california
health
interview
survey

CHIS 2023 Methodology Report Series

Report 3

Data Processing Procedures

September 2024

UCLA Center for Health Policy Research

CALIFORNIA HEALTH INTERVIEW SURVEY

CHIS 2023 METHODOLOGY SERIES

REPORT 3

DATA PROCESSING PROCEDURES

September 2024

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

www.chis.ucla.edu

This report describes the data processing and editing procedures for CHIS 2023 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 2023 Methodology Series: Report 3 - Data Processing Procedures*. Los Angeles, CA: UCLA Center for Health Policy Research, 2024.

The California Health Interview Survey is a collaborative project of the UCLA Center for Health Policy Research with multiple funding sources. Funding for CHIS 2023 came from the following sources: the California Department of Health Care Services, the California Department of Health Care Services (Community Services Division), the California Department of Public Health, The California Endowment, the California Health Benefit Exchange, the California Health Care Foundation, the California Tobacco Prevention Program, the California Wellness Foundation, First 5 California, the California Civil Rights Department, Santa Clara County Public Health Department, City of Long Beach Department of Health and Human Services, and San Diego County Health and Human Services Agency.

PREFACE

Data Processing Procedures is the third in a series of methodological reports describing the 2023 California Health Interview Survey (CHIS 2023). 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 2023 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 2023

The methodological reports for CHIS 2023 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 2023. 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 https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/chis-design-and-methods/chis-methodology-reports-repository. 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

CHIS 2023 SAMPLE DESIGN AND METHODOLOGY SUMMARY

1.1 Overview

A series of five methodology reports is available with more detail about the methods used in CHIS 2023.

- 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 https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/chis-design-and-methods/chis-methodology-reports-repository.

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 https://healthpolicy.ucla.edu/our-work/publications for examples of CHIS studies).

1.2 Sample Additions and Data Collection Methodology Updates

Starting in 2021, the CHIS added a prepaid cell phone sample to the primary ABS sample. A second innovation was altering the envelope for the initial mailing to have a window that would allow the incentive to be seen. The CHIS research team deemed these changes necessary to improve representation of California's diverse population and improve response rates.

For CHIS 2023, respondents in the ABS sample 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, a standard letter, and a final postcard. Where addresses can be matched to a listed telephone number, the nonresponding households are also called up to six times to attempt to complete an interview before the sampled household is considered to be a resolved nonresponse. In addition to the ABS sample frame, CHIS 2023 utilized a supplemental listed prepaid cell phone sample to meet targets in certain stratum.

The prepaid cell phone oversample followed the same dialing protocol of up to six dials before retiring the sample. In addition, the sampled phone number was screened for respondents who were either aged 18 to 24, Hispanic, African American, or would take the survey in one of the non-English languages offered for CHIS 2023.

In addition to the prepaid cell phone oversample, CHIS 2023 included two geographic oversamples:

1) An oversample of households from 11 ZIP codes in the City of Long Beach.
2) An oversample of households in Santa Clara County.

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 https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/access-chis-data/resources .

Additional documentation on constructing the CHIS sampling weights is available in the *CHIS 2023 Methodology Series: Report 5—Weighting and Variance Estimation* posted at https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/chis-design-and-

methods/chis-methodology-reports-repository. 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 2023 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 2023 continued to employ an address-based sample design. For the ABS sample, the 58 counties in the state were grouped into 44 primary 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.

As with CHIS 2021-2022, the address-based sample in CHIS 2023 was stratified into different strata that had higher incidences of individuals with targeted characteristics. For CHIS 2023, these strata were based on predictive models that employed Big Data techniques to identify household attributes such as demographics, spoken languages, and even attitudinal metrics that are correlated with important respondent characteristics. The process begins by taking prior data and building models with those data, and then scoring future samples with the outcomes of those models. In addition to evaluating the predictive models, for CHIS 2023 we also investigated the utility of individual sample flags provided by MSG database information, including the surname flags, child indicator variables, and resident age information as well as PDB block-group characteristics including the density of households with African American residents and households with limited English proficiency.

For CHIS 2023, the following strata were created[1]:

1) Vietnamese
2) Korean
3) Likely Asian-language Interview
4) Likely Spanish-language interview
5) Hispanic
6) Other high-density non-English
7) Other Asian
8) High density African American
9) HH with children
10) Other 65+
11) Residual - Match
12) Residual – No match

This stratification scheme was designed to make use of the most effective predictive variables to target key demographic subgroups in an efficient way that minimizes the impact of the disproportionate sampling on the design effect. Those models that were not sufficiently predictive to add value were excluded. It should be noted that this stratification includes two additional strata: 1) sample records for which none of the variables or models predicted any attribute, but for which auxiliary data could be matched to the address ("Residual - Match" sample) and sample for which no Big Data was found ("Residual - No match" sample). The final step in utilizing the models is to develop sampling fractions by which modeled households will be selected. The final sample fractions balanced the need to increase the frequency of the lowest incidence groups, while accounting for subgroup differences in response propensity and minimizing disproportionate weighting whenever possible.

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 selected parent/guardian were randomly selected. The adolescent was interviewed directly via CATI or Web. The child interview was completed by the selected adult respondent who was the parent or guardian.

[1] The Santa Clara oversample employs a slightly different strata, please refer to Methodology Report 1 – Sample Design for additional details.

Table 1-1. California county and county group strata used in the CHIS 2023 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, 2023 California Health Interview Survey.

In addition to the ABS sample frame, CHIS 2023 utilized a supplemental listed prepaid cellphone sample to meet targets in twelve geographic stratum that were underperforming in completion rate. Listed prepaid cell phones were sampled from the following 12 geographic strata:

1. Los Angeles
    a. SPA1
    b. SPA5
2. San Diego
    a. Central
3. Santa Clara
4. Sacramento
5. Contra Costa

6. Ventura
7. San Joaquin
8. Sonoma
9. Santa Cruz
10. Merced
11. Mendocino
12. San Benito

To better target populations not adequately covered under the ABS frame in CHIS 2023, we utilized a prepaid cell phone oversample of 450 completes to obtain additional in-language interviews, Hispanic and African American samples, and young adults. Prepaid cell phone numbers are associated with cell phones that are "pay-as-you-go" and do not require a contract. Prepaid numbers are more likely to be used by Hispanics, people with lower education and lower income, and other related groups that are often underrepresented in general population samples (e.g., the uninsured)

The CHIS ABS sample and the prepaid oversample were of sufficient size to accomplish the second objective, i.e., 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. These languages were chosen based on analysis of ACS 2021 5-year 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 2023, 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 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.

Table 1-2 shows the number of completed adult, child, and adolescent interviews in CHIS 2023 by mode of interview. Note that these figures were accurate as of data collection completion for 2023 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 https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/chis-design-and-methods/chis-design.

Table 1-2. Number of completed interviews by mode of interview and instrument

	Adult	Child	Adolescent
Totals[1]	23,697	3,650	1,045
Completes by Web	21,101	3,370	989
Completes by phone	2,596	280	56

Source: UCLA Center for Health Policy Research, 2023 California Health Interview Survey.
[1] Includes interviews meeting the criteria of sufficient partial.

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 68 minutes, 20 minutes, and 25 minutes for the adult, child, and adolescent interviews, respectively; the duration of the CAWI interviews averaged around 45 minutes, 13 minutes, and 18 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 76 minutes, 22 minutes, and 25 minutes for the adult, child, and adolescent interviews respectively; the non-English CAWI interviews had an average length of about 54 minutes, 16 minutes, and 18 minutes for the adult, child, and adolescent interviews, respectively.

Nearly 8 percent of the adult interviews were completed in a language other than English, as were about 12 percent of all child (parent proxy) interviews and 2 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 2023 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	✓		
Heart disease, High blood pressure, Cholesterol	✓		
Physical disability	✓		
Mental health	**Adult**	**Adolescent**	**Child**
Mental health status	✓	✓	
Perceived need, Access and utilization of mental health services	✓	✓	
Functional impairment, Stigma	✓		
Suicide ideation and attempts	✓	✓	
Telehealth and mental health services satisfaction, Delays in mental health services	✓	✓	
Climate Change	✓	✓	
Health behaviors	**Adult**	**Adolescent**	**Child**
Moderate physical activity	✓		
Dietary intake	✓	✓	✓
Breastfeeding (younger than 3 years)			✓
Sugar-sweetened beverages	✓	✓	✓
Alcohol use, Cigarette use, E-cigarette use, Marijuana use, CBD use	✓	✓	
CBD Use	✓		
Opioid use, Prescription painkiller use	✓		
Exposure to second-hand smoke/vapor, Exposure to marijuana smoke	✓		
Sexual behaviors, HIV testing, HIV prevention medication	✓	✓	
Caregiving	✓		
Gambling, Financial and mental impacts of gambling	✓		
Gun Violence	**Adult**	**Adolescent**	**Child**
Firearm ownership/presence, Loaded, and secure, Firearm victimization, Quick access to firearm	✓		
Women's health	**Adult**	**Adolescent**	**Child**
Pregnancy status	✓		

(continued)

Table 1-3. CHIS 2023 survey topic areas by instrument (continued)

Dental health	**Adult**	**Adolescent**	**Child**
Last dental visit, Main reason have not visited dentist, Number of dental visits, Location of dental service	✓	✓	✓
Current dental insurance coverage	✓	✓	✓
Source of dental care	✓	✓	✓
Neighborhood and housing	**Adult**	**Adolescent**	**Child**
Safety, Social cohesion	✓	✓	✓
Civic engagement	✓		
Participation in extracurricular activities		✓	
Housing security/stability, Place of residency last year	✓		
Encounters with police	✓		
Adverse Childhood Experiences	**Adult**	**Adolescent**	**Child**
ACES Screener	✓	✓	
Past ACES screener	✓	✓	
Safe and nurtured childhood experiences	✓	✓	
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	✓		✓
Contraception	✓	✓	
Timely appointment	✓	✓	✓
Access to specialist and general doctors	✓		
Telehealth care, Telehealth visit satisfaction and barriers	✓		
Care coordination	✓	✓	✓
Discrimination in healthcare setting	✓		
Difficulty in accessing care, tests, treatment	✓	✓	✓
Voter engagement	**Adult**	**Adolescent**	**Child**
Voter engagement	✓		
Voter attitudes	✓		
Food environment	**Adult**	**Adolescent**	**Child**
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	✓	✓	✓
Partial scope Medi-Cal, Medical debt, Hospitalizations	✓		

(continued)

Table 1-3. CHIS 2023 survey topic areas by instrument (continued)

Public program eligibility	**Adult**	**Adolescent**	**Child**
Household poverty level	✓		
Program participation (CalWORKs, Food Stamps, SSI, SSDI, WIC, TANF)	✓	✓	✓
Assets, Child support, Social security/pension, Worker's compensation	✓		
Medi-Cal eligibility, Medi-Cal renewal, Notice of actions from Medi-Cal	✓		
Reason for Medi-Cal non-participation among potential beneficiaries	✓	✓	✓
Use of public benefits among immigrant residents	✓		
Parental involvement/adult supervision	**Adult**	**Adolescent**	**Child**
Parental involvement			✓
Book ownership, Source of reading materials, Challenges to reading to child			✓
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		✓	✓
Preschool quality			✓
Employment	**Adult**	**Adolescent**	**Child**
Employment status, Spouse's employment status	✓		
Hours worked at all jobs	✓		
Industry and occupation, Firm size	✓		
Paid Family Leave	✓		
Income	**Adult**	**Adolescent**	**Child**
Respondent's and spouse's earnings last month before taxes	✓		
Household income, Number of persons supported by household income	✓		

(continued)

Table 1-3. CHIS 2023 survey topic areas by instrument (continued)

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	✓	✓	✓
COVID-19	**Adult**	**Adolescent**	**Child**
Ever tested positive for COVID-19, Test type	✓		
Experienced long COVID-19 symptoms	✓		
COVID vaccine status, COVID booster status	✓	✓	✓
Future COVID vaccine acceptance, Reasons for COVID vaccine hesitancy	✓		
Challenges experience due to COVID-19 pandemic	✓		
N95 masks, Ability to get N95 masks	✓		
Adolescent Future Preparedness	**Adult**	**Adolescent**	**Child**
Plans for college, Impact of pandemic on college plans		✓	
Discrimination	**Adult**	**Adolescent**	**Child**
Housing discrimination experience, Main reason for discrimination, Housing Choice Section 8 Voucher	✓		
Hate incident experience and witness, Type, Location, Reason for hate incident	✓	✓	

Source: UCLA Center for Health Policy Research, 2023 California Health Interview Survey.

1.5 Response Rates

The overall response rates for CHIS 2023 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 2023, the overall household response rate was 8.5 percent (the product of the screener response rate of 11.8 percent and the extended interview response rate at the household level of 72.1 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 2023 Methodology Series: Report 4 – Response Rates*).

The extended interview response rate for the ABS sample varied across the adult (64.7 percent), child (82.2 percent) and adolescent (27.9 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 2023 (see Table 1-4b).

Table 1-4a. CHIS response rates - Conditional

Type of Sample	Screener[1]	Household (given screened)[1]	Adult (given screened)[1]	Child (given screened & eligibility)[1]	Adolescent (given screened & permission)[1]
Overall	11.8%	72.1%	64.7%	82.2%	27.9%

Source: UCLA Center for Health Policy Research, 2023 California Health Interview Survey.
[1] The prepaid cell, Long Beach, and Santa Clara oversamples are not included in these rates.

Table 1-4b. CHIS response rates - Unconditional

Type of Sample	Screener[1]	Household (given screened)[1]	Adult (given screened)[1]	Child (given screened & eligibility)[1]	Adolescent (given screened & permission)[1]
Overall	11.8%	8.5%	7.7%	9.7%	3.3%

Source: UCLA Center for Health Policy Research, 2023 California Health Interview Survey.
[1] The prepaid cell, Long Beach, and Santa Clara oversamples are not included in these rates.

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 "sufficient partial 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 2023, either a spouse/partner or adult child completed a proxy interview for sixteen 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 https://healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis/chis-design-and-methods/chis-design/chis-2019-2020-redesign.

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 2023 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 under coverage 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 2023 were primarily created from the California Department of Finance's (DOF) 2023 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 2023 weights are based on 2020 Census counts. Please pay close attention when comparing estimates using CHIS 2023 data with estimates using data from CHIS cycles before 2023. 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 2010 Census-based information and 2020 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 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 zipcode 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 2023 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.

DATA EDITING PROCEDURES

Survey data for the CHIS 2023 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 2023 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 2023 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 2023 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 2023 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 require 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, respondents had the option to reach out 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 2023 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 2023 adult extended interview included:

- Gender identity (AD66C);
- Racial/ethnic ancestry (AA5, AA5A, AA5E, AA5E1, AA5F, AA5H, AA5I);
- Tribal names (AA5B, AA5D);
- Sexual orientation (AD46C);
- Country of birth (AH33, AH34, AH35, AI56);
- Citizenship (AG36B);
- Languages spoken at home (AH36);
- COVID-19 (CV17);
- E-cigarettes Use (AC184);
- Marijuana Use (AC125, AC193, AC194);
- CBD Use (AC205);
- Prescription Painkiller Use (AC133, AC222)
- Industry and Occupation (AK5, AK6);
- Health insurance coverage items (AI15, AI15A, AI45, AI45A, AI36, AI24, AL19, AH104, AH105, AH106, AH122, AH140, AH141, AH101h, AH114h, AH121h);
- Child/adolescent health insurance coverage items (CF7, CF18, IA18, CF29, IA29, CF1A, AI115, AI90, AI91, IA1A, IA7, AI94, AI95, AI116);
- Adult/child/adolescent insurance plan names (AI22A, MA2, MA7);
- 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 (AJ203, AJ234);
- Reason for delay in getting needed health care (AJ131, AF80, AJ252, AJ253, AJ254);
- Main birth control method (AJ237, AJ239, AJ243, AJ245);
- Main reason NOT using birth control (AJ170, AJ175);
- Reason for paid leave from work (AK138);

- Reason for missing work (AJ115);
- Medi-Cal non-participation and renewal (AL19, AL91, AL87, AL110);
- Caregiving (AJ194, AJ200, AJ223);
- Reason for not being registered to vote (AP80);
- Hate Incident (AM196, AM200, AM201)

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

- Child Gender (CA73);
- Child condition or disability (CA10A, CE20, CE21, CE22);
- COVID-19 (CCV2, TCV2);
- Child/adolescent race and ethnicity (CH2, CH3, CH3A, CH3B, CH4, CH6, CH7, CH7A, TI1A, TI2, TI2H, TI2I, TI2A, TI2C, TI2D, TI2D1, T12E);
- Teen Gender Identity (TA21B);
- Child/adolescent languages spoken at home (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 health care (CD3, TF2, TE87);
- Child/adolescent reason for delay in getting health care (CD68, TH59. TH60);
- 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 (TG19B, TG35, TG38, TG39);
- Adolescent reason not using birth control (TG20, TG24);
- Adolescent HIV testing (TL48);
- Adolescent use of online mental health tools (TF52, TF53);
- Reason for child not getting dental care (CB23);
- Books and reading to child (CF70, CF68);
- Hate Incidence (TD67, TD68, TD69, TD72, TD73, TD74, TD75)

SSRS data preparation staff reviewed these responses and upcoded 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 to 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 chose 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, 2023 California Health Interview Survey.

GEOGRAPHIC CODING

For CHIS 2023, 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 from the ABS sample 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 2023.

The geocoding for CHIS 2023 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 via CATI and confirmed their sampled address

3 – Respondents who completed the screener via CATI, 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 for screener completes.

Rule	Total Screener Completes
	2023
Address assigned by matching to TomTom Streets Version (T_202209 and T_20202303)	30,619
Address assigned by matching to ToTom Rooftop Version (T_202303_R Rooftop version 202303)	15
Matched to ZIP 9 centroid	550
Total	31,184

Source: UCLA Center for Health Policy Research, 2023 California Health Interview Survey.

Table 3-2. Final distribution of adult extended completed cases by sampling stratum for CHIS 2023

Stratum name	Sampling/Final stratum count[1]
	2023
1 - LOS ANGELES	4,016
2 - SAN DIEGO	2,175
3 - ORANGE	1,205
4 - SANTA CLARA	749
5 - SAN BERNARDINO	761
6 - RIVERSIDE	824
7 - ALAMEDA	659
8 - SACRAMENTO	573
9 - CONTRA COSTA	425
10 - FRESNO	364
11 - SAN FRANCISCO	427
12 - VENTURA	309
13 - SAN MATEO	316
14 - KERN	312
15 - SAN JOAQUIN	249
16 - SONOMA	222
17 - STANISLAUS	246
18 - SANTA BARBARA	254
19 - SOLANO	240
20 - TULARE	252
21 - SANTA CRUZ	232
22 - MARIN	259
23 - SAN LUIS OBISPO	247
24 - PLACER	247
25 - MERCED	236
26 - BUTTE	250
27 - SHASTA	261
28 - YOLO	270
29 - EL DORADO	245

continued

Table 3-2. Final distribution of adult extended completed cases by sampling stratum for CHIS 2023 (continued)

Stratum name	Sampling/Final stratum count[1]
	2023
30 - IMPERIAL	269
31 - NAPA	266
32 - KINGS	240
33 - MADERA	259
34 - MONTEREY	245
35 - HUMBOLDT	296
36 - NEVADA	266
37 - MENDOCINO	224
38 - SUTTER	270
39 - YUBA	238
40 - LAKE	262
41 - SAN BENITO	216
42 - COLUSA, ETC	236
43 - DEL NORTE, ETC	204
44 - AMADOR, ETC	240
Total	20,556

Source: UCLA Center for Health Policy Research, 2023 California Health Interview Survey.

[1] Includes interviews meeting the criteria as partially complete

SCHOOL NAME CODING

CHIS 2023 child and adolescent interviews collected the names of schools attended by selected children or adolescents (CB22 and TA4B, respectively). The adult respondent 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. Further, the California Department of Education's directory information was used to distinguish between Public and Private schools and code respondents.

INDUSTRY AND OCCUPATION CODING

This section describes the CHIS 2023 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 2018 as the classification scheme.

The matched responses percentages rates showed marked increases with the updated auto-coding system. For CHIS 2021, 88.7 % industry and occupation responses matched. For CHIS 2022, 86.3% industry and occupation responses matched. In 2023, 85.8% industry and occupation responses matched. In comparison, the matched responses percentage for CHIS 2019 was 71.1%. For occupation text, 70.5% matched. For CHIS 2020 74.5% industry responses matched. For occupation text, 73.6% matched. For CHIS 2023, 95.7% industry responses matched. For occupation text, 88.7% matched. The new version of NIOCCS used for the CHIS 2021-2023 coding removed the suggested review flag and any record marked as NIOSH-Insufficient Information for either industry or occupation was sent for manual coding.

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. Census coded industry using census codes based on the 2017 North American Industry Classification System. The occupation fields used census codes based on the 2018 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.

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 "upcoded" 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 missing and were later imputed. The imputation procedures are described in *CHIS 2023 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 upcoded 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).

LIMITATIONS FOR DATA PROCESSING PROCEDURES

There is the possibility of unmeasured error associated with data editing procedures such as cleaning and coding, with this or any other population-based survey. While QC steps are in place and are successful at identifying systematic errors, it is challenging to detect and rectify non-systematic errors.

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.