI115, AI94, KAI94, AI95, KAI95, AI116, KAI116);
- Adult/child/adolescent insurance plan names (AI22A, MA2, MA7, KAI22A, KMA2, KMA7);
- Marijuana use (AC125);
- Painkillers/Medicine use (AC133);
- HIV Testing (AD84);
- Reason no longer receiving behavioral health treatment (AF80);
- Use of online mental health tools (AG48);
- Usual source of health care (AH3);
- Language used by doctor to speak to respondent (AJ50);
- Nature of video or telephone conversation with doctor (AJ153b);
- Reason for delay in getting needed health care (AJ131, AF80);
- Main birth control method (AJ154, AJ174, AJ181, AJ182, AJ184, AJ185);
- Main reason NOT using birth control (AJ170, AJ175);

- Medi-Cal non-participation and renewal (AL19, AL91, AL87);
- Caregiving (AJ194, AJ200);
- Reason for not being registered to vote (AP80).

Questions with an "other (specify:)" category in the child and adolescent 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);
- 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, CD28);
- Reasons for using E-cigarettes (TE68);
- Adolescent marijuana use (TE77);
- Adolescent birth control method (TG19, TG23 TG27, TG28, TG30, TG31);
- Adolescent reason not using birth control (TG20, TG24);
- Adolescent HIV testing (TL48);
- Adolescent use of online mental health tools (TF42);
- Reason for child not getting dental care (CB28, CB23, CB26);
- Child dietary intake (CB32);
- Child/adolescent/spouse healthcare coverage (KAH104, KAH105, KAH106, KAI15, KAI15A, KAI45, KAI45A, KAH122, KAI22A, KAI36, KAI24, KAI90, KAI91, KCF7, KCF1A, KAI115, KMA2, KCF18, KCF29, KAI94, KAI95, 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 system responded by placing a message on the screen

and required the response to be re-entered. This system feature gives an opportunity to verify that the response is entered 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.

In a CATI interview, 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. In a CAWI interview, the respondent had an opportunity to reach out to the project staff via a help feature in the program.

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 interview case identification number was recorded to ensure that updates were associated with the appropriate case. The proposed edit was 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, respondent 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 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.

2.4 Special Codes

Respondents may not have a response at a question for several reasons. The following codes (Table 2-1) were assigned to capture the relevant scenarios for each question:

Table 2-1. Special Codes

Code	Label	Description
-1	Inapplicable	Respondent was legitimately skipped out of a question
-3	Web blanks	Respondent choose to leave a question blank. This was only possible in the CAWI mode
-6	Breakoff	Interview breakoff
-7	Refused	Respondent refused to provide a response. This was only possible in the CATI mode.
-8	Don’t know	Respondent did not know how to respond to question Aside from a few select questions, this was only possible in the CATI mode.
-9	Not ascertained	Respondent was skipped erroneously from a question or data did not get recorded correctly due to a system glitch.

Source: UCLA Center for Health Policy Research, 2019-2020 California Health Interview Survey.

3. GEOGRAPHIC CODING

For CHIS 2019-2020, SSRS delivered geo-coded survey data for any household where at least one screener interview had been completed, identifying the approximate (i.e., not "rooftop") location of the respondent's residence. SSRS then prepared and delivered more specific geocodes based on the sampled address and other information. As part of the screening process respondents were required to verify their sampled address. If there were any corrections needed to the sampled address, the respondents could submit it during the screening process. However, if during the geocoding process the sample address was deemed to be significantly different from the respondent provided correction, the case was not determined as ineligible based on address-based sampling, and was not counted as a complete for CHIS 2019-2020.

The geocoding for CHIS 2019-2020 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 were geocoded using an address locator (ArcGIS). The TomTom dataset is updated twice a year and the Census TIGERLine dataset is updated once a year.

At the time of sample generation, all addresses were assigned a longitude, latitude, and census block designation. There were a few instances, less than 1% of sampled cases, when the sampled address could not be used to assign the requisite geocoding information. For these rare cases, we used zip codes at the 9, 7 or 5 digit level.

During the screening process, respondents were asked to verify their sampled address. Based on this verification, respondents were coded into three possible outcomes:

1 – Respondents who completed the screener on the web and confirmed their sampled address

2 – Respondents who completed the screener on the phone and confirmed their sampled address

3 – Respondents who completed the screener on the phone, but asked for minor edits to their sampled address

SSRS staff reviewed cases that fell into the third category, where respondents confirmed their address but with minor edits. During this review, if the edit was deemed to indeed be minor, for instance an edit to the apartment number at the same address, or correction of a typo to the sampled address, the case was geocoded as described above. If, however, the edit was deemed to be substantial, where the

newly provided address did not match the sampled address, the disposition for the case was changed from complete to incomplete and the case was not geocoded.

The frequencies of assigned geocodes by rule and sample type are shown in Table 3-1. With the address-based sampling frame, there were no differences between distributions of the final geocode stratum and the sampling stratum. Table 3-2 provides the distribution of adult completes by stratum.

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

Rule	Screener Completes		Total
	Web	CATI	Total
Address assigned by matching to TomTom dataset	50,643	8,973	59,616
Matched to ZIP 5 centroid	5	-	5
Matched to ZIP 7 centroid	19	1	20
Matched to ZIP 9 centroid	576	89	665
Total	51,243	9,063	60,306

Source: UCLA Center for Health Policy Research, 2019-2020 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 2019-2020

Stratum name	Sampling/Final stratum count[1]		
	2019	2020	Total
1 - LOS ANGELES	4,241	4,314	8,555
2 - SAN DIEGO	2,443	2,297	4,740
3 - ORANGE	1,260	1,253	2,513
4 - SANTA CLARA	777	797	1,574
5 - SAN BERNARDINO	754	839	1,593
6 - RIVERSIDE	967	850	1,817
7 - ALAMEDA	681	738	1,419
8 - SACRAMENTO	645	656	1,301
9 - CONTRA COSTA	482	468	950
10 - FRESNO	438	360	798
11 - SAN FRANCISCO	412	511	923
12 - VENTURA	304	362	666
13 - SAN MATEO	329	323	652
14 - KERN	346	341	687
15 - SAN JOAQUIN	308	283	591
16 - SONOMA	305	276	581
17 - STANISLAUS	307	235	542
18 - SANTA BARBARA	276	258	534

(continued)

Table 3-2. Final distribution of adult extended completed cases by self-reported and original sampling stratum, landline/list sample for CHIS 2019-2020

Stratum name	Sampling/Final stratum count[1]		
	2019	2020	Total
19 - SOLANO	321	254	575
20 - TULARE	221	282	503
21 - SANTA CRUZ	240	301	541
22 - MARIN	287	242	529
23 - SAN LUIS OBISPO	227	289	516
24 - PLACER	256	284	540
25 - MERCED	229	290	519
26 - BUTTE	257	260	517
27 - SHASTA	248	286	534
28 - YOLO	245	285	530
29 - EL DORADO	268	238	506
30 - IMPERIAL	275	249	524
31 - NAPA	266	293	559
32 - KINGS	323	251	574
33 - MADERA	252	273	525
34 - MONTEREY	244	261	505
35 - HUMBOLDT	289	259	548
36 - NEVADA	247	272	519
37 - MENDOCINO	250	259	509
38 - SUTTER	286	264	550
39 - YUBA	271	235	506
40 - LAKE	278	240	518
41 - SAN BENITO	224	274	498
42 - COLUSA, ETC	277	216	493
43 - DEL NORTE, ETC	282	212	494
44 - AMADOR, ETC	322	219	541
Total	22,160	21,949	44,109

Source: UCLA Center for Health Policy Research, 2019-2020 California Health Interview Survey.

[1] Includes interviews meeting the criteria as partially complete

4. SCHOOL NAME CODING

CHIS 2019-2020 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.

5. INDUSTRY AND OCCUPATION CODING

This section describes the CHIS 2019-2020 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 any Spanish, Chinese, Korean, Vietnamese, or Tagalog 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)[2].

After these steps were completed, any records with an open-ended response to either AK5 or AK6 were submitted to NIOSH NIOCCS V3.0. NIOSH 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."[3] 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 2019, 71.1% industry responses matched. For occupation text, 70.5% matched. Although 78.3% of records had either their industry or occupation response match using the NIOCCS system, only 63.2% matched both their industry and occupation responses. For 2020 CHIS, 74.3% industry responses matched. For occupation text, 73.5% matched. Although 73.1% of records had either their industry or occupation response match using the NIOCCS system, only 69.3% matched both their industry and occupation responses. The new version of NIOCCS used for the CHIS 2019-2020 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

[2] https://csams.cdc.gov/nioccs/default.aspx
[3] https://www.cdc.gov/niosh/topics/industries.html

Demographic Survey's Division (DSD) computer-assisted I&O coding system.[4] Census coded industry using census codes based on the 2012 North American Industry Classification System. The occupation 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.

[4] For 2020, Manual CENSUS lookup data is not currently available for 1,060 records. Once available this section of the methodology report will be appended.

6. 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.

6.1 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 2019-2020 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 Technical Documentation (U.S. Census Bureau, 2011). 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).

7. REFERENCES

U.S. Census Bureau. (2011) *Census 2010 Redistricting Data (Public Law 94-171) Summary File – Technical Documentation.* Retrieved from http://www.census.gov/prod/cen2010/doc/pl94-171.pdf

Wells, B. M., Hughes, T., Park, R., CHIS Redesign Working Group, Rogers, T. B., & Ponce, N. (2018). *Evaluating the California Health Interview Survey of the future: Results from a methodological experiment to test an address-based sampling mail push-to-web data collection.* Los Angeles, CA: UCLA Center for Health Policy Research.

Wells, B. M., Hughes, T., Park, R., CHIS Redesign Working Group, & Ponce, N. (2019). *Evaluating the California Health Interview Survey of the future: Results from a statewide pilot of an address-based sampling mail push-to-web data collection.* Los Angeles, CA: UCLA Center for Health Policy Research.