Health Information National Trends Survey 5 (HINTS 5)

Cycle 4 Methodology Report

December 2020

Prepared for
National Cancer Institute
9609 Medical Center Drive
Bethesda, MD 20892-9760

Prepared by
Westat
1600 Research Boulevard
Rockville, MD 20850

This page is deliberately blank.

Table of Contents

Tables **Page**

Appendixes

Cycle 4 Overview 1

The Health Information National Trends Survey (HINTS) is a nationally-representative survey which has been administered every few years by the National Cancer Institute (NCI) since 2003. The HINTS target population is civilian, non-institutionalized adults aged 18 or older living in the United States. The most recent version of HINTS administration (referred to as HINTS 5) included four data collection cycles over the course of four years. The final round of data collection for HINTS 5 (Cycle 4) was conducted from February 24 – June 15, 2020 with a goal of obtaining 3,500 completed questionnaires. A total of 3,865 completed surveys were collected with a response rate of 37%. Unlike Cycle 3, this cycle of HINTS did not include a web option and was conducted completely by mail. This report summarizes the methodology, sampling, data collection protocols, weighting procedures, and response rates for Cycle 4.

Content Focus

HINTS provides NCI with a comprehensive assessment of the American public's access to and use of information about cancer across the cancer care continuum from cancer prevention, early detection, diagnosis, treatment, and survivorship. The content of each HINTS 5 data collection cycle focuses on understanding the degree to which members of the general population understand vital cancer prevention messages. In addition to this standard HINTS content, each round of HINTS 5 has specific topical content for trending in areas of recent developments in the communication environment. For Cycle 4, the topics of special interest focused on genetic testing and clinical trials. Specific content included:

- Genetic testing
  - Knowledge about genetic testing
  - Interpretation and sharing of genetic testing results
  - The use of genetic testing to identify and treat cancer
- Clinical trials
  - Knowledge about clinical trials
  - Sources and trust in information about clinical trials

- Willingness to participate in clinical trials
- History of clinical trial participation for cancer

Impact of COVID-19

HINTS 5 Cycle 4 entered the field on February 24, 2020. While the first mailing and the reminder postcard were sent out on schedule and without any issues, the World Health Organization announcement on March 11 of the COVID-19 pandemic impacted the rest of the Cycle 4 field period. Restrictions from the state of Maryland (where Westat is headquartered) meant that the labor required for sending out survey packets was reduced and the mailing schedule was slightly altered from the original plan, with longer lag times between mailings. Regardless of this impact on the schedule, Cycle 4 continued to be fielded, incoming telephone questions from survey recipients were responded to, and completed questionnaires were processed, albeit at a slower-than-normal pace.

In addition to COVID-19's impact on the mailing schedule, surveys received in the later part of the field period sometimes included COVID-related responses. For example, a number of respondents mentioned COVID when responding to the employment question. To facilitate the comparison of surveys returned early in the field period and late in the field period for possible COVID effects, a specific pandemic return variable was developed. This variable is discussed in more detail in section 4.6 of this report.

Response Rate Calculation Changes

The response rate for Cycle 4 was calculated using a different formula than has traditionally been used for HINTS. The new formula (the American Association of Public Opinion Research formula RR4) is adjusted based on an estimate of the eligibility rate. The details about this change, the formula involved, and the reasoning behind the change, are described in section 6 of this report.

Sample Selection 2

The sampling strategy for the Cycle 4 survey consisted of a two-stage design. In the first stage, a stratified sample of addresses was selected from a file of residential addresses. In the second stage, one adult was selected within each sampled household.

2.1 Sampling Frame

As with prior HINTS iterations, the sampling frame for Cycle 4 consisted of a database of addresses used by Marketing Systems Group (MSG) to provide random samples of addresses. All non-vacant residential addresses in the United States present on the MSG database, including post office (P.O.) boxes, throwbacks (i.e., street addresses for which mail is redirected by the United States Postal Service to a specified P.O. box), and seasonal addresses were subject to sampling.

Rarely are surveys conducted with a sampling frame that perfectly represents the target population. The sampling frame is one of the many sources of error in the survey process. In previous cycles, the sampling frame used for the address sample contained duplicate units because some households receive mail in more than one way and a question about how many different ways respondents receive mail was included on the survey instrument to permit an adjustment for the duplication of households in the sampling frame. However, because this is a rare occurrence, starting in Cycle 3 this question was removed from the questionnaire and the adjustment for duplication was no longer implemented. Cycle 4 followed the procedure established in Cycle 3.

An additional change to the traditional HINTS sampling design was implemented starting with Cycle 3 and continuing in Cycle 4. Certain types of PO Box addresses were excluded from the address frame in an attempt to improve the rate of mail that is able to be delivered. There are two types of PO Box addresses: one type pertains to those that that are linked to city-style addresses and the other type is not. Those that are linked can get mail two ways: by the PO Box address and the city-style address. Those that are not linked can only get mail by the PO Box address. The Cycle 4 sample was limited to PO Box addresses classified as the only-way-to-get mail. Because PO Box addresses tend to have high undeliverable rates, having a smaller number in the sample should result in a lower rate of undeliverable packets compared to past cycles.

In rural areas, some of the addresses do not contain street addresses or box numbers. Simplified addresses contain insufficient information for mailing questionnaires. Consequently, alternative sources of usable addresses were used when a carrier route contained simplified addresses. This partially ameliorated the frame's known undercoverage of rural areas although the actual coverage and undeliverable rates for this portion of the frame is not known.

2.2 Stratification

The sampling frame of addresses was grouped into two explicit sampling strata:

1. Addresses in areas with high concentrations of minority population; and
2. Addresses in areas with low concentrations of minority population.

The high and low minority strata were formed using the census tract-level characteristics from the 2014–2018 American Community Survey data file. Addresses in census tracts that had a population proportion of Hispanics or African Americans that equaled or exceeded 34 percent were assigned to the high-minority stratum. All the remaining addresses were assigned to the low-minority stratum.

The purpose of creating high- and low-minority strata and then oversampling the high-minority stratum is to increase the precision of estimates for minority subpopulations. The gains in precision stem from the increase in sample sizes for the minority subpopulations produced by the oversampling.

2.3 Selection of Address Sample

An equal-probability sample of addresses was selected from within each explicit sampling stratum. The total number of addresses selected for Cycle 4 was 15,350: 11,050 from the high minority stratum and 4,300 from the low minority stratum. The high-minority stratum's proportion of the sampling frame was 26.5 percent and it was oversampled so that its proportion of the sample was 72.0 percent. Conversely, the low minority stratum comprised 73.5 percent of the sampling frame but made up just 28.0 percent of the sample.

Table 2-1 below summarizes the address sample, showing the number of sample addresses, the percent of addresses in the frame and sample, and the percent oversampled/under-sampled relative to a proportional design, by sampling stratum. As part of the deduplication process, Westat checked whether there were households that were selected from Cycles 1, 2, or 3 and determined that there were three such records. To avoid overburdening these households with another HINTS survey in a relatively short time frame, these three households were excluded from data collection, resulting in a final sample size of 15,347.

Table 2-1. Summary by sampling stratum

Stratum	Number of sample addresses	Percent of addresses in the frame	Percent of sample addresses	Percent of sampled addresses oversampled (+) or undersampled (-)
High minority areas	11,050	26.5	72.0	+171.7%
Low minority areas	4,300	73.5	28.0	-61.9%
Total Sampled	15,350			
Deduplication	3			
Sample for Mailing	15,347			

2.4 Within-Household Sample Selection

The second-stage of sampling consisted of selecting one adult within each sampled household. In keeping with previous cycles of HINTS, data collection for Cycle 4 implemented the Next Birthday Method to randomly select the one adult in the household. The within-household selection was conducted by the respondents themselves. Questions were included on the survey instrument to assist the household in selecting the adult in the household having the next birthday (see Page 1 of the survey instrument).

Data Collection 3

Data collection for Cycle 4 started on February 24, 2020 and concluded on June 15, 2020. The survey was conducted exclusively by mail with a $2 pre-paid monetary incentive to encourage participation. The specific mailing procedures and outcomes the data collection effort are described in detail below.

3.1 Mailing Protocol

The mailing protocol for Cycle 4 followed a modified Dillman approach (Dillman, et al., 2009) with a total of four mailings: an initial mailing, a reminder postcard, and two follow-up mailings. All households received the first mailing and reminder postcard, while only non-responding households received the subsequent survey mailings. The second survey mailing was sent via USPS Priority Mail, while all other mailings were sent First Class. All households received one English survey per mailing unless someone from the household contacted Westat to request a Spanish survey, in which case the household received one Spanish survey per mailing for all subsequent mailings.

The contents of the mailings are further described in Table 3-1. The English cover letters and reminder postcard can be found in **Appendix A** and the Spanish cover letters are in **Appendix B**. All cover letters include a list of Frequently Asked Questions (FAQs) on the back. The FAQs in both English and Spanish are in **Appendix C.**

Due to the emerging coronavirus pandemic (COVID-19) in March 2020, a decision was made to be flexible about the mailing date for the second packet. To accommodate the reduction in staff at Westat's offices, the packets prepared for the second mailing went out in smaller batches and on three separate dates, as indicated in Table 3-1.

Table 3-1. Mailing protocol

Mailing	Date(s) mailed	Mailing method	Cycle 3 Materials
Mailing 1	February 24, 2020	1st Class Mail	• English cover letter with FAQs • English Questionnaire • Postage-paid return envelope • $2 bill
Postcard	March 2, 2020	1st Class Mail	Reminder/thank you postcard
Mailing 2	March 20, 2020 March 23, 2020 March 26, 2020	USPS Priority Mail	• English cover letter with FAQs • English questionnaire • Postage-paid return envelope OR (upon request) • Spanish cover letter with FAQs • Spanish questionnaire • Postage-paid return envelope
Mailing 3	May 6, 2020	1st Class Mail	• English cover letter with FAQs • English questionnaire • Postage-paid return envelope OR (upon request) • Spanish cover letter with FAQs • Spanish questionnaire • Postage-paid return envelope

The number of packets sent per mailing is outlined in Table 3-2. Households who sent in completed questionnaires were removed from further mailings. In addition, households with packets that were returned by the Postal Service as undeliverable were removed from any further mailings.

Table 3-2. Number of packets per mailing

Mailing	English	Spanish	Total
Mailing 1	15,347	N/A	15,347
Mailing 2	12,896	20	12,916
Mailing 3	10,789	21	10,810
Total	39,032	41	39,073

3.2 In-bound Telephone Calls

Two toll-free telephone numbers were provided to all respondents: one was used for English calls and one was used for Spanish calls. Both numbers were provided in each mailing. Respondents were told that they could call the number if they had questions, concerns, or if they needed to request materials in Spanish. Each number had a HINTS-specific voicemail message that instructed callers to leave their contact information and the reason for the call and then a study staff member would return their call. The Spanish line was staffed by a native Spanish speaker. When voicemails were received, they were logged into the Study Management System (SMS) and the request was either processed (such as recording their desire for a Spanish questionnaire) or the respondent was called back to ascertain the respondent's need if it was not clear from the message. Callers stating they did not want to participate in the study were coded as "refusal" and removed from any subsequent mailings.

The two toll-free lines together received 94 calls throughout the Cycle 4 field period (see Table 3-3 below). The majority of the in-bound calls were respondents requesting Spanish materials. The rest were respondents calling in with some form of comment or question or refusals. Four calls could not be resolved because they were either hang-ups or non-informative messages and study staff were not able to reach the respondents.

Table 3-3. Telephone calls received

Reason for call	Number of calls received
Request for a Spanish questionnaire	71
Refusal	7
Respondent let the study team know that the survey had been completed	3
Respondent asked a question or made a comment. Topics included: • whether HINTS was a legitimate study • who was sponsoring the study • whether participation was required • they requested more information about the study • whether the survey could be completed online • they made a mistake while filling out the survey and wanted to discuss their options • they misplaced the survey and requested another survey to be mailed to them • they let us know that they would complete the survey as soon as they could get family members to help them complete it	9
Calls that were never resolved due to hang ups or non-informative messages	4
Total	94

3.4 Incoming Questionnaires

Field room staff receipted all returned questionnaires into the SMS using each questionnaire's unique barcode. The SMS tracked each received questionnaire as well as the status of each household. Once a household was recorded as complete, it no longer received any additional mailings. Packages that came back as undeliverable were marked as such in the SMS and those addresses did not receive any further mailings.

In addition to refusing by calling the toll-free line, some respondents also refused by sending a letter stating that they did not wish to participate or asking to be removed from the mailing list. These households were marked in the system as refusals and were removed from subsequent mailings. Respondents who sent back a blank questionnaire were not considered refusals and continued to receive mailings.

The status of all Cycle 4 households at the end of data collection (but before cleaning and editing) can be found in Table 3-4.

Table 3-4. Household status of Cycle 4 at close of data collection

Household status	Total in Cycle 4	
	N	%
Complete	3,890	25.35
Refusal	36	0.23
Undeliverable	1,209	7.88
Nonresponse	10,212	66.54
Total	15,347	100.0

The number of questionnaires returned by date during the field periods for Cycle 4 can be found in Table 3-5. The majority of Cycle 4 returns were early in the field period, with 52 percent of returns coming in after the first mailing of the survey and the mailing of the reminder postcard. The second mailing resulted in an additional 37 percent and the remaining 11 percent were in response to the final mailing.

Table 3-5. Cycle 4 survey response by date

Date of mailing	Period of returns	Number of returns
Mailing 1: February 24	February 25- March 4	518
Postcard: March 2	March 5- March 22	1,506
Mailing 2: March 20, 23, and 26	March 23- May 8	1,440
Mailing 3: May 6	May 9- June 15	426
	Total	3,890

Data Management 4

After being processed and receipted into the SMS, each returned paper questionnaire was scanned, and verified, cleaned, and edited. Imputation procedures were also conducted. These procedures are described below.

4.1 Scanning

All completed paper questionnaires were scanned using a data capture software (TeleForm) to capture the survey data and images were stored in SharePoint. Staff reviewed each form as it was prepared for scanning. The review included:

- Determining if the form was not scannable for any reason, such as being damaged in the mail. Some questionnaires or individual responses needed to be overwritten with a pen that was readable by the data capture software. Numeric response boxes were pre-edited to interpret and clarify non-numeric responses and responses written outside the capture area.
- Reviewing potential problem questionnaires or pertinent comments made by respondents. Comments in Spanish were reviewed by a Spanish-speaking staff member.

The reviewed paper surveys were then sent through the high-speed scanner to capture the responses. TeleForm read the form image files and extracted data according to HINTS 5 Cycle 4 rules established prior to the field period. Scanned data were then subject to validation according to HINTS specifications. If a data value violated validation rules (such as marking more than one choice box in a mark-only-one question) the data item was flagged for review by verifiers who looked at the images and the corresponding extracted data and resolved any discrepancies. Spanish forms were verified by a Spanish-speaking staff member.

Decisions made about data issues as a result of scanning were recorded in a data decision log. The decision log contains the respondent ID, the value triggering the edit, the updated value, and the reason for the update. A total of 50 entries were made into the data decision log during the course of data scanning and processing. The majority of these were attributed to multiple response options selected on a gate question. Additional entries detail the decisions made about numeric entries outside variable parameters (i.e., 2-digit numbers written on single digit question).

A 10 percent quality control check was then conducted on the scanned data and the electronic images of the survey. Quality Assurance (QA) staff compared the hard copy questionnaire to the data captured in the database item-for-item and the images stored in the repository page-for-page to ensure that all items were correctly captured. If needed, updates were made. In addition, QA staff closely reviewed frequencies and cross tabulations of the HINTS raw data to identify outliers and open ended items to be verified. ID reconciliation across the database, images, and the SMS, was completed to confirm data integrity.

4.2 Data Cleaning and Editing

Once the paper questionnaires had been scanned, all survey data were cleaned and edited. General cleaning and editing activities are described briefly below, with more detailed information found in **Appendix D** (Variable Values and Data Editing Procedures).

- Customized range and logical inconsistency edits, following predetermined processing rules to ensure data integrity, were developed and applied against the data.
- Edit rules were created to identify and recode nonresponse or indeterminate responses.
- Missing values were recoded for some responses to questions that featured a forced-choice response format and for filter questions where responses to later questions suggested a particular response was appropriate.
- Derived variables were created to reflect each response recorded for certain “mark-one” type questions (A5, G3, G4, H11, and P11), in order to facilitate the imputation process implemented when respondents did not follow the instruction to mark only one response. For these variables, imputation, as described in Section 4.3, was carried out.

For other "mark-one" type questions where respondents marked multiple responses, editing rules were used to determine which response was retained.

- Variables were designed to summarize the responses for the electronic device, caregiving (who and what conditions), genetic testing (which they've heard of, where they heard about those tests, whether they've had any, with whom they've shared their results, and who helped with understanding test results), Federal tobacco message exposure, cancer, occupation, Hispanic ethnicity, and race questions. These variables, HaveDevice_Cat, CaregivingWho_Cat, CaregivingCond_Cat, HeardGenTest_Cat, TestSource_Cat, HadTest2_Cat, SharedRes2_Cat, UndGenTest_Cat, TobaccoMessages_Cat, Cancer_Cat, Occupation_Cat, Hisp_Cat, and Race_Cat2 indicated each response selected for respondents selecting only one response, and a multiple category for all of the respondents who answered multiple responses.

- A new derived variable was developed to summarize the responses to P4 and P5, the new "Occupation" questions. The derived variable is designed to determine what a respondent's full time occupation is, especially in cases where there may be more than one response option chosen in P5, a "mark all that apply" style question.

- Data cleaning was carried out for the two height variables: Height_Feet and Height_Inches. The rules that were applied minimized the number of out-of-range values by accounting for response measurements in incorrect boxes, responses using metric measures, responses using only one unit of measurement and other response errors.

- "Other, specify" responses were examined, cleaned for spelling errors, categorized, and upcoded into preexisting response codes when applicable. On one of these questions (E3), some of the responses were especially difficult to categorize because they could potentially have been upcoded into multiple categories. In those instances, the response was left as entered in the "Other, specify" field.

4.3 Imputation

The questions for which respondents selected more than one response were recoded to -5 and subject to imputation. A single answer was imputed by selecting one response among those selected by the respondent. The selection of the imputed response was based on the distribution of answers among the single-answer responses. This is the same imputation process as was conducted for the first 3 cycles of HINTS 5. An imputation flag is included on the data-set to indicate imputed values. Imputation occurred as follows:

Table 4-1. Imputation for multiple responses

Question number	Topic	Total imputed
A5	Sources for cancer information	274
G3	Sources for information about clinical trials	247
G4	Most trusted source for information about clinical trials	136
H11	Most important values	93
P11	Sexual orientation	2

In addition, hot-deck imputation was used to replace missing responses for items used in the raking procedure for the weighting. Specifically, this was conducted for items C6, O1, P1, P2, P6, P7, P8, and P9. Hot-deck imputation is a data processing procedure in which a value is assigned with the corresponding value of a "similar" case in the same imputation class. The data record that supplies the imputed value is referred to as the "donor." Under a hot deck approach, the resulting distribution preserves the distribution of values observed for respondents. Imputation classes are defined on the basis of variables that are thought to be correlated with the item with missing values. A donor is then randomly selected within an imputation class to supply the imputed value. Details for items imputed using the hot-deck approach are as follows:

Table 4-2. Imputation for missing response

Question number	Topic	Total imputed
C6	Health Insurance coverage	58
O1	Cancer diagnosis	71
P1	Age	127
P2	Gender	100
P6	Education attainment	143
P7	Marital status	144
P8	Race	291
P9	Ethnicity	355

4.4 Determination of the Number of Household Adults

For the purpose of applying weights, a measure of the number of adults in each household (R_HHAdults) was created using questionnaire responses. The initial measure was taken from responses to demographic section questions asking for the total number of people and the number of children in the household (see items P12-P14). Implausible or missing values that resulted from the answers to those questions were substituted with values to questions on the respondent-selection page of the questionnaire and further substituted with data from the demographic section roster. A detailed list of the steps carried out to identify the number of adults in each household is included in **Appendix D**.

4.5 Survey Eligibility

Returned surveys were reviewed for completion and duplication (more than one questionnaire returned from the same household) to ensure they were eligible for inclusion in the final dataset. Of the 3,977 questionnaires received, 40 were returned blank, 25 more were determined to be incompletely filled out, and 47 additional surveys were identified as duplicates (i.e., the same household returned multiple surveys). The remaining 3,865 surveys were determined to be eligible. The processes for these reviews are detailed below.

Definition of a Complete and Partial Complete Questionnaire

The procedures in HINTS 5 Cycle 4 for determining whether or not a returned questionnaire was the same as for Cycle 3 but slightly modified relative to Cycles 1 and 2. For the first two Cycles of HINTS 5, a complete questionnaire was defined as any questionnaire with at least 80 percent of the required questions answered in Sections A and B. For Cycles 3 and 4, only questions required of every respondent were factored in to the completion rate calculation. Questions that followed skip patterns were excluded from the analysis. A partial-complete was defined as when between 50 percent and 79 percent of the questions were answered in Sections A and B. In Cycle 4, there were 73 partially-completed questionnaires. Both partially-completed and completely-answered questionnaires were retained. The 25 questionnaires with fewer than 50 percent of the required questions answered in Sections A and B were coded as incompletely-filled out and discarded. The 25 incomplete questionnaires represented 0.5% of all surveys with at least one question answered, which was consistent with Cycles 1, 2 and 3 in HINTS 5. Data for 73 partially-completed and 3,792 completed questionnaires were included in the final dataset for Cycle 4 with a total of 3,865 surveys.

Eligibility of Multiple Questionnaires from a Household

47 households returned two filled in questionnaires. The procedures to deal with this issue followed the same guidelines that were used for previous cycles:

- If the same respondent returned multiple questionnaires, the first questionnaire received was retained.
- If the same respondent returned multiple questionnaires on the same day, the first questionnaire to complete the editing process was retained.
- If a return date was unavailable for questionnaires from the same respondent, the questionnaire with fewer substantive questions omitted was retained.
- If different respondents returned a questionnaire and the ages of household members listed in the roster were in agreement (or differed by only one year), the questionnaire that complied with the next birthday rule was retained.[1]
- If, in the above situation, compliance for one or both questionnaires from a household was unclear, the first questionnaire returned was retained.
- If different respondents returned a questionnaire and the ages of household members listed in the roster question were not substantively in agreement, the earliest questionnaire received that complied with the next birthday rule was retained.

4.6 Additional Analytic Variables

Included in the delivery files are four sets of analytical variables: 1) rural-urban commuting area (RUCA) codes that classify census tracts using measures of population density, urbanization, and daily commuting; 2) National Center for Health Statistics (NCHS) urban-rural classification scheme for counties; and 3) Delta Regional Authority service area flag; 4) Urban Influence Codes developed by the Department of Agriculture; and 5) a binary variable indicating whether the household returned their survey before or after the COVID-19 pandemic was declared. These additional variables are described below.

[1] Compliance was determined by whether the person listed in the roster who matched the respondent's age and gender had a month of birth that was the first to follow the month in which the questionnaire was returned.

Rural-Urban Commuting Area (RUCA)

The primary RUCA code (PR_RUCA2010) provides a detailed and flexible way for delineating sub-county components of rural and urban areas. It is based on the 2006-10 American Community Survey (ACS) and has been updated using data from the 2010 decennial census. This primary RUCA code (PR_RUCA2010) delineates metropolitan and nonmetropolitan areas based on the size and direction of primary commuting flows. Previous HINTS datasets have included the secondary code (SEC_RUCA2010) which further subdivides the primary codes to identify areas where classifications overlap based on the size and direction of the secondary, or second largest, commuting flow. This secondary code was excluded from the Cycle 4 datasets to minimize respondent disclosure risk.

Rural-Urban Classification Scheme

The NCHS Urban–Rural Classification Scheme for Counties (NCHSURCODE2013) was developed in 2013 for use in studying associations between urbanization level of residence and health and for monitoring the health of urban and rural residents. The scheme groups counties and county-equivalent entities into six urbanization levels (four metropolitan and two nonmetropolitan), on a continuum ranging from most urban to most rural.

Delta Regional Authority

The Delta Regional Authority is a regional economic development agency serving 252 counties and parishes in parts of eight states: Alabama, Arkansas, Illinois, Kentucky, Louisiana, Mississippi, Missouri, and Tennessee. Its mission is to improve the quality of life for the residents of the Mississippi River Delta Region. The Delta Regional Authority service flag (DRA) identifies the areas served by this agency.

Urban Influence Codes

The 2013 Urban Influence Codes, developed by the United States Department of Agriculture, form a classification scheme that distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas. The standard Office of Management and Budget (OMB) metro and non-metro categories have been subdivided into two metro and 10 non-metro categories, resulting in a 12-part county classification. UIC2013 was dropped from the Cycle 4 datasets to minimize respondent disclosure risk.

Pandemic Return Variable

Because data collection for Cycle 4 started before COVID-19 became an international pandemic and continued after the pandemic was declared by the World Health Organization, there is concern that people who returned the survey early may have responded in different ways to some of the survey questions than people who responded later. For this reason, the variable PANDEMIC was created to flag households whose survey was received at Westat after the World Health Organization declared COVID-19 to be pandemic on March 11, 2020. This variable will facilitate the examination of responses before and after COVID-19 became a widespread issue of concern in the United States.

4.7 Codebook Development

Following cleaning, editing, and weighting (described below), a detailed codebook including frequencies was created for HINTS 5 Cycle 4 for both the weighted and the unweighted data. The codebooks define all variables in the questionnaires, provide the question text, list the allowable codes, and explain the inclusion criteria for each item. The English and Spanish instruments were annotated with variable names and allowable codes to support the usability of the delivery data.

Weighting and Variance Estimation 5

Every sampled adult who completed a questionnaire in HINTS 5 Cycle 4 received a full-sample weight and a set of 50 replicate weights. The full-sample weight is used to calculate population and subpopulation estimates. Replicate weights are used to compute standard errors for these estimates. The use of sampling weights is done to ensure valid inferences from the responding sample to the population, correcting for nonresponse and noncoverage biases to the extent possible.

The computation of the full-sample weights consisted of the following steps:

- Calculating household-level base weights;

- Adjusting for household nonresponse;
- Calculating person-level initial weights; and
- Calibrating the person-level weights to population counts (also known as control totals).

Replicate weights were calculated using the 'delete one' jackknife (JK1) replication method.

5.1 Household Base Weights

The initial step in the weighting process is calculating the household-level base weight for each household in the sample. The household base weight is the reciprocal of the probability of selecting the household for the survey, which depends on the stratum the household was selected from. Generally, base weights for units in the oversampled stratum are smaller than those in the stratum that was not oversampled. In Cycle 4, the base weights for households in the high minority stratum were roughly 1/6 the size of those in the low minority stratum.

5.2 Household Nonresponse Adjustment

Nonresponse is generally encountered to some degree in every survey. The first and most obvious effect of nonresponse is the reduction in the effective sample size, which in turn increases the sampling variance. In addition, if there are systematic differences between the respondents and the nonrespondents, there will be a bias of unknown size and direction. This bias is generally adjusted for in the case of unit nonrespondents (nonrespondents who refuse to participate in the survey at all) with the use of a weighting adjustment term multiplied to the base weights of sample respondents. Item nonresponse (nonresponse to specific questions only) is generally adjusted for through the use of imputation. This section discusses weighting adjustments for unit nonresponse.

The most widely accepted paradigm for unit nonresponse weighting adjustment is the quasi-randomization approach (Oh & Scheuren, 1983). In this approach, nonresponse cells are defined based on those measured characteristics of the sample members that are known to be related to response propensity. For example, if it is known that males respond at a lower rate than females, then sex should be one characteristic used in generating nonresponse cells. Under this approach, sample units are assigned to a response cell, based on a set of defined characteristics. The weighting adjustment for the sample unit is the reciprocal of the estimated response rate for the cell. Any set of response cells must be based on characteristics that are known for all sample units, responding and nonresponding. Thus, questionnaire items on the survey cannot be used in the development of response cells because these characteristics are only known for the responding sample units.

Under the quasi-randomization paradigm, Westat models nonresponse as a "sample" from the population of adults in that cell. If this model is in fact valid, then the use of the quasi-randomization weighting adjustment eliminates any nonresponse bias (see, for example, Little & Rubin (1987), Chapter 4). The weighting procedure for Cycle 4 used a household-level nonresponse adjustment procedure based on this approach. The base weights of the households that did return the questionnaire were adjusted to reflect nonresponse by the remaining eligible households.
A search algorithm[2] was used to identify variables highly correlated with household-level response and these variables were used to create the nonresponse adjustment cells. The variables used to define nonresponse cells for Cycle 4 were:

- Sampling stratum (High Minority; Low Minority)
- Census region (Northeast; South; Midwest; West)
- Route type (Street address; other addresses such as PO Box, Rural Route, etc.)
- Metropolitan Status (county in Metro areas; county in Non-Metro areas)
- High Spanish linguistically isolated area (Yes; No).

Nonresponse adjustment factors were computed for each nonresponse cell *b* using the formula below. This formula is consistent with the RR4 formula of the American Association of Public Opinion Research (AAPOR) for calculating response rates, which is how HINTS is calculating its response rate for Cycle 4. See section 6 for more details.

$$HH\_NRAF(b) = \frac{RESPONSE + NONRESPONSE + UNKNOWN \times e}{RESPONSE},$$

where

- *RESPONSE* is the sum of household base weights for all responding households in nonresponse cell *b*,

[2] An in-house macro WESSEARCH, which calls the Search software, a freeware product developed by the University of Michigan (http://www.isr.umich.edu/src/smp/search/).

- *NONRESPONSE* is the sum of the household base weights for all known nonresponding households in nonresponse cell *b*,
- *UNKNOWN* is the sum of the household base weights for all households that did not return mail whose eligibility is unknown in nonresponse cell *b*, and
- *e* is the estimated percentage of eligible households among the households that did not return mail.

The estimated percentage of eligible households among the households that did not return mail, *e*, was 83.7 percent for Cycle 4 and was calculated using the procedure described in section 6.

The household nonresponse adjustment factors ranged from a low of 2.07 to a high of 5.06, and averaged 3.23 across all nonresponse adjustment cells.

5.3 Initial Person-Level Weights

Each sampled adult in responding households was assigned an initial person-level weight. The initial person-level weight was calculated by multiplying the nonresponse-adjusted household weight by the reciprocal of the sample person's within-household probability of selection. Because only one adult per household was selected to participate in the survey, the reciprocal of the sample person's within-household probability of selection is identical to the number of adults in the household. So, for example, if a household contained three adults and one adult was selected, the initial weight for the selected adult is equal to the nonresponse-adjusted household weight times three.

5.4 Calibration Adjustments

The purpose of calibration is to reduce the sampling variance of estimators through the use of reliable auxiliary information (see, for example, Deville & Sarndal, 1992). In the ideal case, this auxiliary information usually takes the form of known population totals for particular characteristics (called *control totals*). However, calibration also reduces the sampling variance of estimators if the auxiliary information has sampling errors, as long as these sampling errors are significantly smaller than those of the survey itself.

Calibration reduces sampling errors particularly for estimators of characteristics that are highly correlated to the calibration variables in the population. The extreme case of this would be the calibration variables themselves. The survey estimates of the control totals would have considerably higher sampling errors than the "calibrated" estimates of the control totals, which would be the control totals themselves. The estimator of any characteristic that is correlated to any calibration variable will share partially in this reduction of sampling variance, though not fully. Only estimators of characteristics that are completely uncorrelated to the calibration variables will show no improvement in sampling error. Deville and Sarndal (1992) provide a rigorous discussion of these results.

Control Totals

The American Community Survey (ACS) of the U.S. Census Bureau has much larger sample sizes than those of HINTS. The ACS estimates of any U.S. population totals have lower sampling error than the corresponding HINTS estimates, making calibration of the survey weights to ACS control totals beneficial. Westat used the 2018 ACS estimates that are publically available on the Census Bureau web site.

Calibration variables were selected among those that were on the ACS public-use file and were found to be well correlated to important HINTS questionnaire item outcomes (i.e., Westat wanted ACS-available characteristics that tend to have differing mean values for HINTS questionnaire item outcomes). The following ACS characteristics correlate well with HINTS questionnaire items:

- Age
- Gender
- Educational Attainment
- Marital Status
- Race
- Ethnicity
- Census Region

In addition to characteristics from the ACS, two health-related variables were used: *Percent with health insurance* and p*ercent of adults who have ever been diagnosed with cancer*. The *health insurance* variable came from the 2018 National Health Information Survey (NHIS) (Cohen, et al., 2018) and corresponds to the question asked in the HINTS survey (C6, "Are you currently covered by any of the following types of health insurance or health coverage plans?"). The *percent of adults who have ever been diagnosed* *with cancer* variable came from the 2018 National Center for Health Statistics (U.S. Department of Health and Human Services, 2018) and corresponds to the question asked in the HINTS survey (O1, "Have you ever been diagnosed as having cancer?").

Raking to the control totals for these variables (either alone or cross-classified with each other) was then performed. As a result of the raking HINTS weights to the control totals, estimates calculated from HINTS data for the control-total variables agree with those calculated from the source data for the control totals. For example, the national-level estimate of *Percent with health insurance* calculated from HINTS data agrees with the estimate calculated from NHIS 2018 data.

5.5 Replicate Variance Estimation

In addition to the full-sample weight, a set of 50 replicate weights were provided for each adult. These replicate weights are used to calculate standard error of estimates obtained from the HINTS data, using the delete one jackknife (JK1) replication method.

The JK1 jackknife technique is compatible with the sample design and weighting procedures for HINTS. This jackknife variance estimation technique takes carefully selected subsets of the data for each "replicate," and for each respondent in the replicate subset and determines a sampling weight, as if the replicate subset were in fact the responding sample. (This replicate subset is usually almost the entire sample, except for a group of respondents that are "deleted" for that replicate.) The resulting weights are called replicate weights.

The jackknife variance estimator requires the use of replicate weights. For the Cycle 4 data set, a set of 50 replicate weights was assigned to each responding adult. To illustrate how the replicate variance estimates are computed, suppose P is a percentage of adults in the U.S. population having a particular characteristic (e.g., answering one of the HINTS questions in a particular way). A nationally representative estimator p can be computed by aggregating the adult sampling weights of all responding adults with this characteristic (e.g., all responding adults in the survey answering the survey question in a particular way). A JK1 jackknife variance estimator of the sampling variance of p can be computed in two steps:

Step 1. Recompute estimators $p(r)$, r =1,...,50, by aggregating the replicate sampling weights corresponding to replicate r for all responding adults with the characteristic.

Step 2. Compute the JK1 jackknife variance estimator

$$v(p) = \frac{R-1}{R}\sum_{r=1}^{50}(p(r)-p)^2$$

The replicate weights are computed by systematically deleting a portion of the original sample, and recomputing the sampling weights as if the remaining sample (without the deleted portion) were the actual sample. These deleted sample units should be first-stage sampling units, which in HINTS are households. The remainder of the sample with the deleted portion removed is called the replicate subset, and it should mirror the full sample design, as if it were a reduced version of the original sample.

For the purposes of JK1 jackknife variance estimation, each household was assigned to one of 50 replicate "deletion" groups *D(r), r* =1,..., 50. Each replicate sample is the full sample minus the deletion group (i.e., it is roughly 49/50 of the original sample).

The replicate sampling weights were generated in a series of steps that parallel the steps computing the full-sample sampling weights. The replicate base weight for each sampled household or adult and each replicate is either equal to *R*/*(R-1)* times the full sample base weight (if the household is contained in the replicate subset) or equal to 0 (if the household is not contained in the replicate subset, but instead is contained in the "deleted" set for that replicate).

Nonresponse and calibration adjustments were then computed for each set of replicate weights, using the replicate weights in the computation of nonresponse and calibration adjustments in place of the original weights. These calculations generated a set of replicate nonresponse and poststratification adjustments for each responding adult. The final replicate weights were products of the replicate weights, nonresponse adjustments, and calibration adjustments.

5.6 Taylor Series Variance Estimation

Even though replication is the recommended method for variance estimation for HINTS, not all software packages have a replication option to produce variance estimates. For example, SPSS has built-in options for estimating variance using Taylor's Series methods but not replication methods. To accommodate SPSS users or any end user who wants to produce variances using Taylor Series methods, Westat provided the appropriate variables on the HINTS data files to do so.

The full-sample weight (as described in the introduction of Section 5) is used as the weight to compute Taylor's Series variance estimates. The variable VarStratum indicates the variance-estimation stratum and the variable VarCluster indicates the primary sampling unit (PSU) or cluster within the variance-estimation stratum. These variables allow the analyst to produce variance estimates using Taylor's Series.

Response Rates 6

For Cycle 4, response rates were calculated using the RR4 formula of the American Association of Public Opinion Research (AAPOR). HINTS has historically used the RR2 calculation as its official method for computing response rates in HINTS 4 and HINTS 5 (Cycles 1, 2, 3). The difference between the RR2 and RR4 calculation is that the RR4 formula is adjusted based on an estimate of the eligibility rate (e) among the unresolved households (i.e. the households that never return a survey or refuse, or have mailings returned because they were undeliverable).

Incorporating e into HINTS's response rate calculation is appropriate for our address-based sample design where a large proportion of the sampled units' eligibility statuses are never resolved. The RR2 calculation is more conservative than RR4 because it treats all of the unresolved households that never return a survey as eligible for the study (i.e., e = 1). Numerous other federal surveys incorporate an estimate of e in to their response rate calculation, including the CDC's Behavioral Risk Factor Surveillance System (BRFSS), the NCHS's National Household Education Surveys Program (NHES), and the FDA's National Survey of Health Information and Communication (NSHIC). Recently, DeMatteis (2019) developed a method for estimating e in addressed-based samples designs, facilitating the use of RR4 for studies like HINTS.

The formula to calculate e for Cycle 4 is based on DeMatteis (2019):

$$e = \left(\frac{1}{\hat{T}_U}\right)(\hat{T}_{ACS} - \hat{T}_R - \hat{T}_{NR})$$

where $\hat{T}_U$ is the base-weighted number of households with unknown eligibility, $\hat{T}_R$ is the estimated eligible responding households, and $\hat{T}_{NR}$ is the estimated eligible non-responding households (e.g.

refusals and incompletes). $\hat{T}_{ACS}$ is an estimate of the total number of eligible households in the population based on a reliable external source, the 2018 American Community Survey. Table 6-1 summarizes the components of the e calculation.

Table 6-1. Components of e calculation used to compute response rate (RR4)

Household status	Description	Base-weighted sum of households
$\hat{T}_R$	Total responding households	44,546,288
$\hat{T}_{NR}$	Total non-responding, known eligible households	761,295
$\hat{T}_U$	Total households with unknown eligibility	91,075,963
$\hat{T}_{ACS}$	Total households estimated by 2018 ACS	121,520,180

About 2/3 of sampled households in Cycle 4 were unresolved at the end of the field period. We estimated that e was 83.7 percent among these households. Therefore 16.3 percent (1 - e) of the unresolved households were assumed to be ineligible and removed from the denominator of the response rate calculation.

Table 6-2 shows the response rate outcomes overall and by strata based on the RR4 calculation. These data have been weighted to account for the oversampling of addresses in high-minority areas. The overall response rate was 36.7 percent; however this differed significantly by strata. The high-minority strata had the lowest response rate (27.2 percent) and the low-minority had the highest (40.3 percent). The percent of undeliverable households was slightly higher in the high-minority strata (9.3 vs 7.8 percent).

Table 6-2. Response rate calculations by strata based on **RR4** calculation

Response class	High minority	Low minority	Overall
Total sample*	37,374,495	95,007,304	132,381,801
Respondents	9,234,066	35,312,221	44,546,288
Nonrespondents	24,683,378	52,290,513	76,973,892
Undeliverable	3,457,051	7,404,570	10,861,621
Total Households	33,917,444	87,602,734	121,520,180
Percent Undeliverable	9.25%	7.79%	8.20%
Household response rate	**27.23%**	**40.31%**	**36.66%**

*values may not sum to total sample due to rounding of weighted values to nearest single digit

Although Cycle 4 has moved to using the RR4 formula, to enable comparison to prior HINTS 4 and HINTS 5 cycles, the RR2 formula was also calculated. Table 6-3 shows the Cycle 4 response rate overall and by strata based on the RR2 calculation.

Table 6-3. Response rate calculations by strata based on **RR2** calculation

Response class	High minority	Low minority	Overall
Total sample*	42,163,815	105,081,350	147,245,167
Respondents	9,234,066	35,312,221	44,546,288
Nonrespondents	29,472,698	62,364,559	91,837,258
Undeliverable	3,457,051	7,404,570	10,861,621
Total Households	38,706,764	97,676,780	136,383,546
Percent Undeliverable	8.20%	7.05%	7.38%
Household response rate	**23.86%**	**36.15%**	**32.66%**

*values may not sum to total sample due to rounding of weighted values to nearest single digit

Data Suppression for Minimizing Disclosure Risk 7

The overall risk of disclosure with the public-use HINTS file is very low. It does not contain direct identifiers of the respondents or their households. The following recodes are intended to further minimize disclosure risk but not impose restrictions that greatly impede the analytic utility of the data.

To minimize the risk of disclosing HINTS respondents based on their survey answers, the following data suppression rules were applied to the microdata. For variables associated with questions or household characteristics that are potentially identifiable, response categories with fewer than 25 responses were reviewed and, when necessary, recoded in one of three ways:

a) The response option was collapsed with another response option,
b) The response option was set to missing, or
c) The variable with the response option was suppressed entirely.

Variables that are considered potentially identifiable are those that report demographic or geographic[3] information. Forty-five variables were reviewed for disclosure risk and 16 were ultimately recoded or dropped from the public data file. **Appendix E** presents the unweighted distributions of each variable in the review prior to recoding them. Table 7-1 lists the variables that

[3] The following delivered geographic variables are excluded from the public use file and therefore not included in the disclosure review: DMA (Designated Market Area), FIPS (FIPS State/County Code) , FIPST (FIPS State Code)

were reviewed and summarizes the recoding instructions that were applied to the sparse values prior to including them on the public use file. An unperturbed "restricted use" file was also delivered to NCI for restricted use access that protects against disclosure.

Table 7-1. Summary of variables reviewed for disclosure risk and recodes

Variable	Summary
Hisp_cat and all associated dummy variables	-No changes were applied to this variable or the dummy variables. One category of Hisp_cat had fewer than 25 responses (Multiple Hispanic ethnicities selected), however this category is not inherently identifiable.
Race_cat2 and all associated dummy variables	The following changes were applied to the public use file: -Respondents who selected Japanese, Korean or Vietnamese were recoded to 'Selected' for the dummy variable 'Other Asian'. The dummy variables for Japanese, Korean or Vietnamese were dropped from the file. -Respondents who selected Native Hawaiian, Samoan, Guamanian or Chamorro were recoded to 'Selected' for the dummy variable 'Other Pacific Islander.' The dummy variables for Native Hawaiian, Samoan, Guamanian or Chamorro were dropped from the file. -After the dummy variables were recoded as specified above, Race_cat2 was recoded accordingly so there are no categories for the dropped response options.
Birthgender	-No changes were applied to this variable.
GenderIdentity	-This variable was dropped from the public use file. While the non-cis categories are the only ones that are sparsely selected and could be collapsed, the variable could still be crossed with Birthgender to identify non-cis individuals (see crosstab of GenderIdentity and Birthgender in Appendix E).
SexualOrientation	-No changes were applied to this variable.
DRA	-No changes were applied to this variable.
NCHSURCODE2013	-No changes were applied to this variable.
PR_RUCA_2010	-The individual categories were collapsed as follows on the public use file: 1-3 were combined in to the single category (1) and relabeled to 'Metropolitan' 4-6 were combined in to (4) 'Micropolitan' 7-9 were combined in to (7) 'Small town', 10 remained as a single category with the value 'Rural'
RUC2003	-This variable was dropped from the public and restricted use files.
RUC2013	-On the public use file, categories 8 and 9 were collapsed into a single category (8) and re-labeled "Nonmetro - Completely rural or less than 2,500 urban population"
SEC_RUCA_2010	-This variable was dropped from the public and restricted use files.
Stratum	-No changes were applied to this variable.
UIC2013	-This variable was dropped from the public and restricted use files.
APP_REGION	-On the public use file, we recoded 'Northern Appalachia' (N), 'Central Appalachia' (C), and 'Southern Appalachia' (S) into a single category (A) and relabel to 'Appalachia'
CENSDIV	-No changes were applied to this variable.
CENSREG	-No changes were applied to this variable.

References

Cohen, R.A., Martinez, M.E., and Zammitti, E.P. (2018). *Health Insurance Coverage: Early Release of Estimates From the National Health Interview Survey, January – March 2018*. Retrieved from https://www.cdc.gov/nchs/data/nhis/earlyrelease/Insur201808.pdf

DeMatteis, J. (2019). Computing "e" in Self-Administered Address-Based Sampling Studies. *Survey Practice*. Vol 12, Issue 1.

Deville, J.C., and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*, 376-382.

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method.* Hoboken, NJ: John Wiley & Sons.

Little, R., and Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Oh, H., and Scheuren, F. (1983). Weighting adjustments for unit response. In W.G. Madow, I. Olkin, and D. B. Rubin (Eds.), *Incomplete data in sampling surveys, Vol. II: Theory and annotated bibliography*. New York: Academic Press.

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics (2018). *Summary Health Statistics: National Health Interview Survey, 2018*. Retrieved from https://ftp.cdc.gov/pub/Health_Statistics/NCHS/NHIS/SHS/2018_SHS_Table_A-3.pdf

This page is deliberately blank.

Appendix A

Cover Letters in English

This page is deliberately blank.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

FIRST MAILING

Dear {City} Resident:

We are writing to invite you to take part in an important national survey sponsored by the U.S. Department of Health and Human Services - the Health Information National Trends Survey (HINTS). The goal of HINTS is to learn about how people find and use health and medical information. By completing this survey, you will help us learn what health information you need and how to make that information available to you, your family, and your community.

In order to make sure we get responses from a random sample of people, **we ask the adult in your household with the next birthday to complete the survey in the next two weeks**.

Your participation is voluntary and your responses will not be linked to your name. We have enclosed $2 as a token of our appreciation for your participation.

You can find out more about HINTS at hints.cancer.gov. Westat, a research firm, is conducting the survey. If you have any questions about HINTS, please call Westat toll-free at 1-888-738-6805.

Thank you in advance for your participation.

Sincerely,

Kelly D. Blake

Kelly D. Blake, ScD
Director, HINTS
National Institutes of Health
U.S. Dept. of Health and Human Services

Si prefiere recibir la encuesta en español, por favor llame al 1-888-738-6812.

The Health Information National Trends Survey is authorized under 42 USC, Section 285A.

POSTCARD TEXT

A few days ago, you should have received a questionnaire packet asking for your household's participation in the Health Information National Trends Survey (HINTS). By participating in HINTS, you can help the U.S. Department of Health and Human Services determine the best ways of communicating important health information to members of your community.

We are inviting the adult in the household with the next birthday to complete the survey. If that adult has already completed the survey and returned it to us, please accept my sincere thanks. If that adult has not yet completed and returned the survey, we ask that he or she please do so as soon as possible.

Sincerely,

Kelly D. Blake

Kelly D. Blake, ScD
Director, HINTS
National Institutes of Health
U.S. Dept. of Health and Human Services

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

SECOND AND THIRD MAILINGS

Dear {City} Resident:

We recently invited you to participate in an important national survey sponsored by the U.S. Department of Health and Human Services (HHS). The goal of the Health Information National Trends Survey (HINTS) is to learn about how people find and use health and medical information. Your responses will help us keep you, your family, and members of your community better informed on the health issues that matter to you.

We have not yet received your completed survey. To make sure HINTS provides accurate information, we need all the households invited to participate in this year's HINTS to complete the survey. If you did send back your survey and it crossed in the mail with this letter, thank you for the time you took to help make this study a success. In the event that your survey was misplaced, an additional copy is enclosed.

In order to make sure we get responses from a random sample of people, **we ask the adult in your household with the next birthday to complete the survey in the next two weeks**.

Additional information about HINTS is available at hints.cancer.gov. Westat, a research firm, is conducting the survey. If you have any questions about HINTS, please call Westat toll free at 1-888-738-6805.

Thank you in advance for contributing to this important national study.

Sincerely,

Kelly D. Blake

Kelly D. Blake, ScD
Director, HINTS
National Institutes of Health
U.S. Dept. of Health and Human Services

Si prefiere recibir la encuesta en español, por favor llame al 1-888-738-6812.

The Health Information National Trends Survey is authorized under 42 USC, Section 285A.

Health Information National Trends Survey hints

This page is deliberately blank.

Appendix B

Cover Letters in Spanish

This page is deliberately blank.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

FIRST MAILING

Estimado residente de {City}:

Le escribimos para invitarle a participar en una importante encuesta nacional: Encuesta Nacional de Tendencias de Información sobre la Salud (HINTS, por sus siglas en inglés). Esta encuesta está patrocinada por el Departamento de Salud y Servicios Humanos de Estados Unidos.

El objetivo de HINTS es averiguar acerca de cómo las personas encuentran y utilizan la información sobre la salud y la información médica. Complete esta encuesta para ayudarnos a averiguar la información sobre la salud que usted necesita y cómo ponerla a disposición suya, de su familia y de su comunidad.

Para asegurarnos de obtener respuestas que contengan un muestreo aleatorio de la población, **le pedimos que el adulto en su hogar con el próximo cumpleaños complete y devuelva la encuesta en las próximas dos semanas.**

Su participación es voluntaria y sus respuestas no se asociarán con su nombre. Hemos incluido $2 dólares como símbolo de nuestro agradecimiento por su participación.

Usted podrá encontrar más información sobre HINTS en el sitio web hints.cancer.gov. La compañía de estudios de investigación Westat está realizando esta encuesta. Si tiene alguna pregunta sobre HINTS, llame a Westat al siguiente número de teléfono libre de cargo, 1-888-738-6805.

Gracias de antemano por su participación.

Atentamente,

Kelly D. Blake

Kelly D. Blake, ScD
Director, HINTS
Institutos Nacionales de
la Salud
Departamento de Salud y Servicios
Humanos de EE. UU.

La Encuesta Nacional de Tendencias de Información sobre la Salud está autorizada bajo la Sección 285A del USC 42.

Health
Information
National
Trends
Survey
hints

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

SECOND AND THIRD MAILINGS

Estimado residente de {City}:

Recientemente le invitamos a participar en una importante encuesta nacional patrocinada por el Departamento de Salud y Servicios Humanos de Estados Unidos (HHS, por sus siglas en inglés). El objetivo de la Encuesta Nacional de Tendencias de Información sobre la Salud (HINTS, por sus siglas en inglés) es averiguar acerca de cómo las personas encuentran y utilizan la información sobre la salud y la información médica. Sus respuestas nos ayudarán a mantenerlo mejor informado a usted, a sus familiares y a los miembros de la comunidad sobre los temas de salud que les interesan.

Aún no hemos recibido su encuesta completada. Para poder estar seguros de que HINTS provea información acertada, necesitamos que todos los hogares invitados a participar en la encuesta este año, la completen. Si usted ya nos envió de regreso su encuesta y se cruzó con esta carta en el correo, le agradecemos por el tiempo que se tomó para contribuir al éxito de este estudio. En caso de que su encuesta se haya extraviado, adjuntamos una copia adicional.

Para asegurarnos de obtener respuestas que contengan un muestreo aleatorio de la población, **le pedimos que el adulto en su hogar con el próximo cumpleaños complete y devuelva la encuesta en las próximas dos semanas.**

Usted podrá encontrar más información sobre HINTS en el sitio web hints.cancer.gov. La compañía de estudios de investigación Westat está realizando esta encuesta. Si tiene alguna pregunta sobre HINTS, llame a Westat al número libre de cargo, 1-888- 738-6805.

Gracias de antemano por contribuir al éxito de este importante estudio nacional.

Atentamente,

Kelly D. Blake

Kelly D. Blake, ScD
Director, HINTS
Institutos Nacionales de
la Salud
Departamento de Salud y Servicios
Humanos de EE. UU.

La Encuesta Nacional de Tendencias de Información sobre la Salud está autorizada bajo la Sección 285A del USC 42

Appendix C

Frequently Asked Questions (FAQs) English and Spanish

This page is deliberately blank.

Some Frequently Asked Questions about the Health Information National Trends Survey

Q: What is the study about? What kind of questions do you ask?

A: You can find out more about HINTS at hints.cancer.gov. The study concerns health and how people get health information. For example, we will ask how you usually get information about health and what sources of information you most trust. We will also ask about your beliefs on what contributes to good health, how best to prevent cancer, and other health related topics.

Q: How will the study results be used?

A: The results will help the U.S. Department of Health and Human Services promote good health and prevent disease by determining the best ways to communicate accurate health information.

Q: How did you get my address?

A: Your address was randomly selected from among all of the known home addresses in the nation. It was selected using scientific sampling methods.

Q: Why should I take part in this study? Do I have to do this?

A: Getting answers from all the households chosen for the study is the best way to make sure the study results reflect the thoughts and opinions of all Americans. Your participation is voluntary, and you may refuse to answer any questions or leave the study at any time. However, your answers are very important to the success of this study and will represent thousands of others.

Q: Will my answers to the survey be kept private?

A: Yes. Your answers will be kept private under the Privacy Act. Your answers cannot be linked to any information that could identify you or your household, to the extent provided by law. Your completed survey will be stored in a secure file with restricted access. All contact information for your household (such as mailing address) will be destroyed shortly after the research is finalized.

Q: How long will it take to answer the questions?

A: About 20 to 30 minutes.

Q: Who is sponsoring the study?

A: The study is sponsored by the U.S. Department of Health and Human Services.

Q: Who is Westat?

A: Westat is a research company located in Rockville, Maryland. Westat is conducting this survey under contract to the U.S. Department of Health and Human Services.

Preguntas Frecuentes Encuesta Nacional de Tendencias de Información sobre la Salud

P: ¿De qué se trata el estudio? ¿Qué tipo de preguntas contiene?

R: El estudio trata sobre la salud y la manera en que las personas reciben información sobre la salud. Por ejemplo, le preguntaremos cómo obtiene normalmente información sobre cómo mantenerse saludable, el tipo de información en la que más confía y cómo le gustaría obtener dicha información en el futuro. También le preguntaremos sobre lo que cree que contribuye a la buena salud, cómo prevenir mejor el cáncer y su participación en varias actividades afines.

P: ¿Cómo se utilizarán los resultados del estudio? ¿Qué se hará con mi información?

R. Los hallazgos ayudarán al Departamento de Salud y Servicios Humanos de EE.UU. a fomentar la buena salud y prevenir las enfermedades mediante la determinación de formas de comunicar mejor la información sobre la salud a los estadounidenses.

P: ¿Cómo obtuvieron mi dirección?

R: Su dirección fue seleccionada al azar entre todas las direcciones conocidas en la nación usando métodos científicos de muestreo.

P: ¿Por qué debo participar en este estudio? ¿Es obligatorio hacerlo?

R: Su participación es voluntaria y usted puede rehusarse a contestar cualquiera de las preguntas o retirarse del estudio en cualquier momento. Sin embargo, sus respuestas son muy importantes para el éxito de este estudio y representan a miles de personas. El obtener respuesta de todos los hogares escogidos para este estudio es la mejor manera de asegurar que éste refleje los pensamientos y opiniones de todos los estadounidenses.

P: ¿Se mantendrá la privacidad de mis respuestas a la encuesta?

R. Sí. Se mantendrá la privacidad de sus respuestas en virtud de la Ley de Privacidad. Sus respuestas no pueden asociarse a su nombre ni a ninguna otra información que podría identificarlo a usted o a su hogar en la medida de lo permisible por ley. Los cuestionarios completos se almacenarán en un archivo separado con acceso restringido. Las versiones impresas y electrónicas de la información se destruirán poco después de la finalización de la encuesta.

P: ¿Cuánto tiempo tomará responder las preguntas?

R: Cerca de 20 a 30 minutos.

P: ¿Quién patrocina el estudio? ¿Está este estudio aprobado por el Gobierno Federal?

R: El estudio es patrocinado por el Departamento de Salud y Servicios Humanos de EE.UU.

P:¿Quién es Westat?

R. Westat es una compañía de estudios de investigación ubicada en Rockville, Maryland. Westat realiza esta encuesta en virtud de un contrato con el Departamento de Salud y Servicios Humanos de EE.UU.

Appendix D

Variable Values and Data Editing Procedures

This page is deliberately blank.

Missing Value Definitions

Values identifying types of nonresponse or indeterminate responses:

- -1 = Valid skips or appropriately missing data following a dependent question (correctly skipped). Example: If SeekCancerInfo=2 'no' and CancerLotOfEffort was missing, CancerLotOfEffort was assigned the value -1.

- -2 = Question was answered, but respondent should not have answered the question. The question was answered in error by the respondent. Example: If SeekCancerInfo=2 'no' and CancerLotOfEffort was not missing, CancerLotOfEffort was assigned the value -2.

- -4 = Question was answered, but data was removed because the entry of the number or character could not be determined (e.g. unreadable or non-conforming numeric response).

- -5 = Respondent selected more response options than appropriate for the question. Example: If CancerTrustDoctor had values 3 'a little' and 2 'some', CancerTrustDoctor was assigned the value -5. In cases where both -2 and -5 values could be assigned, the -2 value was assigned.

- -6 = Missing data in variables following a missing filter question. Example: If filter question (e.g., SeekCancerInfo) was missing and variables up until the next applicable question (e.g. CancerConfidentGetHealthInf) were missing (e.g., CancerLotOfEffort = missing and CancerFrustrated = missing), variables with missing values were assigned the value -6.

- -9 = Missing data. Not ascertained. Question should have been answered, but no response was recorded. Example: If CancerConfidentGetHealthInf was missing, it was assigned the value -9.

Data Editing Procedures

Variable	Editing Rule	Description of Rule
AdultsInHH	Recoding initial filter/skip question	The value of the following response, MailHHAdults, determined how missing responses to AdultsInHH were re-assigned. As an example, if AdultsInHH was missing and MailHHAdults initially had value 1 (adult in household) then AdultsInHH was assigned the value 2 'no' (indicating not more than 1 adult in the household) and MailHHAdults was assigned the 'missing value' -2 (answered inappropriately). If AdultsInHH was missing and MailHHAdults had value 2 (or greater) then AdultsInHH was assigned the value 1 'yes' (indicating more than 1 adult in the household) and the value for MailHHAdults was retained.
SeekCancerInfo UseInternet WearableDevTrackHealth Smoke100 UsedECigEver SeenFederalCourtTobaccoMessages HeardHPV EverHadCancer	Recoding filter/skip questions	For these filter questions (questions containing a skip instruction associated with the particular response that was selected), response patterns following the question were examined if the filter question was not answered. The 'yes' value (in the majority of cases where a 'yes' response instructed a respondent to continue answering the subsequent questions) was substituted for the missing filter question when any of the subsequent questions were answered. Similarly (when a 'no' response instructed a respondent to skip subsequent questions), the 'no' value was substituted for the missing filter question when all of the subsequent questions that a 'no' response would have directed the respondent to skip were left unanswered and the respondent answered the next applicable question all respondents were supposed to answer. Please note that if neither condition was met, the missing response code values were retained.
StrongNeedCancerInfo_IMP FirstInfoClinTrials_IMP TrustInfoClinTrials_IMP MostImportantValues_IMP SexualOrientation_I	Imputation for multiple responses	Imputation was carried out when multiple responses were selected, resulting in one unique response for these "mark only one" variables. Respondent's multiple answers were replaced with a single imputed answer that had the same distribution over the multiple answers as occurred in the single-answer responses. Imputation was not performed on missing values for this question. The suffixes "_IMP" and "_I" indicate that these variables include imputed values. Flags (indicated by suffix '_IFlag') indicate which values were imputed.

Variable	Editing Rule	Description of Rule
Internet_DialUp Internet_BroadBnd Internet_Cell Internet_WiFi Electronic_SelfHealthInfo Electronic_TalkDoctor Electronic_TestResults Electronic_MadeAppts Tablet_AchieveGoal Tablet_MakeDecision Tablet_DiscussionsHCP WillingShareData_HCP WillingShareData_YourFamily WillingShareData_YourFriends IntRsn_VisitedSocNet IntRsn_SharedSocNet IntRsn_SupportGroup IntRsn_YouTube ProbCare_BringTest ProbCare_WaitLong ProbCare_RedoTest ProbCare_ProvideHist HealthIns_InsuranceEmp HealthIns_InsurancePriv HealthIns_Medicare HealthIns_Medicaid HealthIns_Tricare HealthIns_VA HealthIns_IHS HealthIns_Other	Recoding missing responses for items with forced-choice response formats	Respondents were asked to select 'yes' or 'no' to a series of sub-items, allowing them to select as many responses as would apply. These 'forced-choice' response formats sometimes result in respondents indicating which sub-items apply to them by selecting the 'yes' response option for some and leaving the others unanswered. To allow the data to reflect this practice, if respondents did check one or more 'yes' response options within the group, but did not check a 'no' response option for any sub-item in the question, the sub-items that were missing a response were set to 'no.' However, if a respondent, in addition to leaving other sub-items unanswered, did select a 'no' response option for at least one sub-item, the unanswered sub-items were not assumed to be 'no' responses and instead remained missing.

Variable	Editing Rule	Description of Rule
NotAccessed_SpeakDirectly NotAccessed_NoInternet NotAccessed_NoNeed NotAccessed_ConcernedPrivacy NotAccessed_NoRecord NotAccessed_LogInProb NotAccessed_Uncomfortable NotAccessed_MultipleRec RecordsOnline_ViewResults RecordsOnline_MessageHCP RecordsOnline_DownloadHealth ESent_AnotherHCP ESent_Family ESent_HealthApp MedConditions_Diabetes MedConditions_HighBP MedConditions_HeartCondition MedConditions_LungDisease MedConditions_Depression		
HealthInsurance_I EverHadCancer_I Age_I BirthGender_I MaritalStatus_I Education_I Hisp_Cat_I Race_Cat2_I	Imputation for missing responses	Missing values were imputed for variables that were used in the process of assigning weights. The suffix "_I" indicates that this variable includes imputed values. Flags (indicated by suffix '_IFlag') indicate which values were imputed.
FreqGoProvider AccessOnlineRecord TimesModerateExercise	Recoding filter/skip questions	For these filter questions (questions containing a skip instruction associated with the particular response that was selected), response patterns following the question were examined if the filter question was not answered. The value representing the skip response was substituted for the missing filter question if all of the subsequent questions that the response directed the respondent to skip were left unanswered, and the respondent answered the next applicable question. However, missing values were not substituted with other values if the filter question was not answered but a follow-up question was answered.

Variable	Editing Rule	Description of Rule
Height_Feet Height_Inches	Edits for implausible values	The rules that were applied minimized the number of out-of-range values by accounting for response measurements in incorrect boxes, responses using metric, responses using only one unit of measurement and other response errors. **Rules Applied to Edit Height Variables:** If HEIGHT_Feet was 0 or missing and HEIGHT_Inches>48 and HEIGHT_Inches<=60, then the first digit was taken as the feet value and the second digit was taken as the inches value (to correct for respondents expressing both feet and inches in the inches box). If HEIGHT_Feet was 0 or missing and HEIGHT_Inches>61 and HEIGHT_Inches<=83, then the inches value was converted to its feet-and-inches equivalent (to correct for respondents expressing height in inches, resulting in heights from 5'1" to 6'11"). If HEIGHT_Feet was 1 and HEIGHT_Inches>=3 and HEIGHT_Inches<=9 (or HEIGHT_Inches>=30 and HEIGHT_Inches<=90) then this metric value was converted to feet-and-inches (to correct for respondents using meters and tenths and hundredths of a meter to express height). If HEIGHT_Feet>3 and HEIGHT_Feet<7 and HEIGHT_Inches = 20, 30, etc. thru 90 then the trailing 0 was removed. If HEIGHT_Feet>3 and HEIGHT_Feet<7 and HEIGHT_Inches = 15, 25, etc. thru 95 then the trailing 5 was removed (to correct for respondents expressing values in tenths of an inch). If HEIGHT_Feet>3 and HEIGHT_Feet<7 and HEIGHT_Inches = 12, 23, 34, 45 etc. thru 89 then the first digit was taken (to correct for respondents giving an inch value as a range, e.g., 1-2 or 8-9 inches). If HEIGHT_Feet>3 and HEIGHT_Feet<7 and HEIGHT_Inches = a two digit value whereby the first digit equaled the feet value the second digit was taken as the inches value (to correct for respondents expressing the height in inches as well as in feet, e.g., 5'58" resulted in value 5'8") If HEIGHT_Feet>6 and HEIGHT_Feet<12 and HEIGHT_Inches>3 and HEIGHT_Inches<7, then the values were switched (to correct for respondents

Variable	Editing Rule	Description of Rule
		putting measurements in the wrong boxes, resulting in edited values from 4'7" to <7 feet). If none of the preceding height editing rules were applicable: **<u>Height_Feet (Height in Feet):</u>** Any responses greater than 7 feet were recoded to "-4", which is the code for non-conforming responses. **<u>Height_Inches (Height in Inches):</u>** Any responses greater than 11 inches were recoded to "-4", which is the code for non-conforming responses.
HaveDevice_Cat CaregivingWho_Cat CaregivingCond_Cat HeardGenTest_Cat TestSource_Cat HadTest2_Cat SharedRes2_Cat UndGenTest_Cat TobaccoMessages_Cat Cancer_Cat Occupation_Cat Hisp_Cat Race_Cat2	Summarized distribution of 'mark all that apply' responses	A variable was created to indicate each response selection a respondent made for these 'mark all that apply' variables. The derived variable with the suffix '_cat' summarized the response selected or indicated that multiple responses were selected.
HealthInsurance	Derived variable	Responses to questions asking about different types of health insurance (C7a-h) were compiled into a derived measure of whether or not the respondent was covered by any health insurance.
Education IncomeRanges	Edits for multiple responses	The highest order (e.g., education level or income range) was taken when multiple responses were selected.
R_HHAdults	Derived variable	Responses to questions asking about household size as well as other information about the household (e.g., number of questionnaires returned) were compiled into a derived measure that best represented the number of adults in the household.
HHAdults_Num	Imputation for zero and missing responses	Missing values were imputed for the derived count of household adults when the derived variable had values of zero or missing. A flag (indicated by suffix '_IFlag') indicates which values were imputed.

Variable	Editing Rule	Description of Rule
QDisp	Derived variable	A variable was created to indicate the proportion of items respondents answered in the first two sections. This was used to determine incompletely-filled out questionnaires.
Pandemic	Derived variable	A variable was created to indicate when a respondent's survey return date was after the date the pandemic was declared (3/11/2020).
FullTimeOcc_Cat	Derived variable	A variable was created which combines the responses to P4 and P5, which aims to give a more comprehensive idea of a respondent's full time occupation.
Weight DrinkDaysPerWeek AverageTimeSitting WhenDiagnosedCancer GenderIdentity_OS SexualOrientation_OS Occupation_Other_OS Age SelfAge SelfMOB **HHAdultAge[2-5]** HHAdultMOB[2-5]	Recoding out of range responses	**Weight:** Any responses less than 40 pounds or greater than 500 pounds were recoded to "-4", which is the code for non-conforming responses. **DrinkDaysPerWeek** Any responses greater than 7 days per week were recoded to "-4", which is the code for non-conforming responses. **AverageTimeSitting** Any responses greater than 24 hours were recoded to "-4", which is the code for non-conforming responses. **WhenDiagnosedCancer (Age at Time of Cancer Diagnosis):** Any responses greater than the age of the respondent were recoded to "-4", which is the code for non-conforming responses. **GenderIdentity_OS** Review of verbatim responses - Responses of "none of your business" and other similar phraseology were reviewed for scanning accuracy and recoded to "-4", which is the code for nonconforming responses. **SexualOrientation_OS** Review of verbatim responses - Responses of "none of your business" and other similar phraseology were reviewed for scanning accuracy and recoded to "-4", which is the code for nonconforming responses. **Occupation_OS** Review of verbatim responses - Responses mentioning "COVID-19/Coronavirus" and other similar phraseology were reviewed for scanning

Variable	Editing Rule	Description of Rule
		accuracy and other checked responses. If Employed was checked along with "Other", responses were left alone. Per client decision, if "Employed" was not checked with "Other", then case was updated to show "Less1YUnemployed" along with the existing "Other" and OC response retained. **<u>Age Variables</u>** Responses were examined for out of range or unlikely ages (those listing their age as < 18 and > 105). **<u>SelfMOB (Respondent's Month of Birth):</u>** Any responses less than 1 or greater than 12 months were recoded to "-4", which is the code for non-conforming responses. **<u>HHAdultMOB[2-5] (Second – Fifth Adult in Household Month of Birth):</u>** Any responses less than 1 or greater than 12 months were recoded to "-4", which is the code for non-conforming responses.
HaveDevice_CellPh HaveDevice_None Caregiving_No HeardGenTest_None HadTest2_None	Recoding filter/skip questions	For these "mark all that apply" filter questions ("mark all that apply" type questions where one or more response option contains a skip instruction at the "No" or "None" response), when the "No" or "None" response was selected, all responses within the question group were examined. If other responses were checked, the "No" or "None" response was recoded to "Not selected", and the other responses were retained.
TestSource_NotHeard SharedRes2_NotShared UndGenTest_NoOne NotHisp	Recoding illogical response combinations	For these "mark all that apply" questions ("mark all that apply" type questions where one or more response options do not contain a skip instruction at the "No" or "None" response, but keeping a "No" or "None" response in combination with other responses does not make logical sense), when the "No" or "None" response was selected, all responses within the question group were examined. If other responses were checked, the "No" or "None" response was recoded to "Not selected", and the other responses were retained.

Deriving and Imputing Measure of Household Adults

A program was developed based on the following guidelines in order to develop a single derived indicator for the number of household adults. The derived value is calculated for each household based on three sources of household size information that is solicited in the questionnaire. The guidelines were adapted from the analogous procedures used in cycle 1.

1. Create a composite variable (**RS_HHAdults**) from the raw and edited versions of **MailHHAdults**, resulting in a value of household adults for all households. This will be the raw (unedited) value of **MailHHAdults** for situations when respondents indicate that there are not more than one adult in the household (**AdultsInHH**=2) but enter a value for **MailHHAdults** that is greater than 1.

2. Create a second indicator for the number of adults in the household (**Demo_HHAdults**) based on responses to questions in the demographic section. **Demo_HHAdults** = **TotalHousehold** - **ChildrenInHH**. If **Demo_HHAdults** is negative, then reset the value of **Demo_HHAdults** to be missing.

    a. If **Demo_HHAdults** value is missing, 0, or 11 or greater, then replace value with a value from **RS_HHAdults** if **RS_HHAdults** is between 1 and 10 inclusive; name this new variable **DemoRS_HHAdults**.

    b. If **Demo_HHAdults** is 0 and **RS_HHAdults** is not between 1 and 10 inclusive, retain the value of **Demo_HHAdults** for variable **DemoRS_HHAdults**.

3. Edit/correct the variable **Demo_HHAdults** when its values are implausible by substituting in plausible values of variable **RS_HHAdults**. If **Demo_HHAdults** is between 1 and 10 inclusive or **RS_HHAdults** is not between 1 and 10 inclusive, retain the value of **Demo_HHAdults** for variable **DemoRS_HHAdults**.

4. Create a household size indicator based on the number of adults in the household as listed in the household enumeration roster. This is the sum of household members listed in the table whose ages are between 18 and 115 inclusive (**Roster_HHAdults)**.

5. Edit/correct the variable **DemoRS_HHAdults** using values of variable **Roster_HHAdults** and name the final measure of household size: **R_HHAdults**.

    a. R_HHAdults = DemoRS_HHAdults;

    b. If DemoRS_HHAdults = 0 then R_HHAdults = Roster_HHAdults.

    c. If DemoRS_HHAdults is missing and Roster_HHAdults is greater than 0, R_HHAdults = Roster_HHAdults.

    d. If Roster_HHAdults > DemoRS_HHAdults then R_HHAdults = Roster_HHAdults.

Imputation for the remaining values of zero or missing for R_HHAdults involved replacing these values with the average number of adults in responding households with non-zero or non-missing values of R_HHAdults, resulting in the variable HHAdults_Num. Nine households had missing values of R_HHAdults that needed to be imputed.

Deriving the FullTimeOcc_Cat variable

Fulltimeocc_cat combines responses to P4 (WorkFullTime) and P5 (Occupation_Cat) into a single indicator of occupation status with the response options listed below.

Respondents are assigned to the category they selected in P5 which appears highest in the list below. For participants who chose 'Employed' for P5, their answer to P4 is used to determine whether they are coded as 'Employed full time' or 'Employed part time.' In some instances participants open-ended response to the P5 'Other' category were used to re-categorize them in to a different category than the highest one selected on the list. Respondents who mentioned a COVID-19 related work disruption were assigned to the 'Other' category. Participants who chose both 'Employed' and an Unemployed category in P5 were coded as 'Illogical response combination.'

Category	Value
P4 or P5 are missing	-9
Illogical response combination	-4
Employed full time	1
Employed part time	2
Homemaker	3
Student	4
Retired	5
Disabled	6
Unemployed less than 1 year	7
Unemployed 1 year or more	8
Other	9

Appendix E

Summary of Unperturbed Distributions for Variables Included in Disclosure Risk Assessment

This page is deliberately blank.

Disclosure review
Race and Ethnicity variables

The FREQ Procedure

Derived variable to categorize responses given in O6 (Race)

Race_Cat2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
White only	2606	67.43	2897	74.95
Black only	600	15.52	3497	90.48
American Indian or Alaska Native only	32	0.83	3529	91.31
Multiple races selected	127	3.29	3656	94.59
Asian Indian only	27	0.70	3683	95.29
Chinese only	48	1.24	3731	96.53
Filipino only	38	0.98	3769	97.52
Japanese only	9	0.23	3778	97.75
Korean only	19	0.49	3797	98.24
Vietnamese only	19	0.49	3816	98.73
Other Asian only	20	0.52	3836	99.25
Native Hawaiian only	3	0.08	3839	99.33
Guamanian or Chamorro only	1	0.03	3840	99.35
Other Pacific Islander only	25	0.65	3865	100.00

P9. What is your race? - White?

White	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	2707	70.04	2998	77.57
Not Selected	867	22.43	3865	100.00

P9. What is your race? - Black or African American?

Black	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	649	16.79	940	24.32
Not Selected	2925	75.68	3865	100.00

Disclosure review
Race and Ethnicity variables

The FREQ Procedure

P9. What is your race? - American Indian or Alaska Native?				
AmerInd	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	102	2.64	393	10.17
Not Selected	3472	89.83	3865	100.00

P9. What is your race? - Asian Indian?				
AsInd	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	32	0.83	323	8.36
Not Selected	3542	91.64	3865	100.00

P9. What is your race? - Chinese?				
Chinese	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	61	1.58	352	9.11
Not Selected	3513	90.89	3865	100.00

P9. What is your race? - Filipino?				
Filipino	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	45	1.16	336	8.69
Not Selected	3529	91.31	3865	100.00

Disclosure review
Race and Ethnicity variables

The FREQ Procedure

P9. What is your race? - Japanese?				
Japanese	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	16	0.41	307	7.94
Not Selected	3558	92.06	3865	100.00

P9. What is your race? - Korean?				
Korean	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	26	0.67	317	8.20
Not Selected	3548	91.80	3865	100.00

P9. What is your race? - Vietnamese?				
Vietnamese	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	22	0.57	313	8.10
Not Selected	3552	91.90	3865	100.00

P9. What is your race? - Other Asian?				
OthAsian	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	29	0.75	320	8.28
Not Selected	3545	91.72	3865	100.00

Disclosure review
Race and Ethnicity variables

The FREQ Procedure

P9. What is your race? - Native Hawaiian?				
Hawaiian	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	10	0.26	301	7.79
Not Selected	3564	92.21	3865	100.00

P9. What is your race? - Guamanian or Chamorro?				
Guamanian	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	3	0.08	294	7.61
Not Selected	3571	92.39	3865	100.00

P9. What is your race? - Samoan?				
Samoan	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	2	0.05	293	7.58
Not Selected	3572	92.42	3865	100.00

P9. What is your race? - Other Pacific Islander?				
OthPacIsl	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	291	7.53	291	7.53
Selected	32	0.83	323	8.36
Not Selected	3542	91.64	3865	100.00

Disclosure review
Race and Ethnicity variables

The FREQ Procedure

Derived variable to categorize responses given in O5 (Hispanic ethnicity)				
Hisp_Cat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	355	9.18	355	9.18
Not Hispanic only	2914	75.39	3269	84.58
Mexican only	272	7.04	3541	91.62
Puerto Rican only	69	1.79	3610	93.40
Cuban only	31	0.80	3641	94.20
Other Hispanic only	204	5.28	3845	99.48
Multiple Hispanic ethnicities selected	20	0.52	3865	100.00

P8. Are you of Hispanic, Latino/a, or Spanish origin? - No, not of Hispanic, Latino/a, or Spanish origin.				
NotHisp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	355	9.18	355	9.18
Selected	2914	75.39	3269	84.58
Not Selected	596	15.42	3865	100.00

P8. Are you of Hispanic, Latino/a, or Spanish origin? - Yes, Mexican, Mexican American, Chicano/a.				
Mexican	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	355	9.18	355	9.18
Selected	286	7.40	641	16.58
Not Selected	3224	83.42	3865	100.00

Disclosure review
Race and Ethnicity variables

The FREQ
Procedure

P8. Are you of Hispanic, Latino/a, or Spanish origin? - Yes, Puerto Rican.				
PuertoRican	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	355	9.18	355	9.18
Selected	78	2.02	433	11.20
Not Selected	3432	88.80	3865	100.00

P8. Are you of Hispanic, Latino/a, or Spanish origin? - Yes, Cuban				
Cuban	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	355	9.18	355	9.18
Selected	36	0.93	391	10.12
Not Selected	3474	89.88	3865	100.00

P8. Are you of Hispanic, Latino/a, or Spanish origin? - Yes, another Hispanic, Latino/a, or Spanish origin.				
OthHisp	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	355	9.18	355	9.18
Selected	219	5.67	574	14.85
Not Selected	3291	85.15	3865	100.00

Disclosure review
Gender and Orientation variables

The FREQ Procedure

P11. Do you think of yourself as...				
SexualOrientation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	239	6.18	239	6.18
Multiple responses selected in error	2	0.05	241	6.24
Heterosexual, or straight (Go to question P12)	3402	88.02	3643	94.26
Homosexual, or gay or lesbian (Go to question P12)	81	2.10	3724	96.35
Bisexual (Go to question P12)	82	2.12	3806	98.47
Something else - Specify	59	1.53	3865	100.00

Disclosure review
Gender and Orientation variables

The FREQ
Procedure

P11. Do you think of yourself as...Something else - Specify:				
SexualOrientation_OS	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
(UNREADABLE)	1	0.03	1	0.03
-1	3553	91.93	3554	91.95
-2	13	0.34	3567	92.29
-4	16	0.41	3583	92.70
-6	239	6.18	3822	98.89
-9	7	0.18	3829	99.07
17 YEARS ABSTAIN FROM SEX	1	0.03	3830	99.09
CHRISTIAN	1	0.03	3831	99.12
CRAZY QUESTION, NONE OF THOSE	1	0.03	3832	99.15
FEMALE	6	0.16	3838	99.30
HETEROFLEXIBLE	1	0.03	3839	99.33
HUMAN	4	0.10	3843	99.43
I AM JUST ME	1	0.03	3844	99.46
MALE	2	0.05	3846	99.51
ME	1	0.03	3847	99.53
MORAL	1	0.03	3848	99.56
NON-SEXUAL	1	0.03	3849	99.59
NORMAL	3	0.08	3852	99.66
NORMAL FEMALE	1	0.03	3853	99.69
NORMAL GENDER	1	0.03	3854	99.72
NOT ACTIVE PARTNER	1	0.03	3855	99.74
PANSEXUAL	6	0.16	3861	99.90
QUEER	2	0.05	3863	99.95
REGULAR MALE	1	0.03	3864	99.97
TOO OLD	1	0.03	3865	100.00

Disclosure review
Gender and Orientation variables

The FREQ Procedure

P2. On your original birth certificate, were you listed as male or female?				
BirthGender	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	100	2.59	100	2.59
Male	1561	40.39	1661	42.98
Female	2204	57.02	3865	100.00

P3. What is your current gender identity?				
GenderIdentity	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Missing data (Not Ascertained)	105	2.72	105	2.72
Multiple responses selected in error	1	0.03	106	2.74
Male	1553	40.18	1659	42.92
Female)	2193	56.74	3852	99.66
Transgender	3	0.08	3855	99.74
Gender non-conforming	5	0.13	3860	99.87
Other-Specify	5	0.13	3865	100.00

P3. What is your current gender identity? SPECIFY:				
GenderIdentity_OS	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
-1	3748	96.97	3748	96.97
-2	6	0.16	3754	97.13
-4	4	0.10	3758	97.23
-6	105	2.72	3863	99.95
-9	2	0.05	3865	100.00

Disclosure review
Gender and Orientation variables

The FREQ Procedure

BirthGender	GenderIdentity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Missing data (Not Ascertained)	Missing data (Not Ascertained)	93	2.41	93	2.41
Missing data (Not Ascertained)	Male	1	0.03	94	2.43
Missing data (Not Ascertained)	Female)	6	0.16	100	2.59
Male	Missing data (Not Ascertained)	3	0.08	103	2.66
Male	Multiple responses selected in error	1	0.03	104	2.69
Male	Male	1547	40.03	1651	42.72
Male	Female)	3	0.08	1654	42.79
Male	Transgender	1	0.03	1655	42.82
Male	Gender non-conforming	3	0.08	1658	42.90
Male	Other-Specify	3	0.08	1661	42.98
Female	Missing data (Not Ascertained)	9	0.23	1670	43.21
Female	Male	5	0.13	1675	43.34
Female	Female)	2184	56.51	3859	99.84
Female	Transgender	2	0.05	3861	99.90
Female	Gender non-conforming	2	0.05	3863	99.95
Female	Other-Specify	2	0.05	3865	100.00

Disclosure review Geographic variables

The FREQ Procedure

Sampling Stratum				
Stratum	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High Minority Areas	2420	62.61	2420	62.61
Low Minority Areas	1445	37.39	3865	100.00

Appalachian Subregion				
APP_REGION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	3628	93.87	3628	93.87
Central Appalachia	15	0.39	3643	94.26
Northern Appalachia	97	2.51	3740	96.77
Southern Appalachia	125	3.23	3865	100.00

Mississippi Delta region				
DRA	Frequency	Percent	Cumulative Frequency	Cumulative Percent
In the Mississippi Delta region	153	3.96	153	3.96
Not in the Mississippi Delta region	3712	96.04	3865	100.00

Census division				
CENSDIV	Frequency	Percent	Cumulative Frequency	Cumulative Percent
New England	140	3.62	140	3.62
Middle Atlantic	441	11.41	581	15.03
East North Central	462	11.95	1043	26.99
West North Central	183	4.73	1226	31.72
South Atlantic	966	24.99	2192	56.71
East South Central	211	5.46	2403	62.17
West South Central	551	14.26	2954	76.43
Mountain	261	6.75	3215	83.18
Pacific	650	16.82	3865	100.00

Disclosure review
Geographic variables

The FREQ Procedure

Census region				
CENSREG	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Northeast	581	15.03	581	15.03
Midwest	645	16.69	1226	31.72
South	1728	44.71	2954	76.43
West	911	23.57	3865	100.00

NCHS 2013 Urban-Rural Classification Scheme for Counties				
NCHSURCODE2013	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Metropolitan: large metro	1393	36.04	1393	36.04
Metropolitan: large fringe metro	909	23.52	2302	59.56
Metropolitan: medium metro	811	20.98	3113	80.54
Metropolitan: small metro	322	8.33	3435	88.87
Non-metropolitan: micropolitan	252	6.52	3687	95.39
Non-metropolitan: noncore	178	4.61	3865	100.00

Disclosure review
Geographic variables

The FREQ Procedure

USDA Rural-Urban Commuting Areas				
PR_RUCA_2010	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Metropolitan area core: primary flow within an urbanized area (UA)	3083	79.77	3083	79.77
Metropolitan area high commuting: primary flow 30% or more to a UA	284	7.35	3367	87.12
Metropolitan area low commuting: primary flow 10% to 30% to a UA	20	0.52	3387	87.63
Micropolitan area core: primary flow within an Urban Cluster of 10,000 to 49,999 (large UC)	193	4.99	3580	92.63
Micropolitan high commuting: primary flow 30% or more to a large UC	64	1.66	3644	94.28
Micropolitan low commuting: primary flow 10% to 30% to a large UC	17	0.44	3661	94.72
Small town core: primary flow within an Urban Cluster of 2,500 to 9,999 (small UC)	100	2.59	3761	97.31
Small town high commuting: primary flow 30% or more to a small UC	25	0.65	3786	97.96
Small town low commuting: primary flow 10% to 30% to a small UC	8	0.21	3794	98.16
Rural areas: primary flow to a tract outside a UA or UC	71	1.84	3865	100.00

USDA Rural/Urban Designation (2003)				
RUC2003	Frequency	Percent	Cumulative Frequency	Cumulative Percent
County in metro area with 1 million population or more	2244	58.06	2244	58.06
County in metro area of 250,000 to 1 million population	781	20.21	3025	78.27
County in metro area of fewer than 250,000 population	347	8.98	3372	87.24
Nonmetro county with urban population of 20,000 or more, adjacent to a metro area	165	4.27	3537	91.51
Nonmetro county with urban population of 20,000 or more, not adjacent to a metro area	59	1.53	3596	93.04
Nonmetro county with urban population of 2,500-19,999, adjacent to a metro area	156	4.04	3752	97.08
Nonmetro county with urban population of 2,500-19,999, not adjacent to a metro area	64	1.66	3816	98.73
Nonmetro county completely rural or less than 2,500 urban population, adjacent to	25	0.65	3841	99.38
Nonmetro county completely rural or less than 2,500 urban population, not adjacent to	24	0.62	3865	100.00

Disclosure review
Geographic variables

The FREQ Procedure

USDA 2013 Rural-Urban Continuum Codes				
RUC2013	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Metro - Counties in metro areas of 1 million population or more	2293	59.33	2293	59.33
Metro - Counties in metro areas of 250,000 to 1 million population	822	21.27	3115	80.60
Metro - Counties in metro areas of fewer than 250,000 population	320	8.28	3435	88.87
Nonmetro - Urban population of 20,000 or more, adjacent to a metro area	131	3.39	3566	92.26
Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area	43	1.11	3609	93.38
Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area	154	3.98	3763	97.36
Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area	62	1.60	3825	98.97
Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area	17	0.44	3842	99.40
Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area	23	0.60	3865	100.00

Disclosure review
Geographic variables

The FREQ
Procedure

USDA 2010 SECONDARY RURAL-URBAN COMMUNITY AREA CODE				
SEC_RUCA_2010	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Metropolitan area core: primary flow within an urbanized area (UA), No additional code	2669	69.06	2669	69.06
Metropolitan area core: primary flow within an urbanized area (UA), Secondary flow 30% to	414	10.71	3083	79.77
Metropolitan area high commuting: primary flow 30% or more to a UA, No additional code	268	6.93	3351	86.70
Metropolitan area high commuting: primary flow 30% or more to a UA, Secondary flow 30% to	16	0.41	3367	87.12
Metropolitan area low commuting: primary flow 10% to 30% to a UA, No additional code	20	0.52	3387	87.63
Micropolitan area core: primary flow within an Urban Cluster of 10,000 to 49,999	193	4.99	3580	92.63
Micropolitan high commuting: primary flow 30% or more to a large UC, No additional code	54	1.40	3634	94.02
Micropolitan high commuting: primary flow 30% or more to a large UC, Secondary flow 30%	10	0.26	3644	94.28
Micropolitan low commuting: primary flow 10% to 30% to a large UC, No additional code	17	0.44	3661	94.72
Small town core: primary flow within an Urban Cluster of 2,500 to 9,999 (small UC),	100	2.59	3761	97.31
Small town high commuting: primary flow 30% or more to a small UC, No additional code	19	0.49	3780	97.80
Small town high commuting: primary flow 30% or more to a small UC, Secondary flow 30% to	4	0.10	3784	97.90
Small town high commuting: primary flow 30% or more to a small UC, Secondary flow 30%	2	0.05	3786	97.96
Small town low commuting: primary flow 10% to 30% to a small UC, No additional code	8	0.21	3794	98.16
Rural areas: primary flow to a tract outside a UA or UC, No additional code	22	0.57	3816	98.73
Rural areas: primary flow to a tract outside a UA or UC, Secondary flow 30% to 50% to	49	1.27	3865	100.00

Disclosure review
Geographic variables

The FREQ Procedure

URBAN INFLUENCE CODES (2013)				
UIC2013	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Large-in a metro area with at least 1 million residents or more	2293	59.33	2293	59.33
Small-in a metro area with fewer than 1 million residents	1142	29.55	3435	88.87
Micropolitan adjacent to a large metro area	57	1.47	3492	90.35
Noncore adjacent to a large metro area	27	0.70	3519	91.05
Micropolitan adjacent to a small metro area	123	3.18	3642	94.23
Noncore adjacent to a small metro with town of at least 2,500 residents	83	2.15	3725	96.38
Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents	12	0.31	3737	96.69
Micropolitan not adjacent to a metro area	72	1.86	3809	98.55
Noncore adjacent to micro area and contains a town of 2,500-19,999 residents	22	0.57	3831	99.12
Noncore adjacent to micro area and does not contain a town of at least 2,500 residents	11	0.28	3842	99.40
Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents	12	0.31	3854	99.72
Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500	11	0.28	3865	100.00