California Health Interview Survey

Making California's Voices Heard on Health

CHIS 2001 Methodology Series

Report 1

Sample Design

CALIFORNIA HEALTH INTERVIEW SURVEY

CHIS 2001 METHODOLOGY SERIES

REPORT 1

SAMPLE DESIGN

August 2002 *(rev 1/9/03)*

This report was prepared for the California Health Interview Survey by Ismael Flores-Cervantes and J. Michael Brick of Westat, Inc.

www.chis.ucla.edu

This report provides analysts with information about the sampling methods used for CHIS 2001, including both the household and person (within household) sampling. This report also provides a discussion on achieved sample size and how it compares to the planned sample size.

Suggested citation:

California Health Interview Survey. *CHIS 2001 Methodology Series: Report 1 - Sample Design*. Los Angeles, CA: UCLA Center for Health Policy Research, 2002.

The California Health Interview Survey is a collaborative project of the UCLA Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. Funding for CHIS 2001 came from multiple sources: the California Department of Health Services, The California Endowment, the California Children and Families Commission, the National Cancer Institute, the Centers for Disease Control and Prevention, and the Indian Health Service.

PREFACE

Sample Design for CHIS 2001 is the first in a series of methodological reports describing the 2001 California Health Interview Survey (CHIS 2001). The other reports are listed below.

CHIS is a collaborative project of the University of California, Los Angeles (UCLA) Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. Westat was responsible for the data collection and the preparation of five methodological reports from the 2001 survey. The survey examines public health and health care access issues in California. The CHIS telephone survey is the largest state health survey ever undertaken in the United States. The plan is to monitor the health of Californians and examine changes over time by conducting periodic surveys in the future.

Methodological Reports

The first five methodological reports for the 2001 CHIS are as follows:

- Report 1: Sample Design for CHIS 2001
- Report 2: Data Collection Methods in CHIS 2001
- Report 3: Data Processing Procedures in CHIS 2001
- Report 4: Response Rates in CHIS 2001
- Report 5: Weighting and Variance Estimation for CHIS 2001

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

This report describes the procedures used to design and select the sample from CHIS 2001. An appropriate sample design is a feature of a successful survey, and CHIS 2001 presented many issues that had to be addressed at the design stage. This report explains why the design features of CHIS were selected and presents the alternatives that were considered.

The primary purpose of this report is to provide analysts information about the sampling methods used for CHIS 2001, including both the household and person (within household) sampling. In general terms, once a household was sampled, an adult within that household was sampled. In some cases, adults were sampled with differential sampling rates. If there were children and/or adolescents in the household, one child and/or one adolescent was eligible for sampling. This report also provides a discussion on achieved sample size and how it compares to the planned sample size.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

List of Tables

TABLE OF CONTENTS (continued)

List of Tables

List of Figures

1. CHIS 2001 DESIGN AND METHODOLOGY SUMMARY

1.1 Overview

The 2001 California Health Interview Survey (CHIS 2001) is a collaborative project of the UCLA Center for Health Policy Research, the California Department of Health Services, and the Public Health Institute. The focus of the survey is on a variety of public health topics, including access to health care and health insurance coverage. CHIS 2001 is the largest state health survey ever undertaken in the United States. It is a random digit dialing (RDD) telephone survey of California households designed to produce reliable estimates for the whole state, for large- and medium-sized population counties in the state, and for groups of the smallest population counties. Three California cities that have their own health departments were also sampled as part of CHIS 2001.

The survey design supports study of California's major race and ethnic groups, and a number of smaller ethnic groups within the state. Adults, parents of children below age 12, and adolescents (ages 12-17) residing in California households are the eligible respondents to the survey. CHIS 2001 collected data between November 2000 and October 2001. The plans are to conduct independent cross-sectional surveys of the California population on a biannual basis to monitor important health-related indicators and potentially track changes over time. CHIS 2001 is the first of these planned surveys.

CHIS 2001 collected information on if, where, and how people get health care in California. The goal is to provide health planners, policymakers, state, county, and city health agencies, and community organizations with information on the health and health care needs facing California's diverse population. For example, the number and characteristics of adults, children, and adolescents without access to care and lacking health insurance can be estimated from the data collected in CHIS 2001. Other key estimates on the prevalence of cancer screening, diabetes, asthma, and other health conditions can also be produced. The survey includes major content areas, such as health status and conditions, health-related behaviors, access to health care services, and health insurance coverage.

1.2 Sample Design Objectives

The CHIS 2001 sample is designed to meet two objectives: (1) provide local-level estimates for counties and groupings of counties with populations of 100,000 or more; and (2) provide statewide estimates for California's overall population and its larger race/ethnic groups, as well as for several smaller ethnic groups. To address these objectives, the sample was allocated by county and aggregates of smaller counties, with supplemental samples of selected populations and cities. Table 1-1 shows the sampling strata (i.e., counties and groups of counties that were identified in the sample design as domains for which separate estimates would be produced). A sufficient amount of sample was allocated to each of these domains to support the first sample design objective.

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

1. Los Angeles	15. San Joaquin	29. El Dorado
2. San Diego	16. Sonoma	30. Imperial
3. Orange	17. Stanislaus	31. Napa
4. Santa Clara	18. Santa Barbara	32. Kings
5. San Bernardino	19. Solano	33. Madera
6. Riverside	20. Tulare	34. Monterey, San Benito
7. Alameda	21. Santa Cruz	35. Del Norte, Humboldt
8. Sacramento	22. Marin	36. Lassen, Modoc, Siskiyou, Trinity
9. Contra Costa	23. San Luis Obispo	37. Lake, Mendocino
10. Fresno	24. Placer	38. Colusa, Glen, Tehama
11. San Francisco	25. Merced	39. Sutter, Yuba
12. Ventura	26. Butte	40. Plumas, Nevada, Sierra
13. San Mateo	27. Shasta	41. Alpine, Amador, Calaveras, Inyo,
14. Kern	28. Yolo	Mariposa, Mono, Tuolumne

Source: UCLA Center for Health Policy Research, 2001 Califomia Health Interview Survey.

Samples were also drawn from each of the three California cities that have their own local health department. In addition, supplemental samples were developed for three counties that contracted for additional sample to enhance their overall estimates. These city and supplemental county samples were in the following locations:

- The cities of Berkeley, Long Beach, and Pasadena; and
- The counties of San Francisco, Santa Barbara, and Solano.

The three city samples and the Solano county supplemental sample were implemented with and incorporated in the original statewide RDD sample. The separate San Francisco and Santa Barbara supplemental samples were subsequently added to the statewide RDD sample prior to constructing the sample weights and are part of the final CHIS 2001 RDD sample file.

To accomplish the second objective, larger sample sizes were allocated to the more urban counties where a significant portion of the state's African American and Asian ethnic populations reside. Additionally, supplemental samples were used to improve the sample size and precision of the estimates for specific ethnic groups. The supplemental ethnic group samples in CHIS 2001 were as follows:

- South Asian, Cambodian, Japanese, Korean, and Vietnamese;
- American Indian/Alaska Natives in urban and rural areas; and
- Latinos residing in Shasta County (a sample requested by the local health department).

1.3 Data Collection

To capture the rich diversity of the California population, interviews were conducted in six languages: English, Spanish, Chinese (Mandarin and Cantonese dialects), Vietnamese, Korean, and Khmer (Cambodian). These languages were chosen based on research that identified these as the languages that would cover the largest number of Californians in the CHIS sample design that either did not speak English or did not speak English well enough to otherwise participate.

Westat, a private firm that specializes in statistical research and large-scale sample surveys, conducted the CHIS 2001 data collection for the CHIS project. Westat staff interviewed one randomly selected adult in each sampled household. In those households with children (under age 12) or adolescents (ages 12-17), one child and one adolescent were randomly sampled, so up to three interviews could have been completed in each sampled household. The sampled adult was interviewed, and the parent or guardian who knew the most about the health and care of the sampled child was interviewed. The sampled adolescents responded for themselves, but only after a parent or guardian gave permission for the interview. Since adolescents were not reliable sources concerning their own health insurance coverage, the parents of sampled adolescents were interviewed about this topic separately.

One criterion for the adolescent and child to be selected for the survey is that they had to be "associated" with the selected adult. This meant that in most cases the interviewed adult had to be either the parent or guardian. The CHIS 2001 sample weights adjust for this selection criterion so as not to bias estimates based on the adolescent and child surveys. Table 1-2 shows the number of completed adult, child, adolescent, and adolescents' health insurance interviews in CHIS 2001, by the type of sample (RDD or supplemental sample).

Table 1-2. Number of completed interviews by type of sample, instrument

Type of sample	Adult	Child	Adolescent	Adolescent insurance
Total RDD + supplemental cases	57,848	13,276	6,058	8,302
RDD (includes 3 cities + Solano county supplemental cases)	54,122	12,392	5,733	7,809
Santa Barbara supplemental cases	206	49	22	31
San Francisco supplemental cases	1,100	151	46	79
Total CHIS 2001 RDD file	*55,428*	*12,592*	*5,801*	*7,919*
Other supplemental samples:				
South Asian	443	158	39	65
Cambodian	126	44	37	44
Japanese	330	51	18	33
Korean	326	95	30	44
Vietnamese	540	124	34	60
American Indian/Alaska Native	351	106	51	71
Shasta Latinos	304	106	48	66

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

The interviews done in English were administered using Westat's computer-assisted telephone interviewing (CATI) system. Spanish and Vietnamese language interviews were also conducted entirely in CATI, while interviews conducted in Cantonese, Mandarin, Korean, and Khmer used English CATI screens and paper translations in tandem. The average adult interview took around 32 minutes to complete. The average child and adolescent interviews took 14 minutes and 19 minutes, respectively. Interviews in the non-English languages generally averaged longer to complete. Approximately 12 percent of the adult interviews were completed in a language other than English, as were 21 percent of all child (parent proxy) interviews and 9 percent of all adolescent interviews.

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

Table 1-3. Survey topic areas by instrument

Adult interview	Child interview	Adolescent interview
Age, sex, race, ethnicity	Age, sex, race, ethnicity	Age, sex, race, ethnicity
Physical activity		Physical activity
	Bike helmet use	Bike helmet, seatbelt use
	Recent serious injury	Recent serious injury
Health status	Health status	Health status
Women's health	Child care	
Chronic health conditions	Asthma, ADD	Asthma, diabetes
Cancer history, screening		
Skin cancer prevention	Skin cancer prevention	Skin cancer prevention
Health care use and access	Health care use and access	Health care use and access
Alcohol, tobacco use		Alcohol, tobacco, drug use
Mental health		Mental health
Health insurance	Health insurance	Health insurance
Diet (fruit-vegetable intake)	General diet	General diet
Dental health	Dental health	Dental health
Employment		Employment
Gun access, training		Gun access, violence
Income		
	Family interaction	Parental involvement
	Video games, computer use	Video games, computer use
Sexual orientation		Sexual behavior, orientation
		Future plans

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

1.4 Response Rate

The overall response rate for CHIS 2001 is a composite of the screener completion rate (i.e., success in introducing the survey to a household in order to select a respondent), and the extended interview completion rate (i.e., success in getting the selected respondent to complete the full interview). For the adult survey, the screener completion rate was 59.2 percent and the extended interview completion rate was 63.7 percent. This gives an overall response rate of 37.7 percent. To maximize the survey's response rate, an advance letter (in five languages) was mailed to all sampled telephone numbers for which an address could be obtained from reverse directory services. Approximately 66 percent of the sample was mailed an advance letter. Response rates varied by sampling stratum and were slightly higher in households that received an advance letter.

To assist in achieving sample size goals, respondents that completed 80 percent of the questionnaire (i.e., through Section I on health insurance) after all followup attempts were exhausted to complete the full questionnaire were counted as "complete." This resulted in 397 "partial completes" being included in the final adult survey data. Employment and income information as well as potential public program eligibility and food insecurity information would be missing from these cases.

Proxy interviews were allowed for frail and ill persons over the age of 65. The reason is that health estimates made for elderly persons could be biased if this is not allowed. Eligible selected persons were recontacted and offered a proxy option and 316 had a proxy interview completed by either a spouse/partner or adult child. Only a subset of questions identified as appropriate for a proxy respondent were administered. (Note: The questions not administered are identified in their response set as being skipped (denoted by a value of "-2") because a proxy is responding for the selected person.)

1.5 Weighting the Random Digit Dial Sample

To produce correct population estimates for the RDD CHIS results, weights are applied to the sample data to compensate for a variety of factors, some directly resulting from the design and administration of the survey. Sample weighting was carried out in CHIS 2001 to accomplish the following objectives:

- Compensate for differential probabilities of selection for households and persons (Note: households with listed addresses and thus eligible for an advance letter were assigned a probability of selection of 1.25 over unlisted households);
- Reduce biases occurring because nonrespondents may have different characteristics than respondents;
- Adjust, to the extent possible, for undercoverage in the sampling frames and in the conduct of the survey; and
- Reduce the variance of the estimates by using auxiliary information.

As part of the weighting process for the RDD samples (each stratum is an independent sample), a household weight was created for all households that completed the screener interview. This household weight is the "base weight" computed as the inverse of the probability of selection of the sample telephone number adjusted for each of the following:

- Subsampling for listed address/advance letter status;
- Unknown residential status;
- Screener interview nonresponse;
- Multiple telephone numbers; and
- Household poststratification.

A "poststratified household weight" was then used to compute a person-level weight. This person-level weight incorporates the within-household probability of selection of the sampled person and adjusts for nonresponse, plus an adjustment resulting from raking the data to person-level control totals. Each of these adjustments corresponds to a multiplicative weighting factor.

Raking can be thought of as a multidimensional poststratification procedure because the weights are basically poststratified to one set of control totals (a dimension), then these adjusted weights are poststratified to another dimension. After all dimensions were adjusted, the process was iterated until the control totals for all the dimensions were simultaneously satisfied (within a specified tolerance).

There are 11 dimensions used in CHIS 2001. The first 10 dimensions are created by combining demographic variables (age, sex, race, and ethnicity) and different geographic areas (city, county, group of counties, and state). The 11th dimension is created to adjust the weights for households without a telephone number.

The control totals used in the raking were derived from the *Census 2000 Summary File 1* (SF1). Population items in SF1 include sex, age, race, ethnicity (Latino/non-Latino), household relationships, and group quarters. The race classification in SF1 include six groups: White, African American, American Indian/Alaska Native, Asian, Native Hawaiian/Pacific Islander, and a category of Other Race. Since a person could report multiple races, the SF1 provided counts for each of 63 possible race combinations a person could report.

One of the limitations of using the SF1 for the control totals is the inability to produce counts that exclude the fraction of the population living in "group quarters" (e.g., nursing homes, prisons) for some dimensions used in CHIS 2001. The group quarter population represented 2.4 percent of the total population in California. As a result, the number of persons living in group quarters was estimated for some of the raking dimensions, and the SF1 totals were reduced by these estimated amounts prior to raking.

1.6 Imputation Methods

Three different imputation procedures were used in CHIS 2001 to fill in missing responses that were essential for weighting the data or for such basic descriptive purposes as income categories. The first imputation technique is deterministic or non-stochastic in nature. Deterministic imputation was used to fill in the missing items for self-reported county of residence (item AH42). These imputations required no randomization because other geographic data are available that can be used to determine the respondent's county of residence with a relatively high level of probability of being correct although not with 100 percent certainty in all cases.

The second imputation technique is a completely random selection from the observed distribution. This method is used only when a very small percentage of the items are missing. For example, when imputing the missing values for self-reported age, the distributions of the responses for age by type of interview (adult, child, or adolescent) were used to randomly assign an age using probabilities associated with these distributions.

The third technique is hotdeck imputation. Hotdeck imputation was used to impute race, ethnicity, and household income in CHIS 2001. The hotdeck approach is probably the most commonly used method for assigning values for missing responses in large-scale household surveys.

With a hotdeck, a value reported by a respondent for a particular item is assigned or donated to a "similar" person who did not respond to that item. To carry out hotdeck imputation for CHIS 2001, the respondents to an item form a pool of donors, while the nonrespondents are a group of recipients. A recipient is matched to the subset pool of donors, with the same household structure. The recipient is then randomly imputed the same household income, ethnicity/race (depending on the items that need to be imputed) from one of the donors in the pool. Once a donor is used, it is removed from the pool of donors.

Imputation flags are used in the data file to identify all imputed values.

1.7 Methodology Report Series

A series of five methodology reports are available with more detail about the methods used in CHIS 2001:

- Report 1 – Sample Design
- Report 2 – Data Collection Methods
- Report 3 – Data Processing Procedures
- Report 4 – Response Rates
- 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 www.CHIS.ucla.edu or contact CHIS at CHIS@ucla.edu.

2. TELEPHONE SAMPLING METHODS

This chapter describes the two general sampling methods used in the CHIS 2001 telephone surveys. CHIS 2001 consisted of an RDD sample supplemented by geographic and race-ethnic supplemental samples. The RDD sample and geographic supplemental samples were drawn using a list-assisted RDD approach, whereas the supplemental race-ethnic samples were drawn from separate lists of telephone numbers. The first section below describes the list-assisted RDD sampling and the procedures implemented in CHIS 2001 to save costs by reducing the number of calls to ineligible telephone numbers in this sample. The second section reviews the sampling alternatives that were considered for supplementing the RDD sample to increase the sample size for specified race and ethnic subgroups of interest. This section also gives the rationale for deciding that the best approach to the supplemental samples was the use of special lists of telephone numbers.

Households without a telephone were not sampled for CHIS 2001, which could give rise to bias in the estimates. The bias is related to the percentage of households without telephones and the difference in characteristics of the telephone and nontelephone households. Approximately 5 percent of households in California are without telephones. Recent evidence (Ford 1998; Anderson Nelson, and Wilson 1998) shows that the health characteristics of those with and without telephones are not as different as they had been in the past. Based on these factors, it is unlikely that most estimates from CHIS will have substantial bias because nontelephone households are not sampled. However, some estimates that are very directly correlated to income may be subject to greater biases due to this form of undercoverage. To mitigate the effects of excluding households without telephones, special weighting procedures were used and these are described in Report 5: Weighting and Variance Estimation.

In many in-person and telephone household surveys, persons who do not speak English and in some cases who do not speak Spanish are sampled, but never interviewed because of language difficulties. While technically we prefer to treat this as a nonresponse problem (the language cases are considered nonrespondents), it could easily be thought of as a coverage problem since none of the persons are interviewed. In CHIS 2001, significant efforts were expended to limit this source of bias. As mentioned in the previous section, the multiple languages used to interview for CHIS should eliminate a large source of the bias that might result from conducing interviews only in English.

2.1 List-Assisted Random-Digit-Dial Sampling

List-assisted sampling is a sampling procedure for telephone surveys made possible by recent technological developments (Casady and Lepkowski, 1993). In list-assisted sampling, the set of all telephone numbers in operating telephone prefixes is considered as composed of 100-banks. Each 100-bank contains the 100 telephone numbers with the same first eight digits (i.e., the identical area code, telephone prefix, and first two of the last four digits of the telephone number). All 100-banks with at least one residential number listed in a published telephone directory are identified. The sampling frame is restricted to these 100-banks. A simple random or a systematic sample of telephone numbers is selected from this frame.

List-assisted RDD sampling is currently the standard method of choice for telephone surveys. It results in an unclustered sample that can be released to interviewers once the sample of telephone numbers is chosen. These are both important features not shared by the Mitofsky-Waksberg method that used to be the standard RDD sampling technique (Brick and Waksberg, 1991). Furthermore, the working residential rate among sampled numbers (critically important in determining the cost of an RDD sample) is comparable to the Mitofsky-Waksberg technique. The only disadvantage is a small amount of undercoverage because telephone numbers in 100-banks with no listed telephone numbers are not sampled. Studies have been carried out on the potential losses associated with this truncated form of list-assisted sampling. The studies show only about 2 to 4 percent of households are excluded by this method. Furthermore, the households that are excluded do not appear to be very different from those included in the frame (Brick, et al., 1995; Giesbrecht, et al., 1996). As a result, the bias due to this method of sampling is considered negligible for most estimates.

When using a list-assisted approach, special procedures can be implemented prior to the beginning of the data collection period to reduce costs. These procedures take advantage of developments in technology and linking of data sources. Three such procedures were used in CHIS 2001. In the first procedure, every sampled telephone number is classified as listed, unlisted, or nonresidential by matching the sample to computerized files from the White Pages (residential numbers) and Yellow Pages (business numbers). Telephone numbers listed only in the Yellow Pages are eliminated as nonresidential numbers. In CHIS 2001 about 4.6 percent of the sampled numbers were eliminated by this method.

The second procedure is to use a computer to dial the telephone numbers automatically if they are not listed in either the White or Yellow Pages. The numbers in the White Pages are not included

because they are very likely to be residential. The computerized dial device detects if a tritone signal is produced when a telephone number is dialed. A tritone is a signal that the telephone company has not assigned the number to a household or a business. The device works quickly and the tritone is usually detected before any audible ring of the phone at a number. This operation resulted in eliminating an additional 19.7 percent of the sampled telephone numbers. A total of 24.3 percent of the sampled numbers was eliminated by the tritone and business purge.

The third procedure is somewhat different than the first two in that it involves subsampling rather than eliminating sampled telephone numbers. First, every telephone is classified by whether a mailing address can be associated with the telephone number[1] (i.e. mail status). We refer to these as telephone numbers that have a "mailable" address. Telephone numbers classified as listed and/or mailable in CHIS 2001 were then subsampled at differential rates. Since listed and/or mailable telephone numbers are much more likely to be residential, all of these telephone numbers were retained in the sample. The unlisted and/or not mailable telephone numbers are less likely to be residential so the cost of finding a residential number is greater in this substratum. For CHIS 2001, about 80 percent of the unlisted telephones and/or not mailable addresses were retained in the subsampling procedure.

Genesys[2] provided the sampling frame or list of banks used for sampling in CHIS 2001. While the list of banks is continuously updated, the sampling frame is updated quarterly. Because the data collection period for CHIS was rather long (beginning in November 2000 and ending in October 2001), the RDD sample was drawn at two points in time (July 2000 and February 2001). Sampling at these two times enabled us to reflect the dynamic nature of the population in California over the data collection period. Over the two selections, a total of 345,136 telephone numbers was selected for the RDD sample and the geographic supplemental samples. Details of the sampling, including the sampling by listed/mailable status, are given in the next chapter.

2.2 Supplemental Sampling

As noted in the first chapter, CHIS 2001 included both geographic and race-ethnic supplemental samples. The geographic supplemental samples were list-assisted samples for the specific

[1] Several companies provide services of this type in which a telephone number is matched to commercially-available files of addresses.

[2] Genesys Sampling Systems is a service sampling company that provides a wide variety of services to the survey research community. Among these services, Genesys maintains databases for sample selection in telephone surveys.

areas and used the methods described in the previous section for the RDD sample. The only difference is that some of the supplemental samples were selected with the entire RDD sample (the Solano County supplement) and others were selected after the RDD sample because of funding issues. Since the method of sampling for the geographic supplemental samples is the same as the RDD sample, the remainder of this section is devoted to the sampling method for the race-ethnic supplemental samples.

An important goal of CHIS 2001 was to produce reliable estimates for the specific racial and ethnic subgroups in California identified in Table 1-2. These subgroups are important for analytic reasons but constitute a small proportion of the total population and are dispersed throughout the state. As a result, the expected sample yield in even a large survey like CHIS 2001 is too small to support making inferences for the subgroups at the desired level of precision. Because these subgroups are a small percentage of the population, and are geographically dispersed, and no single list of all the members of the subgroup is available, sampling methods for rare populations were considered for including them in CHIS 2001. Kalton and Anderson (1986) and Sudman, Sirken, and Cowan (1988) are general references for sampling rare populations.

Several sampling strategies were considered to increase the sample size for racial and ethnic subgroups in CHIS 2001. The sampling strategies include household screening, use of auxiliary information to classify telephone numbers, network sampling, and the use of special lists. Each of these strategies is described below. At the end of the section we summarize the reasons for choosing the list sampling option in CHIS 2001.

The first sampling strategy considered is screening households. The procedure works by increasing the sample size for the survey until it is large enough to support the smallest or rarest subgroup in the population. When a telephone number is sampled, the household is classified by which group it contains. If the household contains a member of the rarest subgroup, it is retained. Otherwise, it is subsampled. For example, suppose there is one rare group and it is 5 percent of the population. In this case, the sample size would be increased by a factor of 20 to obtain enough of the rare population. If a telephone number is sampled and there is a member of the rare group, it is retained. Otherwise, it is subsampled and has a chance of about 5 percent of being included in the sample. This screening of households for a member of the rare subgroup is probably the simplest method, but it is also the most costly. This strategy is relatively simple to implement and has good statistical properties, with the exception of the measurement error associated with screening households on the basis of asking a

question about the racial composition in the beginning of a telephone interview. Because the data collection costs for this scheme are very large, this method was not explored extensively for CHIS 2001.

A second sampling strategy considered is essential in a stratification approach. Under this scheme auxiliary information is used to classify telephone exchanges (or banks of telephone numbers) by the proportion of members of the groups residing in these exchanges. After classifying the exchanges into strata, the telephone numbers in the exchanges with a high proportion of members would be sampled at a higher rate than the numbers in the other strata. If the data used to stratify the numbers is accurate, then the telephone numbers in the exchanges sampled at higher rates are more likely to result in interviews with members of the rare subgroup. While this procedure has been used in other RDD surveys to improve the precision of estimates of African Americans and Latinos, the method is not currently feasible for groups of interest in CHIS 2001 because no data for these race-ethnic groups is available to stratify the telephone numbers.

A third sampling strategy we considered for improving the precision of estimates of the race-ethnic groups is called multiplicity or network sampling. In this approach, each household sampled that belongs to the group is asked to identify other households. These households are then interviewed. The advantage of this method is that it provides an inexpensive method of locating and interviewing a larger sample. The identification of other households is a key feature of the process because it is part of the sample selection process. These linkages to the other households must be unambiguously defined to compute unbiased estimates in accordance with the requirements of probability sampling. For example, most households in California can only be sampled once in the RDD sample through the household's telephone number. Using multiplicity sampling a household could be sampled not only by selecting the household's own telephone number, but also as a result of linking the household to other households in California. The links in multiplicity sampling are usually clearly defined in terms of immediate relatives. Most often, sampled households are asked about their children, parents, or siblings and these households constitute a network. The probability of each household in this network (including the household with the sampled telephone number) is then computed using the reported size of the network. If all the households have the same probability of selection (assuming that they are all telephone households in the same county), then the probability of selection for each is the probability of sampling a telephone number divided by the number of households in the network. The most serious impediment to the successful application of multiplicity sampling in CHIS is nonresponse, in several manifestations. The first and most obvious issue is the willingness of race-ethnic groups to identify all their siblings who live in California

and provide enough information to the interviewers so that they can be contacted. A related issue is the willingness of the linked siblings to respond to the interview.

An alternative version of this form of sampling is sometimes called "snowball" sampling. The mechanism is the same as that mentioned above, in that sampled households act as referents for other households of the same race-ethnic group. The method is different because the probabilities of selection associated with the links are not evaluated, leading to a nonprobability sample. This approach can increase the sample yield but does not allow for unbiased estimation of the number and characteristics of groups using the data collected in the snowballing. This approach was not acceptable since it is not a probability sample.

A fourth scheme considered is based on the concept of a dual frame design. In this design, the regular CHIS sample selected from the RDD survey is supplemented using a much less expensive sample from a list of telephone numbers of the race-ethnic groups. The list frame does not have to be complete to be useful, although the more complete the list the greater the potential for increasing the precision of the estimates. The composition of the list affects its efficiency, but not the ability to produce unbiased estimates. Unbiased estimates can be produced if the list membership of every sampled person from the RDD sample can be determined. Of course, if the list only contains members of one subgroup of the entire race-ethnic group, the efficiency for many types of analysis may be adversely affected. For example, in the case of American Indian/Alaska Natives, the precision of comparisons of the characteristics of rural and urban American Indian/Alaska Natives would not be substantially improved if only rural American Indian/Alaska Natives were members of the list. In most applications, the cost of data collection using a list is dramatically lower than the cost for screening for members of the rare population. See Report 2: Data Collection Methods for CHIS 2001, for a comparison of per-completed interview costs for the RDD, and supplemental samples.

In a dual frame approach, the characteristics of the list are very important and worth reviewing in some detail. The first characteristic is that the list must contain the telephone number for members of the race-ethnic subgroup so the sample from the list can be interviewed. The telephone numbers are also needed for estimation purposes, as described in Report 5: Weighting and Variance Estimation. A second important property of the list is its completeness in terms of containing a large percentage of the subgroup of interest. Lists that are more complete make the sampling process more efficient. A third property of the list is the need to cover a relatively broad spectrum of types of members of the race-ethnic groups. The example above for American Indian/Alaska Natives illustrates the

importance of this property. Finally, the accuracy of the lists in identifying the members of these groups is important. A list is accurate if the telephone numbers on the list actually do contain members of the rare subgroup targeted. If the list is inaccurate, then a larger screener cost is incurred.

After evaluating the different sampling strategies, the dual frame or list supplemental sampling method was chosen for all the race-ethnic supplemental samples. The screening approach was too costly; the stratification method could not be implemented because no data were available. The costs and yields for the multiplicity or network sampling approach could not be estimated in advance and this made the approach unacceptable. Furthermore, the measurement and nonresponse problems could not be tested in the time available before fielding the sample. Thus, the dual frame approach using lists was deemed to be the one most likely to succeed within the time and cost constraints of the survey.

The lists for the supplemental samples were created using surnames of each of the race-ethnic groups for the state of California. The only exception is the sampling for American Indian/Alaska Natives discussed below. Genesys maintains lists of surnames associated with certain ethnic subgroups. The target subgroups for CHIS 2001 were among those subgroups. By matching the surnames for the subgroup against the listed surname in the White Pages for the state, a sample was selected for each subgroup. For the Shasta Latino sample, the list of Latino surnames was restricted to Shasta County. For all other subgroups, the sampling was done over the entire state. Simple random samples were drawn from the surname lists. Table 2-1 shows the size of the surname lists.

Table 2-1. Number of records in the surname lists

Surname list	Number of records
South Asian	56,335
Cambodian	9,941
Japanese	100,854
Korean	208,315
Vietnamese	216,036
American Indian/Alaska Native	39,591
Shasta Latinos	1,906

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

The American Indian/Alaska Native supplemental sample used a different procedure to obtain a list. The list was developed by UCLA in coordination with American Indian/Alaska Native tribes and organizations. A large fraction of the listed telephone numbers was obtained from the U.S. Department of Indian Health Services (IHS). The list was stratified into rural and urban strata and simple

random samples were selected from each of the two strata. The number of records in each stratum is shown in Table 2-2.

Table 2-2. Number of records in the American Indian/Alaska Native list by stratum

Stratum	County	Number of records
Urban	Alameda, Contra Costa, Kern (the city of Bakersfield only), Los Angeles, Marin, Merced, Monterey, Napa, Orange, Sacramento, San Benito, San Diego (the city of San Diego only), San Francisco, San Joaquin, San Luis Obispo, San Mateo, Santa Barbara (the city of Santa Barbara only), Santa Clara, Santa Cruz, Solano, Stanislaus, and Ventura.	10,161
Rural	Alpine, Amador, Tuolumne, Butte, Calaveras, Colusa, Del Norte, El Dorado, Fresno, Glenn, Humboldt, Imperial, Inyo, Kern (excluding the city of Bakersfield), Kings, Lake, Lassen, Madera, Mariposa, Mendocino, Modoc, Mono, Nevada, Placer, Plumas, Riverside, San Bernardino, San Diego (excluding the city of San Diego), Santa Barbara (excluding the city of Santa Barbara), Shasta, Sierra, Siskiyou, Sonoma, Sutter, Tehama, Trinity, Tulare, Yolo, and Yuba.	29,430
Total		39,591

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

Table 2-3 summarizes CHIS 2001 samples by type of sample, source of the sample or sampling frame, and the geographic area covered by each sample.

Table 2-3. Summary of the CHIS 2001 samples

Sample		Type of sample	Frame	Area
1	RDD sample	Main sample	1+ bank list	State
2	Pasadena City	Geographic	1+ banks list	City
3	Long Beach City	Geographic	1+ banks list	City
4	Berkeley City	Geographic	1+ banks list	City
5	Solano County	Geographic	1+ banks list	County
6	San Francisco County	Geographic	1+ banks list	County
7	Santa Barbara County	Geographic	1+ banks list	County
8	South Asian	Race-Ethnic	Surname list	State
9	Cambodian	Race-Ethnic	Surname list	State
10	Japanese	Race-Ethnic	Surname list	State
11	Korean	Race-Ethnic	Surname list	State
12	Vietnamese	Race-Ethnic	Surname list	State
13	American Indian/Alaska Native	Race-Ethnic	Special list	State
14	Shasta Latinos	Race-Ethnic	Surname list	County

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

3. SAMPLING HOUSEHOLDS

This chapter describes the sample design and selection of households for CHIS 2001. We begin by clearly identifying the target population and the persons included and excluded in the survey. The goals of the survey in terms of the desired number of completed adult interviews by county are described in the second section. The goals for both the RDD sample and supplemental samples are included in this discussion. The remainder of the chapter describes how the sample of telephone numbers was selected in order to achieve the stated goals. The last section reviews the statistical issues considered in arriving at the allocation of the RDD sample by county.

3.1 Population of Interest

As discussed in Chapter 1, CHIS 2001 consisted of sampling and interviewing a randomly selected adult from every sampled telephone number that was residential. Any person 18 years or older is defined as an adult for the survey. In addition, one randomly selected child (under 12 years) and one adolescent (12 to 17 years) were eligible to be sampled if there were any persons in these ages in the household.

The eligibility rules used in CHIS 2001 are typical of those made in other telephone surveys. Eligible residential households included houses, apartments, and mobile homes occupied by individuals, families, multiple families, or extended families or occupied by multiple unrelated persons, provided that the number of unrelated persons is less than nine. Persons living temporarily away from home were eligible and enumerated at their usual residences. These include college students in dormitories, patients in hospitals, vacationers, business travelers, and so on.

The survey excluded dwelling units without a telephone number or group quarters. A group quarter is any unit occupied by nine or more unrelated persons (e.g., communes, convents, shelters, halfway houses, or dormitories). Institutionalized persons (i.e., those living in prisons, jails, juvenile detention facilities, psychiatric hospitals and residential treatment programs, and nursing homes for the disabled and aged), the homeless, persons in transient or temporary arrangements, and those in military barracks were also excluded.

3.2 Sample Allocation

In this section we describe the targeted number of completed interviews for CHIS 2001. We begin by discussing the RDD sample and then deal with the supplemental samples.

Two of the goals of CHIS were to produce reliable statewide estimates for the total population and for subgroups and to produce reliable estimates at the county level for as many counties as possible. These goals required a compromise in terms of the allocation of the sample. To achieve the most reliable statewide estimates, the optimal design is to allocate the sample to counties proportional to their population. On the other hand, the optimal allocation for producing as many county level estimates of high precision is to assign each county an equal sample size. In this section we present the final compromise that was reached that allowed for both precise statewide estimates and reliable county level estimates for most of the counties in California. We also discuss the rationale for the allocation, but we leave the more detailed statistical issues until a later section.

The 58 California counties were arranged into 41 strata as shown in Table 3-1. Thirty-three of the 35 counties with a population of 100,000 or more are separate sampling strata. The two remaining counties with over 100,000 persons are each combined with an adjoining smaller county to form a stratum. The 23 remaining counties with populations of less than 100,000 were placed into six strata for analytic reasons (including geography). The minimum target sample size of 800 completed adult interviews was set for each stratum. The target sample size for the counties with larger populations were greater than the minimum, with the largest target sample size of 11,292 for Los Angeles. The target sample sizes are given in Table 3-1. CHIS 2001 had a goal of completing 51,364 adult interviews, between 4,000 and 5,000 adolescents (depending on compliance since parental consent and adolescent ascent are required), and from 12,000 to 13,000 children by adult proxy for the RDD sample.

Because of the need for producing reliable estimates for the counties, the sample allocation is not in all cases proportional to the population across counties (proportional allocation was done among San Diego, Orange, Santa Clara, San Bernardino, Riverside and Alameda counties). With a proportional allocation, the estimates from the moderate and smaller counties would be based on small sample sizes and would not be adequate for the envisioned analysis. To achieve the goal of producing local or county estimates, the sample sizes from the largest counties are re-distributed to the smaller counties.

Table 3-1. Number of targeted complete adult interviews for the RDD sample by county

	Sampling stratum	Targeted number of adult interviews	Population Size
1	Los Angeles	11,292	Over 9,000,000
2	San Diego	2,660	1,200,000 or greater
3	Orange	2,540	
4	Santa Clara	1,554	
5	San Bernardino	1,527	
6	Riverside	1,359	
7	Alameda	1,232	
8	Sacramento	1,200	800,000 to 1,200,000
9	Contra Costa	1,200	
10	Fresno	1,000	500,000 to 800,000
11	San Francisco	1,000	
12	Ventura	1,000	
13	San Mateo	1,000	
14	Kern	1,000	
15	San Joaquin	1,000	
16	Sonoma	800	100,000 to 500,000
17	Stanislaus	800	
18	Santa Barbara	800	
19	Solano	800	
20	Tulare	800	
21	Santa Cruz	800	
22	Marin	800	
23	San Luis Obispo	800	
24	Placer	800	
25	Merced	800	
26	Butte	800	
27	Shasta	800	
28	Yolo	800	
29	El Dorado	800	
30	Imperial	800	
31	Napa	800	
32	Kings	800	
33	Madera	800	

Table 3-1. Number of targeted complete adult interviews for the RDD sample by county (continued)

	Sampling stratum	Targeted number of adult interviews	Population Size
34	Monterey (pop. >100,000) San Benito (pop. <100,000)	800	Small and medium counties combined
35	Humboldt (pop. >100,000) Del Norte (pop. <100,000)	800	
36	Siskiyou Trinity Lassen Modoc	800	Less than 100,000 population per county
37	Mendocino Lake	800	
38	Tehama Colusa Glenn	800	
39	Sutter Yuba	800	
40	Nevada Sierra Plumas	800	
41	Tuolumne Mariposa Calaveras Mono Amador Alpine Inyo	800	
	Total of 41 Strata	51,364	

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

At the beginning of the study, different allocations of the sample consistent with the budget constraints were evaluated. The UCLA CHIS staff consulted with the counties and other analytic groups to define the relative importance of particular types of estimates. Westat statistical staff helped evaluate each alternative and examined the consequences of the sample allocations. The main statistical issues were communicated by computing effective sample size for the main groups for the alternative designs. The expected effective sample size computations are discussed in Section 3.4.

The RDD main sample was augmented with supplemental samples for some geographic areas and race-ethnic groups. Supplemental samples were planned for three cities in California with their own health departments. In addition, each county had the opportunity to increase the sample size for their area if funding could be arranged. Three counties provided funds for larger sample sizes (San Francisco, Solano, Santa Barbara). The geographic supplemental samples included in CHIS 2001 are listed in Table 3-2. The table also includes the targeted number of adult interviews from both the RDD sample and the supplemental sample.

Table 3-2. Number of targeted complete adult interviews for the geographic supplemental samples

City/County	Targeted number of adult interviews		
	RDD	Supplemental	Total
Long Beach	582	218	800
Pasadena	178	622	800
Berkeley	98	702	800
Solano County	800	800	1,600
San Francisco County	1,000	1,000	2,000
Santa Barbara County	800	200	1,000
Total Sample	3,458	3,542	7,000

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

The sample sizes for the cities of Pasadena, Long Beach, and Berkeley were increased to 800 per city. The supplemental sample for Solano County doubled the RDD sample allocation to a total of 1,600 completed adult interviews. The supplemental samples for San Francisco and Santa Barbara were targeted at specific areas defined by ZIP Codes. The targeted sample sizes for these two geographic samples by ZIP Code are shown in Table 3-3. The target San Francisco supplemental sample was later increased to 1,100 because of a shortage in the RDD sample.

Table 3-3. Number of targeted complete adult interviews for the San Francisco and Santa Barbara supplemental samples by ZIP Code

Area	ZIP Code	Neighborhood/Area	Estimated number of adult interviews
San Francisco	94124	Bayview Hunters Point	82
	94134	Visitation Valley-Portola	107
	94108	Chinatown	45
	94133	North Beach-Chinatown	85
	94112	Ingleside-Excelsior	198
	94121	Outer Richmond	126
	94103	South of Market	49
	94102	Tenderloin-Western Addition	89
	94110	Inner Mission	219
Total			1,000
Santa Barbara	93454, 93434	Santa Maria and Guadalupe CA	200
Total			1,200

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

The second type of supplemental sample included in CHIS 2001 is the race and ethnic supplemental samples. The targeted number of completed adult interviews for these supplemental samples, along with the expected sample size from the RDD sample are shown in Table 3-4. The supplements for the five Asian subgroups were sampled using lists of surnames across the entire state. The supplemental sample for Latinos was restricted to Shasta County.

The American Indian/Alaska Native supplemental sample was designed to produce reliable urban and rural estimates using a definition of urban and rural at the county level that is used by the IHS in California. The main RDD sample was expected to interview approximately 450 American Indian/Alaska Natives. Approximately 60 percent of the households were expected to be in rural areas and 40 percent in urban areas. The supplement was designed to achieve at least 400 completed adult interviews in each of the urban and rural areas.

Table 3-4. Number of targeted complete adult interviews for the race and ethnic supplemental samples by subgroup

Subgroup	Targeted number of adult interviews		
	RDD	Supplement	Total
South Asian	314	486	800
Cambodian	102	498	600
Japanese	506	294	800
Korean	310	490	800
Vietnamese	475	325	800
American Indian/Alaska Native in urban areas	187	213	400
American Indian/Alaska Native in rural areas	257	143	400
Shasta Latinos	800	378	1,178
Total	2,951	2,827	5,778

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

Note: The urban/rural definition is one used by the IHS in California.

The race-ethnic supplemental sample targets were adjusted during data collection as the actual RDD yield became known.

3.3 Sampling Frames and Sample Selection

In this section we describe the actual steps used in selecting the sample of telephone numbers for CHIS 2001. These steps include stratifying the telephone numbers by sampling strata, selecting the sample of numbers after adjusting for expected losses due to nonresponse, and subsampling the numbers based on listed/mailable status to improve the efficiency of the sample.

Since CHIS 2001 RDD sample is a stratified sample, the first step is stratifying the sampling frame of 100-banks with one or more listed telephone numbers into non-overlapping strata corresponding to a city, county, or a group of counties. The required geographic information for stratification is available only at the exchange level[3], so 100-banks cannot be assigned directly to a single stratum. All banks within an exchange are stratified indirectly by mapping the exchanges to a county represented by the stratum. However, some telephone exchanges actually service households in more than one county.

To solve the stratification problem, Genesys produced coverage reports for each county in California. The coverage reports list all the exchanges in the county. For each exchange, the report gives the total number of listed households in the exchange and the proportion of listed households that are within the county. After combining the information of the coverage reports for all 52 counties, we created a frame of exchanges with variables for the number of listed households in each county that the exchange covers. Each exchange was assigned to the county that contains the most listed households.

A second step in sampling for CHIS 2001 involved dealing with uncertainties in the overall sample size that was dependent on funding and accounting for changes in the frame of telephone numbers over the data collection period. Data collection began in late November 2000 and the RDD sample ended data collection in early September 2001. During this time the population experienced changes with respect to people moving into the state, others moving out of the state, as well as people moving to different areas within the state and changing telephone numbers. The frame of telephone numbers also changed during this period, with new exchanges and 100-banks appearing and others dropping out of the frame.

As a result of these concerns, the CHIS 2001 RDD sample was selected at two points in time: July 2000 and February 2001. Table 3-5 shows the number of exchanges in the July 2000 and

[3] A telephone exchange consists of 10,000 consecutive telephone numbers with the same first six digits including area code. An exchange is a set of area codes and prefixes serving the same geographic area.

February 2001 sampling frames. The table shows that the vast majority of exchanges were in both sampling frames. Furthermore, the percentage of listed households in the portions of the sampling frame that were not present at both times was very small. Therefore, while the sampling did account for the dynamic nature of the sampling frame, there were only small changes within the time period.

The CHIS sample was designed so that the sample selected from the July 2000 sampling was targeted to achieve 17,092 completed adult extended interviews, approximately one-third of the total CHIS targeted sample size. The remainder of the sample was selected from the February 2001 sampling frame. In addition to simply being selected at different times, the samples drawn from the two frames also differed in other respects that are discussed below.

Table 3-5. Number of exchanges in the July 2000 and February 2001 RDD frames

Exchange in the July 2000 Frame	Exchange in the February 2001 frame		
	Yes	No	Total
Yes	5,818	236	6,054
No	293		293
Total	6,111	236	6,347

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

Note: The urban/rural definition is one used by the IHS in California.

The number of telephones selected in any RDD survey has to be greater than the targeted number of completed interviews to account for a variety of factors. For example, a substantial percent of the sampled telephone numbers is not residential. For CHIS 2001 the sample of telephone numbers was increased to deal with the losses due to following sources:

- Nonworking, nonresidential, and never answered numbers;
- Subsampling by listed/mailable status;
- Nonresponse to screening interview; and
- Nonresponse to extended adult interview.

The first, third, and fourth sources noted above are typical of all RDD surveys. To deal with these losses, we estimated the percentage of the telephone numbers that would not be residential and the percentage that would not respond, and increased the sample size accordingly. The only source of loss that requires additional discussion is the subsampling by listed/mailable status. As mentioned in Chapter

2, to increase the efficiency of the CHIS 2001 RDD sample we first stratified the telephone numbers within county by listed or mailable status and then subsampled those numbers that were not listed/mailable. The subsampling rate was 80 percent, meaning that about 20 percent of the not listed/mailable telephone numbers were removed from the sample. In sampling from the July 2000 sampling frame, we used the listed status to stratify and subsample the numbers. Before sampling from the February 2001 frame, we completed additional research that showed that mailable status was a slightly preferred method of stratification. As a result, when we selected the sample in February 2001 we first stratified the sample within county by mailable status. At both times, 80 percent of the not listed or not mailable telephone numbers were retained in the sample and 20 percent were subsampled out.

Another difference between the sampling at the two points in time involved evolutions in the sample design. As with the July 2000 sample, the number of telephones required was increased to account for losses due the factors mentioned above. However, in February we used the rates observed in the July sample to improve our estimates. The February sample was also expanded to include a reserve of telephone numbers in each stratum in case the targeted sample size goals were not met after exhausting the February sample. Both selections (July and February) and the reserve sample were split into sub-samples for a phased release over the data collection period. Table 3-6 gives the number of telephones selected for the July and February samples by strata. The data collection procedures are discussed in Report 2: Data Collection Methods.

Table 3-6. Number of telephones sampled by sampling frame and stratum

Strata	Description	July 2000 sample size	February 2000 sample size
1.1	Long Beach		4,080
1.2	Pasadena	31,350	5,281
1.3	Remainder of Los Angeles		45,247
2	San Diego	7,350	8,276
3	Orange	7,200	10,861
4	Santa Clara	4,200	6,679
5	San Bernardino	4,200	4,407
6	Riverside	3,750	3,957
7.1	Berkeley	3,450	4,719
7.2	Remainder of Alameda		4,663
8	Sacramento	3,450	3,634

Table 3-6. Number of telephones sampled by sampling frame and stratum (continued)

Strata	Description	July 2000 sample size	February 2000 sample size
9	Contra Costa	3,300	4,054
10	Fresno	2,850	4,341
11	San Francisco	2,700	5,388
12	Ventura	2,850	2,975
13	San Mateo	3,150	3,788
14	Kern	2,850	3,171
15	San Joaquin	2,850	2,559
16	Sonoma	1,800	2,165
17	Stanislaus	1,800	2,141
18	Santa Barbara	1,800	2,516
19	Solano	1,800	6,161
20	Tulare	1,800	3,222
21	Santa Cruz	1,800	3,006
22	Marin	1,800	3,181
23	San Luis Obispo	1,800	2,335
24	Placer	1,800	2,480
25	Merced	1,650	2,444
26	Butte	1,800	1,678
27	Shasta	1,800	1,966
28	Yolo	1,800	1,954
29	El Dorado	1,650	2,775
30	Imperial	1,800	2,147
31	Napa	1,800	2,757
32	Kings	1,800	2,506
33	Madera	1,800	2,291
34	Monterey, San Benito	1,650	3,783
35	Del Norte, Humboldt	1,800	3,234
36	Lassen, Modoc, Siskiyou, Trinity	1,800	3,456
37	Lake and Mendocino	1,800	2,923
38	Colusa, Glen, Tehama	1,650	2,457
39	Sutter, Yuba	1,800	2,557
40	Plumas, Nevada, Sierra	1,800	2,735
41	Alpine, Amador, Calaveras, Inyo, Mariposa, Mono, and Tuolumne	1,800	3,352
	State Total	131,700	200,302

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

When the sample was selected from the July 2000 sampling frame, it was not yet determined if sufficient funds would be available for increasing the sample sizes for the geographic supplemental samples of Los Angeles county and Alameda county. As a result, the July 2000 sample was selected for these counties without subsampling by the cities as denoted in Table 3-6. By the February 2001 sample selection it had been resolved to include these supplemental samples. The strata for Los Angeles County (stratum 1) and Alameda County (stratum 7) were split into separate strata corresponding to the cities as indicated in Table 3-7. The numbers of telephone numbers selected from these substrata in February 2001 are shown in Table 3-6.

Table 3-7. February 2001 stratum definition for Los Angeles and Alameda Counties

July 2000 sampling stratum	Description	February 2001 sampling stratum	Description
1	Los Angeles County	1.1	Long Beach
		1.2	Pasadena
		1.3	Remainder of Los Angeles County
7	Alameda County	7.1	Berkeley
		7.2	Remainder of Alameda County

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

The last samples selected were the other supplemental samples. These samples were selected after issues related to a variety of factors were resolved, including the determination of which languages would be used for interviewing, the translation of the interviews into these languages, funding, and acquisition of the lists or samples for the surname supplemental samples.

The number of telephone numbers needed to meet the targeted goals for each supplemental sample required estimating the losses due to the same sources as in the RDD sample plus additional sources. The most influential additional reason for needing a larger sample of telephone numbers was an issue with eligibility. For example, with the Korean sample the targeted goal was completed interviews with adults who said they were Korean when asked the race items in the interview. However, not all the persons with a Korean surname were actually Korean and not all the surname sampled telephone numbers were residential. The sample size was increased for these sources of loss[4], but the estimates were not based on any previous empirical evidence since the surname lists had not been used for the purpose previously.

[4] The sample size for Latinos in Shasta County could not be increased to the level desired because all of the listed persons with Hispanic surnames in the county were included in the sample.

Table 3-8 summarizes the sample sizes for each type of sample and for the time of sampling. The total sample size for CHIS 2001 was 365,308 telephone numbers. Some of the supplemental samples, like the Solano geographic supplemental sample, were included with the regular RDD sample as shown in the table. Other supplements were selected later. The sample of telephone numbers in San Francisco and Santa Barbara was augmented in August 2001 because the observed response rates in these two counties were lower than projected. This addition took place relatively late in the data collection period.

Table 3-8. Number of telephone numbers sampled by type of sample [a]

	Sample	July 2000	February 2001	May 2001	July 2001	August 2001	Total
1	RDD sample [b]	122,848	172,157				295,005
2	Pasadena City	1,431 [c]	4,080				5,511
3	Long Beach City	823 [c]	5,281				6,104
5	Berkeley City	298 [c]	4,719				5,017
6	Solano County	1,800	6,161				7,961
7	San Francisco County	2,700	5,388			12,241 [d]	20,329
8	Santa Barbara County	1,800	2,516			893 [d]	5,209
10	Cambodian			2,565			2,565
11	South Asian			3,670			3,670
12	Japanese			2,463			2,463
13	Korean			3,635			3,635
14	Vietnamese			2,983			2,983
15	Shasta Latinos				1,903		1,903
16	American Indian/ Alaska Native				2,953		2,953
Total		131,700	200,302	15,316	4,856	13,134	365,308

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

[a] Duplicate telephone numbers were removed. RDD numbers sampled in other subsamples were counted as RDD records. San Francisco supplemental records sampled in other subsamples were counted as San Francisco records.

[b] The sample sizes for Solano, Pasadena, Long Beach, Berkeley, San Francisco, and Santa Barbara are excluded from these counts. We typically refer to all of the numbers sampled in July 2000 and February 2001 as the RDD and geographic supplemental samples excluding the late supplements in San Francisco and Santa Barbara.

[c] Not sampled by separate strata.

[d] Sample in selected ZIP Codes.

3.4 Expected Design Effect

Section 3.2 described the allocation of the sample of telephone numbers by county or stratum and noted that it was a compromise between two goals. The CHIS 2001 sample was designed to produce reliable estimates for the entire state and for counties. If the sample were allocated proportional to the population in the counties, this would be approximately optimal for statewide estimates. For county estimates, an equal allocation would be more efficient. In this section, we describe the statistical methods used to examine the efficiency of the sample under different allocations. These methods were used to help guide the sample allocation for CHIS 2001.

If CHIS 2001 had been a simple random sample, then it would be relatively simple to predict the precision of the estimates. Under the assumption of simple random sampling, suppose we wish to estimate a proportion of adults with a characteristic, say p. If the sample size is large enough, then the standard $(1-\alpha)\cdot 100\%$ confidence interval of the estimated proportion is

$$\left(p - z_{1-a/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{1-a/2}\sqrt{\frac{p(1-p)}{n}} \right) \quad (1)$$

where $z_{1-a/2}$ is the critical value from the standard normal distribution, and n is the number of completed interviews. This form of the confidence interval is not appropriate for CHIS 2001 for several reasons. The main reason we discuss below is because the allocation of the sample to the counties does not produce a simple random sample. The other reasons that (1) is not fully appropriate are because of sampling within households and other adjustments to the estimation weights. These issues are covered in Report 5: Weighting and Variance Estimation.

To adjust (1) to account for the sample allocation to the counties or strata we introduce the concept of a design effect. Kish (1992) discusses the design effect in some detail. Here we simply note that in stratified designs like CHIS, the design effect measures the departures with respect to a sample proportionally allocated among the strata. A sample with proportional allocation has a design effect of unity. Departures from proportional allocation result in design effects greater than one.

The design effect due to departures from proportional allocation can be computed as

$$D = \left(\sum_{h=1}^{H} W_h k_h \right) \left(\sum_{h=1}^{H} \frac{W_h}{k_h} \right), \tag{2}$$

where W_h is the proportion of the population in sampling stratum h ($W_h = N_h \left(\sum N_h \right)^{-1}$) where N_h is the population total in stratum h, and k_h is the relative sampling rate for strata h. More specifically, k_h is defined as $k_h = \frac{n_h}{N_h} \frac{N_1}{n_1}$, where n_h is the sample size in stratum h and the reference stratum is set to be stratum 1 so that $k_1 \equiv 1$ (the choice of the reference stratum does not affect the computations since the relative rates are the only factor involved).

Using the design effect computed in this way, we can estimate the effective sample size for a stratified sample with a given allocation. The effective sample size is the number of cases needed from the stratified sample to produce estimates with the same precision that would be expected from a simple random sample design. The effective sample design is computed as

$$n_{eff} = \frac{n}{D} \tag{3}$$

where n_{eff} is the effective sample size and the other terms are as defined above.

In CHIS 2001, we expected to complete 52,805 adult interviews when we add together the expected RDD and geographic supplemental samples for Solano County and the cities of Long Beach, Pasadena, and Berkeley. The other race and ethnic supplemental samples were not included in this evaluation. The nominal sample sizes (the actual number of adult interviews expected), the expected design effects due to the sample allocation to the strata using (2), and the expected effective sample sizes using (3) are given in Table 3-9. The expected design effects and effective sample sizes are given for the entire state and for domains defined by race and ethnicity. It is important to remember that the design effects are computed at the household level and they do not include any adjustments for nonresponse, within a household sampling, or other weighting adjustments.

Table 3-9. Expected design effects and effective sample size associated with the sample allocation

	Domain	Nominal sample size	Design effect	Effective sample size
1	White	34,109	1.35	25,193
2	Latino	10,559	1.21	8,735
3	Asian American/Pacific Islander	4,340	1.40	3,096
4	African American	3,348	1.17	2,858
5	American Indian/Alaska Native	449	1.49	300
6	Total	52,805	1.29	40,795

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

For example, the expected sample yields from the CHIS 2001 sample for American Indian/Alaska Native was 449 adults (excluding the supplemental sample of American Indian/Alaska Natives). Due to the allocation of the sample, the effective sample size is 300. The 95 percent confidence interval for an estimated proportion can be computed by using the entries in this table and replacing n in (1) by n_{eff} . For example, for estimating a proportion of p = 0.5 for American Indian/Alaska Natives, the 95 percent confidence interval is

$$\left(0.5-1.96\sqrt{\frac{0.5^2}{300}},\ 0.5+1.96\sqrt{\frac{0.5^2}{300}},\right)=\left(0.44342\ ,\ 0.5565\right)$$

As the UCLA CHIS staff consulted with various groups in California to evaluate the data needs that CHIS could help to support, they developed different allocation schemes for distributing the sample to the counties. The effects of these allocations were examined by using the methods presented above. The UCLA CHIS staff then chose the sample allocation that best satisfied the needs of users of the survey.

4. WITHIN-HOUSEHOLD SAMPLING

Once the sample of telephone numbers is selected, interviewers call the numbers and conduct interviews with sampled persons within the household. This chapter describes the procedures for selecting the sample of persons within households for CHIS 2001. Person subsampling was done primarily to reduce respondent burden at the household level. Samples of adults, children, and adolescents within the household were selected using different sampling procedures, but one adult, and up to one child and adolescent were sampled within each household. The within household sampling procedures were developed to maximize the analytic utility of the data collected from the respondents. The next section describes the within-household sampling alternatives we evaluated to accomplish this and the reasons for choosing the specific method of sampling. The second section describes sampling adults within sampled households. The third section gives the sampling methodology used for sampling children and adolescents. The last section details the specifics of how the sampling was implemented in CHIS 2001 interviews.

4.1 Sampling Alternatives

The general idea for the CHIS 2001 sample design was to randomly sample one adult from all the adults in every sampled household. In addition, in those households with adolescents (ages 12-17) and/or children (under age 12), one adolescent and one child were to be sampled and interviewed (a parent of the child was interviewed about the child). One approach to accomplishing this goal is to simply list all the persons in the age group (adult, child, and adolescent) in the household and select one randomly person from each group. We call this the *completely random* sampling method.

The completely random sampling method is not a problem in most households because most households have only one family. However, in households with two or more families, the completely random method could result in selecting persons from the different age groups who were not members of the same family. This situation is undesirable because the adult interview collected data about the family of the sampled adult. The data from the adult are of great value for the analysis of the data from the child and adolescent interviews. If the sampled child and/or sampled adolescent were not members of the same family as the sampled adult, then the data collected about them would be of very limited utility.

To illustrate this type of household consider Figure 4-1. It shows the familial relationships in a household with two families (*F1* and *F2*). In the figure, family *F1* consists of 3 adults, (*AD1*, *AD2* and *AD3*) and one adolescent (*TN1*); *AD3* is a young adult (18 or older) child of *AD1* and *AD2*. A second family, *F2*, shares the same household but the members of *F2* are not related to the family *F1*. Family *F2* consists of one adult *AD4* and one adolescent *TN2*.

If one adult and one adolescent were selected using the completely random method, one possible outcome is the selection of adult *AD4* and adolescent *TN1*. In this case, the family data collected from the *AD4* would not be useful for describing the family circumstances of *TN2* because they are not members of the same family.

To resolve this analytic problem, a second sampling alternative was adopted for CHIS 2001. We call this method the *linked* sampling approach. In this approach, the children and adolescents in the household were linked to the adults. Children and/or adolescents for whom an adult (or the spouse of the adult) was a blood, adoptive, foster parent, or other legal guardian were considered as linked or "associated" with the adult.

In the linked sampling method persons are sampled in two phases. In the first phase, an adult is randomly sampled from all the adults in the household. In the second phase, a child and/or adolescent is sampled from all the children/adolescents associated with the sampled adult. In the example in Figure 4-1, if adult *AD4* is sampled, then the only adolescent eligible for sampling is *TN2* and that adolescent would be selected. Since the sampling of adolescents (and children) is a two-phase procedure, the probability of sampling the adolescent is the product of the probability of sampling the adult (phase one) and the probability of sampling the adolescent from the all the adolescents associated with that adult (phase two).

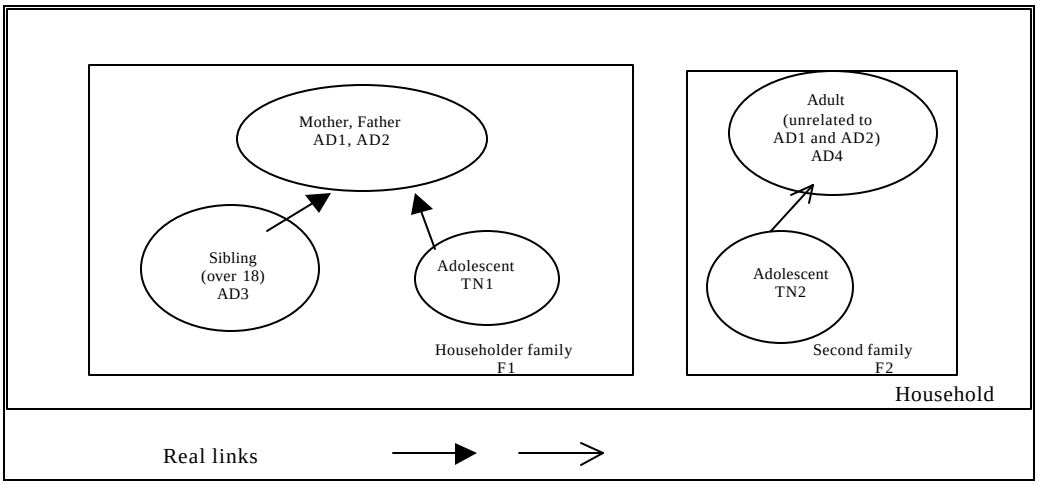

Figure 4-1. Illustrative household with two families

To use the linked sampling method, data are needed linking each child and adolescent to adults within the households. These data were collected in the adult interview. However, we expected that it would not be possible to link or associate a child or adolescent to an adult within a very few sampled households because of unusual household structures. If a child or adolescent were not associated with an adult, then the unassociated person would not have any chance of being selected. Even though the potential bias due to this unusual situation was expected to be small, a provision was made in the linked sampling method to remove any possibility of bias. If any child or adolescent was not associated with at least one adult using the usual relationships, then a link was formed to a randomly selected adult. The randomly linked child or adolescent was then eligible for sampling in the second phase if the adult with whom they were randomly associated was selected. Only 17 of 16,523 households with children had any randomly associated children; of the 10,867 households with at least one adolescent only 37 had at least one randomly linked adolescent.

Although the linked sampling method resolved the problem of sampling members of the same family, it did raise the possibility that no child or adolescent would be selected in households where they were present. Returning again to Figure 4-1 as an example, if *AD3* is sampled, then no adolescent interview is conducted in the household despite the fact there are two adolescents. None of the adolescents in the household are associated to *AD3*. This situation was allowed for two reasons. First, the method avoids the bias as mentioned above. Second, in many of these cases a sampled adult who was not associated with the child or adolescent could not be the source of reliable data required for the child or adolescent analysis. A provision in the sampling procedure was made to reduce the probability of sampling adults such as *AD3*. This method of sampling is discussed in the next section.

4.2 Sampling Adults

An adult is defined as any person 18 years or older residing in the household. One adult was randomly selected from the roster of adults created during the screener interview. If the age of a person on the roster was not known, the person was assumed to be an adult and was eligible for sampling as an adult with unknown age. The main reason only one adult was sampled in a household was to reduce the response burden on the household. Even with this restriction, up to three interviews could be conducted in the same household.

In most households adults were selected randomly from the list of all enumerated adults within the household. For example, in a household with 2 adults one adult was sampled and each adult had a 50 percent chance of being the sampled adult. Because the linked sampling method was used, the sampling plan was modified to reduce the probability of sampling an adult who was not associated with a child or adolescent. The procedure involved sampling adults with different probabilities of selection in households with both adults under 24 years old and adults 40 years or older (and no adults with unknown age). In these households, adults aged 40 or older were given a chance of selection twice as large as the chance of selection for adults younger than 40. Increasing the probability of selecting adults aged 40 or older in these types of households reduced the chance of selecting adult children. For example, suppose there are three adults in the household, two 55 years old and one 19 years old. Under the unequal probability sampling method, the 19 year old is sampled 20 percent of the time rather than 33 percent that would have occurred under the equal probability sampling scheme.

Table 4-1 shows the number of the adults who were sampled with equal and unequal probabilities of selection within a household. The counts are for all sampled households that completed the screener interview in the RDD and geographic supplemental samples. The unequal probability scheme was used in only about 12 percent of the total sampled households.

Table 4-1. Number adults sampled by method of adult sampling

Type of adult sampling	Presence of children in household	Presence of adolescents in household	Total*
Equal probability	No**	No	46,074
	No	Yes	5,516
	Yes	No	15,933
	Yes	Yes	6,795
Total			74,318
Differential probability	No	No	4,418
	No	Yes	2,497
	Yes	No	1,418
	Yes	Yes	1,400
Total			9,733
Grand Total			84,051

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

* Includes RDD sample and geographic supplemental samples

** "No" includes unknown presence of children or adolescents in the household

4.3 Child and Adolescent Sampling

The sampling for children and adolescents took place after the adult was sampled and completed the enumeration of all persons under 18 years old in Section H of the adult extended interview. If there were any children under 12 in the household that were associated with the sampled adult, then exactly one child was sampled and each associated child had an equal probability of selection. The same procedure was followed for sampling exactly one adolescent with equal probability from all the adolescents associated with the sampled adult.

As described above, if there were children or adolescents in the household that were not associated with the sampled adult, they were not eligible to be selected in this second phase of sampling. Consequently, some households with a child or adolescent had no child or adolescent sampled. To evaluate the percent of households in which no child or adolescent was sampled, we tabulated the results of the sampling procedures for children in Table 4-2 and for adolescents in Table 4-3. Table 4-2 shows that in only 1,279 (774+505) households had children present but none were sampled. Thus, the situation arose in 2.3 percent of all households. Similarly, Table 4-3 shows that in 1,088 (501+587) households or 2.0 percent of all households had adolescents present in the household but none was sampled.

Table 4-2. Percentage of households by method of adult sampling, presence of child, and child sampling

Type of adult sampling	Child present	Child sampled	Total	Percent
Equal probability	No	No	35,651	64.1%
	Yes	Yes	13,380	24.0%
	Yes	No	774	1.4%
Total			49,805	89.5%
Differential probability	No	No	4,385	7.9%
	Yes	Yes	958	1.7%
	Yes	No	505	0.9%
Total			5,848	10.5%
Total both types			55,653	100.0%

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

Note: The counts are for the RDD and geographic supplemental samples

Table 4-3. Percentage of households by method of adult sampling, presence of adolescent, and adolescent sampling

Type of adult sampling	Adolescent present	Adolescent sampled	Total	Percent
Equal probability	No	No	41,766	75.0%
	Yes	Yes	7,538	13.5%
	Yes	No	501	0.9%
Total			49,805	89.5%
Differential probability	No	No	3,554	6.4%
	Yes	Yes	1,707	3.1%
	Yes	No	587	1.1%
Total			5,848	10.5%
Total both types			55,653	100.0%

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

Note: The counts are for the RDD and geographic supplemental samples

Before the within-household sampling plan was accepted, we evaluated the consequences of using the linked sampling approach using data from the March 1999 Current Population Survey (CPS) file. This analysis was done using only the records for California. Based on the CPS data, we computed an expected percentage of households in which a child or adolescent would be sampled. Table 4-4 shows this expected percentage and the percentage actually observed in CHIS 2001. The observed percentages are very close to the expected values. Note that the percentages in Table 4-4 refer to the number of sampled children or adolescents, not to the number of child or adolescent extended interviews that were completed.

Table 4-4. Expected and observed percentage of households with a child or adolescent

Type	Expected (CPS)	Observed
Households with children	90.4%	91.81%
Households with adolescents	88.6%	89.5%

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

Note: The counts are for the RDD and geographic supplemental samples

4.4 Enumeration, Assignment, and Sampling Procedures

As described in the previous section, the sampling of persons in CHIS 2001 was conducted in two phases, with adult sampling in the first phase and child and adolescent sampling in the second phase. The adult was sampled during the screener interview and the child and adolescent were sampled after the persons under 18 years old were enumerated at the end of Section H in the adult extended interview. We begin by giving the specific sampling procedures used and then conclude the section by discussing the overall probability of selection for each sampled person.

The following steps are the details of the selection process used in CHIS 2001.

1. The respondent enumerates all adults (18 years or older) in the household screener.

2. An adult respondent (AR) is randomly selected from the roster of all adults in the household. All adults in the household have equal probabilities of selection, unless there is at least one adult 18 to 24 years old **and** there is at least one adult over 40 years old. In such households, adults 40 or older are assigned a measure of size so that they have twice the chance of being sampled as those under 40.

3. As part of the adult extended interview with the AR, the adult identifies if they have a spouse or partner (ARSP) living in the household. They also enumerate all children (ages 0 through 11) and adolescents (ages 12 through 17) in the household.

4. The AR is asked in the adult interview if either the AR or the ARSP is the parent or guardian for each adolescent and child. Adolescents and children for whom the AR or ARSP is the parent/guardian are associated with both the AR and the ARSP. This step completes the linking process if all the children and adolescents are associated with the AR or the ARSP.

5. If there are any children or adolescents not associated with the AR or ARSP, then the AR is asked to identify an adult (including AR) responsible for each remaining child and adolescent. Adolescents and children for whom there is a responsible adult in the household are associated to that adult and that adult only.

6. Adolescents who are still not associated with an adult after Steps 4 and 5 are associated to an adult in the household by using a randomly generated number in the computer system.

7. If there are adolescents in the household above age 14 and remaining unassociated children, the AR is asked for each unassociated child if the child has an adolescent parent in the household. If the unassociated child has an adolescent parent in the household, the child is associated with the same adult(s) as the adolescent parent. This procedure was added to capture children of children who might otherwise not be eligible for sampling (since they are not associated with an adult in the household). The procedure also increases the utility of the data because in many cases data are collected on the parent of the adolescent, the adolescent, and the child of the adolescent.

8. Children who remain unassociated after Step 7 are randomly associated to an adult in the household. At this point all children and adolescents must be associated with one or more adults in the household.

9. If any adolescents are associated with the AR, then exactly one of these associated adolescents is randomly selected. Each associated adolescent has an equal probability of selection.

10. If any children are associated with the AR, then exactly one of these associated children is randomly selected. Each associated child has an equal probability of selection.

The last step is to compute the probability of selection for each sampled person. If an adult is selected with equal probability, then the probability of selection is just the inverse of the number of adults in a household. If the unequal probability of selection method of sampling adults is used, then the probability of sampling the adult is the adult measure of size (1 or 2 depending on household composition) divided by the sum of the measures of size for all adults in the household.

Since children and adolescents are sampled in two phases, the probability of selection for a child or adolescent is the product of the probability of selection of the adult and the conditional probability that the child or adolescent is selected given that the associated adult is selected. If the child or adolescent is associated with two adults (the AR and the ARSP), the probability of selection is the sum of the probabilities calculated in this way for each adult. In other words, you would compute the probability of sampling the person through the AR and add to that the probability of sampling the person through the ARSP.

For example, consider the following hypothetical situation. A married couple has one child of their own (assigned to both the AR and the ARSP in Step 4) and there is one child who is not related to the couple but is the child of a friend of the AR. This child is associated with the AR (but not to the spouse of the AR) in Step 5. The within-household probability of sampling the child of the AR is 0.75. This is the sum of the probability of selecting the child via the AR (0.5 * 0.5) plus the probability of sampling the child via the ARSP (.5 * 1). The within-household probability of sampling the child of the AR's friend is 0.25, since the only way this child can be sampled is via the AR (0.5 * 0.5).

These probabilities are also discussed in Report 5: Weighting and Variance Estimation. In that report, the inverse of the probability of selection is the initial weight for the adults, children, and adolescents.

5. ACHIEVED SAMPLE SIZES

This chapter summarizes the number of completed interviews in CHIS 2001 for the reporting areas and the relationship between the targeted and the achieved numbers. As mentioned in the previous chapters, the targeted goals for CHIS 2001 were stated in terms of the total number of completed adult interviews obtained at the end of the data collection period. The actual number of completed interviews is a function of the number of telephones sampled, the within-household person sampling, and estimates of different reasons for attrition. These reasons are discussed in more detail in Chapter 3. Detailed information about the response rates is presented in Report 4: Response Rates.

5.1 Comparison to Goals

Table 5-1 gives the number of completed adult interviews by two methods of classifying the geographic areas in which the sampled adults reside. The first column of completed interviews in the table uses the data on the county that was available at the time of sampling. As noted in Chapter 3 on sampling households, each telephone number is assigned to exactly one stratum for sampling purposes, but the adult may actually live in a different county. The third column in the table uses the self-reported county of the adult respondent. This classification is based on the county and ZIP Code data collected in the adult interview. It is the classification that is most appropriate for analysis of the CHIS 2001 data. Report 3: Data Processing Procedures describes how the self-reported data were processed and how reporting discrepancies were resolved.

The table also gives these completes as percentages of the targeted number of adult interviews set at the time of the design. The targeted goals by county for the RDD sample are given in Table 3-1 and the targeted goals for the geographic supplemental samples are given in Table 3-2. Since Table 5-1 is based on the RDD and geographic supplemental samples (excluding the supplemental samples in San Francisco and Santa Barbara), the overall targets are just the sum of the numbers in those two tables. A 100 percent indicates the targeted number of adult interviews was achieved in the county.

Table 5-1 shows that CHIS 2001 came very close to the targeted goals with all but 3 of the 47 rows of the table exceeding 95 percent of the target goal. The achieved sample sizes were on target

Table 5-1. Number of completed adult interviews by sampling and self reported strata

Area	Sampling		Reported	
	Completed interviews	Percent of target	Completed interviews	Percent of target
State Total	54,122	100.8	54,122	100.8
Los Angeles	12,215	100.7	12,196	100.5
Long Beach	819	102.4	913	114.1
Pasadena	814	101.8	671	83.9
Remainder of Los Angeles	10,582	100.5	10,612	100.8
San Diego	2,666	100.2	2,672	100.5
Orange	2,495	98.2	2,454	96.6
Santa Clara	1,514	97.4	1,508	97.0
San Bernardino	1,547	101.3	1,554	101.8
Riverside	1,386	102.0	1,391	102.4
Alameda	1,985	102.6	1,974	102.1
Berkeley	794	99.3	809	101.1
Remainder of Alameda	1,191	105.0	1,165	102.7
Sacramento	1,238	103.2	1,230	102.5
Contra Costa	1,199	99.9	1,214	101.2
Fresno	1,041	104.1	1,053	105.3
San Francisco	893	89.3	886	88.6
Ventura	971	97.1	1,015	101.5
San Mateo	925	92.5	945	94.5
Kern	1,096	109.6	1,093	109.3
San Joaquin	1,052	105.2	1,058	105.8
Sonoma	771	96.4	776	97.0
Stanislaus	819	102.4	794	99.3
Santa Barbara	798	99.8	795	99.4
Solano	1,587	99.2	1,553	97.1
Tulare	827	103.4	826	103.3
Santa Cruz	793	99.1	791	98.9
Marin	750	93.8	752	94.0
San Luis Obispo	799	99.9	807	100.9
Placer	784	98.0	764	95.5
Merced	832	104.0	849	106.1
Butte	825	103.1	835	104.4
Shasta	826	103.3	827	103.4
Yolo	834	104.3	844	105.5
El Dorado	780	97.5	807	100.9
Imperial	798	99.8	794	99.3

Table 5-1. Number of completed adult interviews by sampling and self reported strata (continued)

Area	Sampling		Reported	
	Completed interviews	Percent of target	Completed interviews	Percent of target
Napa	806	100.8	833	104.1
Kings	843	105.4	837	104.6
Madera	824	103.0	820	102.5
Monterey, San Benito	790	98.8	794	99.3
Del Norte, Humboldt	861	107.6	855	106.9
Lassen, Modoc, Siskiyou, Trinity	846	105.8	841	105.1
Lake, Mendocino	813	101.6	808	101.0
Colusa, Glen, Tehama	839	104.9	839	104.9
Sutter, Yuba	822	102.8	801	100.1
Plumas, Nevada, Sierra	814	101.8	824	103.0
Alpine, Amador, Calaveras, Inyo, Mariposa, Mono, Tuolumne	818	102.3	813	101.6

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

using both the sampling strata classification (the one that was known during data collection) and the final self-reported classification (the one that will be used for analysis). The discrepancies between the two classifications are greatest for the city of Pasadena. The differences are largely a function of how well the sampling classification matched with the self-reported classification. For smaller geographic areas, the sampling classification tends to be less precise but this varies by specific location. These types of differences are discussed in more detail in Report 3: Data Processing Procedures.

Table 5-2 shows the number of completed child and adolescent interviews for the RDD sample and geographic supplemental samples (again excluding the supplemental samples in San Francisco and Santa Barbara). For these interviews, the targets were set overall rather than by county. The self-reported area is used in this table. The CHIS 2001 targeted goals were between 12,000 and 13,000 completed child interviews in the state and between 4,000 and 5,000 completed adolescent interviews in the state. In both cases, the achieved samples for children and adolescents were very close to the expected numbers.

Table 5-2. Number of completed child and adolescent completed interviews by self-reported areas

Areas	Completed child interviews	Completed adolescent interviews
State Total	12,392	5,733
Los Angeles	2,820	1,121
Long Beach	231	65
Pasadena	129	36
Remainder of Los Angeles	2,460	1,020
San Diego	585	270
Orange	604	212
Santa Clara	354	140
San Bernardino	440	211
Riverside	378	157
Alameda	362	141
Berkeley	97	36
Remainder of Alameda	265	105
Sacramento	295	135
Contra Costa	258	121
Fresno	270	136
San Francisco	124	35
Ventura	242	112
San Mateo	153	77
Kern	322	147
San Joaquin	282	135
Sonoma	161	90
Stanislaus	191	92
Santa Barbara	174	71
Solano	400	174
Tulare	226	110
Santa Cruz	173	103
Marin	134	67
San Luis Obispo	152	68
Placer	177	84
Merced	236	117
Butte	170	71
Shasta	165	87
Yolo	203	94
El Dorado	176	106
Imperial	226	154
Napa	166	85
Kings	275	151
Madera	183	104
Monterey, San Benito	210	101
Del Norte, Humboldt	168	109

Table 5-2. Number of completed child and adolescent completed interviews by self-reported areas (continued)

Areas	Completed child interviews	Completed adolescent interviews
Lassen, Modoc, Siskiyou, Trinity	152	82
Lake, Mendocino	140	82
Colusa, Glen, Tehama	196	108
Sutter, Yuba	174	81
Plumas, Nevada, Sierra	148	92
Alpine, Amador, Calaveras, Inyo, Mariposa, Mono, Tuolumne	127	100

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

Table 5-3 shows the number of completed adult, child and adolescent interviews for the supplemental samples. The second column shows the revised target of the number of completed adult interviews. The supplemental sample targets were revised during the data collection period as experience was gained on the actual RDD sample yield. In contrast to the RDD sample, the targets were set overall rather than by county. As with the RDD sample, the targets were generally achieved. The only supplemental sample size that is far from the target is the sampling of Latinos in Shasta county. The reason for this shortfall is that all Latino surnames in the county were sampled and there was no way to increase the sample size for this group without major changes in the sampling scheme.

Table 5-3. Number of completed adult, child, and adolescent by supplemental sample

Supplemental sample	Revised Target	Adult	Percentage of target	Child	Adolescent
Cambodian	130	126	96.9	44	37
South Asian	426	443	104.0	158	39
Japanese	325	330	101.5	51	18
Korean	322	326	101.2	95	30
Vietnamese	503	540	107.4	124	34
American Indian/Alaska Native Urban	256	251	98.0	69	33
American Indian/Alaska Native Rural	100	100	100.0	37	18
San Francisco	1,100	1,100	100.0	151	46
Santa Barbara	200	206	103.0	49	22
Shasta Latinos	378	304	80.4	106	48

Source: UCLA Center for Health Policy Research, 2001 California Health Interview Survey.

The tables confirm that the sampling procedures achieved the goals for both the main RDD sample and the supplemental samples of CHIS 2001.

REFERENCES

Anderson, J.E., Nelson, D.E., and Wilson, R.W. (1998). Telephone coverage and measurement of health risk indicators: Data from the National Health Interview Survey. *American Journal of Public Health*, 88, 1392-1395.

Brick, J.M., and Waksberg, J. (1991). Avoiding sequential sampling with RDD. *Survey Methodology*, 17(1), 27-41.

Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in list-assisted telephone surveys. *Public Opinion Quarterly*, 59(2) 218-235.

Casady, R., and Lepkowki, J. (1993). Stratified telephone survey designs. *Survey Methodology*, 19, 103-113.

Ford, E.S. (1998). Characteristics of survey participants with and without a telephone: Findings from the Third National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, 51, 55-60.

Giesbrecht, L.H., Kulp, D.W., and Starer, A.W. (1996). "Estimating coverage bias in RDD samples with Current Population Survey Data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 503-508.

Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, A, 149, 65-82.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8, 183-200.

Sudman, S., Sirken, M.G., and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.

The California Health Interview Survey (CHIS) is a collaboration of:

UCLA Center for
Health Policy Research

California Department
of Health Services

Public Health
Institute

California Health Interview Survey (CHIS)
UCLA Center for Health Policy Research
10911 Weyburn Avenue, Suite 300
Los Angeles, California 90024-2887
Phone 310-794-0925
Toll Free 1-866-275-2447
Fax 310-794-2686
chis@ucla.edu
www.chis.ucla.edu