Health Information National Trends Survey 4 (HINTS 4)

Cycle 2 Methodology Report

March 2013

Prepared for
National Cancer Institute
6120 Executive Boulevard
Bethesda, MD 20892-7195

Prepared by
Westat
1600 Research Boulevard
Rockville, MD 20850

Table of Contents

Appendices

The Health Information National Trends Survey (HINTS) is a nationally-representative survey which has been administered every few years by the National Cancer Institute since 2003. The HINTS target population is all adults aged 18 or older in the civilian non-institutionalized population of the United States. The most recent version of HINTS administration (referred to as HINTS 4) includes four mail-mode data collection cycles over the course of three years. The second of these cycles (Cycle 2) was conducted from October 2012 through January 2013. This report summarizes the methodology, sampling, and procedures of Cycle 2. Data cleaning and weighing procedures as well as response rates are discussed.

Cycle 2 Overview   **1**

Cycle 2 was conducted by mail using a protocol similar to that used in Cycle 1with a goal of obtaining 3,500 completed questionnaires. Based on results of methodological experiments conducted in Cycle 1, for Cycle 2 some changes were made to the layout of the Spanish language instrument and an embedded experiment on mailing procedures was conducted.

1.1   Respondent Selection

Based on the results of the embedded experiment in Cycle 1, it was decided that respondent selection would be conducted uniformly for all households in Cycle 2 using the Next Birthday Method. This method asks that the adult with the next birthday complete the questionnaire and therefore results in one completed questionnaire per household. A detailed description of the sampling methods can be found in Chapter 2.

1.2   Spanish Instrument Changes

During the course of cognitive testing for the Cycle 2 instrument, it was clear that Spanish-speaking respondents were struggling with the complexity of the survey instrument, particularly the grid items. An examination of missing data from Cycle 1 provides some confirmation of this observation, with significantly more grid items left blank in Spanish questionnaires than in English. Therefore, for Cycle 2 the layout of the Spanish language instrument was changed from 2 columns

to a single column. It was anticipated that this change would make it easier for Spanish-speaking respondents to complete all the questionnaire items.

In addition, a short version of the Spanish questionnaire was developed. This short version, which included only items considered "core," replaced the long version of the Spanish questionnaire in the final mailing.

1.3   Embedded Experiment

The embedded experiment included in Cycle 2 of HINTS was focused on trying to increase the participation of Spanish-speaking respondents. As in Cycle 1, households were flagged as potentially Spanish-speaking in one of three ways:

- Linguistically isolated areas: The US Census Bureau defines linguistically isolated households as those in which everyone over 14 years old speaks a language other than English and does not speak English very well. Sampled households from Census tracks with relatively high proportions of linguistically isolated Spanish households were flagged as potentially Spanish-speaking.

- Hispanic surname match: The surnames provided by the address vendor were compared to typical Hispanic surnames. Households identified as having a Hispanic surname were flagged as potentially Spanish-speaking.

- Respondent request: Respondents who called Westat to request Spanish materials were flagged as Spanish-speaking.

The experiment, based on results from the embedded experiment in Cycle 1 as well as from other studies, had two levels. One was mailing both a Spanish and English questionnaire only to Spanish surname and Linguistically Isolated households (control treatment group). The second was to mail the two questionnaires to all households (experimental treatment group). As shown in Table 1-1 below, the majority of the sample was in the control group. This minimized the risk that the new mailing procedure would significantly impact the overall response rate.

Table 1-1. Sampled households in each mailing category

Mailing Experiment	
Control Treatment: Spanish questionnaire mailed only to Hispanic surname and Linguistically Isolated households	10,045
Experimental Treatment: Spanish questionnaire mailed to all addresses in sample	2,010
Total	**12,055**

A description of the operationalization of the mailing procedures can be found in Chapter 3.

Sample Selection 2

The sample design for the HINTS Cycle 2 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

The sampling frame 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. The sampling frame used for the address sample contained duplicate units because some households receive mail in more than one way. To permit adjustment for this duplication of households in the sampling frame, a question about how many different ways respondents receive mail was included on the survey instrument (see question 026).

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 three explicit sampling strata: the first stratum consisted of addresses in areas with high concentrations of minority population, the second stratum consisted of addresses in areas with low concentrations of minority population, and the third stratum consisted of addresses located in counties comprising Central Appalachia regardless of minority population. The high and low minority strata were formed using the block group level characteristics from the 2010 Decennial Census Summary File. Addresses in census block groups that had a population proportion of Hispanics or African Americans that equaled or exceeded 40 percent were assigned to the high-minority stratum. All the remaining addresses were assigned to the low-minority stratum. Addresses in counties comprising Central Appalachia were assigned to the Central Appalachia stratum regardless of minority status.

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. Central Appalachia is an add-on area with separate precision requirements. It is in a separate stratum to ensure an adequate sample size to meet these precision requirements.

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 2 was 12,055: 7,490 from the high minority stratum, 4,350 from the low minority stratum, and 215 from the Central Appalachia stratum. The high-minority stratum's proportion of the sampling frame was 24.6 percent, and it was oversampled so that its proportion of the sample was 62.1 percent. The Central Appalachia stratum's proportion of the sampling frame was 0.8 percent, and it was moderately oversampled so that its proportion of the sample was 1.8 percent. Conversely, the low minority stratum comprised 74.6 percent of the sampling frame, but made up just 36.1 percent of the sample.

To carry out the data collection experiment described in Chapter 1, the address sample was divided up into two groups with one out of every six sample addresses assigned to the experimental treatment group and the remaining 5/6 of the sample addresses assigned to the control treatment group. The experimental treatment group represented the sample that received both English and Spanish versions of the questionnaire, and the control treatment group represented the sample where only likely Hispanic households (those that matched to the Hispanic surname file or resided in a high Spanish linguistically isolated area) received both versions of the questionnaire. All other households received the English version. This assignment rate was made uniform across all three strata. The table below summarizes the address sample for Cycle 2, showing the number of sample addresses by sampling stratum and treatment group.

Table 2-1. Cycle 2 Sample Sizes by Stratum and Treatment Group

Stratum	Total	Control Treatment Group	Experimental Treatment Group
High minority areas	7,490	6,241	1,249
Low minority areas	4,350	3,625	725
Central Appalachia	215	179	36
Total	12,055	10,045	2,010

2.4 Within-Household Sample Selection

The second-stage of sampling consisted of selecting one adult within each sampled household. As noted in Chapter 1, data collection for Cycle 2 implemented only the Next Birthday Method to select the one adult in the household. Questions were included on the survey instrument to assist the household in selecting the adult in the household having the next birthday.

Data Collection 3

Data collection for Cycle 2 started on October 9, 2012 and concluded on January 11, 2013. The survey was conducted exclusively by mail with a $2 monetary incentive to encourage participation. The specific mailing procedures and outcomes are described in detail below.

3.1 Mailing Protocol

A total of four mailings were sent out as part of Cycle 2. All households in the sample received the first mailing, while only non-responding households received subsequent mailings. Most households received one survey per mailing (in English), while households that were flagged as potentially Spanish-speaking, or who were part of the experimental treatment group, received two surveys per mailing (one English and one Spanish).

The mailing protocol 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. The second mailing was sent via USPS Priority Mail, while all other mailings were sent First Class. The final mailing contained the shortened version of the Spanish questionnaire. Unlike in Cycle 1, the Cycle 2 mailings targeted potentially Spanish-speaking households starting with the initial mailing. The contents of the mailings are further described in Table 3-1 below. Cover letters in English can be found in Appendix A and cover letters in Spanish are in Appendix B. All cover letters include a list of Frequently Asked Questions (FAQs) on the back. These FAQs in both English and Spanish are in Appendix C.

Table 3-1. Mailing Protocol

Mailing	Date(s) Mailed	Mailing Method	Materials	Type of Recipients	
				Control Treatment Group	Experimental Treatment Group
Mailing 1	Oct 9-12, 2012	1st Class Mail	English cover letter with FAQs English questionnaire Return envelope $2 bill	All sampled households that were not identified as possibly Spanish-speaking	N/A
			English and Spanish cover letters with FAQs English and Spanish questionnaires Return envelope $2 bill	All sampled households that were identified as possibly Spanish-speaking	All sampled households
Reminder	Oct 15-17, 2011	1st Class Mail	Reminder/thank you postcard	All sampled households	All sampled households
Mailing 2	Nov 9, 2012	USPS Priority Mail	English cover letter with FAQs English questionnaire Return envelope	Non-responding households that were not identified as possibly Spanish-speaking	N/A
			English and Spanish cover letters with FAQs English and Spanish questionnaires Return envelope	Non-responding households that were identified as possibly Spanish-speaking	Non-responding households
Mailing 3	Dec 4, 2012	1st Class Mail	English cover letter with FAQs English questionnaire Return envelope	Non-responding households that were not identified as possibly Spanish-speaking	N/A
			English and Spanish cover letters with FAQs English and Spanish questionnaires (short version Spanish) Return envelope	Non-responding households that were identified as possibly Spanish-speaking	Non-responding households

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

Table 3-2. Number of Packets per Mailing

Mailing	Number of Packets Mailed		
	English Only	English and Spanish	Total
Mailing 1	8,556	3,501	12,057
Mailing 2	5,859	2,687	8,546
Mailing 3	5,098	2,382	7,480
Total	19,513	8,570	28,083

3.2   Spanish Language Households

In Cycle 2, extra efforts were made to elicit responses from Spanish-speaking households compared to Cycle 1. Mailings that included both English and Spanish materials were sent to households who:
- Were flagged as living in a linguistically-isolated area;
- Were flagged as having a potentially Hispanic surname;
- Were part of the experimental treatment group; or
- Called to request materials in Spanish.

Spanish materials were sent to households starting with the first mailing. The outer envelope for these households included a message stating 'Please return within two weeks' in both English and Spanish so these households would know without opening the package that the contents were multi-lingual. Any household that had received only English materials but called to request Spanish materials then received materials in only Spanish for all subsequent mailings.

3.3   In-bound telephone calls

Two toll-free telephone numbers were provided to respondents -- one was used for English calls and one was used for Spanish calls. These numbers were both provided in each mailing. Respondents were told that they could call the number if they had comments, 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 45 calls throughout the Cycle 2 field period (Table 3-3). A majority of the in-bound calls were to request Spanish materials. The rest were refusals or respondents calling in with some form of a comment or a question. Six calls were never able to be resolved because the study staff were never able to reach the respondent.

Table 3-3. Telephone Calls Received

Reason for Call	Number of Calls Received
Request for Spanish questionnaires	13
Refusal	10
Calls to say he/she had returned the questionnaire but received another	5
Wanting to know where we had gotten their contact information	2
Call to say he/she would be filling it out or had filled it out	3
Call asking if they were required to fill it out	2
Other	4
Calls that were never resolved	6
Total	**45**

3.4 Incoming Questionnaires

Field room staff receipted into the SMS all received questionnaires using each questionnaire's unique barcode. The SMS tracked each received questionnaire as well as the status of each household (nonresponsive or complete). 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 our 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 final status of all households can be found in Table 3-4 below. Incompletes, those who answered less than 50 percent of the required items, are coded as nonrespondents. See section 4.5 for a description of how this coding was conducted.

Table 3-4. Final Household Status

Household-level Results	English	Spanish	TOTAL	
			N	%
Complete	3,509	121	3,630	30.11
Refusal			90	0.75
Undeliverable			1,469	12.19
Nonresponse			6,866	56.95
TOTAL			12,055	100.0

The number of questionnaires returned by date during the field period can be found in Table 3-5 below.

Table 3-5. Survey Response by Date

Date of Mailing	Period of Returns	Number of Returns
Mailing 1: October 9-12	October 10 - October 17	353
Postcard: October 15-17	October 18 - November 10	2,089
Mailing 2: November 9	November 11 - December 7	798
Mailing 3: December 6	December 8 - January 11	390
	Total	3,630

Data Management 4

After being processed and entered into the SMS, each returned questionnaire was scanned, verified, cleaned, and edited. Imputation procedures were also conducted. These procedures are described below. A list of the variables where initial values were edited, imputed, or recoded can be found in Appendix D (Variable Values and Edits).

4.1   Scanning

All completed questionnaires were electronically scanned to capture the survey data and images. 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; and

- Documenting potential partial complete questionnaires or pertinent comments made by respondents in a decision log. Comments in Spanish were reviewed by a Spanish-speaking staff member.

The reviewed surveys were then sent through the high-speed TeleForm scanner to capture the responses. TeleForm read the form image files and extracted data according to HINTS Cycle 2 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 were recorded in a data decision log. The decision log contained the respondent ID, the value triggering the edit, the updated value, and the reason for the update. A total of 165 entries were made into the data decision log during the course of data processing. The majority of these were attributed to illogical responses on a numeric question.

A 10% quality control check was then conducted on the scanned data and the electronic images of the data forms. 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 scanned, the data were cleaned and edited. General cleaning and editing activities are described briefly below, with more detailed information found in Appendix E (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. Appendix D (Variable Values and Edits) provides a list of the values and their definitions.
- 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. Appendix D (Variable Values and Edits) provides details about when and how these recodings took place.
- Variables were designed to summarize the response for the Hispanic ethnicity, race and cancer questions. These variables, hisp_cat, race_cat2 and cancer_cat, indicated each response selected for respondents selecting only one response, and a multiple category for all of the respondents who answered multiple responses.
- Derived variables were created to reflect each response recorded for the occupation question in order to capture this information when respondents did not follow the instruction to mark only one response. For some variables for which a respondent was expected to select only one response, rules, as described in Appendix D (Variable Values and Edits), were used to determine which response was retained. For other variables, imputation, as described below, was carried out.
- 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, responses using only one unit of measurement and other response errors. A list of the rules applied to clean these variables is included in Appendix E (Data Editing Procedures).
- "Other, specify" responses were examined, cleaned for spelling errors, categorized, and upcoded into preexisting response codes when applicable.

4.3   Imputation

For two variables featuring a mark-only-one response instruction (WhereSeekHealthInfo and StrongNeedCancerInfo), imputation was carried out for the questionnaires in which multiple responses were recorded. The imputation process used was the same as that carried out for the HINTS 2007 and HINTS 4 Cycle 1 surveys. Responses where a missing value of -5 (multiple responses selected) was applied were imputed. This occurred for 358 respondents for WhereSeekHealthInfo and 149 respondents for StrongNeedCancerInfo. A 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.

In addition, hot-deck imputation was used to replace missing responses with imputed data for items used in the raking procedure of the Cycle 2 weighting process. Hot-deck imputation is a data processing procedure in which a case with a missing value for a specific variable is assigned 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. Items imputed using the hot-deck approach were those involving the following characteristics: age, gender, educational attainment, marital status, race, ethnicity, health insurance coverage, and cancer diagnosis.

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. 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. Edits were carried out to reconcile different values reported within households and correct differences with the receipted number of returned questionnaires. A detailed list of the steps carried out to identify the number of adults in each household is included in Appendix E (Data Editing Procedures).

4.5   Survey Eligibility

Of the 3,689 questionnaires received, 55 were incompletely-answered and 4 were determined to be ineligible since they were duplicates from the same household (n=3), or they provided clearly subverted data (n=1). A total of 3,630 questionnaires were coded as eligible questionnaires. Details about the eligibility rules are below.

Definition of a Complete and Partial Complete Questionnaire
The procedures for determining whether or not a returned questionnaire was complete were similar to those used for the HINTS 2007 and HINTS 4 Cycle 1 surveys. A complete questionnaire was defined as any questionnaire with at least 80 percent of the required questions answered in Sections A and B. A partial complete was defined as when between 50 percent and 79 percent of the questions were answered in Sections A and B. There were 48 partially complete questionnaires. Both partially-completed and completely-answered questionnaires were retained. The 55 questionnaires with fewer than 50 percent of the required questions answered in Sections A and B were coded as incompletely-filled out and discarded.

Eligibility of Multiple Questionnaires from a Household
Three households returned more than the requested number of completed questionnaires. The procedures to deal with this issue followed the same guidelines that were used for households using the Next Birthday method in Cycle 1:
- If the same respondent returned multiple questionnaires, the first questionnaire received was retained unless it was a reduced version and a subsequent return was a full version.
- If a return date was unavailable for questionnaires from the same respondent that were the same version, questionnaires with fewer substantive questions omitted were 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.

---

[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.

4.6   Codebook Development

Following cleaning and editing, a detailed codebook including frequencies was created for both the weighted and unweighted data. The codebooks define all variables in the dataset, provide the question text, list the allowable codes, and explain the inclusion criteria for each item. The English and both versions of the 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 Cycle 2 received a full-sample weight and a set of 50 replicate weights. The full-sample weight is the weight which is used to calculate population and subpopulation estimates from the HINTS data collected in Cycle 2. Replicate weights are used to compute standard errors for these estimates.

The weighting process encompasses the procedures used to create the final full-sample and replicate weights for the survey respondents. 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 for Cycle 2 consisted of the following steps. Each of these steps is described in more detail in the sections 5.1 through 5.4.

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

The replicate weights were calculated using the delete one jackknife (JK1) method. The steps used to derive these weights were aimed at reflecting the features of the sample design, so that when the JK1 jackknife variance estimation procedure is implemented, approximately unbiased estimates of sampling variance are obtained. In addition, the various weighting procedures carried out on the full-sample weight were repeated on each set of replicate weights to appropriately reflect the impact

of the weighting adjustments on the sampling variance of a survey estimate. Details of replication used for variance estimation can be found in section 5.5.

5.1   Household Base Weights

The initial step in the weighting process was 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 oversampled strata are smaller than those in strata that were not oversampled. In Cycle 2, the base weights for households in the high minority and Central Appalachia strata were roughly 1/5 the size of those in the low minority stratum.

If two different addresses led to the same household – for example, if a household receives mail via both a street address and a post office box – that household had twice the chance of selection of a household with only one address (and should therefore receive half the normal weight). An additional adjustment was made to the base weights of households that had multiple ways of receiving mail (as determined by the answer to survey question O26).

5.2   Household Nonresponse Adjustment

Nonresponse is generally encountered to some degree in every survey. The first and most obvious effect of nonresponse is to reduce the effective sample size, which increases the sampling variance. In addition, if there are systematic differences between the respondents and the nonrespondents, there also 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 2 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 2 were:
- Sampling stratum (High Minority; Low Minority; Central Appalachia)
- Treatment group (Only likely-Hispanic addresses receive Spanish questionnaire; All addresses receive Spanish questionnaire)
- 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 as follows:

$$HH\_NRAF(b) = \frac{\sum_{S(b)} HH\_BWT_i}{\sum_{C(b)} HH\_BWT_i},$$

where HH_BWT_i is the base weight for sampled household i, $S(b)$ is the set of all eligible sampled households) in nonresponse cell b, $C(b)$ is the set of all cooperating sampled households in cell b, and $HH\_NRAF(b)$ is the household nonresponse adjustment factor for nonresponse cell b.

---

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

The household nonresponse adjustment factors ranged from a low of 1.86 to a high of 5.45, and averaged 2.92 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. Since in Cycle 2 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.

<u>Control Totals</u>

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. The ACS estimates are available via the Internet. Westat used the 2011 ACS estimates that were 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. These variables came from the 2011 National Health Information Survey (NHIS) and correspond to questions asked in the HINTS survey. They were:

- Percent With Health Insurance
- Percent Ever Had 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 Ever Had Cancer calculated from HINTS data agrees with the estimate calculated from NHIS 2011 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 2 data set, a set of 50 replicate weights was assigned to each responding adult. To illustrate how the replicate weights 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.

Response Rates   6

Response rates were calculated using the RR2 formula of the American Association of Public Opinion Research (AAPOR).

6.1   Overall Response Rate

Table 6-1 shows the response rate calculation. These data have been weighted to account for the oversampling of addresses in high-minority areas. The overall response rate was 39.97 percent; however this differed significantly by strata. The High Minority strata had the lowest response rate (28.67 percent) and the Low Minority had the highest (43.82 percent). The percent of undeliverable households ranged from 11 percent to 14 percent across strata, with the Central Appalachia strata having the highest undeliverable rate.

Table 6-1. Response rate calculations by Strata

Response class	High minority	Low minority	Central Appalachia	Overall
Total sample*	34,876,461	102,222,537	1,111,484	138,210,482
Respondents	8,712,131	39,902,039	315,351	48,929,521
Nonrespondents	21,680,214	51,158,267	635,872	73,474,354
Undeliverable	4,484,116	11,162,231	160,260	15,806,608
Total Households	30,392,345	91,060,306	951,224	122,403,874
Percent Undeliverable	12.86%	10.92%	14.42%	11.44%
Household response rate	**28.67%**	**43.82%**	**33.15%**	**39.97%**

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

6.2   Experimental Effects on Response Rates

Table 6-2 and Table 6-3 summarize the response rate outcomes by strata for each experimental treatment group. Households that were in the control treatment group (n = 10,045) were sent Spanish questionnaires in every mailing only if their household was linked to a linguistically isolated area or Hispanic surname according to the sampling frame. All households in the experimental treatment group (n = 2,010) were sent Spanish questionnaires along with all mailings.

The control group yielded a higher response rate than the experimental group with the difference being about 2.5 percentage points. The experimental treatment appeared to slightly suppress response within the Low Minority strata where the response rate was about 3 percentage points lower than in the control treatment. Response rates appear to be higher in the Central Appalachia region in the experimental treatment; however this response rate is estimated based on a small number of households relative to the other two strata and therefore is less reliable.

Table 6-2. Response rate calculations for Control Treatment by Strata

Response class	High minority	Low minority	Central Appalachia	Overall
Total sample*	29,060,613	85,185,448	925,375	115,171,436
Respondents	7,296,584	33,721,687	253,315	41,271,586
Nonrespondents	18,066,845	42,298,981	532,478	60,898,304
Undeliverable	3,697,184	9,164,779	139,582	13,001,545
Total Households	25,363,429	76,020,668	785,793	102,169,891
Percent Undeliverable	12.72%	10.76%	15.08%	11.29%
Household response rate	**28.77%**	**44.36%**	**32.24%**	**40.40%**

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

Table 6-3. Response rate calculations for Experimental Treatment by Strata

Response class	High minority	Low minority	Central Appalachia	Overall
Total sample*	5,815,848	17,037,090	186,109	23,039,046
Respondents	1,415,547	6,180,351	62,036	7,657,934
Nonrespondents	3,613,369	8,859,287	103,394	12,576,049
Undeliverable	786,932	1,997,452	20,679	2,805,063
Total Households	5,028,916	15,039,638	165,430	20,233,983
Percent Undeliverable	13.53%	11.72%	11.11%	12.18%
Household response rate	**28.15%**	**41.09%**	**37.50%**	**37.85%**

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

References

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.

Appendix A

English Cover Letters and Postcards

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 what health information people want to know and where they try to find it. By completing this questionnaire, 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 that <u>the adult in your household with the next birthday</u> complete and return this questionnaire 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, will conduct the survey. If you have any questions about HINTS {or if you need more questionnaires}, or if you would like to complete this survey in a language other than English or Spanish, please call Westat toll-free at 1-888-738-6805.

Thank you in advance for your cooperation.

Sincerely,

Bradford W. Hesse, Ph.D.

HINTS Project Officer
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.
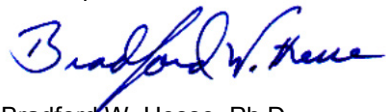
POSTCARD TEXT

A few days ago you should have received a questionnaire packet asking for your household's participation in the Health Information National Trends Survey. By completing the questionnaire, 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 questionnaire. If that adult has already completed the questionnaire and returned it to us, please accept my sincere thanks. If that adult has not yet completed and returned the questionnaire, we ask that he or she please do so as soon as possible.

Your household's participation is important to the study's success.

Sincerely,

Bradford W. Hesse, Ph.D.
HINTS Project Officer
National Institutes of Health
U.S. Dept. of Health and Human Services

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 what health information people want to know and where they go to find it. 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 questionnaire. 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 questionnaire was misplaced, an additional copy is enclosed.

In order to make sure we get responses from a random sample of people, we ask that <u>the adult in your household with the next birthday</u> complete and return this questionnaire in the next two weeks.

Additional information about HINTS is available at: www.hints.cancer.gov. If you have any questions, or would like to complete this survey in a language other than English or Spanish, please call Westat toll free at 1-888-738-6805.

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

Sincerely,

Bradford W. Hesse, Ph.D.

HINTS Project Officer
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.

Appendix B

Spanish Cover Letters

FIRST MAILING

Estimado residente de {City}

Le escribimos para invitarlo 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 qué información sobre la salud les interesa saber a las personas y dónde tratan de buscarla. Complete este cuestionario para ayudar 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 <u>el adulto en su hogar con el próximo cumpleaños</u>, complete y devuelva este cuestionario 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 o le gustaría completar esta encuesta en otro idioma distinto al inglés o español, llame a Westat al siguiente número de teléfono libre de cargo, 1-888-738-6812.

Gracias de antemano por su cooperación.

Atentamente,

Bradford W. Hesse, Ph. D.

Oficial del Proyecto 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.

6

SECOND MAILING

Estimado residente de {City}:

Recientemente lo invitamos a participar en una importante encuesta nacional patrocinada por el Departamento de Salud y Servicios Humanos de Estados Unidos. El objetivo de la Encuesta Nacional de Tendencias de Información sobre la Salud (HINTS, por sus siglas en inglés) es averiguar cuál es la información sobre la salud que las personas quieren saber y dónde van a buscarla. 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 cuestionario completado. 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 que su cuestionario 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 este cuestionario en las próximas dos semanas.

Usted podrá encontrar más información sobre HINTS en el sitio web hints.cancer.gov. Si usted tiene preguntas o le gustaría completar esta encuesta en otro idioma distinto al inglés o español, llame a Westat al número libre de cargo, 1-888-738-6812.

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

Atentamente,

Bradford W. Hesse, Ph. D.

Oficial del Proyecto 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.

THIRD MAILING

Estimado residente de {City}:

Recientemente lo invitamos a participar en una importante encuesta nacional patrocinada por el Departamento de Salud y Servicios Humanos de Estados Unidos, la Encuesta Nacional de Tendencias de Información sobre la Salud (HINTS, por sus siglas en inglés). El completar esta encuesta nos ayudará a mantenerlos mejor informados en asuntos de salud que usted y su familia consideran importantes.

Si usted ya 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.

Si aún no ha tenido la oportunidad de completar la encuesta, comprendemos que a veces es difícil encontrar el tiempo para participar en un estudio como HINTS. Para reducir la cantidad de tiempo que va a tomarle, hemos incluido una versión más corta de la encuesta. Esta versión corta se concentra en asuntos que nos informan la manera como personas como usted buscan y utilizan información sobre la salud.

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 este cuestionario en las próximas dos semanas.

Usted podrá encontrar más información sobre HINTS en el sitio web hints.cancer.gov. Si usted tiene preguntas o le gustaría completar esta encuesta en otro idioma distinto al inglés o español, llame a Westat al número libre de cargo, 1-888-738-6812

Gracias de antemano por contribuir con este importante estudio nacional.

Atentamente,

Bradford W. Hesse, PhD.

Oficial del Proyecto HINTS
Institutos Nacionales de la Salud
Departamento de Salud y Servicios Humanos de EE.UU.

La Encuesta de 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

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: The study concerns health and how people receive health information. For example, we will ask how you usually get information about how to stay healthy, the sources of information you most trust, and how you might like to get such information in the future. We will also ask about your beliefs on what contributes to good health, how best to prevent cancer, your participation in various health-related activities, and related topics.

Q: How will the study results be used? What will be done with my information?

A: Findings will help the U.S. Department of Health and Human Services promote good health and prevent disease by determining ways of better communicating accurate health information to Americans.

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: Your participation is voluntary, and you may refuse to answer any questions or withdraw from the study at any time. However, your answers are very important to the success of this study and will represent thousands of others. Getting an answer 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.

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 connected to your name or any other information that could identify you or your household, to the extent provided by law. The completed questionnaires will be stored in a separate file with restricted access. Both the paper and electronic versions of the information 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? Is this study approved by the Federal Government?

A: The study is sponsored by the U.S. Department of Health and Human Services. The study has been approved by the Office of Management and Budget (OMB), the office that reviews all federally-sponsored surveys. The OMB approval number assigned to this study is 0925-0538.

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. El estudio ha sido aprobado por la Oficina de Administración y Presupuesto (OMB, por sus siglas en inglés). Esta oficina revisa todas las encuestas patrocinadas federalmente. El número de aprobación asignado por la OMB para este estudio es 0925-0538.

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 Edits

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 SeekHealthInfo=2 'no' and WhereSeekHealthInfo was missing, WhereSeekHealthInfo 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 SeekHealthInfo=2 'no' and WhereSeekHealthInfo was not missing, WhereSeekHealthInfo 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 CancerTrustRadio had values 3 'a little' and 2 'some', CancerTrustRadio 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., SeekHealthInfo) was missing and variables up until the next question respondents who skipped were instructed to answer (e.g. CancerConfidentGetHealthInf) were missing (e.g., LookElsewhere= 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 SeekHealthInfo was missing, it was assigned the value -9.

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.

Variable	Editing Rule	Description of Rule
SeekHealthInfo SeekCancerInfo UseInternet Internet_Other InternetCancerInfoSite Smoke100 FamBetween9and27 DrShouldPSATest EverHadCancer UndergoCancerTreatment BornInUSA	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.
WhereSeekHealthInfo_IMP StrongNeedCancerInfo_IMP	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. Missing values were not imputed for. Flags (indicated by suffix '_IFlag') indicate which values were imputed and which values were original.
WhoLookingFor	Edits for multiple responses	Multiple responses (e.g., 'myself', 'someone else') were recoded to the logically applicable third response option ('both myself and someone else').
Internet_DialUp Internet_BroadBnd Internet_Cell Internet_WiFi MedConditions_Diabetes MedConditions_HighBP MedConditions_HeartCondition MedConditions_LungDisease MedConditions_Arthritis MedConditions_Depression ProstateCa_PSATest ProstateCa_SlowGrowing ProstateCa_SideEffects CancerTx_Chemo CancerTx_Radiation CancerTx_Surgery CancerTx_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 not check a 'no' response option for any sub-item in the question, the sub-items that were missing a response were re-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
HealthInsurance_I GenderC_I EverHadCancer_I Age_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. Flags (indicated by suffix '_IFlag') indicate which values were imputed and which values were original.
TimesModerateExercise FreqGoProvider UseMenuCalorieInfo HowLongFinishTreatment_Cat	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 'never', 'none' or '0 days' 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 all respondents were supposed to answer. However, missing values were not substituted with other values if the filter question was not answered but the follow-up question was answered.
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. A list of the rules applied to clean the variables is included in 'AppendixF – Data Editing Procedures.'
Cancer_cat Race_cat2 Hisp_Cat	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.
Employed Unemployed Homemaker Student Retired Disabled OtherOCC	Derived variables for multiple responses	For the variable OccupationStatus, derived variables were created to indicate each response selected, showing the unique response for respondents selecting one occupation, and showing each response for respondents who did not follow the 'mark only one' response instruction.

Variable	Editing Rule	Description of Rule
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 and which values were original.
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.

Appendix E

Data Editing Procedures

Rules Applied to Edit Height Variables:

1. 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).
2. 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").
3. 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).
4. If HEIGHT_Feet>3 and HEIGHT_Feet<7 and HEIGHT_Inches = 20, 30, etc. thru 90 then the trailing 0 was removed
5. 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).
6. 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).
7. 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")
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 putting measurements in the wrong boxes, resulting in edited values from 4'7" to <7 feet)

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. Only four households had missing values of R_HHAdults that needed to be imputed.