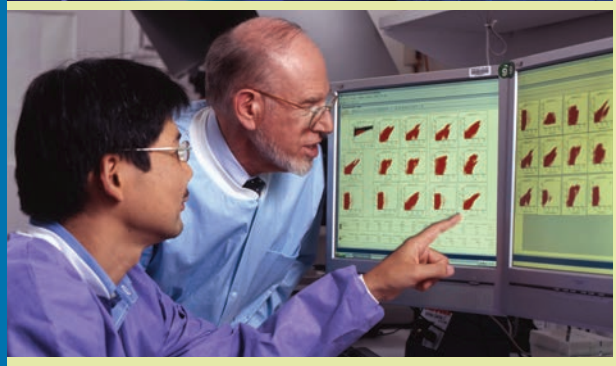
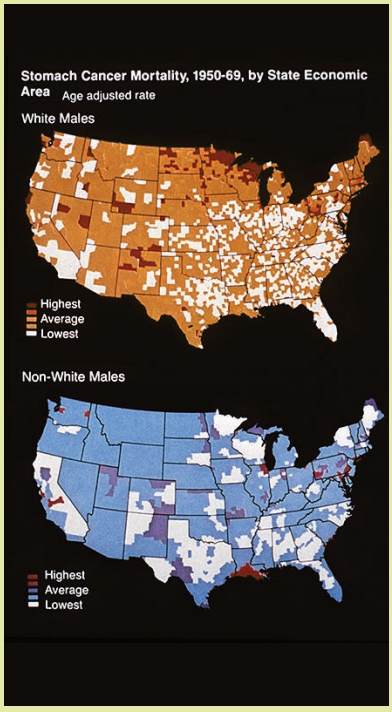
National Cancer Institute

Health Information National Trends Survey

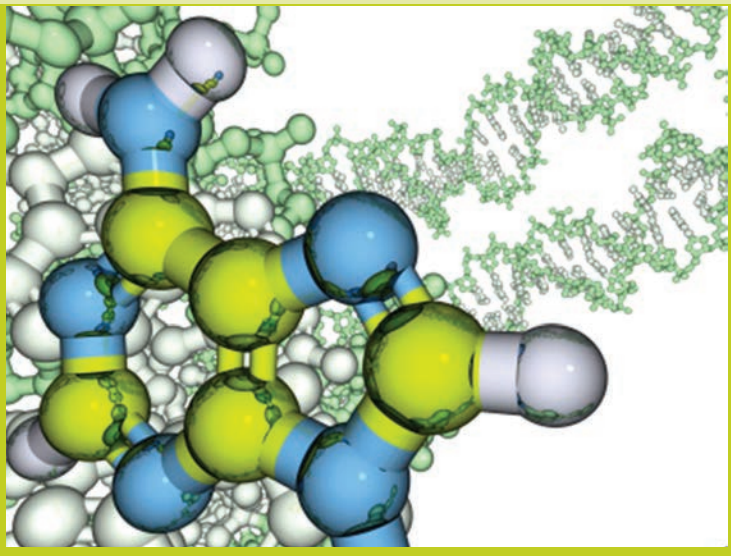
Integrative Analytic Methods Using Population-Level Cross-Sectional Data

June 2013





Richard P. Moser, PhD
Sana Naveed, MPH
David Cantor, PhD
Kelly D. Blake, ScD
Lila J. F. Rutten, PhD, MPH
A. Susana Ramírez, PhD, MPH
Benmei Liu, PhD
Mandi Yu, PhD



U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Health Information National Trends Survey hints

Table of Contents

Foreword: Celebrating the 10-year Anniversary of Health Information National Trends Survey (HINTS)

On July 27, 2012, the Centers for Disease Control and Prevention in the United States published a vision for public health surveillance in the 21st Century. The report was premised on the notion that data about disease and wellness, and the environmental conditions that lead to both, are essential for public health action to occur. "In public health, we can't do anything without surveillance," said U.S. Surgeon General David Satcher in a quote leading off the report; "that's where public health begins." Much of the hard, labor-intensive work of the 20th Century went into building sustainable data systems that could inform public health research and policy.

The Health Information National Trends Survey (HINTS) took its place within the pantheon of federally funded surveillance systems in 2001, with its first data collection completed in 2003. Its objective was to gather information on how the information environment was itself changing through the introduction of new communications technologies. Its goal was to offer insights to communication scientists, to public health practitioners, and to the general public on ways that this new environment could be leveraged to achieve national health goals. Always just a little bit ahead of its time, the program took the bold step of using changes within the communication environment to enable "crowdsourcing" of HINTS analysis. In a way, the program was offering a foreshadowed glimpse of what the President's Council of Advisors on Science and Technology called the "promise of a digital future" for science, citizens, and entrepreneurs. It would become an era of accelerated discovery through team science; an era of "Big Data."

Today, 12 years after the HINTS program began and 10 years after the first data collection was completed, enthusiasm for connecting data systems for the benefit of the public health community has increased. HINTS data can be found among the many publicly available datasets made available through the Department of Health and Human Services' data.gov. What is needed now, though, is an analytic strategy that will help connect these datasets in meaningful and actionable ways. That is the purpose of this analytic guide for integrative data analysis using the HINTS datasets. The guide offers invaluable insights, along with practical suggestions, for how to connect a decade's worth of survey data over time and how to leverage the benefits of a complementary and divergent set of public health surveillance resources to gain new insights into the health information environment.

I especially want to acknowledge the extraordinary leadership and dedication of the HINTS data analytic team for providing all of us with this much needed guide at a time of expanding analytic

opportunities. Dr. Richard Moser's vision to provide the community with tools for analyzing the HINTS data in new, integrative ways will elevate the productivity of the entire HINTS community over the next decade. In addition, the contributions from David Cantor and Teri Davis, who have been faithful stewards of the HINTS data collection since its inception, along with the programmatic insights of cancer control leaders Lila Rutten, Ellen Beckjord, Kelly Blake, and Sana Naveed, help anchor these tools on real world examples of extreme public health importance. Similarly, statistical insights from Benmei Liu and Mandi Yu from the National Cancer Institute's Surveillance Research Program help ensure that the tools are rigorous in their application and forward-facing in helping to unravel the future of public health surveillance.

Through the labor and passion of each of these contributors, the guide offers a practical roadmap for navigating the new environs of an integrative approach to data analysis. Armed with the techniques described in this guide, users of the HINTS data should be well equipped: (a) to make the conceptual leap from analyses focused on one point in time to an integrative set of analyses aimed at detecting macro trends; (b) making the technical leap of merging multiple datasets both for boosting sensitivity and for detecting trends; (c) mapping geographic distributions of HINTS knowledge constructs; (d) creating model-based state estimates; and (e) for using imputation techniques for transcending some of the limitations of the national survey.

To be sure, movement into an integrative data analytic environment may provoke anxiety in many of us who were trained under the more reductionistic approaches to data analysis; and, as always, there are limits to the types of questions that can be asked of public health surveillance systems with these new approaches. Nevertheless, many of us are equally intrigued by the new analytic capacities that a "disruptive innovation" in our public health and biomedical systems can offer. We knew as a community that we crossed that Rubicon as soon as our interconnected data systems logged the three billionth base pair of the human genome. The human genome was only the beginning, though. Many of us with roots in the population sciences understand that the molecular determinants of health account for only a small amount of overall variance when compared to the influence of the social and physical environment. New scientific exigencies will demand equal attention to the mysteries of a surrounding "exposome." Our hope is that this guide will serve as an insightful – and pragmatic – map for integrating HINTS analyses into our broader understanding of the communication pathways that can be strengthened to improve public health across many levels.

Bradford W. Hesse, PhD
Chief, Health Communication and Informatics Research (NCI) and
Project Director, the Health Information National Trends Survey

Introduction

The Health Information National Trends Survey (HINTS) is a nationally representative cross-sectional survey that has been administered every few years by the National Cancer Institute since 2003. The HINTS target population is all civilian non-institutionalized adults aged 18 or older in the United States. HINTS is unique in that it collects data on the American public's need for, access to, and use of health-related information and health-related behaviors, perceptions and knowledge (Hesse, Moser, Rutten, & Kreps, 2006; Nelson, Kreps, Hesse, Croyle, Willis, Arora et al., 2004). The primary goal of HINTS is to monitor changes in the rapidly evolving field of health communication.

The most recent version of HINTS administration (referred to as HINTS 4) involves four separate mail-mode data collection cycles in a three-year field period that began late in 2011 and will extend into 2014. The first cycle of HINTS 4 data were made available to the public in early 2012 and will be one of the data sets used to demonstrate different analytic techniques in this report (Cycle 2 data will be available in June, 2013). For more information about HINTS, see a more detailed methodology report (http://hints.cancer.gov/docs/HINTS4_Cycle1_Methods_Report_revised_Jun2012.pdf), and to download the latest iteration of the HINTS data, please visit http://hints.cancer.gov/. Previous iterations used random-digit-dial (RDD) samples and telephone interviews (HINTS 1 [2003], HINTS 2 [2005]) or a mixed-frame (RDD/Postal Address) and mixed- mode (telephone interview/self-administered mail survey) approach (HINTS 3 [2008]).

Final sample weights were created and assigned for each respondent to account for all of the stages of selection and for attrition from noncontacts, nonresponse, and noncoverage. These weights are designed to provide approximate unbiased estimators of population totals. Replicate weights are also provided to obtain accurate variance estimation for statistical testing (i.e., any analysis that involves a *p* value or a confidence interval). The replicate weights are based on the jackknife replication method with R=50 replicate weights for each survey year (note that HINTS 3 provides separate sample and replicate weights to analyze the RDD sample, the mail sample, or a combination of both). The replicate weights are formed by deleting a selected portion of the original sample (approximately $1/50^{th}$) and reweighting the remaining sample to match the population totals. For more information about how the weights were constructed, see the respective methodology reports for each HINTS iteration found on the HINTS website.

Data Integration

There is a growing awareness that to understand the most intractable health problems vexing this country (e.g., tobacco use, obesity, cancer incidence, heart disease), our analytic methods need to be applied to original data that have been integrated from multiple studies. The scientific process will be more efficient, and discovery can advance more quickly by combining data to create a cumulative knowledge base, as opposed to the traditional way of doing science, where scientists work independently and do not integrate information across studies. These datasets can then be analyzed using statistical techniques that have been termed integrative data analysis (IDA) (Curran & Hussong, 2009). By combining across datasets, researchers can increase sample sizes (increasing statistical power), which can be especially useful when trying to get robust estimates for hard-to-reach populations. Data integration also allows for replication of previous results, comparisons across time (where possible), or assessing for differences between samples. There are, of course, multiple statistical and measurement issues that need to be addressed in order to ensure that data integration is possible. For example, analyses must account for different between-study heterogeneity from the databases (e.g., sampling; geography; mode) and perhaps most importantly, must address measurement issues. Integration is only possible when the data have common data elements (the same or comparable items/scales across databases) that are measuring the same construct or underlying concept. For survey data, this means making sure that the items/scales ask the same questions with the same response options or ones that can be made comparable (e.g., combining a categorical and continuous item assessing household income). Without measurement comparability, integration is not possible. Fortunately, there are many techniques, such as item-response theory (IRT) that can be used to assess or create comparable data elements that can then be combined with confidence (see Bauer & Hussong [2009] for more information on this subject). Testing for comparability across measures and performing data management to allow for proper integration of data is a non-trivial pursuit and can be time-consuming. However, once the integrated database is created, both traditional and novel statistical techniques can be applied as if one were analyzing a single dataset.

Goals of this Report

This report was designed to educate survey researchers about integrative analytics methods that can be used with HINTS and other similar cross-sectional survey data. It is the authors' hope that, by providing details of different types of analyses and providing statistical code where appropriate in the appendices, survey researchers will begin to apply these techniques to their own work.

This report is divided into different sections to demonstrate several of these integrative techniques:

1. Bridging Across HINTS Iterations That Use Multiple Survey Modes (Section 1);
2. Merging and Analyzing Multiple Iterations of HINTS Mainland Data (Section 2);
3. Merging HINTS Mainland and HINTS Puerto Rico Data (Section 3);
4. Multilevel Determinants of Smoking Behavior: An Integrated Analysis (Section 4);
5. Model-Based State Level Estimates for Cancer Related Knowledge Variables Using HINTS Data (Section 5);
6. Using Imputataion to Augment Multiple Iterations of HINTS Data (Section 6)

Bridging Across HINTS Iterations That Use Multiple Survey Modes

1

The Health Information National Trends Survey (HINTS) has been administered on an ongoing basis since 2003 (HINTS 1 [2003], HINTS 2 [2005], HINTS 3 [2008], and HINTS 4 [2011-2012]). This time period spans a dramatic change in the climate of administering general population surveys, especially those using an RDD sample frame. Response rates for RDD surveys began a steady decline in the 1990s and have continued to drop. During this same time, many households have moved away from the use of traditional landline telephones in favor of mobile phones. As of the end of 2011, approximately 33 percent of adults in the U.S. live in households with access to only a mobile phone (Blumberg & Luke, 2012). While it is possible to administer surveys of users of mobile phones in similar ways as landline phones, this introduces additional complications. The migration to mobile phones made it more difficult to maintain response rates, as well as increased the overall expense of conducting telephone surveys.

At the same time, a new national address sample frame became available. This list is provided by the United States Postal Service (USPS) and contains all addresses to which mail is delivered. Evaluations of this frame made it apparent that it is reasonably complete (Iannacchione, Staab, & Redden, 2003). The primary sources of error are those units that do not have a standard city-style address (e.g., post office boxes). The result is undercoverage in highly rural areas where city-style addresses are less common. The availability of this list made it possible to conduct national surveys by mail, which had not previously been possible (Link, Battaglia, Frankel, Osborn, & Mokdad, 2008).

In 2008, HINTS transitioned to a mail survey using the USPS-based address frame. For this administration, both the mail and RDD surveys were administered to half the sample respectively. The RDD-telephone method was used to allow users the ability to link prior HINTS surveys with the HINTS 3 collection. The mail survey established a baseline for the methodology that would be used in future HINTS administrations. This transition was necessary because the two methodologies are not entirely comparable. By maintaining an RDD component, users are able to examine trends with prior years. In HINTS 4, the entire sample was administered by mail using the address frame. Trends between HINTS 3 and 4 can be estimated using the HINTS 4 mail component.

The use of two modes does introduce a layer of decisions that analysts have to make. However, two modes provide users with the ability to take advantage of the strengths of each methodology. When one method is not clearly better, analysts are able to conduct analyses using both modes to test the robustness of their results. The purpose of this section is to discuss how analysts should approach the decision on which mode to use when estimating trends across survey iterations involving the RDD and mail components.

Sources of Differences Between Modes

The mode of data collection can affect who is approached and who cooperates with a survey request. It also affects how survey respondents understand, process, and answer survey questions (Dillman, 2007; De Leeuw, Hox, & Dillman, 2008). The choice of mode affects multiple sources of survey error. When analyzing HINTS data, there are three general sources of differences that should be considered: 1) coverage, 2) non-response, and 3) mode of communication. With respect to coverage, the RDD sample only included those with a landline telephone. At the time that HINTS was in the field, this was estimated to be approximately 15 percent of all adults, with a heavy concentration of those in the younger age groups. In addition, there is some evidence that the RDD sample also excludes households that are not included with the telephone banks identified in the list-assisted sampling methodology (Fahimi, Kulp, & Brick, 2008). The coverage of the USPS frame is considered more complete. It includes all households that have a mailing address. As noted above, evaluations of this frame have found it to be reasonably complete, with the exception of households that do not have a city-style mailing address (e.g., P.O. box; drop box).

The non-response mechanisms between a mail and telephone survey are also different. When the surveys were run in parallel in HINTS 3, the mail survey had a slightly higher response rate (31% vs. 24%). Increasingly, the general public is reluctant to cooperate on telephone calls from individuals they cannot clearly identify. The mail survey included an incentive as part of the survey request, as well as an express package to follow-up with nonrespondents. For the telephone, an incentive was in the pre-notification letter to those for whom there was a mail address. The survey procedures were also different. On the telephone, a screening interview was completed with an adult in the household and a household member was randomly selected. On the mail survey, all adults were asked to complete the survey. It was left to the person who opened the mail to communicate this to everyone in the household.

The HINTS 3 public use file provides a set of weights for the telephone survey, a set of weights for the mail survey and a weight that combines the telephone and mail survey. Each set of weights are

adjusted for non-response and calibrated to the population totals by age, race, gender, marital status, education, health insurance coverage, and cancer. These adjustments make the final distributions very similar. However, for particular measures of interest, it is possible that these adjustments do not account for differential coverage and/or nonresponse across the modes. The effectiveness of these adjustments depends on the extent the group that was not observed (i.e., not covered or not responding) within the weighting adjustment classes differs from respondents. For example, prior research has shown that those who do not have a landline telephone are significantly different from those with a landline phone for a number of HINTS measures (Han & Cantor, 2008). This implies that even though the telephone weights were adjusted for exclusion of mobile-only households, certain measures may still reflect undercoverage in the telephone sample.

A similar caution should be noted for Hispanics. The telephone mode included an interview translated into Spanish and administered by Spanish-speaking interviewers. For the mail survey, there was a note on the advance material for those wishing to do the interview in Spanish to call a toll free number. This procedure resulted in very few Spanish interviews from the address-frame (n=11). To the extent Hispanics who are Spanish speakers differ from those who speak English, there will be a difference between the two modes. More generally, there may be other differences between the modes that are due to coverage and non-response. As will be discussed below, analysts have the ability to test whether differences between the two modes are important by conducting analyses with both samples.

In addition to differences in coverage and non-response, the two surveys differ in important ways with respect to the channel of communication. In particular, several differences may arise from:

- Sensitive and socially desirable questionnaire items, and
- Aural vs. visual stimulus presented to respondents in each mode.

Sensitive and Socially Desirable Items

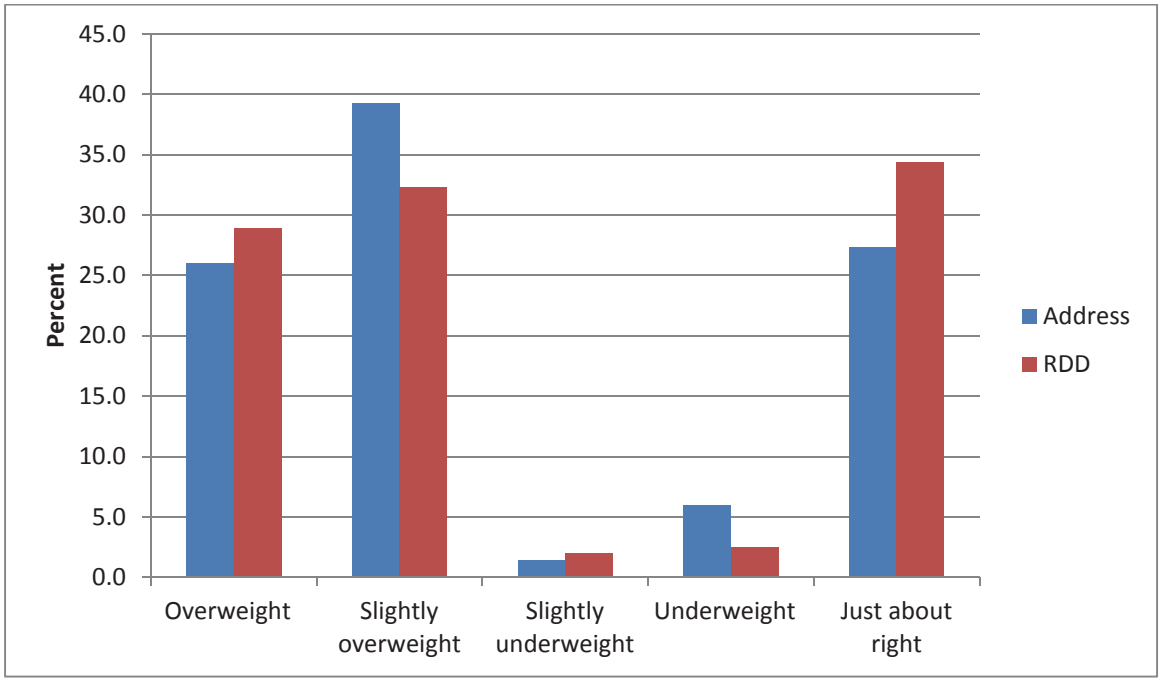
Estimates from questions asking about sensitive content, or content with a socially defined standard for a desirable response may differ by mode of collection. In particular, self-response methods tend to result in higher estimates of sensitive behavior and lower estimates of socially desirable behaviors or attitudes (e.g., Tourangeau & Smith, 1996). HINTS 3 included several questions that fall into this category. Although it should be noted that for many items it is difficult to assess from just the wording what is sensitive or desirable to most individuals (Paulhaus, 2003).

One set of questions that might be affected by social desirability on HINTS 3 are perceptions related to weight. Question BR12 asks respondents:

> There are so many different messages about whether being overweight is harmful to one's health that it is hard to know what weight one should maintain to be healthy. Would you say you are overweight, slightly overweight, underweight, or just about right for you?

There was a small tendency for the respondents to the RDD survey to report they were "just about right" when compared to the address frame (Figure 1-1).

Figure 1-1. HINTS 3 self-reports of perception of weight status*



* Difference between distributions statistically significant at p<.01

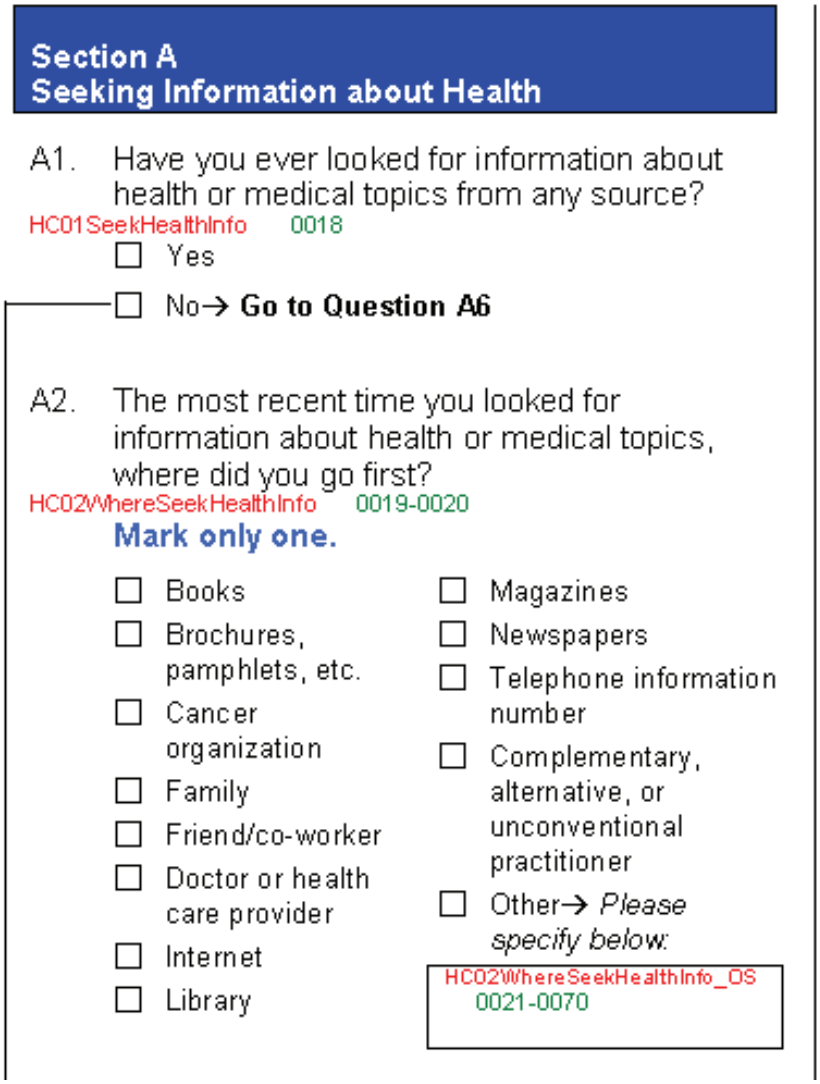
Aural vs. Visual Stimulus by Mode

The presentation of the question, or the stimulus, also can affect responses to a particular survey item. Respondents may respond to a visual presentation of the question differently than an aural presentation for many reasons. One is the extent the visual mode provides greater detail about the question. An example of this type of question is when the answer categories listed on the mail questionnaire provide additional cues to the respondent. These cues might assist the respondent in

understanding the intent of the question, assist in recall, or provide information on what types of answer categories are expected by the investigators.

The visual nature of a mail survey also neutralizes, to some extent, the effect of question order. Because respondents are able to look ahead to other questions, their understanding of the question may be different than in a method where items are presented one at a time. An illustration of both the effects of order and cueing are the two questions that ask about looking for information on health or medical topics. The questions for the mail survey are shown in Exhibit 1-1.

Exhibit 1-1. Mail survey version of seeking information on health and medical topics



Section A
Seeking Information about Health

A1. Have you ever looked for information about health or medical topics from any source?

HC01SeekHealthInfo 0018

☐ Yes

☐ No→ **Go to Question A6**

A2. The most recent time you looked for information about health or medical topics, where did you go first?

HC02WhereSeekHealthInfo 0019-0020

Mark only one.

☐ Books
☐ Brochures, pamphlets, etc.
☐ Cancer organization
☐ Family
☐ Friend/co-worker
☐ Doctor or health care provider
☐ Internet
☐ Library
☐ Magazines
☐ Newspapers
☐ Telephone information number
☐ Complementary, alternative, or unconventional practitioner
☐ Other→ *Please specify below:*

HC02WhereSeekHealthInfo_OS
0021-0070

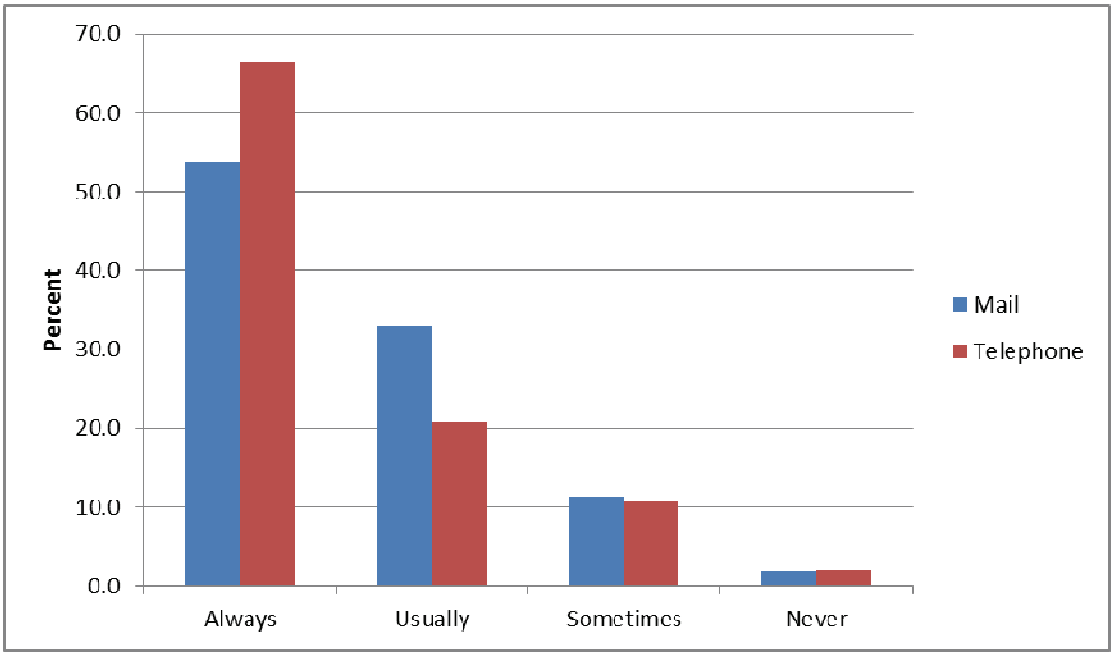
On the mail survey, respondents can answer A1 after they have read through the list of sources in A2. On the telephone survey, respondents answer the first question without this knowledge. As a result, the number of people who answer the first question affirmatively is much different between the two modes: 77 percent on the mail vs. 61 percent on the telephone.

A second consequence of differing aural and visual stimuli are response order effects. This refers to respondents selecting response options in different ways, depending on the channel of communication. For lists of unrelated categories (e.g., see A2 in Exhibit 1-1), there may be a tendency for respondents to the mail survey to pick categories toward the beginning of the list. If the telephone interviewer reads the categories for the respondent, there may be a tendency for respondents to pick the latter categories. It should be noted, however, that the HINTS telephone survey did not have many questions like this. On the telephone, questions with lists like in A2 were open-ended questions (i.e., categories were not read out). The interviewer coded the responses into one of the categories. In these situations, interviewers may use the categories to probe respondents who are having a difficult time coming up with an answer. If that happens, respondents may be more likely to report those items that are probed the most often. For example, McBride et al. (2010) report that telephone interviewers were most likely to probe with the "internet" category to question A2 above. Perhaps as a result, this category is more frequently reported on the telephone survey.

A second type of response order effect is for ordinal response scales, such as "strongly agree, agree, disagree, strongly disagree." HINTS has a large number of these types of questions using various scales (approximately 50 questions). Some survey methodologists claim that telephone respondents provide more extreme answers to these items (Groves & Kahn, 1979; Tarnai & Dillman, 1992; Christian Dillman, & Smyth, 2008; Dillman, Phelps, Tortora, Swift, Kohrell, & Berck, 2009). Others have argued that telephone respondents tend to acquiesce, due to the presence of an interviewer (e.g., see Ye Fulton, & Tourangeau, 2011). Regardless of the argument, this points to potential differences for questions with these types of scales between the mail and telephone questionnaires.

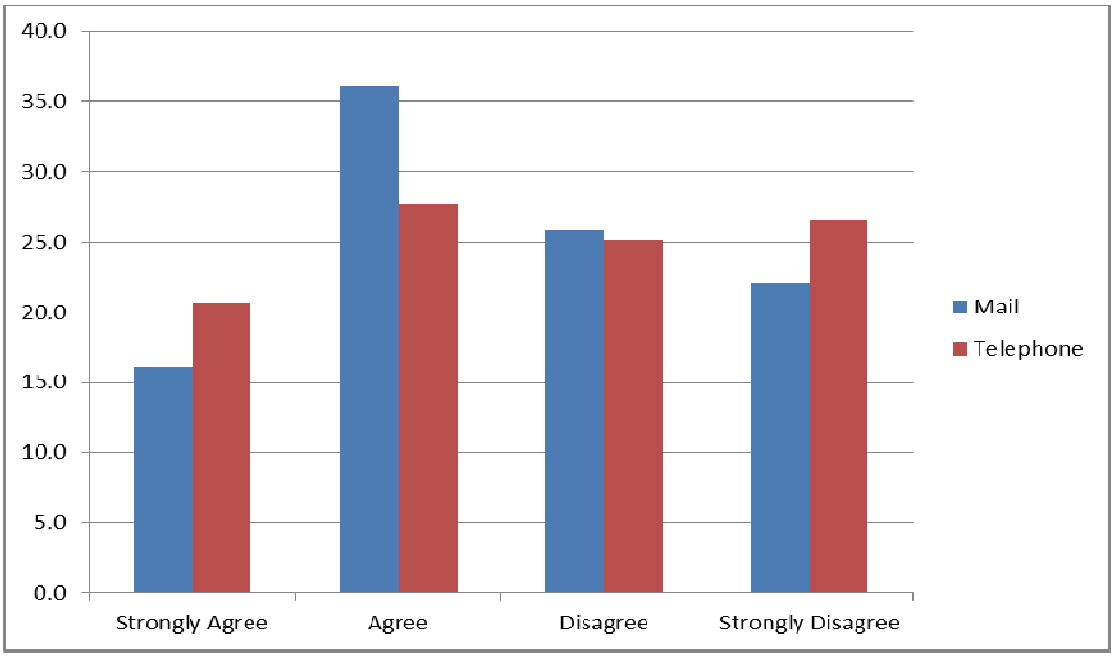
The HINTS data do exhibit different response patterns across the two modes for these items. This is illustrated in Figure 1-2 below, which shows the telephone having more responses in the "Always" category and fewer responses in the "Usually" category. For bi-directional scales, there are also instances where telephone respondents are more likely to choose both extremes (Figure 1-3). When analyzing ordinal scales on HINTS 3 one could mitigate some of the differences by collapsing extreme categories. For example, collapsing "Always" and "Usually" in Figure 1-2 produces essentially the same estimates for both frames.

Figure 1-2. How often did they make sure you understood the things you needed to do to take care of your health?*



* Difference between distributions statistically significant at p<.01

Figure 1-3. Based on most recent search about health or medical topics: You were concerned with the quality of the information*



* Difference between distributions statistically significant at p<.01

The mode of communication can also affect the number of Don't Know (DK) responses. Neither the mail nor telephone surveys provided explicit DK response categories. For the telephone survey, interviewers can record a DK if the respondent provides this response. In the mail form, a DK is captured only when the respondent chooses to write "don't know" rather than just leave the question blank. This may lead to a difference in the number of DK responses that are recorded by mode. For example, there are a series of knowledge questions on HINTS 3 for which the telephone respondents had a greater tendency to answer "Don't know" than on the mail. The question on the effects of sunlight on Vitamin D is a specific example of this ("Do you agree or disagree that sunlight helps produce Vitamin D naturally?"[1]). Respondents to the telephone survey were more likely to say DK than the mail survey respondents.

The above discussion highlights a number of reasons the two surveys may yield different results. This discussion was not meant to be exhaustive of all the possible differences. The purpose was to illustrate why the results between the two modes may differ. When there are large differences, such as noted for question A1 above, this allows the analyst to choose between the two modes when concentrating on HINTS 3. For A1 above, for example, one could argue that the address frame provides the best measure because respondents are more informed on the context of the question.

As will be noted below, when analyzing trends it is not absolutely necessary to understand why the results differ. When selecting which estimates to rely on for HINTS 3, the choice only depends on whether there are meaningful differences between the two estimates. However, as will be shown below, if the two methods produce radically different estimates it is useful to try to understand the reasons. In the next section we discuss what analysts should do to account for the change in mode when analyzing trends that span the change in methodology.

Linking Between Iterations with Different Modes

In this section we provide a strategy for calculating trends involving the HINTS 3 data. The discussion below concentrates on comparing trends involving HINTS 1 and/or 2 with HINTS 3 and later. We do not discuss computing trends with data prior to HINTS 3 because this does not involve any change in mode and is covered in other HINTS publications (e.g., Rizzo, Moser, Waldron, Wang, & Davis, 2008).

[1] Question D16 on the mail survey and BR-16 on the telephone survey.

In HINTS 3 the RDD and address samples were each representative of the national population. Each has a set of weights that can be used to carry out separate, nationally representative analyses. The data file also includes a set of weights when combining the two different surveys. This offers the analyst a number of options with respect to examining trends. In brief, the user should consider the following steps when conducting trend analyses:

1. Examine whether there are differences between RDD and address samples for HINTS 3
2. If no differences, then use both samples for HINTS 3 and use the respective combined weights
3. If there are differences,
    a. Use the RDD sample (and weights) for change up to HINTS 3.
    b. Use the Address sample (and weights) for change from HINTS 3 to later administrations.

This strategy controls for the change in methodology when it substantively makes a difference. If there are no substantive differences, then combining the two modes increases the precision of the HINTS 3 estimate. To provide some perspective on these choices, Table 1-1 provides the three estimates of internet use for the different combinations of sample. In this case the estimate for the address frame is 5 percentage points below the RDD frame. The estimate that combines the two is between the two. The weights for combining the two frames were formed by essentially averaging the two estimates together based on sample size. This essentially produces an estimate that is close to the simple average of the two frame-specific estimates[2]. By combining the two frames together, the standard error is reduced. For change between pre-HINTS 3 surveys, the user would decide whether the increased precision for the combined estimate is preferable. The example in Table 1-1 suggests using the mode-specific estimates since the difference in point estimates (5%) is quite a bit larger than any gains in precision on the point estimate.

Table 1-1. Percent using the internet for HINTS 3 Address, RDD, and Combined estimates

	Address frame	RDD frame	Combined estimate
% using internet	66%	71%	69%
Standard Error	1.1%	.9%	.65%
1.96*standard error	2.2%	1.8%	1.3%

[2] The telephone sample has a slightly larger sample size. This will tend to produce a combined estimate that is somewhat closer to the RDD estimate.

Nonetheless, the substantive differences in the estimates are relatively small. The estimate for internet use for HINTS 1 was 61 percent. Comparing this estimate to HINTS 3 RDD or combined estimate produces essentially the same result. Both would show a statistically significant difference, with estimates of change being 10 (ie using RDD = 71 - 61) or 8 (ie using combined = 69 - 61) percentage points, depending on the HINTS 3 estimate. Similarly, the estimate for HINTS 4 (Cycle 1) for internet use was 78 percent. This would also be highly significant with estimates of change ranging from 12 (using address = 78 - 66) to 9 (ie using combined = 78 – 69) percentage points.

A more dramatic example is illustrated with the question on how much the respondent trusts family members for health or medical information (Table 1-2). In this case, there is a very large difference, with the address sample producing a much lower estimate than the RDD sample. This difference could be due to the acquiescence bias described above or it could be social desirability, with respondents wanting to show they trust their family. Regardless of the reason, in this instance the mode-specific estimates are clearly appropriate when developing trends. In HINTS 2 the estimate was 23 percent, while in HINTS 4 the estimate was 7.1 percent. In terms of trends, this would mean a change of -1 percent between HINTS 2 and HINTS 3 and a change of -2.2 percent between HINTS 3 and HINTS 4.

Given the large differences between the address and RDD samples, it is questionable they are measuring the same attitude. Note that while the absolute change is about the same, the relative change is much different. Between HINTS 2 and 3 the percentage change for the RDD sample is -4 percent (1/23), while it is -23 percent (2.2/9.3) for the address sample. This may mean that one could not get two comparable measures of change for the two types of samples. As noted above, one way to mitigate some of this difference is to combine adjacent categories of the two scales. In this case, this combines the "A lot" category with the "Some" category. This reduces the magnitude of the differences between the different frames. For the Address frame in HINTS 3, this estimate is 59 percent, while for the RDD frame it is 66 percent. The estimates are still different, but not nearly as large on a relative basis. In HINTS 2 and 4, the comparable estimates were 66 percent and 57 percent for the RDD and address frames, respectively. Using these collapsed measures indicates essentially no change for either frame, which would indicate no change between HINTS 2 and 4.

Table 1-2. Percent reporting "a lot" of trust in medical information from friends and family by frame

	Address frame	RDD frame	Combined estimate
% trust a lot	9.3%	22%	15.5%
Standard Error	.6%	.9%	.6%
1.96*standard error	2.2%	1.8%	2.2%

The above analyses are carried out by use of the three different weights available on the HINTS 3 public use file. Table 1-3 is taken from the public use documentation and provides the variable names for the final sample weights and associated replicate weights needed to produce the frame-specific and combined estimates.

Table 1-3. Final sample weight and replicate weight variable names

Survey mode	Final sample weight	Replicate weights
RDD Only	rwgt0	rwgt1 thru rwgt50
Mail Only	mwgt0	mwgt1 thru mwgt50
Combined RDD and Mail	cwgt0	cwgt1 thru cwgt50

When developing formal significance tests across years, the first decision is which weights should be used for HINTS 3. Once making this decision, the methods described in Rizzo et al. (2008) to estimate differences and significance tests can then be applied.

The above discussion has illustrated methods for analyzing trends that involve questions that are included on HINTS 3. This methodology does not apply when comparing pre- and post-HINTS 3 surveys if the item is not included on the HINTS 3 survey. In these situations, it is difficult to conduct trend analyses, unless one can argue that the measures across the different modes are equivalent.

Merging and Analyzing Multiple Iterations of HINTS Mainland Data 2

Before considering trending on an item, there are several factors to consider in determining whether an item can even be trended across time. These factors include confirming that the survey question was asked in the same way in each iteration, with the same response options, and the same universe of respondents (i.e., who makes up the denominator). For a list of HINTS items that can be used for trend analysis, please visit All Hints Questions (http://hints.cancer.gov/questions.aspx).

To illustrate how to merge all existing iterations of HINTS together to test for a trend across time, an item that is found in all four iterations will be used. This question, asked of those who say they access the internet, asks, "Have you ever used e-mail or the internet to communicate with a doctor or doctor's office?" with response options of Yes or No.

Methods

When testing for trends across HINTS iterations, it is vital that the analyst thinks carefully about how to "bridge" across iterations that use different modes. The goal is to identify the best way to merge the data and apply the most appropriate weights to ensure the validity of the results. (See Section 1 of this report for more information.) Prior to merging the data, variable names and response options should be identical across all HINTS datasets, and if using the HINTS 3 data, it is important to first test for mode effects to decide which weights to use when merging the data. (See Appendix A section 1 for code to test for mode effects.)

The analyst will also need to create a combined dataset that contains the appropriate final sample and replicate weights (see Table 2-1 "Construction of Statistical Weights for a Combined Data File"). We will demonstrate using all four survey years. Since each HINTS survey contains one final sample weight and 50 replicate weights, the combined dataset will contain one final sample weight and 200 replicate weights (to account for the 50 replicate weights from each survey). For this example, the final sample weight of the combined dataset will be named NFWGT0 and the replicate weights will be named NFWGT1 through NFWGT200. For the first 50 replicate weights in the combined dataset (NFWGT1, …, NFWGT50), we copy over replicate weights FWGT1, …, FWGT50 from HINTS 1, and use the respective final sample weight, FWGT, for replicate weights 51-200 in the combined datasest. For the next 50 replicate weights (NFWGT51, …, NFWGT100),

we copy over replicate weights FWGT1, …, FWGT50 from HINTS 2, and use the respective final weight, FWGT, for replicate weights 1-50, and 101-200 in the combined dataset. For the next 50 replicate weights (NFWGT101, …, NFWGT150), we copy over replicate weights from HINTS 3, where the analyst will need to decide which weights to incorporate (i.e., for the RDD sample, the mail sample, or the combined sample) and use the respective final sample weight, for replicate weights 1-100, and 151-200 in the combined dataset. Finally, for the last 50 replicate weights (NFWGT151,…..,NFWGT200), we copy over the replicate weights from HINTS 4 (PERSON_FINALWT1,….PERSON_FINALWT50) and use the respective final sample weight, PERSON_FINALWT0, for replicate weights 1-150 in the combined dataset. When the sums of squares for all 200 replicates are combined, the result is a sum of HINTS 1, HINTS 2, HINTS 3, and HINTS 4 variances, as desired (as the surveys are in fact independent). Note that from Table 2-1 one can see that the replicate weights for each respective iteration only contributes variance for that iteration (see Cochran 1977, for formula to estimate variance). Please refer to Appendix A section 2 for the code to construct these statistical weights for a combined dataset. After merging, it is always good practice to re-run frequencies to confirm that the data are intact.

All HINTS replicate weights were created using a jackknife minus one replication method. Thus, the proper denominator degrees of freedom (ddf) should be 49 when one iteration of HINTS data are being analyzed. Once merged together, the combined data will contain a set of 50*k replicate weights, where they can be viewed as being created using a stratified jackknife method with k as the number of strata and 49*k as the appropriate ddf, where k is the number of iterations of HINTS data used in the analysis.

Measures

Outcome: The following question (found across all four HINTS iterations) assessed whether internet users electronically communicated with their health care provider: "Have you ever used e-mail or the internet to communicate with a doctor or doctor's office?" with possible response options of Yes or No.

Sociodemographic Variables: Sociodemographic variables—(comparable across all iterations) included gender, age in categories (18-34, 35-39, 40-44, 45+), and education (Less than high school, High school graduate, Some college, and College graduate).

Survey Year: A survey year variable was created to indicate each HINTS iteration: 1 = HINTS 2003; 2 = HINTS 2005; 3 = HINTS 2008; 4 = HINTS 2011-2012.

Table 2-1. Construction of statistical weights for a combined data file

	Final sample weights	Replicate weights 1-50	Replicate weights 51-100	Replicate weights 101-150	Replicate weights 151-200
HINTS 1 (2003)	HINTS 1 Final Weight (fwgt)	HINTS 1 Replicate Weights (fwgt1-fwgt50)	HINTS 1 Final Weight (fwgt)	HINTS 1 Final Weight (fwgt)	HINTS 1 Final Weight (fwgt)
HINTS 2 (2005)	HINTS 2 Final Weight (fwgt)	HINTS 2 Final Weight (fwgt)	HINTS 2 Replicate Weights (fwgt1-fwgt50)	HINTS 2 Final Weight (fwgt)	HINTS 2 Final Weight (fwgt)
HINTS 3 (2008*)	HINTS 3 Final Weight	HINTS 3 Final Weight	HINTS 3 Final Weight	HINTS 3 Replicate Weights	HINTS 3 Final Weight
HINTS 4 (2011-2012)	HINTS 4 Final Weight (person_finalwt0)	HINTS 4 Final Weight (person_finalwt0)	HINTS 4 Final Weight (person_finalwt0)	HINTS 4 Final Weight (person_finalwt0)	HINTS 4 Replicate Weights (person_finalwt1-person_finalwt50)
Combined Data	Final Weight (nfwgt0)	Final Replicate Weights (nfwgt1-nfwgt50)	Final Replicate Weights (nfwgt51-nfwgt100)	Final Replicate Weights (nfwgt101-nfwgt150)	Final Replicate Weights (nfwgt151-nfwgt200)

* HINTS 3 allows for utilizing the RDD Weights (rwgt0), the Mail weights (mwgt0), or the composite weights (cwgt0).

Statistical Analysis

For descriptive purposes, a crosstabulation table was created to obtain population estimates of the outcome for each HINTS iteration. To decide which weights to use for the HINTS 3 iteration, a t-test was conducted to assess for significant differences in the outcome across mode. Analysis of the B7d outcome revealed no mode of survey administration effects between the RDD and mail survey. Therefore, the composite sample and replicate weights (CWGT0, and CWGT1 to CWGT50) were utilized when creating the combined data file. A multivariable logistic regression was conducted, regressing the outcome on the set of variables including age, education, and gender to assess for the odds (and log-odds) of ever having communicated electronically with one's doctor. Because we are using four iterations of data, we can test for three orthogonal trends: cubic, quadratic, and linear (coefficients to code for these trends were taken out of a standard statistics book). Predicted marginals—also known as model-adjusted risks—were also computed to estimate the probability of internet users who reported using email or the internet to communicate with a doctor or doctor's office. An interaction term between gender and survey year was included to assess for differential change over time by gender. See Appendix A section 3 for code to conduct a logistic regression model.

Results

The percent of respondents who used the internet to communicate with their doctor or doctor's office increased from 7 percent in HINTS 1 (2003) to 19 percent in HINTS 4 (2011-2012) (Table 2-2).

Table 2-2. Population estimates of internet users who used the internet to communicate with their healthcare provider since 2003

	HINTS1	HINTS2	HINTS3	HINTS4
In the last 12 months, have you used email or the internet to communicate with a doctor or doctor's office?				
Yes	7.00%	9.62%	13.59%	19.11%
No	93.00%	90.38%	86.41%	80.89%

For the multivariable logistic model that adjusts for sociodemographic variables, Table 2-3 shows the results of testing for the three trends across time. It can be seen that only the linear trend is statistically significant at alpha < .05. Figure 2-1 graphs the predicted marginals illustrating this trend.

Table 2-3. Test of trend

Trend	F	P-value
Cubic Trend	0.14	0.7104
Quadratic Trend	0.00	0.9558
Linear Trend	99.36	0.0000

Figure 2-1. Graph showing the percent of internet users who emailed their healthcare provider since 2003

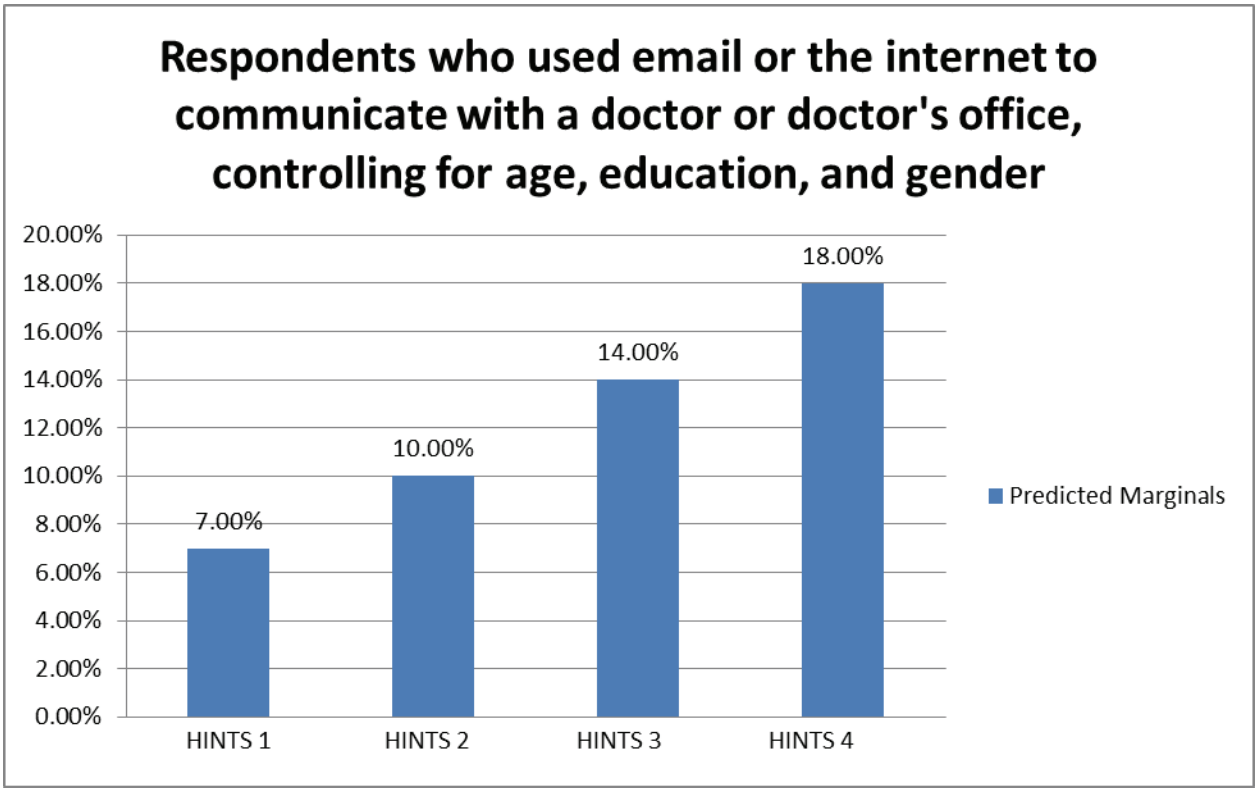
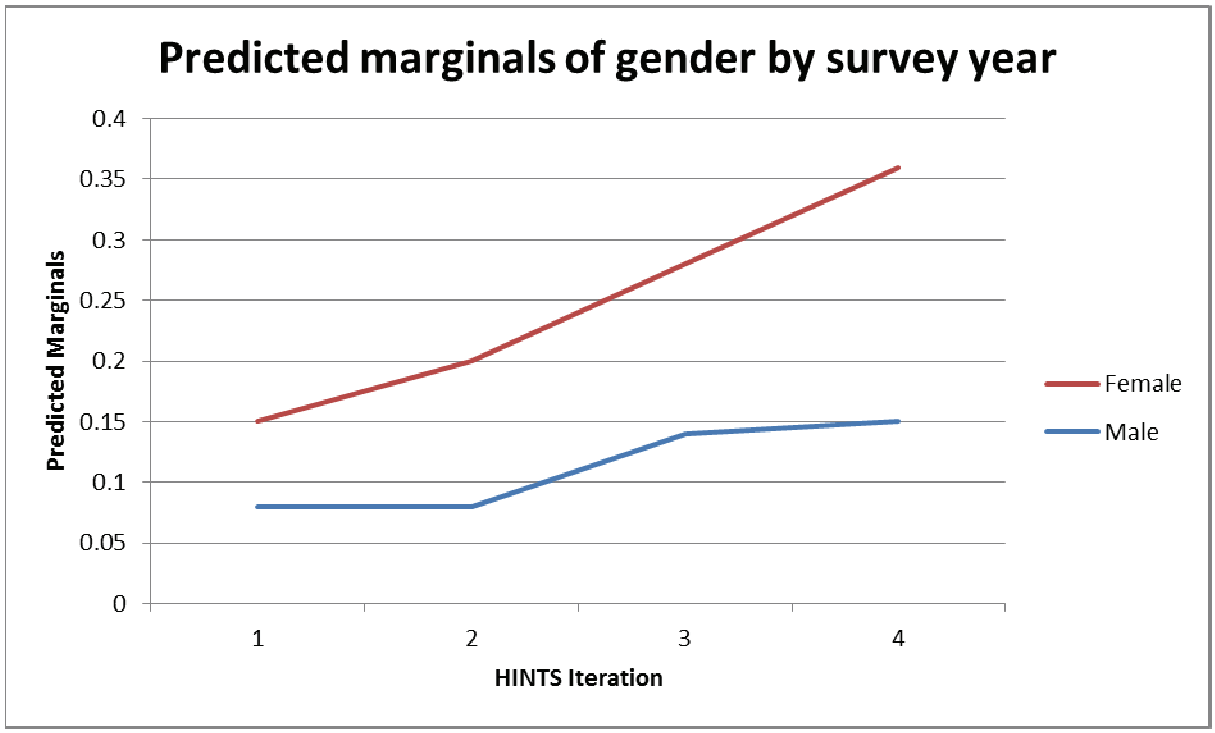


Table 2-4 shows that respondents who used the internet in 2012 had 2.14 times (95% CI: 1.48 to 3.09) higher odds (compared to 2003) of using email to communicate with their healthcare provider, after controlling for age, education, and gender. Testing for the interaction between gender and survey year, the overall F-test (Wald F = 309.95, P-value < 0.0001) shows that this interaction was statistically significant. To interpret this interaction, predicted marginals were plotted (see Figure 2-2). It can be seen that females seemed to have the largest change over time while rates for males increased at a much lower rate.

Table 2-4. Multivariable logistic regression of emailing a healthcare provider since 2003, controlling for age, education, and gender, and testing an interaction between survey year and gender

Variable	OR	95% CI	P-value
Survey Year			—
2003	1.00	—	
2005	1.02	0.70 - 1.48	
2008	1.91	1.42 - 2.57	
2011-12	2.14	1.48 - 3.09	
Education			0.0000
Less than High School	1.00	—	
High School Graduate	1.02	0.61 - 1.71	
Some College	1.64	0.99 - 2.71	
College Graduate	2.57	1.58 - 4.16	
Gender			—
Male	1.00	—	
Female	0.82	0.61 - 1.09	
Survey Year * Gender			0.0122
2003, Male	1.00	1.00 - 1.00	
2003, Female	1.00	1.00 - 1.00	
2005 Male	1.00	1.00 - 1.00	
2005, Female	1.82	1.17 - 2.85	
2008, Male	1.00	1.00 - 1.00	
2008, Female	1.20	0.83 - 1.75	
2011-12, Male	1.00	1.00 - 1.00	
2011-12, Female	1.82	1.17 - 2.83	

Figure 2-2. Predicted marginals of gender by survey year



Merging HINTS Mainland and HINTS Puerto Rico Data 3

In 2009, the University of Puerto Rico Comprehensive Cancer Center, the Puerto Rico Behavioral Risk Factors Surveillance System, and the U.S. National Cancer Institute, fielded a Spanish translation of the HINTS 3 survey in Puerto Rico (HINTS PR). This demonstration project was conducted to assess the feasibility of adapting the national HINTS survey to a local setting, in addition to better understanding the health information and education needs of the population in Puerto Rico. See HINTS Brief #18 on the HINTS website to get more information about the implementation of the HINTS PR survey: http://hints.cancer.gov/brief_18.aspx.

The goal of this section is to 1) demonstrate how to merge and analyze HINTS 3 and HINTS PR data, and 2) demonstrate comparisons across groups.

For this analysis, the following question—common to both datasets—will be used as the main outcome of interest: "Have you ever looked for information about cancer from any source?" with possible response options of Yes or No.

Methodology

Data were collected using RDD and computer-assisted telephone interview (CATI) by experienced bilingual Puerto Rican interviewers. For more information about how the data were collected and other technical aspects, see the Final Report (Davis, Dipko, & Sigman, 2009) on the HINTS website: http://hints.cancer.gov/instrument.aspx.

Similar to merging multiple iterations of Mainland HINTS data, it is vital that the analyst thinks carefully about how to combine across iterations that use different modes (see previous sections for more information). To keep mode consistent between iterations, only the RDD sample and respective weights of HINTS 3 will be used for the present analysis.

Weights and merging: Since the HINTS 3 dataset contains 50 replicate weights and the HINTS PR dataset contains 48 replicate weights, the combined dataset should contain 98 replicate weights. For the first 50 replicate weights in the combined dataset (TWGT1,..., TWGT50), we copy over replicate weights RWGT1,…,RWGT50 from HINTS 3 , and use the respective final sample weight

(RWGT0) for replicate weights 51-98 in the combined dataset. For the last 48 replicate weights in the combined dataset (TWGT51,…,TWGT98), we copy over replicate weights from HINTS-PR (R12WGT1,…, R12WGT48), and use the respective final sample weight (R12WGT0) for replicate weights 1-50 in the combined dataset. See Table 3-1, which illustrates this process, and Appendix B Section 1 for more detail.

The number and type of replicate weights differs between HINTS 3 and HINTS PR. HINTS 3 used jackknife 1 (JK1) technique, while HINTS PR used the JKn replication method with 8 sampling strata formed, thus, each dataset is analyzed with a different jackknife multiplier: 0.98 in HINTS 3 and 0.83 in HINTS PR. Therefore, additional syntax is required to properly apply the correct jackknife multiplier to each replicate weight in the combined data.

The proper denominator degrees of freedom (ddf) to use when analyzing a merged HINTS 3/PR dataset is 98-1-8=89.

Measures

Outcome: The following question assessed whether users sought cancer information from any source: "Have you ever looked for information about cancer from any source?" with possible response options of Yes or No.

Sociodemographic Variables: Sociodemographic variables included gender, age in categories (18-34, 35-39, 40-44, 45+), and education (Less than high school, High school graduate, Some college, and College graduate), and a derived variable coding for ethnicity (U.S. Mainland Hispanics, U.S. Mainland Non-Hispanics, and Puerto Rico Hispanics).

HINTS Iteration: A HINTS iteration variable was created to flag whether an item was asked in the U.S. Mainland survey or in the Puerto Rico survey.

Statistical Analysis

A crosstabulation table and a chi-square test of association were conducted to compare the percent of U.S. Mainland respondents vs. Puerto Rico respondents who indicated whether or not they sought cancer information from any source. Another crosstabulation table and chi-square test was conducted to compare the percent of Hispanics on the U.S. Mainland vs. Non-Hispanics on the U.S. Mainland vs. Hispanics in Puerto Rico who sought information about cancer from any source.

Finally, a crosstabulation table and chi-square test was conducted to compare the percent of Hispanics on the U.S. Mainland vs. Hispanics in Puerto Rico who sought information about cancer from any source.

Two multivariable logistic regression models were conducted. The first model sought to determine the odds of seeking cancer information from any source between the two HINTS iterations, after controlling for age, gender, and education. The second model was conducted to determine the odds of seeking cancer information from any source between different ethnic groups, after controlling for age, gender, and education. Note that due to extremely low sample size, non-Hispanics in Puerto Rico (n = 8) were excluded from analysis involving ethnicity.

Table 3-1. Construction of statistical weights for a combined HINTS 3 and HINTS PR dataset

	Final sample weights	Replicate weights 1-50	Replicate weights 51-98
HINTS 3	HINTS 3 Final Weight (rwgt0)	HINTS 3 RDD Replicate Weights (rwgt1-rwgt50)	HINTS 3 Final Weight (rwgt0)
HINTS PR	PR Final Weight (r12wgt0)	PR Final Weight (r12wgt0)	PR Replicate Weights (r12wgt1-r12wgt48)
Combined Data	Final Weight (twgt0)	Final Replicate Weights (twgt1-twgt50)	Final Replicate Weights (twgt51-twgt98)

Results

A significantly higher percentage of adults on the U.S. Mainland reported seeking information about cancer from any source compared to adults in Puerto Rico (39.40% vs. 28.11%; Table 3-2).

Table 3-2. Comparing U.S. Mainland vs. Puerto Rico in seeking cancer information from any source

Seek information about cancer	U.S. Mainland		Puerto Rico		Chi-Square	P-value
	N	Weighted %	N	Weighted %	26.38	0.0000
Yes	1911	39.40%	181	28.11%		
No	2162	60.60%	458	71.89%		
Total	4073	100.00%	639	100.00%		

From Table 3-3 it can be seen that non-Hispanic adults on the U.S. Mainland reported a significantly higher percentage of seeking information about cancer from any source (42.78%), compared to Hispanic adults on the U.S. Mainland (21.19%), and Hispanic adults in Puerto Rico (27.55%).

Table 3-3. Comparing percent of Hispanics on the U.S. Mainland vs. Non-Hispanics on the U.S. Mainland vs. Hispanics in Puerto Rico who sought information about cancer from any source

Seek information about cancer	Non-Hispanics in U.S.Mainland		Hispanics in U.S.Mainland		Hispanics in PR		Chi-Square	P-value
	N	Weighted %	N	Weighted %	N	Weighted %	30.15	0.0000
Yes	1683	42.78%	90	21.19%	167	27.55%		
No	1718	57.22%	207	78.81%	428	72.45%		
Total	3401	100.00%	297	100.00%	595	100.00%		

Table 3-4 shows that there was no statistically significant difference (alpha = .05) in seeking information about cancer between Hispanic adults on the U.S. Mainland (21.19%) and Hispanic adults in Puerto Rico (27.55%).

Table 3-4. Comparing percent of Hispanics on the U.S. Mainland vs. Hispanics in Puerto Rico who sought information about cancer from any source

Seek information about cancer	Hispanics in U.S. Mainland		Hispanics in PR		Chi-Square	P-value
	N	Weighted %	N	Weighted %	3.32	0.0717
Yes	90	21.19%	167	27.55%		
No	207	78.81%	428	72.45%		
Total	297	100.00%	595	100.00%		

We conducted two multivariable logistic regression models. The first regression showed that adults in Puerto Rico had 0.64 times (95% CI: 0.50 – 0.82) lower odds, compared to adults on the U.S. Mainland, in seeking information about cancer from any source, after controlling for age, gender, and education (Table 3-5). Refer to Appendix B section 2 for logistic regression code.

Table 3-5. Odds of seeking cancer information from any source, after controlling for age, gender, and education: Comparing U.S. Mainland vs. Puerto Rico

	Odds of seeking cancer information		
	OR	95% CI	P-value
HINTS Iteration			0.0005
U.S. Mainland	1.00	—	
Puerto Rico	0.64	0.50 - 0.82	
Age			0.0000
18-34	1.00	—	
35-39	1.78	1.06 - 2.98	
40-44	1.60	1.05 - 2.44	
45+	2.02	1.52 - 2.69	
Gender			0.0001
Male	1.00	—	
Female	1.56	1.27 - 1.92	
Education			0.0000
Less than HS	1.00	—	
HS Graduate	2.12	1.39 - 3.24	
Some College	3.71	2.43 - 5.67	
College Graduate	5.82	3.82 - 8.86	

Tables 3-6 and 3-6a provide results for the second multivariable logistic model that included the three-group ethnicity variable and controlled for the same sociodemographic variables. It can be seen from Table 3-6 that the ethnicity was a significant predictor of the seeking cancer information. Compared to Hispanic adults on the U.S. Mainland, non-Hispanic adults on the U.S. Mainland had a 1.64 (95% CI: 1.11 – 2.42) times greater odds of seeking information about cancer from any source, after controlling for age, gender, and education. Table 3-6a shows user-defined comparisons of ethnic groups and shows that there was no significant difference between U.S. Mainland Hispanics and PR Hispanics in the odds of seeking cancer information. However, there was a significant difference between U.S. Mainland Hispanics and U.S. Mainland non-Hispanics in the odds of seeking cancer information (see Appendix B section 3 for logistic regression code).

Table 3-6. Odds of seeking cancer information from any source, after controlling for race/ethnicity, age, gender, and education

	Odds of seeking cancer information		
	OR	95% CI	P-value
Ethnic Group			0.0004
Hispanics in the U.S.	1.00	—	
Non-Hispanics in the U.S.	1.64	1.11 - 2.42	
Hispanics in Puerto Rico	0.99	0.66 - 1.47	
Age			0.0003
18-34	1.00	—	
35-39	1.80	1.06 - 3.03	
40-44	1.62	1.06 - 2.48	
45+	1.94	1.44 - 2.60	
Gender			0.0001
Male	1.00	—	
Female	1.55	1.26 - 1.91	
Education			0.0000
Less than HS	1.00	—	
HS Graduate	1.91	1.22 - 3.00	
Some College	3.35	2.19 - 5.13	
College Graduate	5.21	3.33 - 8.16	

Table 3-6a. Comparing different ethnic groups

	Wald F	P-value
Hispanics in the U.S. Mainland vs. Hispanics in PR	< 0.01	0.9490
U.S. Mainland Hispanics vs. U.S. Mainland Non-Hispanics	6.36	0.0133

Multilevel Determinants of Smoking Behavior: An Integrated Data Analysis

4

Introduction

Tobacco use is the leading preventable cause of disease, disability, and death in the United States, accounting for 443,000 deaths annually (U.S. Department of Health and Human Services, 2000; Centers for Disease Control and Prevention, 2005). Beyond those who directly smoke tobacco products, secondhand smoke—a mixture of exhaled smoke with gases and particles from burning cigarettes, cigars, or pipes—causes approximately 50,000 U.S. deaths annually (Centers for Disease Control and Prevention, 2005; Centers for Disease Control and Prevention, 2006). Secondhand smoke exposure can cause heart disease and lung cancer in nonsmoking adults, and also is associated with asthma, ear infections, and bronchitis in children (National Cancer Institute, 1999).

The *Population Strategy of Prevention*, coined by Geoffrey Rose in 1985, posits that we can achieve greater gains in overall health and risk reduction at a population level (compared to focusing efforts at the individual level) by controlling the underlying determinants of disease incidence and lowering the mean level of disease risk factors, thereby shifting the entire distribution of exposure in a favorable direction (Rose, 1985). As per this approach in a growing number of states, legislators are addressing the tobacco problem by enacting smoke-free laws or smoking restrictions for public places. In August 2007, the President's Cancer Panel released its report, *Promoting Healthy Lifestyles: Policy, Program, and Personal Recommendations for Reducing Cancer Risk*, where members urged the leadership of the nation to "summon the political will to address the public health crisis caused by tobacco use" (President's Cancer Panel, 2007). Increasingly, media coverage, communication campaigns, and advocacy efforts illuminating the negative effects of tobacco and the dangers of secondhand smoke have led to policy efforts such as indoor air laws and tobacco tax increases, among other measures, aimed at denormalizing tobacco use and protecting nonsmokers in the U.S. and around the world (Hammond, Fong et al., 2006; Arnott, Dockrell et al., 2007).

Research shows that benefits of implementing smoke-free laws include: significant declines in hospital admissions for heart attacks, reduced exposure to secondhand smoke, increased smoking cessation rates, and either a positive effect or no decline in total restaurant or bar revenues (Huang & McCusker, 2004; Centers for Disease Control and Prevention, 2009). Similar measures are being implemented to raise the price of cigarettes, particularly through state cigarette taxes. Numerous

economic studies have documented that cigarette tax or price increases reduce both adult and underage smoking. The general consensus, derived from this research, is that every 10 percent increase in the real price of cigarettes reduces overall cigarette consumption by approximately 3 to 5 percent, reduces the number of young-adult smokers by 3.5 percent, and reduces the number of kids who smoke by 6 to 7 percent (Chaloupka, 1999). Research studies have also found that cigarette price increases and tax increases are particularly effective in reducing smoking among males, Blacks, Hispanics, and lower-income smokers (Chaloupka & Pacula, 1998; U.S. Centers for Disease Control and Prevention [CDC], 1998).

Despite strong gains in tobacco control policy implementation, the comprehensiveness of state and local indoor air ordinances and cigarette excise tax varies across the U.S. Moreover, despite recent reductions in smoking prevalence over the past several decades, smoking prevalence is not evenly distributed in the population. In particular, individuals with low socioeconomic status (SES) are significantly more likely to smoke (Finney Rutten, Augustson et al., 2008).

Purpose

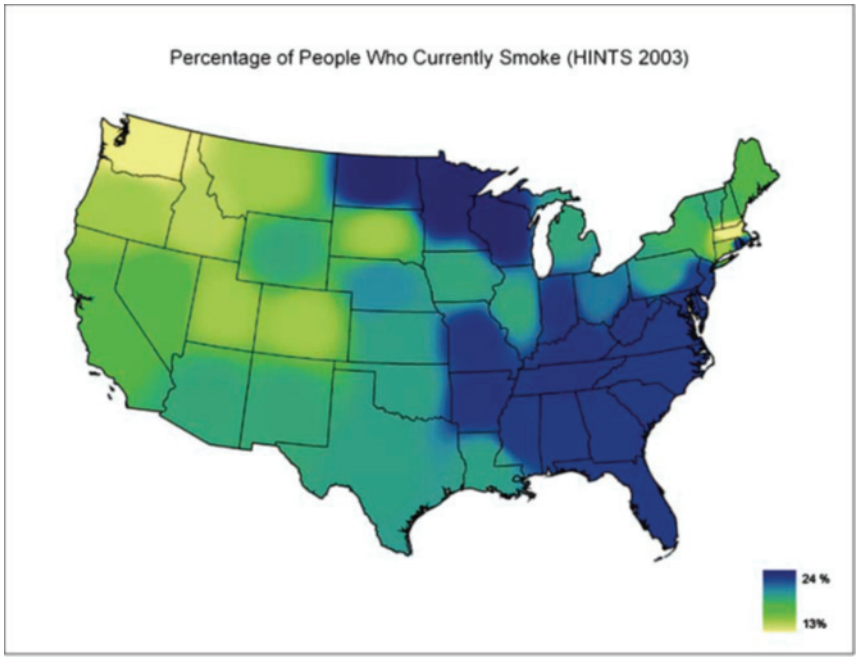
The purpose of this integrated data analysis exercise is two-fold. First, we explore the geographic distribution of smoking behavior and tobacco control policies using publicly available, existing maps that depict spatial gradients of tobacco use and tobacco control policies, as a hypothesis-generating exercise. Second, we undertake a cross-sectional, multilevel, integrated data analysis to examine the independent statistical associations of individual-level SES and state-level tobacco-control policies on smoking behavior using data from three distinct sources. The purpose of this multilevel, integrated data analysis is to examine factors that may account for observed spatial variation in smoking behavior. These analyses can inform efforts to target policy strategies to decrease tobacco use at the population level, by disentangling individual versus policy determinants of adult smoking behavior.

In a 2008 publication using HINTS 1 data, Geographic Information System (GIS) isopleth maps were generated to demonstrate the geographic distribution of adult smokers in the United States. GIS isopleth maps visually represent a large number of data points in a "weather-map" fashion. Crude prevalence maps were generated to provide visual data to explore possible geographic relationships with HINTS cancer-related knowledge variables and to generate hypotheses through comparison of the geographic distribution of knowledge variables with smoking behavior (Finney Rutten, Augustson et al., 2008). The percentage of self-reported current smokers in the United States, as represented in HINTS 1, was summarized geographically; varying geographic regions

ranged from 13 percent to 24 percent (Figure 1). Smoking prevalence was relatively lower in the Pacific Northwest and in a small pocket in southern New England. Smoking prevalence was relatively higher in the northern Plains and in the southeastern quadrant of the United States (i.e., the "tobacco belt").

Publicly available maps, provided by the Americans for Nonsmokers' Rights Foundation (ANR) and the National Conference of State Legislatures (NCLS), depict geographic gradients of comprehensiveness of state indoor air laws (2013 data) and state cigarette excise tax rates (2010 data), respectively (Figures 2 and 3). Considering these two maps, we can begin to hypothesize about state-level factors that influence smoking behavior, as regions with fewer indoor air laws and lower tax rates also generally have higher smoking prevalence. To test these observed associations while taking into account individual factors such as SES and other known individual-level predictors of smoking behavior, a multilevel random intercepts model is required.

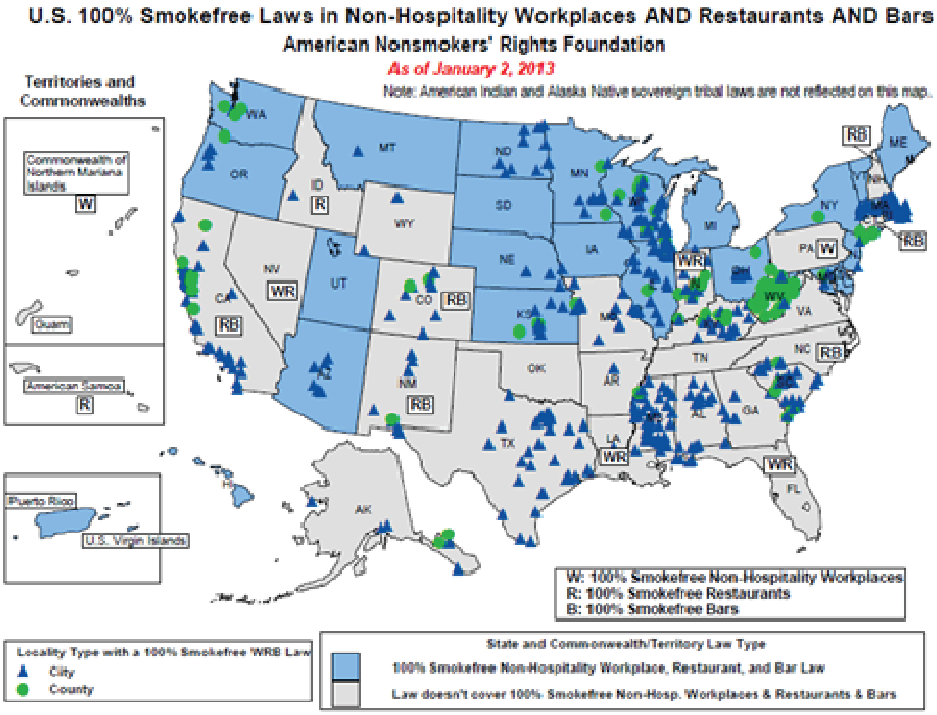
Figure 4-1. GIS isopleth map depicting crude prevalence of current smokers in the U.S., HINTS 1



Finney Rutten, L. J., Augustson, E. M., Moser, R. P., Beckjord, E. B., and Hesse, B. W. Smoking knowledge and behavior in the United States: sociodemographic, smoking status, and geographic patterns. *Nicotine Tob Res.* 2008 Oct.10(10):1559-70.

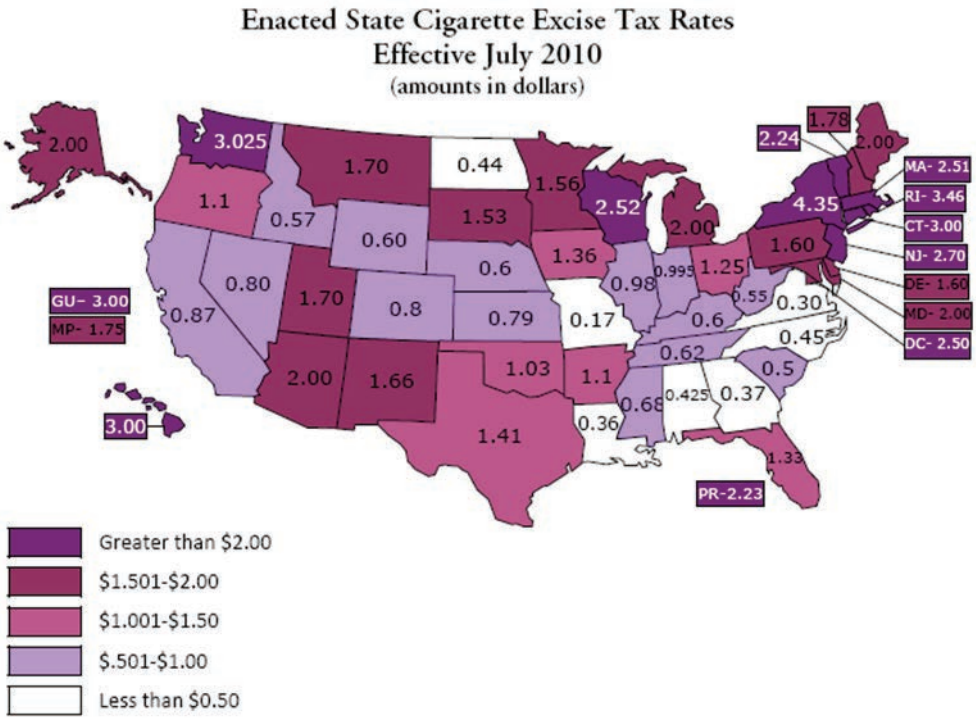
HINTS maps available at hints.cancer.gov and statecancerprofiles.gov.

Figure 4-2. U.S. map depicting 100 percent comprehensive state indoor air laws, ANR 2013



Americans for Nonsmokers' Rights: http://www.no-smoke.org/goingsmokefree.php?id=519

Figure 4-3. U.S. map depicting gradients of state cigarette excise tax rates, NCSL 2010



National Conference of State Legislatures: http://www.ncsl.org/issues-research/health/2010-state-cigarette-excise-taxes.aspx

Data

Health Information National Trends Survey (HINTS)

The Health Information National Trends Survey (HINTS) program was developed by the National Cancer Institute and first fielded in 2003 to track changes in the rapidly evolving health communication and information technology landscape and to assess the impact of health communication and health information technology on health outcomes, healthcare quality, and health disparities. HINTS is a nationally representative survey of the U.S., non-institutionalized, adult population that collects data on the American public's need for, access to, and use of health-related information (Nelson, Kreps et al., 2004). Data for our analyses are from HINTS 4 Cycle 1, collected from October 2011 to February 2012 (n=3959) and HINTS 4 Cycle 2, collected from October 2012 to January 2013 (n=3,630), through mailed questionnaire. The sample design was a two-stage, stratified sample with addresses selected from a comprehensive United States Postal Service national residential file, and individual respondents were selected per each household in the sample. The final response rate for HINTS 4 Cycle 1 (2011-2012) was 36.7 percent and the final response rate for HINTS 4 Cycle 2 (2012-2013) was 40%. Further details on survey design and sampling strategies have been published elsewhere (Finney Rutten, Davis et al., 2012).

Americans for Nonsmokers' Rights (ANR), 2011

The Americans for Nonsmokers' Rights Foundation (ANR) collects data on the presence and comprehensiveness of clean indoor air ordinances across the U.S and at the state and county level. Data include information on smoke-free workplaces, bars, and restaurants, as well as effective dates for each policy. Data for the current analysis are from October 2011, and were therefore in effect before smoking behavior was assessed in HINTS 4 Cycle 1.

Campaign for Tobacco Free Kids (CTFK), 2011

The Campaign for Tobacco Free Kids regularly provides data to the public on state cigarette excise taxes and state rankings based on levels of cigarette taxation. Data for the current analysis were derived for CTFK from Orzechowski & Walker's *Tax Burden on Tobacco*, 2011 (http://www.taxadmin.org/fta/tobacco/papers/Tax_Burden_2011.pdf), and are publicly available on the CTFK Web site: http://www.tobaccofreekids.org/research/factsheets/pdf/0097.pdf. Data for the current analysis are from Fiscal Year 2011, and were therefore in effect before smoking behavior was assessed in HINTS 4 Cycle 1.

Data Integration

To create the integrated HINTS 4 dataset to include both HINTS 4 Cycle 1 and HINTS 4 Cycle 2, as well as state-level tobacco policy data from ANR and CTFK, we first combined the HINTS 4 Cycle 1 and Cycle 2 data using a similar method as illustrated in Chapter 2 of this document, so that 100 replicate weights were created in the combined HINTS file. We multiplied a factor of 0.5 to each full sample and replicate weights so the sum of each set of weights represent only one population. The file also contains state of residence for each respondent [3]. We then sorted the HINTS data by state, then matched and entered the respective state cigarette excise taxes (as a continuous variable) and state tax rank information, as well as information on 100 percent comprehensive indoor smoking policies (Yes/No), to each state. The newly created variables were:

- STATETAXRATE2011—State excise tax rate
- TAXRANK—Ranking of tax rates by state
- COMPREHENSIVE—Yes/No to 100 percent comprehensive indoor air policy
- STATENUM—State coded as a numeric variable

Methods

Measures

Smoking Behavior (primary outcome): Using the standard classification method for establishing current smoking status, two questions from HINTS 2011-2012 were used to create a dichotomous outcome for smoking status: "Have you ever smoked 100 cigarettes in your life?" (Yes) and "How often do you now smoke cigarettes?" (Everyday and Some days).

Sociodemographics: Using HINTS 4 Cycle 1 and Cycle 2 demographic questions, categorical and dichotomous variables were created for: age, marital status, education, race/ethnicity, and household income. We also included an indicator variable to distinguish between the two cycles.

[3] State information is generally restricted but may be available for HINTS data users who provide justification through a request to the HINTS program at the National Cancer Institute: http://hints.cancer.gov/contact-us.aspx.

Cigarette Excise Taxes: Using the CTFK data added to the HINTS 4 Cycle 1 and 2 combined dataset, we used the continuous STATETAXRATE2011 variable for the purpose of this analysis. In 2011, the lowest state cigarette tax was $0.30 (Virginia) and the highest state cigarette tax was $4.35 (New York). Data do not include additional county or local tax ordinances that may have been in place in states, nor do they include the nationwide federal cigarette tax rate of $1.01 per pack.

Comprehensive Indoor Air Laws: Using the ANR data added to the HINTS 4 Cycle 1 and 2 combined dataset, we created a dichotomous variable representing states with 100 percent comprehensive clean indoor air laws (in workplaces and bars and restaurants) versus all others. In October 2011, there were 23 states plus the District of Columbia with 100 percent comprehensive clean indoor air laws.

Statistical Analysis

Sample weights were developed to compensate for differential selection probabilities, nonresponse, and undercoverage of the target adult U.S. population. For variance estimation, replicate weights were generated using the jackknife replication method (Wolter, 1985). To address the issue of nonindependence of responses from members of the same household, all respondents from the same household were assigned to the same replicate weight, which accounts for clustering within the primary household sampling unit. We conducted our descriptive analyses using SAS-callable SUDAAN (SUDAAN Language Manual, 2008) and developed our own algorithm for multilevel modeling analysis to account for the complex survey data.

We used multilevel statistical procedures to model the variation in adult tobacco use according to individual-level socioeconomic status (fixed parameters) and state-level tobacco control policies (random parameters). We hypothesized that individual-level demographic characteristics would not fully account for geographically patterned differences in smoking behavior, and that state-level policy differences (aggregate exposures) would account for much of the variation in smoking behavior in the population.

We first conducted weighted descriptive analyses for the HINTS 4 sample as a whole and for current smokers in HINTS, by socioeconomic status and other known individual-level predictors of smoking behavior. Second, we ran a series of weighted multilevel logistic regression models using SAS PROC GLIMMIX (SAS 9.2 User's Guide; see Appendix C, Section 1 of this report) and the full sample weight to obtain the population-level point estimates for the parameters of interest including the odd ratios for the predictors and the random effect variance. To compute the correct

standard errors for each parameter of interest accounting for the complex design, we ran PROC GLIMMIX iteratively using each replicate weight and then combined the results from each replicate using the jackknife variance estimation formula (Wolter, 1985). Data were analyzed using a multilevel structure with respondents (level 1) nested within states (level 2). We focused on the fixed effects of both individual and state-level variables, but allowed for heterogeneity between states in order to let the average relationship between smoking and SES to vary between states. We fit four 2-level random intercepts models in a stepwise fashion: The first was a null model including no fixed effects, the second model (model 1) included only individual-level variables, the third (model 2), only state-level policy variables (indoor air policies and tobacco taxes), and the fourth model (model 3) included both the state-level variables and the individual-level SES variables. Statistical significance was evaluated at $p < .05$.

We excluded the replicate from the computation of standard errors if the model didn't converge with the specific replicate weights.

Results & Discussion

In HINTS 4 Cycles 1 and 2 (N=7589), the unweighted frequency of smokers was 1201 (16.11%) and the weighted frequency was 42.67 million (18.26%), reflecting the similar average U.S. smoking prevalence commonly represented by BRFSS and TUS-CPS.

State Cigarette Tax Rate

Results presented in our full model (Table 4-1, model 3) reveal no significant effects for a one unit (one dollar) increase in state cigarette taxes being associated with a decrease in odds of smoking. This finding is consistent with what Chahine et al (2011) found: that state cigarette tax was among the factors that explained a larger proportion of state variance in smoking behavior, but that individually, taxes had no statistical significance. In a 2006 study, Osypuk et al found that higher state excise tax was at least marginally associated with lower individual smoking odds for both Hispanic women and men, at approximately 0.94–0.95 odds of smoking for each $.10 higher tax rate. However, the associations presented by Osypuk were unexpectedly reversed for blacks and not significant for whites. Black men and women had significantly higher odds of smoking in states with higher tax rates (odds of 1.08 and 1.03 for women and men respectively, for each $.10 increased tax rate).

Comprehensive Smokefree Laws

In our study, individuals in states with comprehensive indoor air laws had neither an increase or decrease in odds of smoking. Our nonsignificant findings for the impact of indoor air laws is consistent with other studies. In a 2011 study, Chahine et al found that indoor air laws were among the factors that explained a larger proportion of state variance in smoking behavior, but that individually, these laws had no statistical significance.

Individual SES and other Sociodemographic Variables

At the individual level, probabilities of smoking were strongly associated with known compositional predictors of smoking behavior, such as SES. We saw gradations of effect by education and income, as well as significant differences by age and marital status (Table 4-1, model 3).

Random Effect Variance

Table 4-2 presents the random effect variance for the influence of smokefree air laws and cigarette tax rates on smoking behavior, showing the unexplained variation in smoking at the state level after including both compositional and contextual variables. Using only individual-level variables reduced the unexplained state-level variance to 2.138, a 4.8% reduction from the null. Including only the contextual variables barely reduced the unexplained state-level variance to 2.245, only a 0.04% reduction in unexplained between-state variance from the null model. Due to the insignificance of both state-level variables, including them in the model with individual-level variables added some noise to the full model (model 3) compared to the model which included individual-level variables only, thus a smaller state-level reduction was found from the full model (3.7% change from the null) compared to the model with individual-level variables only (model 1). The standard errors of the random effect estimates are all unexpectly large, indicating a large variation due to the weighting.

Table 4-1: Adjusted Odd Ratios from Random Intercepts Two-level Logistic Models of Smoking Prevalence: Influence of State-Level Smokefree Air Laws and Cigarette Tax Rates and Individual-Level SES

Variable	Model 1 Individual variables only OR (95% CI)	Model 2 State variables only OR (95% CI)	Model 3 Full model OR (95% CI)
State Level			
Cigarette Tax Rate		0.66 (0.29, 1.48)	0.71 (0.31, 1.60)
Comprehensive Smokefree Law			
No (Ref)		1.00	1.00
Yes		1.83 (0.95, 3.15)	1.80 (0.96, 3.37)
Individual Level			
Age Group			
75+ (Ref)	1.00		1.00
18-34	**3.15 (1.78, 5.59)**		**3.15 (1.78, 5.59)**
35-49	**3.61 (2.14, 6.10)**		**3.61 (2.14, 6.09)**
50-64	**2.91 (1.87, 4.54)**		**2.91 (1.86, 4.57)**
65-74	1.41 (0.85, 2.34)		1.41 (0.84, 2.36)
Marital Status			
Married (Ref)	1.00		1.00
Others	**1.41 (1.07, 1.87)**		**1.41 (1.07, 1.87)**
Education			
College Graduate or more (Ref)	1.00		1.00
Less than high school	**3.5 (2.22, 5.51)**		**3.5 (2.23, 5.50)**
High school graduate	**3.19 (2.25, 4.51)**		**3.19 (2.24, 4.52)**
Some college	**2.38 (1.71, 3.30)**		**2.38 (1.71, 3.31)**
Racial/Ethnicity			
Non-Hispanic White (Ref)	1.00		1.00
Non-Hispanic Black	0.73 (0.51, 1.06)		0.73 (0.51, 1.06)
Non-Hispanic Asian	0.5 (0.25, 1.00)		0.5 (0.25, 1.01)
Non-Hispanic other	0.86 (0.41, 1.80)		0.86 (0.38, 1.94)
Hispanic	0.76 (0.47, 1.24)		0.76 (0.47, 1.24)
Household income			
$75K+ (Ref)	1.00		1.00
Less than $20K	**2.54 (1.57, 4.10)**		**2.54 (1.57, 4.11)**
$20K - <$35K	**1.77 (1.15, 2.72)**		**1.77 (1.15, 2.72)**
$35K - <$50K	1.39 (0.87, 2.24)		1.39 (0.87, 2.23)
$50K - <$75K	1.22 (0.78, 1.91)		1.22 (0.78, 1.91)
Cycle			
Cycle 2(Ref)	1.00		1.00
Cycle 1	0.87 (0.68, 1.12)		0.87 (0.68, 1.12)

Notes:

Model 1 only includes the individual-level predictors

Model 2 only includes the state-level predictors

Model 3 includes all the predictors

*Three replicate weights were excluded from model 1 due to nonconvergence
*Five replicate weights were excluded from model 2 due to nonconvergence
*One replicate weight was excluded from model 3 due to nonconvergence

Table 4-2: Random Effect Variance at the State Level: Influence of Smokefree Air Laws and Cigarette Tax Rates on Smoking Behavior

Null Model	Model 1 Individual variables only		Model 2 State variables only		Model 3 Full model	
Variance (SE)	Variance (SE)	% change from Null	Variance (SE)	% change from Null	Variance (SE)	% change from Null
2.246 (4.416)	2.138 (2.520)	4.8	2.245 (2.331)	0.04	2.162 (2.390)	3.7

*Three replicate weights were excluded from model 1 due to nonconvergence
*Five replicate weights were excluded from model 2 due to nonconvergence
*One replicate weight was excluded from model 3 due to nonconvergence

Model-Based State Level Estimates for Cancer Related Knowledge Variables Using HINTS Data 5

GIS maps using HINTS data can provide a visual representation of possible geographic relationships in HINTS cancer-related knowledge variables. However, due to instability in some state values from relatively small sample sizes, the GIS maps that have been developed cannot provide specific state-level estimates of HINTS variables. Rather, they can mainly illustrate regional differences. The goal of this section is to produce model-based state level estimates for the knowledge variables using small area estimation (SAE) techniques.

We will estimate the state level proportions of people who answered "YES" to the following cancer-related knowledge questions using HINTS:

- Does smoking increase your chance of cancer a lot? (CK13 in HINTS 1)
- Does lung cancer cause the most deaths? (CK15 in HINTS 1)
- Have you ever heard of a sigmoidoscopy or colonoscopy? (CC15 in HINTS 1)
- Have you ever heard of a stool blood test? (CC4 in HINTS 1)
- At what age are people supposed to start having sigmoidoscopy or colonoscopy exams? Proportion of people whose answer is 50. (CC24 in HINTS 1)
- Have you ever heard about HPV? (CV11 in HINTS 2)
- Have you ever looked for cancer information from any sources? (HC09 in HINTS 1, CA08 in HINTS 2 and HC08 in HINTS 3)
- Have you ever looked for information about health or medical topics from any source? (HC01 in HINTS 3)

Brief Background of Small Area Estimation Techniques

A considerable amount of methodological research on SAE has been conducted in recent years. The key idea in SAE is to combine information from a variety of relevant sources to form model-based estimates that generally increase the effective sample size thus increasing precision. These model-based estimates are based on mathematical models that supplement the direct estimates with information from other sources, such as administrative or census records. A comprehensive account of the range of SAE methods can be

found in the recent definitive book on this subject by Rao (2003). Recent uses of large-scale survey data to produce small-area proportions can be found in the National Cancer Institute's recently launched website on "Small Area Estimates for Cancer Risk Factors & Screening Behaviors," and the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program (Citro & Kalton, 2000; Maples & Bell, 2005), etc.

Proposed Small Area Estimation Model

Let N_i denote the population size in state i of the target finite population ($i = 1,...,m;$). Let y_{ik} be the binary response for the characteristic of interest for unit k in state i $(k = 1,..., N_i)$. The parameters to be estimated are the small area proportions $P_i = \Sigma_k y_{ik} / N_i$.

Let n_i denote the sample size in state i and w_{ik} denote the sampling weight for sampling unit k in state i. The standard direct survey estimator for P_i is:

$$p_{iw} = \frac{\sum_{k=1}^{n_i} w_{ik} y_{ik}}{\sum_{k=1}^{n_i} w_{ik}}, i = 1,...,m.$$

The variance of p_{iw} can be expressed as

$$VAR_{st}(p_{iw}) = \frac{P_i(1-P_i)}{n_i} DEFF_i,$$

where $DEFF_i$ is the design effect reflecting the effect of the complex sample design (Kish, 1965).

The problem is that p_{iw} is very imprecise when the sample size n_i is small or even cannot be computed if the sample size is zero. Small area estimation procedures can be used to address this problem.

Let $z_i = \arcsin\left(\sqrt{p_{iw}}\right)$. The following small area model is applied:

The sampling model: $z_i|\theta_i \sim N\left(\theta_i, \frac{DEFF_i}{4n_i}\right)$; (5.1)

The linking model: $\theta_i = x_i'\beta + v_i$; where $v_i \sim N(0, A)$. (5.2)

The sampling model takes account the sampling error for the direct estimate of z_i. The linking model assumes the model parameter θ_i is related to a set of auxiliary variables θ_i. Our goal is to estimate $P_i =$

$sin^2(\theta_i)$. We use the hierarchical Bayesian (HB) method based on the following commonly used prior assumptions for the hyper-parameters β and A:

$$\beta \propto 1;\ A \sim unif(0,\ 100)$$

The HB estimates of P_i are produced using the Markov Chain Monte Carlo (MCMC) technique (Robert & Casella, 1999; Rao, 2003, Sec. 10.2) implemented in WinBUGS software (Lunn et al., 2002). See Appendix D, Section 1 for the WinBUGS code.

For each outcome, an estimate of the design effect $DEFF_i$ is required for state i in the sampling model. We use the Kish formula to estimate the design effect, which is defined as the ratio of the variance under the complex design over the variance under simple random sampling. Due to the small sample size, design effect computed for each state is not very precise. We, therefore, compute the design effect at census region levels first, which are more reliable, then let the individual state level design effect equal to the corresponding regional level design effect for smoothing purposes.

Auxiliary Variables

Finding a good set of auxiliary variables is the key for model-based SAE approaches. For this study, the pool of the auxiliary variables includes the following state-level demographic and socioeconomic variables obtained from Census 2000 and other administrative sources:

- % people in urban areas among total Pop
- % Hispanics among Pop 18+
- % Blacks among Pop 18+
- % American Indian and Alaska Native among Pop 18+
- % Asian among Pop 18+
- % Native Hawaiian and Other Pacific Islander among Pop 18+
- % two or more races among Pop 18+
- % males among Pop 18+
- % 65+ among Pop 18+
- % 1 person household
- % family with own kids under 18

- % married but separate among Pop 15+
- % widowed among Pop 15+
- % divorced among Pop 15+
- % foreign born among Total Pop
- % living in an MSA/PMSA in 2000 among total Pop
- % commute time to work >=30 mins among Pop 16+
- % less than high school among Pop 25+
- % high school graduates among Pop 25+
- % college graduates among Pop 25+
- % graduate school degree among Pop 25+
- % at least Bachelor's degree among Pop 25+
- % unemployment among Pop 16+
- % below 150 percent poverty in 1999 among whose poverty status is determined
- % white collar workers among Pop 16+
- % households that are linguistically isolated
- % households with social security income

Since we only have 51 small areas (i.e., m=51), including all the auxiliary variables listed above would potentially overfit the model. For each outcome, a backward model selection procedure was therefore applied to select a reduced set of auxiliary variables. Logarithm transformation was applied to those auxiliary variables.

Model Evaluation

Typical model evaluation procedures are applied to evaluate the model fitting. Three measures were computed to assess the goodness of fit (Rao, 2003, Chapter 10):

- Global measure that compares two discrepancy measures, one based on the difference between the model-based and direct-state estimates, and the other based on the difference between the model-based estimates and estimates simulated from the posterior normal distributions for the model-based state estimates;

- State-level measure computed as the proportion of the final MCMC samples that had a smaller simulated value based on the sampling model with the estimated θ_i (as opposed to direct estimates); and

- State-level measure that is computed as the difference between the mean of the simulated values based on the sampling model with the estimated θ_i and the direct estimate, divided by the standard deviation of the simulated values, where the mean and standard deviation of the simulated values are computed across the final MCMC samples.

For each outcome, the three diagnostic measures show that the model fit the state-level data well. We also plot the ratio of the direct estimate over the model-based estimate against the state-level sample size (see Figure 5-1 below as an example). As expected, the ratio converges to 1 as the sample size gets larger.

Final Results

We present the final estimates for all 10 outcomes in two tables. The comparison of the estimated percentages of people who sought cancer information from any source in HINTS 1, HINTS 2, and HINTS 3 is displayed in Figure 5-2. We also demonstrate how those estimates look on U.S. maps.

Table 5-1 presents the state level model-based estimates along with the 95 percent confidence intervals for the following five knowledge variables: HC09, CA08, and HC08 (Have you ever looked for cancer information from any source? [HINTS 1, 2, and 3]), HC01 (Have you ever looked for information about health or medical topics from any source? [HINTS 3]), CK13 (Does smoking increase your chance of cancer a lot? [HINTS 1]).

Figure 5-1. The ratio of the direct estimate over the model-based estimate against the state-level sample size

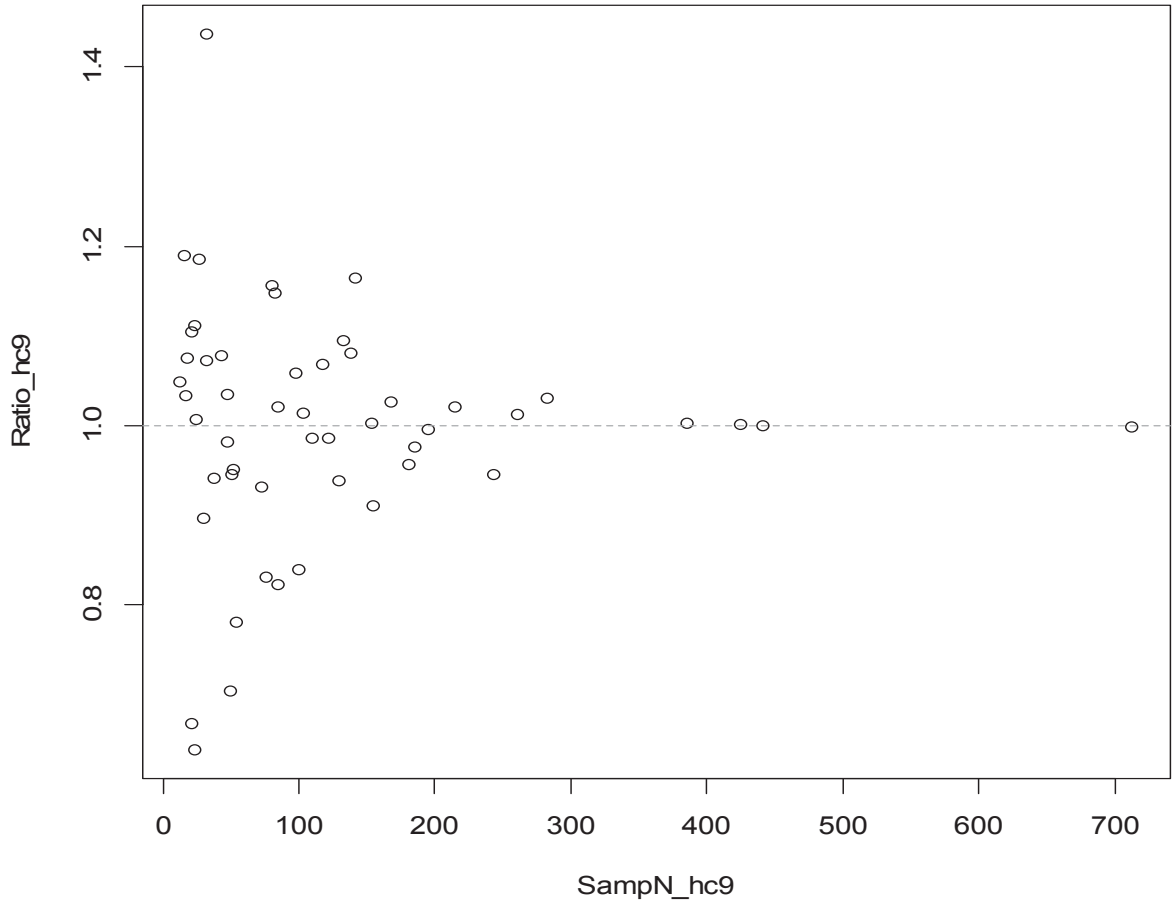


Table 5-2 presents the state level model-based estimates along with the 95% confidence intervals for the remaining five knowledge variables: CK15 (Does lung cancer cause the most deaths, HINTS 1) CC15 (Have you ever heard of a sigmoidoscopy or colonoscopy, HINTS 1), CC4 (Have you ever heard of a stool blood test, HINTS 1), CC24 (At what age are people supposed to start having a sigmoidoscopy or colonoscopy? Proportion of people whose answer is 50, HINTS 1), and CV11 (Have you ever heard about HPV, HINTS 2).

Table 5-1. Model-based estimates and 95 percent confidence intervals for HC09, CA08, HC08, HC01, and CK13

State	HC9	95% CI	CA08	95% CI	HC08	95% CI	HC01	95% CI	CK13	95% CI
AL	49.0	(40.2,56.5)	49.1	(40.2,58.6)	37.6	(29.5,45.8)	69.4	(61.7,76.6)	77.8	(69.9,85.3)
AK	42.1	(25,59.9)	38.6	(20.5,58.2)	36.0	(19.1,53.8)	65.2	(49.8,79.7)	97.3	(88.1,100)
AZ	47.9	(40.4,55.5)	48.2	(39.5,56.5)	38.5	(30.9,46.5)	70.3	(63.7,78.2)	86.3	(78.9,92.4)
AR	41.3	(33,49.2)	43.1	(33.7,53)	36.8	(27.7,45.3)	65.4	(57.5,72.3)	80.0	(71.4,87.4)
CA	44.4	(40.6,48.2)	42.8	(38.4,47.3)	37.5	(34.1,41.1)	66.9	(62,71.4)	86.5	(81.3,91.1)
CO	47.0	(38.9,55.3)	52.7	(43.4,61.6)	37.5	(29.4,45.8)	75.2	(67.6,82.1)	92.9	(85.9,98.1)
CT	53.9	(44.5,62.3)	51.5	(41.4,62.5)	39.4	(27.9,49.8)	72.5	(65.7,78.1)	85.3	(79,91.5)
DE	50.3	(42.2,58.6)	52.7	(42.9,62.1)	42.3	(32.9,52.2)	68.9	(61.2,75.8)	76.8	(68.6,84.9)
DC	41.8	(24.2,59.8)	56.9	(29.5,83.1)	19.0	(4.5,38.5)	65.4	(52.2,77.3)	97.7	(89,100)
FL	42.1	(37.1,47.2)	50.2	(42.6,58.2)	35.9	(30.5,41.4)	69.3	(63.8,74.5)	82.9	(76.4,88.6)
GA	40.6	(34.6,46.8)	49.1	(40.7,57.9)	40.0	(33,47.1)	71.2	(64.5,77)	89.1	(83.2,94.4)
HI	28.6	(12.8,46.6)	33.6	(13.3,56.5)	50.0	(33.1,67.3)	75.5	(66.4,84.3)	97.6	(88,100)
ID	40.9	(31,51)	60.4	(49.5,70.9)	42.3	(31.2,53.6)	68.3	(58,78.1)	84.2	(75.2,91.8)
IL	39.7	(34,45.1)	53.4	(46.3,60.7)	39.8	(34,45.7)	72.9	(68.2,77.5)	84.8	(78.9,89.4)
IN	49.6	(43.3,56.6)	45.6	(38.2,52.9)	42.1	(33.9,49.4)	70.9	(64.7,76.4)	80.6	(73.9,86.2)
IA	36.1	(27.9,44.1)	44.7	(35.6,53.6)	43.0	(34.6,51.9)	69.8	(62.9,76.4)	84.9	(77.3,90.8)
KS	48.0	(39.7,56.7)	45.6	(36.6,54.7)	41.4	(32.4,50.3)	73.0	(66.4,79.4)	87.3	(80.5,93.8)
KY	46.2	(38.8,54)	47.4	(37.6,57.8)	36.6	(27.7,45.4)	64.8	(56.8,72.4)	79.9	(71.2,87.2)
LA	46.6	(38.9,55.2)	48.8	(38.6,58.4)	36.5	(28.1,45.3)	67.0	(59.9,74.8)	85.2	(78.1,92)
ME	43.8	(31.9,55.6)	47.7	(32.7,62.4)	45.4	(32.3,58.9)	73.5	(63.8,82)	97.4	(91,100)
MD	46.9	(39.6,54.8)	61.8	(52,72)	37.9	(29.5,46.6)	69.0	(61.4,76.3)	81.2	(73.6,88.5)
MA	49.0	(42.4,55.5)	53.8	(45.1,61.8)	39.9	(32.9,47.4)	71.3	(64.9,76.9)	88.3	(82.5,93.5)
MI	48.6	(42.7,54.5)	53.3	(46.5,60.3)	40.9	(34.5,47.5)	69.9	(64.9,74.7)	84.1	(78.7,89.3)
MN	38.2	(31.4,45)	52.9	(44.1,61.1)	37.0	(29.8,44.4)	74.0	(68.8,80)	90.6	(84.6,96)
MS	40.5	(31,50.4)	56.5	(44.2,68.3)	41.5	(31.9,51.4)	69.1	(58.9,78.3)	72.5	(62,82.4)
MO	40.7	(34.1,46.6)	55.6	(48.4,62.7)	35.9	(28.6,42.6)	68.1	(61.3,73.1)	82.6	(75.7,88.4)
MT	45.1	(32.3,58.4)	62.1	(48.3,75.5)	33.2	(20,46.9)	71.6	(62.1,80.4)	93.9	(87,98.7)
NE	43.9	(33.6,55.6)	40.3	(29,52.3)	42.7	(31.5,54.4)	72.1	(65,79.1)	91.6	(85.7,97.4)

Table 5-1. Model-based estimates and 95 percent confidence intervals for HC09, CA08, HC08, HC01, and CK13 (continued)

State	HC9	95% CI	CA08	95% CI	HC08	95% CI	HC01	95% CI	CK13	95% CI
NV	35.2	(23.8,46.8)	23.1	(12.7,34.5)	45.7	(33.6,58)	71.7	(61.4,81.1)	97.7	(92.2,100)
NH	46.6	(37.5,56.2)	58.9	(47.6,70.2)	44.8	(34,55.9)	74.9	(67.2,81.9)	89.7	(82.1,95.9)
NJ	43.6	(36.6,50.5)	50.9	(41.3,60.3)	39.8	(32.2,47.4)	72.7	(67.8,77.7)	82.7	(75.3,89.2)
NM	57.0	(46.7,67.4)	47.9	(36.7,59.6)	41.3	(30.8,52.1)	67.0	(59.2,74.6)	83.6	(72.9,93)
NY	46.9	(42.2,51.5)	50.1	(43.3,56.7)	39.3	(33.7,45.1)	72.0	(67.5,76.6)	83.8	(78.8,88.2)
NC	46.0	(40.4,52)	41.0	(33.4,48.8)	43.6	(37.2,50.6)	73.0	(67.5,78.5)	85.3	(79.6,89.9)
ND	38.5	(26,51.7)	36.8	(22.3,52.5)	29.2	(16.4,43.3)	64.7	(53.3,75.3)	98.4	(93.1,100)
OH	45.3	(40.2,50.5)	51.5	(45.1,57.8)	41.4	(35.5,47.6)	69.5	(64.3,74.3)	82.6	(76,87.5)
OK	42.4	(32.4,52.5)	45.5	(32,59.5)	35.7	(24.9,47.4)	67.5	(60.5,73.3)	84.9	(78.4,91.1)
OR	49.2	(42.1,57.3)	49.2	(41.2,57.7)	45.7	(38.3,54.2)	74.2	(68.2,80.9)	92.9	(86.8,97.2)
PA	47.8	(42.8,53.1)	50.3	(43.8,56.5)	42.0	(36.3,48.2)	70.2	(65.8,74.6)	83.1	(78.1,87.9)
RI	52.2	(41.8,62.7)	42.8	(30.5,56.5)	45.7	(33.1,58.2)	74.3	(66.9,81.6)	87.9	(79.8,95.1)
SC	50.1	(42.3,58)	43.7	(33.8,53.3)	41.4	(32.9,49.5)	73.7	(66.8,80.3)	82.7	(75.2,89.6)
SD	38.4	(28.3,48.7)	43.3	(31.6,55)	33.5	(21.8,45.3)	65.3	(53.6,75.8)	95.8	(89,99.7)
TN	45.2	(39.3,51.8)	48.8	(40.8,56.1)	38.6	(31.8,45.1)	70.0	(63.8,75.8)	79.3	(72.1,86.6)
TX	44.2	(39.5,48.9)	44.1	(37.1,51.4)	35.0	(30,40)	67.6	(62.8,72.6)	82.2	(76.1,87.6)
UT	52.8	(39.6,65.4)	57.9	(45.5,69.7)	58.2	(45.4,70.7)	77.5	(69,86)	82.8	(64.4,96.2)
VT	49.3	(36.8,62.6)	62.0	(46.3,76.4)	45.8	(31.8,60.8)	76.4	(64.6,86.6)	90.2	(82.9,96.3)
VA	46.8	(40.1,53.5)	49.4	(40.3,58.4)	37.0	(30.1,44.1)	70.8	(64.5,77.2)	82.2	(75.8,88.2)
WA	44.0	(36.9,52)	53.1	(44.7,61.7)	37.5	(30.2,44.6)	71.2	(64.9,76.8)	95.6	(90.3,99.1)
WV	41.3	(30.2,52.3)	53.6	(38.3,67.7)	28.6	(17.1,41.2)	59.1	(47.3,70.7)	78.3	(67.1,88.3)
WI	45.4	(39.5,51.4)	52.4	(45.5,60.5)	41.2	(34.8,47.9)	72.9	(67.9,78.6)	82.4	(75.2,88.6)
WY	36.8	(23.9,50.5)	59.3	(45,73)	43.9	(29.7,59.3)	74.2	(60.1,87.1)	90.0	(78.4,98.3)

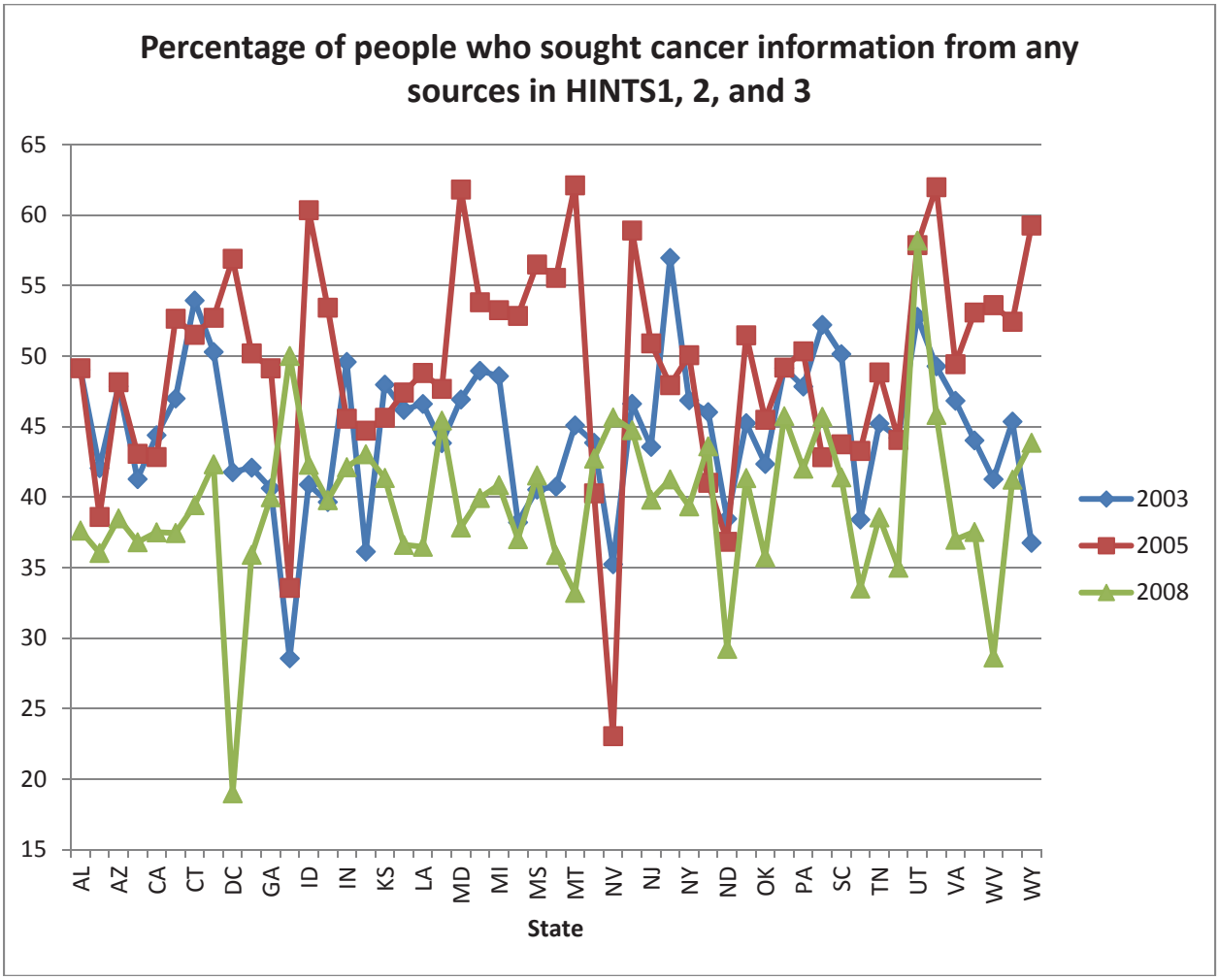
Table 5-2. Model-based estimates and 95 percent confidence intervals for CK15, CC15, CC4, Cc24, and CV11

State	CK15	95% CI	CC15	95% CI	CC4	95% CI	CC24	95% CI	CV11	95% CI
AL	29.4	(24.6,34.7)	73.9	(65.4,81.4)	58.5	(52.5,63.9)	26.5	(19.1,33.8)	39.3	(31,47.9)
AK	28.8	(16.5,43)	64.7	(50.1,79.2)	55.3	(43.2,67.1)	36.4	(21.5,51.9)	23.5	(7,43)
AZ	30.3	(24.4,36.2)	68.6	(61.3,75.6)	53.6	(47.5,61)	28.0	(20.3,36.3)	44.9	(35.7,52.9)
AR	29.1	(23.3,34.3)	75.6	(66.1,84.4)	55.1	(48.5,62.2)	31.3	(23.1,39.9)	33.1	(23.6,41.6)
CA	32.3	(28.8,36)	68.1	(64.8,71.3)	46.8	(42.9,50.7)	29.7	(25.2,34.3)	40.0	(34.3,45.6)
CO	26.6	(21.2,32.7)	71.2	(64.2,77.9)	60.2	(53.4,66.6)	29.8	(21.4,38.4)	49.3	(38.5,59.6)
CT	32.1	(25.4,38.3)	90.8	(84.6,96.3)	62.8	(57.3,69.6)	35.8	(29.1,43)	35.8	(26.5,45.9)
DE	28.6	(23.3,33.6)	80.0	(68.7,88.7)	61.7	(55.2,67.6)	37.4	(27.9,47.2)	33.3	(24.7,43.7)
DC	31.2	(20.9,41.6)	67.4	(48.4,84.3)	58.2	(48.4,67.7)	19.2	(9,31.5)	40.5	(15.3,67.7)
FL	26.6	(22.8,30.7)	74.4	(69.3,79.1)	57.0	(52,61.7)	29.5	(24.5,34.6)	41.7	(34.3,50.1)
GA	28.1	(24,32.4)	71.7	(64.9,78.2)	57.7	(52,62.9)	27.9	(22.1,34.4)	42.9	(35.7,51.1)
HI	55.0	(36.6,72.9)	65.1	(49.8,78.1)	53.1	(45.8,60.4)	21.6	(6.6,39.4)	58.2	(43.4,73.1)
ID	29.1	(23,34.7)	77.0	(68.1,85.4)	48.8	(40.7,57.3)	38.2	(28.5,48.2)	38.9	(29.2,48.6)
IL	30.0	(25.7,34)	76.7	(69.7,83.3)	59.0	(54.4,63.7)	27.7	(22.4,33.1)	43.7	(36.8,51.4)
IN	31.9	(27,37.6)	76.8	(68.5,84.2)	60.0	(54.5,65)	39.3	(32.5,46.5)	33.6	(25.8,42)
IA	36.3	(30.6,42.4)	83.0	(74.1,90.3)	62.6	(56.2,69.2)	38.0	(29.2,46.2)	35.5	(26.4,44.7)
KS	35.0	(29,40.6)	78.9	(68.5,87.6)	60.7	(54.4,67.5)	35.0	(26.5,44.3)	42.6	(33.5,52.4)
KY	28.2	(22.7,33.8)	73.4	(65.3,81.3)	55.0	(48.8,61.2)	38.3	(30.6,47)	38.7	(30.6,47.3)
LA	28.9	(23.5,34.9)	67.6	(57.8,76.2)	54.4	(48,60.4)	32.5	(24.5,41.9)	28.1	(18.5,36.5)
ME	32.3	(26.2,38.4)	90.3	(82.8,96.7)	62.1	(55,69.5)	37.9	(27.6,48.6)	43.7	(34.6,54.2)
MD	25.2	(20.5,30.6)	83.5	(76.8,89.4)	65.6	(60.2,71.9)	39.1	(32,47.2)	38.3	(28.8,49.4)
MA	29.4	(24,34.7)	84.2	(77.7,89.8)	62.6	(56.6,68)	37.6	(31.2,44.3)	38.4	(29.5,47.5)
MI	27.3	(22.6,32)	81.7	(74.9,88.2)	62.4	(57.8,67.6)	35.4	(29.5,41.3)	36.7	(29.4,43.5)
MN	33.3	(27.8,39.5)	82.4	(74.4,89.4)	62.4	(56.7,68.1)	37.3	(30.6,44.1)	44.6	(36.5,52.8)
MS	25.9	(18.5,33.6)	68.0	(56.7,78.1)	49.0	(39.7,58)	22.1	(13.2,31.5)	35.0	(24.5,45.7)
MO	30.6	(25.8,34.9)	76.5	(67.9,83.7)	58.3	(52.5,63)	29.3	(22.5,35.7)	38.2	(30.6,45.3)
MT	36.5	(29,44.8)	79.5	(69.7,88.3)	61.7	(53.8,70.3)	30.8	(20.1,42.3)	55.9	(41.1,71.2)
NE	36.7	(30.1,43.1)	83.1	(73.4,91.8)	60.8	(54,67.7)	35.8	(26.1,45.9)	34.3	(24,44.9)

Table 5-2. Model-based estimates and 95 percent confidence intervals for CK15, CC15, CC4, Cc24, and CV11 (continued)

State	CK15	95% CI	CC15	95% CI	CC4	95% CI	CC24	95% CI	CV11	95% CI
NV	31.6	(22.2,41.4)	49.5	(38.7,59.9)	52.9	(45.5,59.1)	25.9	(15.7,36.8)	33.4	(18.4,50.3)
NH	28.2	(22.3,34.6)	87.1	(78.8,94.1)	64.4	(57.9,71.5)	38.7	(29,48.7)	44.3	(33.3,55.1)
NJ	31.5	(25.8,37.1)	78.7	(72.3,84.4)	61.0	(55,66.1)	25.2	(19.8,30.7)	38.5	(27.7,48.4)
NM	32.3	(23.9,42)	64.2	(53.7,73.3)	45.3	(37.9,52.3)	34.9	(25.6,45.2)	31.9	(21.9,43.7)
NY	26.7	(22.7,31)	81.0	(76.7,85)	54.3	(49.7,58.5)	29.7	(25.4,33.9)	30.9	(23.6,38.1)
NC	24.6	(20.3,29.2)	74.7	(68.3,80.5)	59.5	(54.6,64.3)	32.8	(26.7,39.6)	32.4	(24.5,40.3)
ND	43.9	(32.6,55.7)	69.9	(56.5,81.9)	56.5	(48.6,63.8)	26.2	(14.8,39.3)	51.7	(37,67)
OH	29.0	(25,33.2)	83.2	(77.1,88.8)	64.6	(60.3,69.7)	35.7	(30,41.4)	33.4	(27.2,40.1)
OK	29.7	(22.3,37.3)	76.7	(68.4,84.7)	56.2	(49.7,61.5)	32.2	(25,39.4)	40.3	(32.3,49.1)
OR	31.2	(25.1,36.6)	84.1	(77.8,90.1)	60.1	(53.5,65.6)	38.7	(30.7,47.5)	40.9	(33,48.3)
PA	33.2	(28.4,38.3)	84.6	(80.1,88.9)	64.9	(60.4,69.6)	36.4	(31.7,41.5)	37.8	(31.4,44.3)
RI	30.9	(23.3,38.7)	79.5	(69.2,88.4)	56.1	(49.4,62.9)	40.6	(30.9,51.4)	23.5	(13.2,35.1)
SC	25.3	(20.4,30.4)	72.7	(63.8,80.6)	59.5	(53.5,65.1)	24.0	(16.4,31.2)	36.3	(26.8,47.9)
SD	42.1	(33,52.1)	72.7	(60.1,83.4)	51.9	(42.3,60.9)	30.1	(19.3,41.8)	43.0	(30.6,56.4)
TN	26.5	(22,31.3)	80.4	(73.2,87.2)	61.2	(55.6,68.3)	31.9	(25,38.5)	36.5	(28.7,44.2)
TX	23.4	(20.1,26.7)	62.5	(57.3,67.5)	48.4	(44.3,52.4)	27.1	(22.5,31.9)	35.7	(29.3,43.4)
UT	32.7	(23.2,42.7)	81.8	(72.7,89.8)	47.3	(39,55.4)	40.0	(27.1,53.2)	28.6	(15.5,43.1)
VT	31.6	(25.2,39.3)	85.9	(77.3,93.3)	56.1	(46.5,64.9)	39.1	(27.8,51.2)	52.1	(40.2,64.3)
VA	23.3	(19,27.6)	77.5	(71.1,83.5)	62.0	(56.9,67)	31.7	(25.5,38)	44.0	(35,53.3)
WA	33.6	(28,39.3)	71.7	(64.9,78)	59.4	(52.6,65)	32.6	(24.7,40.5)	44.8	(36.7,53.4)
WV	31.7	(24.8,39)	75.1	(65.1,84.2)	62.6	(55.3,69.9)	33.4	(23.2,44.2)	57.1	(42.2,71.6)
WI	33.1	(28.5,38.1)	76.5	(68.2,84)	61.0	(56,66.5)	38.1	(31.5,45.2)	36.9	(29.2,44.4)
WY	29.6	(21.9,37.6)	81.7	(71.7,91.2)	58.9	(50,67.9)	34.7	(24.5,45.5)	45.1	(34.2,56.4)

Figure 5-2. A plot of the estimated percentages of people who sought cancer information from any source in HINTS 1, 2, and 3. From the chart, we can see that in general more people sought cancer information in HINTS 2 and fewer people sought cancer information in HINTS 3



Figures 5-3 to 5-12 present the final estimates on U.S. maps.

Figure 5-3. State level model-based estimates for percentage of people who have ever looked for cancer information from any source based on HINTS 1

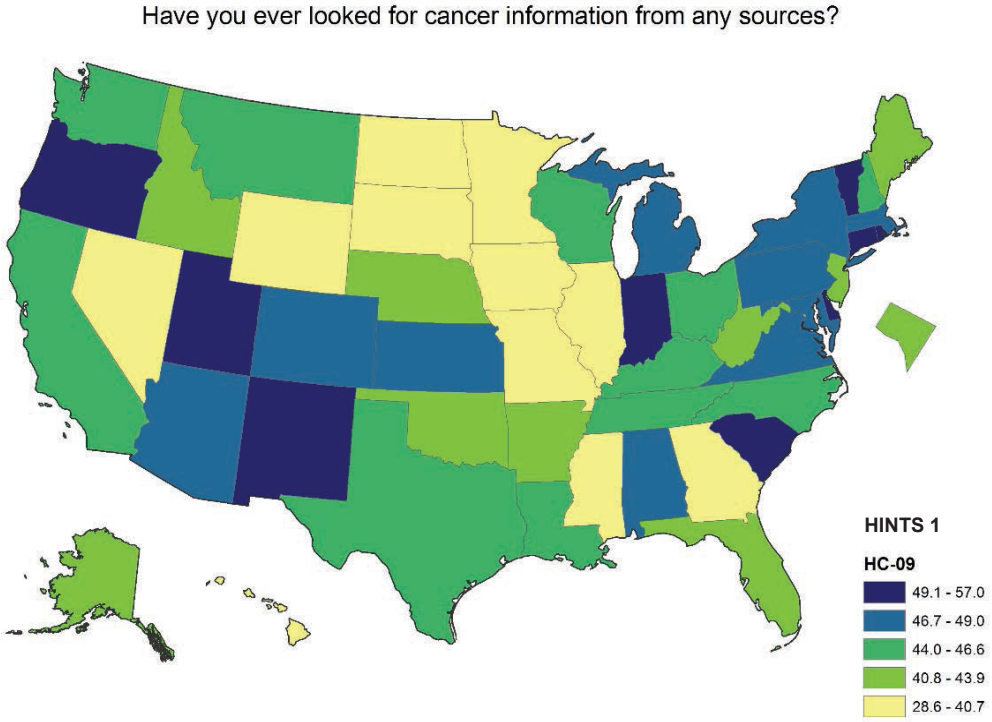


Figure 5-4. State level model-based estimates for percentage of people who have ever looked for cancer information from any source based on HINTS 2

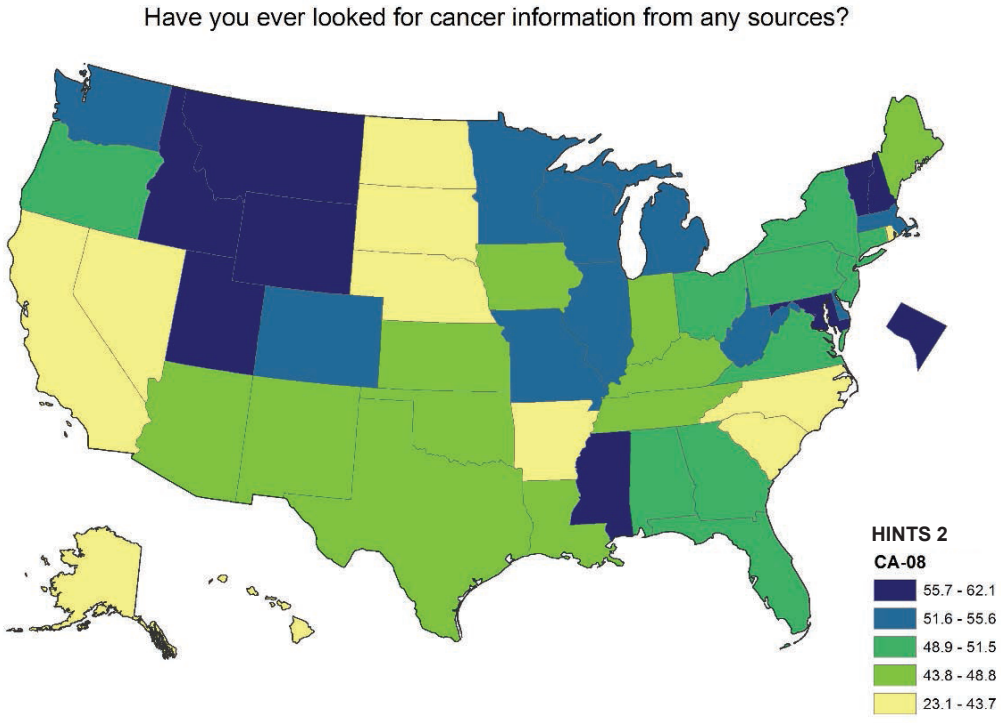


Figure 5-5. **State level model-based estimates for percentage of people who have ever looked for cancer information from any source based on HINTS 3**

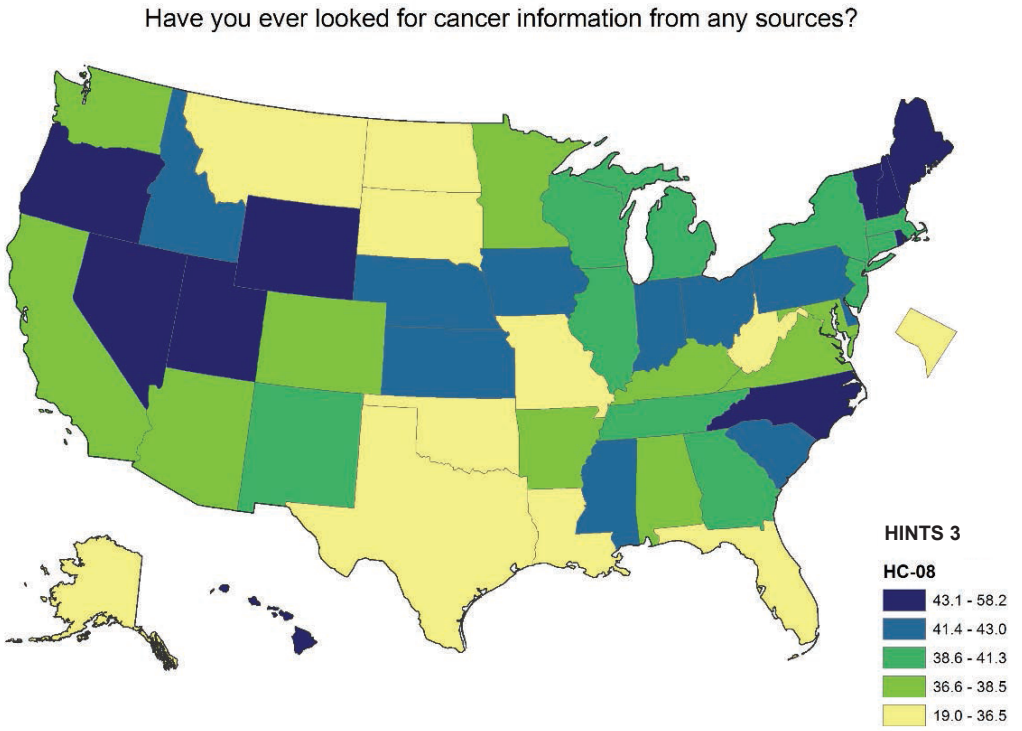


Figure 5-6. **State level model-based estimates for percentage of people who have ever looked for information about health or medical topics from any source based on HINTS 3**

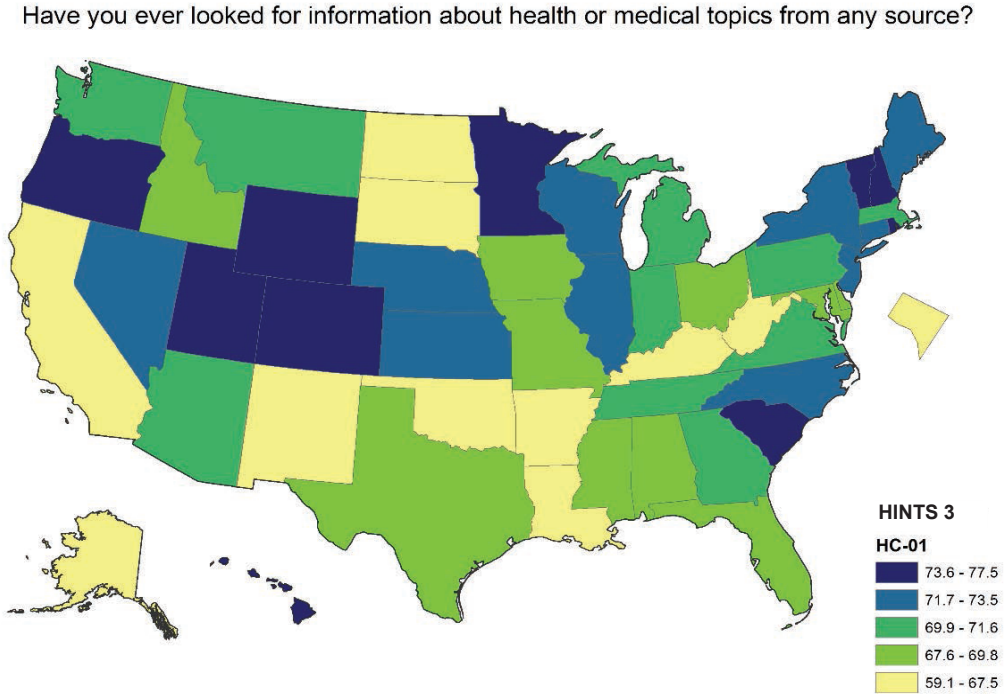


Figure 5-7. State level model-based estimates for percentage of people who believe smoking increases the chance of cancer a lot based on HINTS 1

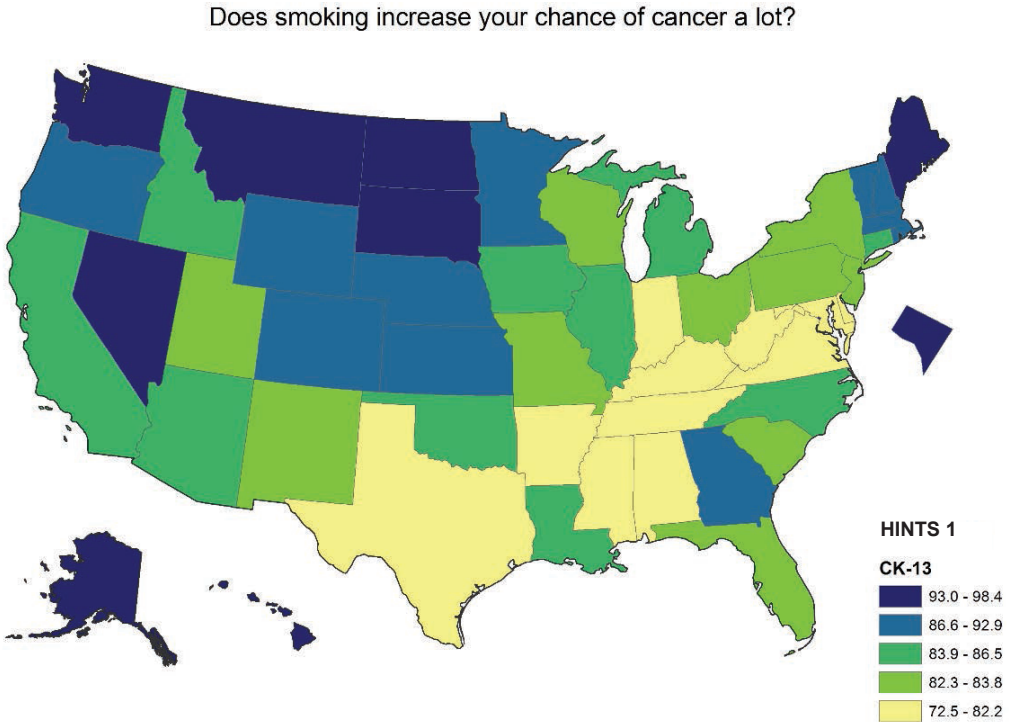


Figure 5-8. State level model-based estimates for percentage of people who believe lung cancer will cause the most deaths based on HINTS 1

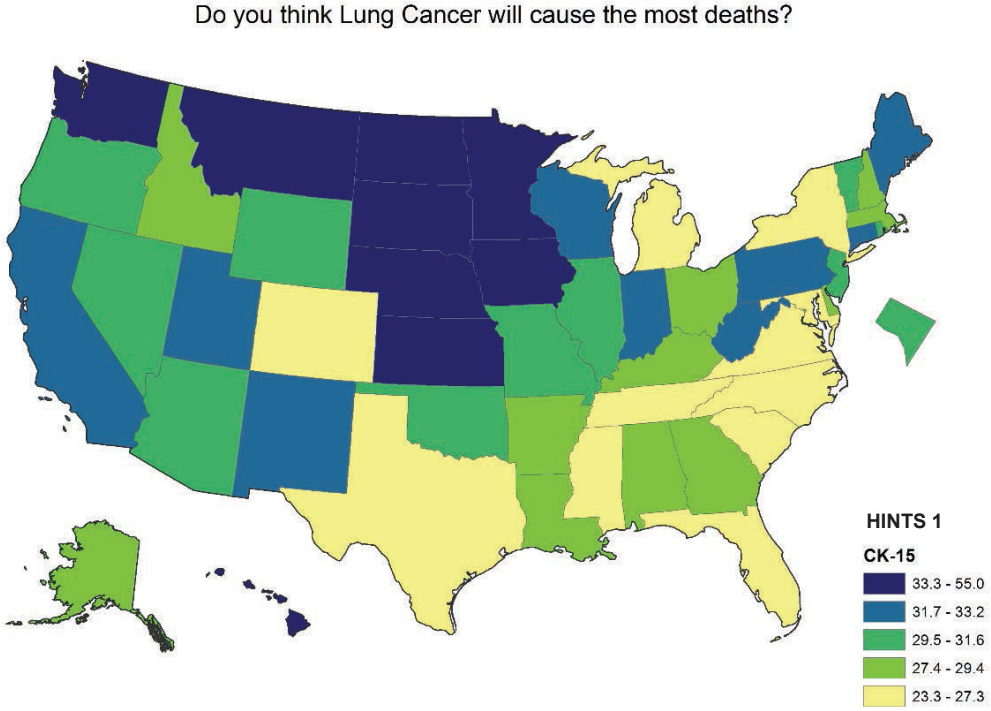


Figure 5-9. State level model-based estimates for percentage of people who have ever heard of a sigmoidoscopy or colonoscopy based on HINTS 1

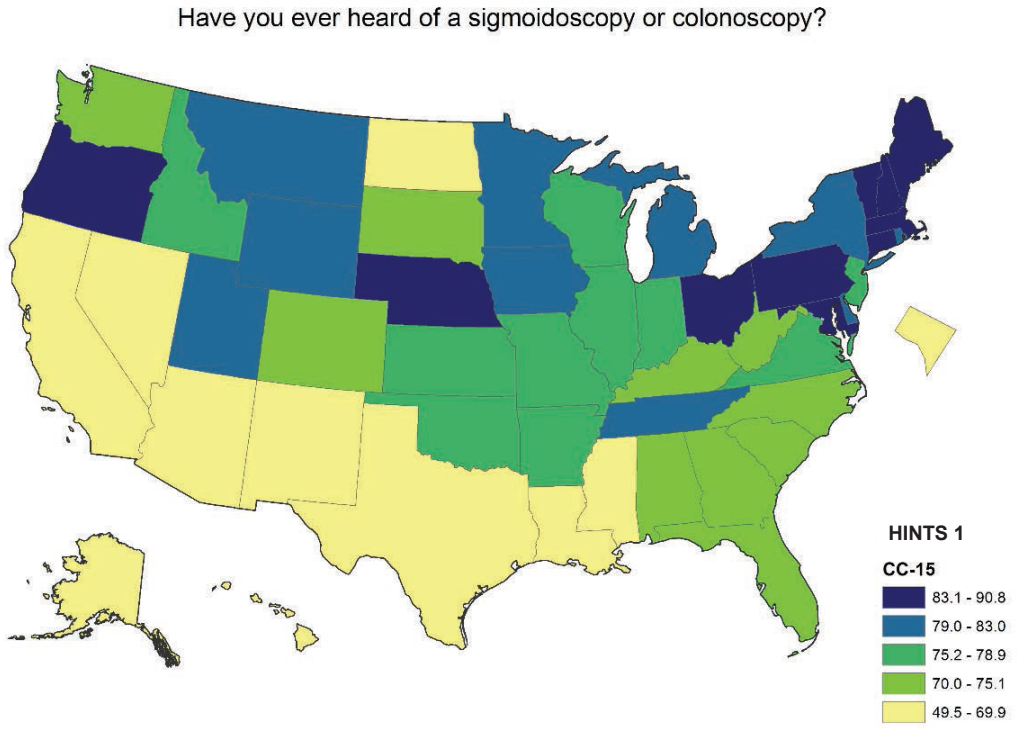


Figure 5-10. State level model-based estimates for percentage of people who have ever heard of a stool blood test based on HINTS 1

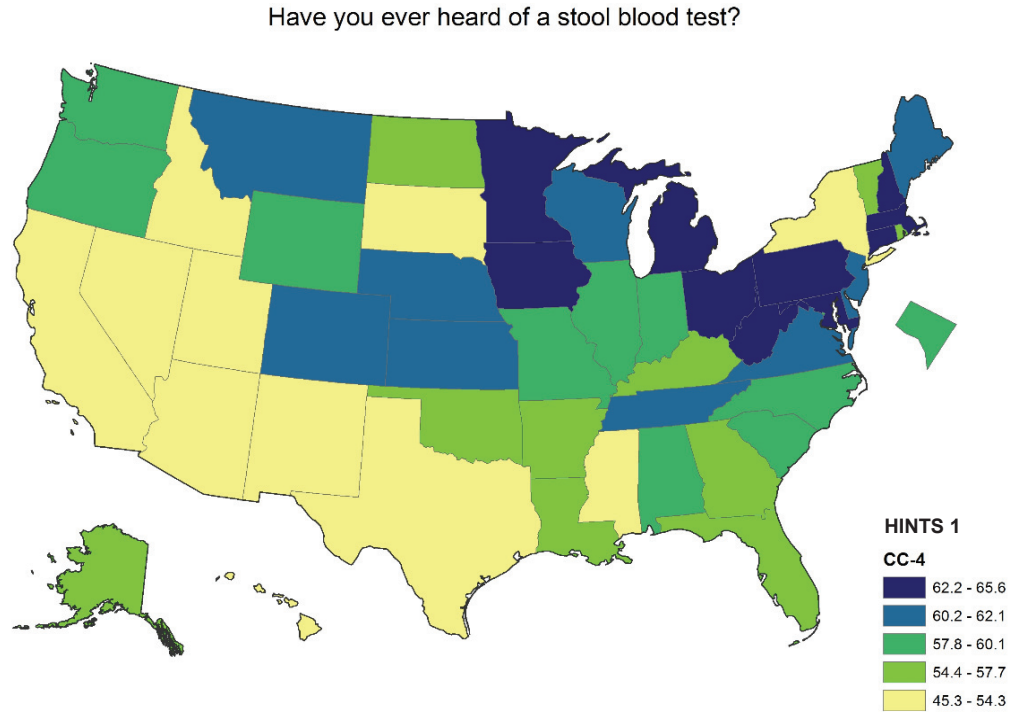


Figure 5-11. State level model-based estimates for percentage of people who think 50 is the age at which people are supposed to start having sigmoidoscopy or colonoscopy exams based on HINTS 1

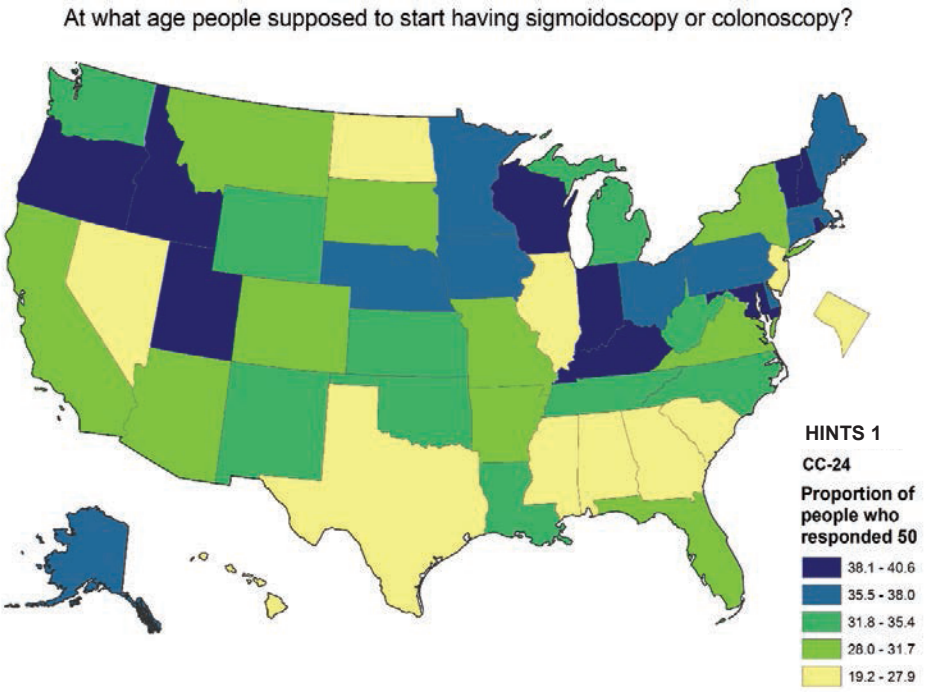
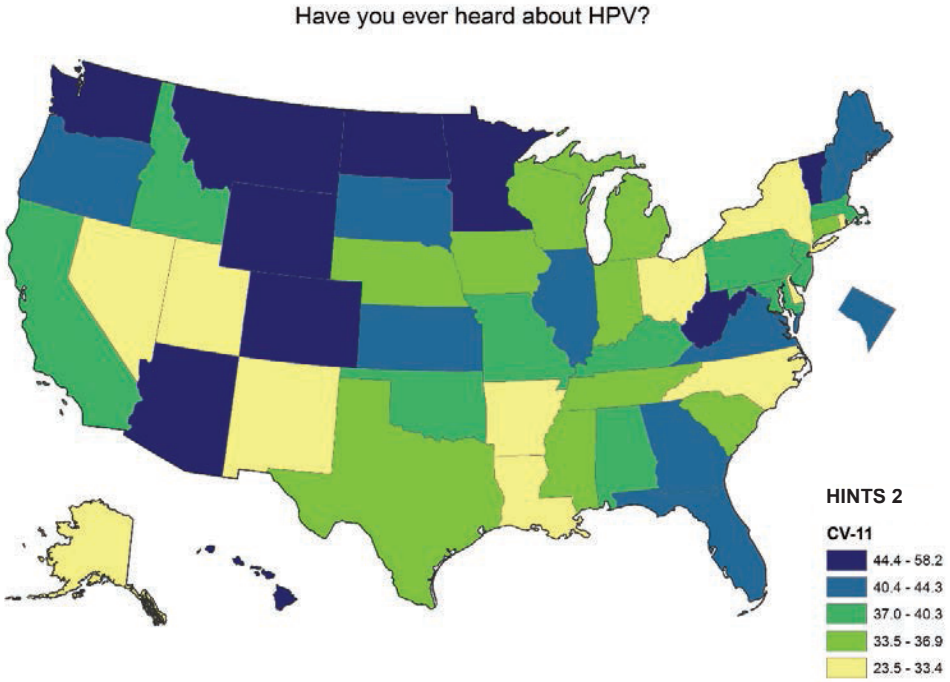


Figure 5-12. State level model-based estimates for percentage of people who have ever heard about HPV based on HINTS 2



Using Imputation to Augment Multiple Iterations of HINTS Data 6

The topic of missing data has gained considerable attention in the last decade. This section distinguishes between two types of missing data and provides examples for handling it using imputation techniques.

Item-Level Missing Data

The first type of missing data is caused by item nonresponse, which occurs when a respondent is asked a survey question, but fails to provide an answer. This type of missing data is very common despite the efforts to improve the completeness of data collection. Given the considerable time and expense of performing surveys, and the desire to make generalizable inferences, it is a waste of resources and may lead to the potential for systematic errors when the researchers discard observations with missing values or include indicators for missing data. For an overview of missing data analysis including data missing mechanisms and methods of missing data treatments, see Allison (2001). For more technical discussions on statistically sophisticated methods, such as single and multiple imputation, likelihood methods, and Bayesian method, see Little and Rubin (2002).

Overview of Imputation Method

Imputation is a flexible method for handling missing data. Imputations could be means or random draws from a predictive distribution of missing data. Mean imputation leads to bias in both the point estimate of a parameter, for example, a simple linear regression coefficient, and the standard error of the estimate (in the direction of underestimation). Draw imputation, on the other hand, can have unbiased point and variance estimates, if carefully implemented. Because of these reasons, this application only considers draw imputation. A second issue concerns which models to use to generate the predictive distribution of missing data. The models can be explicit, as in the case of stochastic regression imputation, or implicit, as in the case of hot deck imputation. A third issue is how many imputations to generate, whether it is one as in a single imputation, or a small number, (i.e., 5, 10, or 20) as in a multiple imputation. Both single and multiple imputations lead to valid point estimates. Valid variance estimates can be achieved for single imputation with the use of resampling procedures, such as bootstrap (Efron, 1979) or jackknife (Miller, 1974). For multiple

imputation, the standard errors are obtained by combining the between-imputation and the within-imputation variances (Rubin, 1987).

Single Imputation of Missing Income Data in HINTS 4 (Cycle 1)

In this section, we illustrate the use of single hot deck imputation to deal with missing data for income in HINTS 4 (Cycle 1). The core questionnaire of HINTS 4 was used to collect data on 134 items from 3,959 respondents from October 2011 through February 2012. In addition to the routinely collected socio-demographics, this survey also collects information on topics related to cancer and health. An examination of the sociodemographic data suggests that the missing rate ranges from 1.2 percent for U.S. born to 10.1 percent for income (Table 6-1).

Table 6-1. Item missing rates for major sociodemographic variables, HINTS 4 (Cycle 1)

Variable	Missing (N)	Missing data rate (%)
Age	68	1.7
Occupation	160	4.0
Marital Status	111	2.8
Education	85	2.1
Race/Ethnicity	220	5.6
Rent or Own Home	89	2.2
Comfortable Speaking English	155	3.9
Born in USA	47	1.2
Income	401	10.1

Considering the fact that income is a commonly used measure of socioeconomic status and its relatively high rate of missing data, the NCI has taken on the challenge and addressed this issue by implementing single hot deck imputation for income. The basic idea of hot deck imputation is to replace a missing observation with the value of a respondent who matches the individual on a set of other covariates. This procedure assumes that the missing income is not related to the missing portion of income, but it can be associated with observed income and other observed variables, for example, education, and gender. This missing mechanism assumption is called missing at random (MAR) (Rubin, 1987). Other missing mechanisms include missing completely at random (MCAR) and not missing at random (NMAR). What hot deck imputation actually does is create MAR situations by forming cells using variables, which are considered to be related to income or the likelihood of having missing income, so that within each cell, the missing data can be treated as similar to the observed data.

The NCI chose the hot deck imputation method because it has several strengths: it imputes real values, it avoids strong parametric assumptions, it can incorporate covariates, and it can provide valid inference for both linear and nonlinear statistics given that imputation uncertainty is properly incorporated.

The procedure, more specifically, the Cox-lannacchlone Weighted Sequential Hot Deck (WSHD) (Cox, 1980; Cox & Folsom, 1981), is implemented using *proc hotdeck* in SUDAAN. For more discussions on WSHD, see Andridge and Little (2010). The covariates selected are education, race/ethnicity, rent or own home, comfortable speaking English, and born in USA. Table 6-2 shows the comparison on estimated percents of each income category before and after imputation. There is still a small percent of missing data because respondents with missing data for any covariates are excluded from the imputation procedure. This comparison shows that respondents with missing income are not equally likely to be imputed into each income category, evidence of not MCAR, and some categories are more likely to be the donors than others, for example the categories of $10K-15K, $35K-$50K, and $50K-$75K. See Appendix E, section 1 for more information.

Single Versus Multiple Imputation

Despite its variance estimation not being proper, single imputation it is more attractive than multiple imputation because it is less confusing to users who are not familiar with the concept and technique of multiple imputation. In addition, when the fraction of missing data is low (i.e. less than 10 percent), the underestimation of variance in single imputation is negligible. We demonstrate this point by implementing a multiple imputation analysis with five imputations and compare it to a single imputation situation, as shown in the last four columns of Table 6-2. The relative efficiency of having five imputations is very close to one, which suggests multiple imputation is nearly fully efficient compared to the case of having complete information. The comparison in estimated percents and their standard errors between the single and multiple imputed data also suggests that the amount of underestimation with single imputation is very small. See Appendix E, section 2 for more information.

Table 6-2. Before and after imputation frequency comparison for income, HINTS 4 (Cycle 1)

	Before Imputation			After Imputation (Single)				After Imputation (Multiple=5)			
Income Ranges ($)	N	%	S.E. of %	N	%	S.E. of %	Relative % change*	%	S.E. of %	Relative efficiency**	Ratio of variance ***
$0 to $9,999	319	9.3	1.0	344	9.9	1.1	0.06	10.0	1.0	1.00	1.08
$10,000 to $14,999	269	6.2	0.8	294	7.1	0.9	0.15	6.7	0.9	0.99	1.02
$15,000 to $19,999	241	6.4	0.6	258	6.8	0.6	0.06	6.8	0.7	0.98	0.91
$20,000 to $34,999	584	15.5	1.1	640	16.8	1.1	0.08	16.7	1.1	1.00	1.06
$35,000 to $49,999	520	11.3	0.7	573	12.7	0.7	0.13	12.7	0.8	0.99	0.81
$50,000 to $74,999	594	15.2	0.9	650	16.8	1.2	0.11	17.0	1.1	1.00	1.04
$75,000 to $99,999	415	10.2	0.7	440	10.8	0.7	0.05	11.0	0.7	1.00	0.92
$100,000 to $199,999	463	12.3	0.9	509	13.1	0.9	0.07	13.2	0.9	1.00	0.91
$200,000 or more	153	3.6	0.3	165	3.8	0.3	0.06	3.8	0.3	1.00	0.97
Missing	401	9.8	0.7	86	2.2	0.4	-	2.2	0.4	1.00	-
Total	3959	100.0	-	3959	100.0	-	-	100.0	-	-	-

Note:

* Relative percent change is calculated as (Single imputed percent-Before imputation percent)/(Before imputation percent);

**Relative efficiency, see Page 114, Rubin (1987) for the formula;

***Ratio of variance is calculated as the square of the ratio of single and multiple imputation standard errors.

Multiple Imputation for Data Enhancement

Due to various limitations, including the concern of burden on the respondents, not all variables are asked in every iteration of HINTS. However, this may limit the usefulness of the data for any temporal trend analyses, which requires the presence of the same variables in all iterations. This type of missing data is caused when a variable is asked in some cycles of data collection, but not in all cycles. We illustrate the use of parametric imputation models to fill in these missing data, thus enabling a trend analysis over the whole span of survey iterations. This multiply imputed complete dataset also allows for other types of statistical analyses that may benefit from having a full span of data.

The outcome variable is whether the respondent has looked for health or medical information for his or herself using the internet during the past 12 months (referred to as "healthinfoself" hereafter, coded as Yes or No). Respondents who reportedly use the internet were asked this question in HINTS 1, 2, and 4, but not in HINTS 3. The idea is to treat the HINTS 3 data from this question as missing and fill in with imputed values. Multiple imputation, as opposed to single imputation, is more suitable for this application because the percent of missing data is large, thus more imputations are needed to account for the estimation uncertainty. To create a MAR situation, variables associated with the outcome are included as covariates. Because most variables have item missing data, we use sequential regression imputation method (SRIM), implemented using IVEware (http://www.isr.umich.edu/src/smp/ive/), to simultaneously fill in these item missing data. The specific imputation models are multiple linear regressions for continuous variables, logistic regressions for binary variables, and polynomial regressions for categorical variables.

Table 6-3 shows the estimated percentages of people who have looked for health information for him or herself online during the past 12 months. Complex survey features are incorporated in these figures. There seems to be a steadily increasing trend for the likelihood of using the internet to look for health information for oneself. However, there is a large gap of seven years between the last two estimates (HINTS 2 and 4). It is useful to know what has happened during this period and this can be realized by borrowing information from HINTS 3, a separate survey on the same population. The idea is to assume the relationship between the outcome and covariates in this survey is the same as in the other three iterations, which can be estimated and used to impute the missing data in HINTS 3.

Table 6-3. Sample size and characteristics of Healthinfoself, HINTS 1-HINTS 4 (Cycle 1)

	N		% of Look for Health Info. For Self	
Year of Data Collection	Use Internet	Missing	Point	S.E.
HINTS 1 (2003)	3,982	8	50.6	1.0
HINTS 2 (2005)	3,244	95	56.9	1.3
HINTS 3 (2008)	5,078	-	-	-
HINTS 4 (2011-2012)	2,914	14	77.6	1.3

To validate the imputation model, we conducted a simulation study using the data from HINTS 1, 2, and 4. We set the outcome to be missing for the HINTS 2 sample and imputed them using the proposed algorithm. A comparison between the estimated outcome for HINTS 2 using the imputed data and the estimate using the actual observed data provides evidence in support of the validity of the models. Specifically, variables collected in all cycles are included as the covariates to make MAR assumption more likely. The covariates include age, gender, race/ethnicity, Spanish speaker, education, employment status, marital status, household income, overall health status, BMI, residence rurality, health insurance coverage, internet use, use the internet for medication purchase, use the internet for participating in online health groups, use the internet for communications with the doctors, seek health care for self, seek cancer information, ever had cancer, and family member ever had cancer. In addition to the main effect of each covariate, interaction effects between data collection year and health-related variables are also included to capture the potential temporal differences in online health behaviors. Table 6-4 shows the multiple imputation (M=5) results from the simulation study. Despite that fact that the HINTS 2 estimated percentage is slightly higher than the actual data, the difference is not statistically significant and the overall increasing trend is preserved. Part of the inflation in point estimate may also be due to the imputation of item missing data for HINTS 2. As expected, the HINTS 2 estimates are less stable than the other two years because of the large fraction of missing data. Based on these findings, we feel the models are sufficiently adequate to capture the temporal relationships between the outcome and covariates.

Table 6-4. Comparison of estimated percent of Healthinfoself from multiply imputed and the actual data by data collection year, simulation study

	Actual data (%)		Multiply imputed data (%)			
Iteration	Estimate	S.E.	Estimate	Within Var.	Btwn Var.	Combined S.E.
HINTS 1	50.6	1.0	50.7	0.9	0.0	1.0
HINTS 2	56.9	1.3	62.7	2.2	1.9	2.1
HINTS 4	77.6	1.3	78.0	1.7	0.0	1.3

We implemented the same imputation model on the full dataset with all four cycles included and generated 10 imputed data. Results in Table 6-5 suggest that there is a steady linear increasing trend of using the internet to look for health information for oneself in the U.S. from HINTS 1 (2003) to HINTS 4 (2011-2012). Again, as expected, the standard error for the HINTS 3 estimate is slightly higher because of the large fraction of missing data. See Appendix E, section 3 for more information.

Table 6-5. Comparison of estimated percent of Healthinfoself from the multiply imputed enhanced and the actual data by data collection year, final application

Iteration	Actual data (%)		Multiply imputed data (%)			
	Estimate	S.E.	Estimate	Within Var.	Btwn Var.	Combined S.E.
HINTS 1	50.6	1.0	50.7	0.9	0.0	1.0
HINTS 2	56.9	1.3	58.4	1.9	0.0	1.4
HINTS 3	-	-	63.2	1.4	1.0	1.6
HINTS 4	77.6	1.3	78.0	1.7	0.0	1.3

Acknowledgment

The authors would like to thank Robin Rinker for her stewardship in completing this report.

References

Allison, P. D. (2001). *Missing Data*. SAGE Publications, Inc.

Andridge, R. R. & Little, R. J. A. (2010). "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review* 78(1): 40-64.

Arnott, D., Dockrell, M., Sandford, A., & Willmore, I. (2007). "Comprehensive smoke-free legislation in England: How advocacy won the day." *Tobacco Control, 16*, 423-428.

Bauer, D. J., & Hussong, A. M. (2009). "Psychometric Approaches for Developing Commensurate Measures Across Independent Studies: Traditional and New Models." *Psychological Methods*, 14(2), 101-175.

Blumberg, S. J., & Luke, J. V. (2009). "Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2008." National Center for Health Statistics. May. Available from: http://www.cdc.gov/nchs/nhis.htm

Blumberg, S. J., & Luke, J. V. (2012). "Wireless substitution: Early release of estimates from the National Health Interview Survey, July–December 2012." National Center for Health Statistics. June. Available from: http://www.cdc.gov/nchs/nhis.htm

Centers for Disease Control and Prevention. (2005a). "Tobacco use, access, and exposure to tobacco in media among middle and high school students—United States, 2004." *MMWR Morb Mortal Wkly Rep., 54*(12), 297-301.

Centers for Disease Control and Prevention. (2005b). "Cigarette smoking among adults—United States, 2004." *MMWR Morb Mortal Wkly Rep, 54*(44), 1121-1124.

Centers for Disease Control and Prevention. (2006). "The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General." Atlanta, GA.

Centers for Disease Control and Prevention. (2009). "Reduced hospitalizations for acute myocardial infarction after implementation of a smoke-free ordinance—city of Pueblo, Colorado, 2002–2006." *Morbidity and Mortality Weekly Report, 57*, 1373-1377.

Chahine T., Subramanian S.V., Levy J.I. (2011). "Sociodemographic and geographic variability in smoking in the U.S.: A multilevel analysis of the 2006-2007 Current Population Survey." *Tobacco Use Supplement. Social Science & Medicine. 73*, 752-758.

Chaloupka, F. J. (1999). "Macro-social influences: The effects of prices and tobacco-control policies on the demand for tobacco products." *Nicotine & Tobacco Research, 1*(Suppl 1), S105-109.

Chaloupka, F., & Pacula, R. (1998). "An Examination of Gender and Race Differences in Youth Smoking Responsiveness to Price and Tobacco Control Policies." (Working Paper 6541).

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2008). "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." In Advances in Telephone Survey Methodology, eds. James M. Lepkowski, N. Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, 250–75 New York: Wiley-Interscience.

Citro, C., & Kalton, G. (Eds.). (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Washington, DC: National Academy Press.

Cochran, W. G. (1977). *Sampling Techniques*. Third Edition, New York: John Wiley & Sons, Inc.

Cox B. G. (1980). *The Weighted Sequential Hot Deck Imputation Procedure*. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Cox B. G., Folsom R. E. (1981). An evaluation of weighted hot deck imputation for unreported health care visits. ASA Proc Section on Survey Res Methods. 412–417.

Curran, P. J., & Hussong, A. M. (2009). "Integrative data analysis: The simultaneous analysis of multiple data sets." *Psychological methods*, 14(2):81–100.

Davis, T., Dipko, S., & Sigman, R. (2009). HINTS Puerto Rico Final Report. Westat: Rockville, MD.

de Leeuw, E. D. (2005). "To Mix or Not to Mix Data Collection Modes in Surveys" *Journal of Official Statistics*, vol. 21, No. 2.

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). "Mixed Mode Surveys: When and Why" Chapter 16 in *International Handbook of Survey Methodology*, Lawrence Erlbaum Associates: London.

Dillman, D. A. (2007). *Mail and Internet Surveys: The Tailored Design Method, 2nd edition*. John Wiley & Sons, Inc.

Dillman, D. A., Phelps, G., Tortora, R. D., Swift, K., Kohrell, J., & Berck, J. (2009). "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR), and the Internet." *Social Science Research* 38:1–18.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7(1): 1-26.

Fahimi, M., Kulp, D., & Brick, J. M. (2008). "Bias in List-Assisted 100-Series RDD Sampling." *Survey Practice*, September 28, 2008.

Finney Rutten, L., Augustson, E., Moser, R., Beckjord, E., & Hesse, B. (2008). "Smoking knowledge and behavior in the United States: Sociodemographic, smoking status, and geographic patterns." *Nicotine & Tobacco Research, 10*(10), 1559-1570.

Finney Rutten, L., Davis, T., Beckjord, E., Blake, K., Moser, R., & Hesse, B. (2012). "Picking up the pace: Changes in method and frame for the health information national trends survey (2011-2014)." *Journal of Health Communication, 17*(8), 979-989.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). "How many imputations are really needed? Some practical clarifications of multiple imputation theory." *Provention Science* 8(3): 206-213.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews.* New York: Academic Press.

Hammond, D., Fong, G. T., Zanna, M. P., Thrasher, J. F., & Borland, R. (2006). "Tobacco denormalization and industry beliefs among smokers from four countries." *Am J Prev Med, 31*(3), 225-232.

Han, D. & Cantor, D. (2008). "Cell Phone Only Households in a National Mail Survey – Who are they?" ASA Proceedings of the Survey Research Methods, Alexandria, VA.

Hesse, B. W., Moser, R. P., Rutten, L. J., & Kreps, G. L. (2006). "The health information national trends survey: Research from the baseline." *J Health Commun*, 11 Suppl 1, vii-xvi.

Huang, P., & McCusker, M. (2004). "Impact of a Smoking Ban on Restaurant and Bar Revenues—El Paso, Texas, 2002." *Morbidity and Mortality Weekly Report, 53*, 150-152.

Iannacchione, V. G., Staab, J. M., & Redden, D. T. (2003). "Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey." *Public Opinion Quarterly* 76:202-210.

Kish, L. (1965). *Survey Sampling*, New York: John Wiley and Sons.

Krosnick, J. A. & Alwin, D. (1987). "An evaluation of a cognitive theory of response-order effects in survey measurement" *Public Opinion Quarterly*, 51, 201-209.

Krosnick, J. A. (1991). "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L. & Mokdad, A. H. (2008). "A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) For General Population Surveys." *Public Opinion Quarterly* 72: 6-27.

Link, M. W., Battaglia, M. P., Giambo, P., Frankel, M. R., Mokdad, A. H., Rao, S. R. (2005). "Assessment of Address Frame Replacements for RDD Sampling Frames," Paper prepared for the Annual Meetings of the American Association for Public Opinion Research, Miami, FL.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* New York, John Wiley.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modeling framework: Concepts, structure, and extensibility, Statistics and Computing, Vol.10, No.4, pp.325-337

Maples, J., & Bell, W. R. (2005). "Evaluation of school district poverty estimates: Predictive models using IRS income tax data." *Proceedings of the American Statistical Association.*

McBride, B., Cantor, D & Kerwin, J. (2010). "Using recordings of telephone interactions to understand mode differences." Presentation at HINTS data user workshop, Silver Spring, MD, September 24. Last accessed February 7, 2010.

Miller, R. G. (1974). "The jackknife—a review." *Biometrika* 61(1): 1-15.

National Cancer Institute. (2011). "The Health Information National Trends Survey (HINTS) Brief 18: Implementation of HINTS in Puerto Rico." Accessed on February 1, 2013. http://hints.cancer.gov/brief_18.aspx

National Cancer Institute. (1999). "Smoking and Tobacco Control Monograph 10: Health Effects of Exposure to Environmental Tobacco Smoke." Accessed on March 4, 2013. http://cancercontrol.cancer.gov/brp/tcrb/monographs/10/

Nelson, D. E., Kreps, G. L., Hesse, B. W., Croyle, R. T., Willis, G., Arora, N. K., et al. (2004). "The Health Information National Trends Survey (HINTS): Development, design, and dissemination." *J Health Commun*, 9(5), 443-460.

Nelson, D. E., Kreps, G. L., Hesse, B. W., Croyle, R. T., Willis, G., Arora, N. K., et al. (2004). "The Health Information National Trends Survey (HINTS): Development, design, and dissemination." *Journal of Health Communication, 9*(5), 443-460.

Osypuk, T. L., Kawachi, I., Subramanian, S.V., & Acevedo-Garcia, D. (2006). "Are State Patterns of Smoking Different for Different Racial/Ethnic Groups? An Application of Multilevel Analysis." *Public Health Rep. 121*(5), 563–577.

Paulhus, D. L. (2003). "Self-presentation measurement." In *Encyclopedia of Psychological Assessment*, ed. R. Fernandez-Ballesteros, 858–61. Thousand Oaks, CA: Sage.

President's Cancer Panel. (2007). "Promoting Healthy Lifestyles: Policy, Program, and Personal Recommendations for Reducing Cancer Risk." National Cancer Institute. Retrieved from http://deainfo.nci.nih.gov/advisory/pcp/annualReports/pcp07rpt/pcp07rpt.pdf

Rao, J. N. K. (2003). *Small area estimation*. New York: John Wiley and Sons.

Rizzo, L., Moser, R.P., Waldron, W., Wang, Z. & Davis, W.W. (2008). "Analytic Methods to Examine Changes Across Years Using HINTS 2003 & 2005 Data." NCI, NIH publication No. 08-6435. http://hints.cancer.gov/docs/HINTS_Data_Users_Handbook-2008.pdf, last accessed January 30, 2013.

Robert, C. P., & Casella, G. (1999). *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Rose, G. (1985). "Sick Individuals and Sick Populations." *International Journal of Epidemiology, 14*(1), 32-38.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, Wiley & Sons.

Schwarz, N., Hippler, H. & Noelle-Neumann, E. (1991). "A cognitive model of response-order effects in survey measurement." In N. Schwarz & S. Sudman (eds) *Context effects in social and psychological research* (pp. 187-201). New York: Springer-Verlag.

Tarnai, J., & Dillman, D. A. (1992). "Questionnaire Context as a Source of Response Differences in Mail versus Telephone Surveys." In Context Effects in Social and Psychological Research, eds. Norbert Schwarz and Seymour Sudman, 115–29 New York: Springer Verlag.

Tourangeau, R., & Smith, T. W. (1996). "Asking Sensitive Questions: The Impact of Data Collection, Mode, Question Format, and Question Context." *Public Opinion Quarterly* 60:275–304.

U.S. Centers for Disease Control and Prevention (CDC). (1998). Responses to Cigarette Prices By Race/Ethnicity, Income, and Age Groups – United States 1976-1993. *Morbidity and Mortality Weekly Report, 47*(29), 605-609.

U.S. Department of Health and Human Services. (2000). *Reducing Tobacco Use: A Report of the Surgeon General.* Atlanta: Retrieved from http://profiles.nlm.nih.gov/NN/B/B/L/Q/_/nnbblq.pdf.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Ye, C., Fulton, J. & Tourangeau, R. (2011). "More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice." *Public Opinion Quarterly* 75: 349-365.

Appendix A
Merging and Analyzing Multiple Iterations of HINTS Mainland Data

1. Testing for significantly different responses in a bi-modal administration of HINTS – creation of weights combining RDD and mail weights

```
data h07mergewts; **User-defined dataset names;
set c.hints2007;

array h07mwts[50] mwgt1-mwgt50; *Mail replicate weights;
array h07rwts[50] rwgt1-rwgt50; *RDD (Phone) replicate weights;
array h07twts[100] twgt1-twgt100; *Combined replicate weights;

**Note: Sampflag should be used to distinguish between mode;
if sampflag = 1 then do i = 1 to 50;*Address (Mail) sample;
 twgt0 = mwgt0;
 h07twts[i] = h07mwts[i];
 h07twts[i+50] = mwgt0;
 end;
else if sampflag = 2 then do i = 1 to 50;***RDD (Phone) sample;
 twgt0 = rwgt0;
 h07twts[i] = rwgt0;
 h07twts[i+50] = h07rwts[i];
 end;
run;


***T Tests of differences in outcome by mode ***;
proc descript data=h07mergewts design=jackknife ddf = 98;
weight twgt0;
jackwgts twgt1-twgt100 / adjjack=.98;
class sampflag;
var talkdoctor; **Outcome of interest;
contrast sampflag = (1 -1);
run;
```

2. Merging all HINTS iterations into one dataset

```
data hintsmerge; ***User-defined dataset names;
set h03 h05 h07 h4c1;

***Set new weight variables for the combined dataset;
array h03weights [50] fwgt1-fwgt50; ***Replicate weights from HINTS 2003;
array h05weights [50] fwgt1-fwgt50; ***Replicate weights from HINTS 2005;
array h07weights[50] cwgt1-cwgt50; ***Composite replicate weights from HINTS
2008;
```

```
array h4c1weights [50] person_finwt1-person_finwt50; ***Replicate weights
from HINTS 4 (2011-2012);
array nfwgt[200] nwgt1-nwgt200;***Final set of combined replicate weights;

*HINTS1 - 2003;
if survyear = 1 then do i = 1 to 50;***Replicate Weights 1 - 50;
      nfwgt0 = fwgt;
      nfwgt[i] = h03weights[i]; *1 - 50;
      nfwgt[i+50] = fwgt; *51 - 100;
      nfwgt[i+100] = fwgt; *101-150;
      nfwgt[i+150] = fwgt; *151-200;
      end;

*HINTS2 - 2005;
else if survyear = 2 then do i = 1 to 50;***Replicate Weights 51 - 100;
      nfwgt0 = fwgt;
      nfwgt[i] = fwgt; *1 - 50;
      nfwgt[i+50] = h05weights[i]; *51 - 100;
      nfwgt[i+100] = fwgt; *101-150;
      nfwgt[i+150] = fwgt; *151-200;
      end;

*HINTS3 - 2008;
else if survyear = 3 then do i = 1 to 50;***Replicate Weights 101 - 150;
      nfwgt0 = cwgt0;
      nfwgt[i] = cwgt0; *1 - 50;
      nfwgt[i+50] = cwgt0; *51 - 100;
      nfwgt[i+100] = h07weights[i]; *101-150;
      nfwgt[i+150] = cwgt0; *151-200;
      end;

*HINTS4 - 2011/2012;
else if survyear = 4 then do i = 1 to 50;***Replicate Weights 151 - 200;
      nfwgt0 = person_finwt0;
      nfwgt[i] = person_finwt0; *1 - 50;
      nfwgt[i+50] = person_finwt0; *51 - 100;
      nfwgt[i+100] = person_finwt0; *101-150;
      nfwgt[i+150] = h4c1weights[i]; *151-200;
      end;

run;
```

3. SUDAAN logistic regression procedure using combined data file from all HINTS iterations, including an interaction term between survey year and gender and predicted marginals

```
proc rlogist data = hintsmerge design = jackknife ddf = 196;
weight nfwgt0;
jackwgts nfwgt1- nfwgt200 / adjjack = 0.98;
class survyear agegrpa educa gender;
model talkdoctor = survyear agegrpa educa gender survyear*gender;
```

```
reflev survyear = 1 educa = 1 gender = 1;
predmarg survyear survyear*gender;
effects survyear = (-1 3 -3 1)/name = "Cubic trend";
effects survyear = (1 -1 -1 1)/name = "Quadratic trend";
effects survyear = (-3 -1 1 3)/name = "Linear trend";
run;
```

Appendix B
Merging HINTS Mainland and HINTS Puerto Rico Data

1. Code to combine and analyze HINTS 2008 and HINTS PR data

```
data hintsmerge; **User defined dataset names;
set h07 hintspr;

array h07weights[50] rwgt1-rwgt50; **RDD replicate weights from HINTS 2008;
array hprweights [48] R12WGT1 - R12WGT48; **RDD replicate weights from HINTS
PR;
array twgt[98] twgt1-twgt98; **Combined replicate weights;

if survyear = 1 then do; ***HINTS 2008;
twgt0 = rwgt0;

 do i = 1 to 50;
 twgt[i] = h07weights[i];
 end;

 do j = 1 to 48;
 twgt[j+50] = rwgt0;
 end;

end;

else if survyear = 2 then do; ***HINTS PR;
twgt0 = R12WGT0;

 do p = 1 to 48;
 twgt[p+50] = hprweights[p];
 end;

 do q = 1 to 50;
 twgt[q] = R12WGT0;
 end;
end;

**Creating derived ethnicity variable;
if survyear = 1 and sampflag = 2 and raceeth = 1 then ethnicity = 1;
*Hispanics in U.S.;
if survyear = 1 and sampflag = 2 and raceeth = 2 then ethnicity = 2; *non-
Hispanics in U.S.;
if survyear = 2 and raceeth = 1 then ethnicity = 3; *Hispanics in PR;

run;
```

2. Logistic regression model for comparing HINTS 2008 and HINTS PR

```
proc rlogist data = hintsmerge design = jackknife ddf = 89;
weight twgt0;
jackwgts twgt1-twgt98;
jackmult 50*0.98 48*0.83; **Applying different multipliers to each respective
dataset;
class survyear agegrpa educa gendern/nofreq;
model HC08SeekCancerInfo = survyear agegrpa gendern educa;

run;
```

3. Logistic regression model for comparing ethnicity

```
proc rlogist data = hintsmerge design = jackknife ddf = 89;
weight twgt0;
jackwgts twgt1-twgt98;
jackmult 50*0.98 48*0.83; **Applying different multipliers to each respective
dataset;
class ethnicity agegrpa educa gendern/nofreq;
model HC08SeekCancerInfo = ethnicity agegrpa educa gendern ;
reflev ethnicity = 1 gendern=1 agegrpa=1 educa=1;
effects ethnicity = (1 0 -1); **Comparing U.S. Hispanics vs. Puerto Rico
Hispanics;
effects ethnicity = (1 -1 0); **Comparing U.S. Mainland Hispanics vs. U.S.
Mainland non-Hispanics;
run;
```

Appendix C
Multilevel Determinants of Smoking Behavior: An Integrated Analysis

SAS- PROC GLIMMIX code:

```
/*SAS code to compute the point estimates using the full model */

Title "Model 3: The full model";

PROC GLIMMIX NOCLPRINT NOITPRINT DATA=hints4;
CLASS statenum  AGEGRPB_REC maritalstatus_rec EducA_Rec RaceEthn_rec
HHInc_Rec comprehensive2011;
 MODEL SMOKE (DESCENDING) = taxdollar comprehensive2011 AGEGRPB_REC
maritalstatus_rec EducA_Rec RaceEthn_rec HHInc_Rec/DIST=BINARY ODDSRATIO;
 RANDOM statenum;
WEIGHT person_finwt0;
ods output oddsratios=M3Oddratio covparms=M3Covparm;
RUN;

/*SAS code to compute the standard errors incorporating the complex survey
design*/

data M3Oddratio;
set M3Oddratio(rename=(estimate=OddsRatio) drop=DF ALPHA Lower Upper);
Index=_N_;
run;

data M3Covparm;
set M3Covparm(rename=(estimate=Covariance) drop=StdErr);
run;

ods listing close;

/*Macro to compute the point estimate using each jackknife replicate
weights*/

%macro REPGLIMMIX(R);
%do REP=1 %to &R;
Title "Model 3: The full model";

PROC GLIMMIX NOCLPRINT NOITPRINT DATA=hints4;
CLASS statenum  AGEGRPB_REC maritalstatus_rec EducA_Rec RaceEthn_rec
HHInc_Rec comprehensive2011;
 MODEL SMOKE (DESCENDING) = taxdollar comprehensive2011 AGEGRPB_REC
maritalstatus_rec EducA_Rec RaceEthn_rec HHInc_Rec/DIST=BINARY ODDSRATIO;
 RANDOM statenum;
WEIGHT person_finwt&REP;
ods output oddsratios=M3Oddratio2 covparms=M3Covparm2;

RUN;
```

```
data M3Oddratio2;
set M3Oddratio2(rename=(estimate=OddsRatio&REP) drop=DF ALPHA Lower Upper);
Index=_N_;
run;

data M3Covparm2;
set M3Covparm2(rename=(estimate=Covariance&REP) drop=StdErr);
run;

data M3Oddratio;
merge M3Oddratio(in=in1) M3Oddratio2(in=in2);
by index;
if in1 and in2;
run;

data M3Covparm;
merge M3Covparm M3Covparm2;
by CovParm;
run;

%end;
 %mend REPGLIMMIX;
 %REPGLIMMIX(R=100)

/*compute the standard error using the Jackknife method*/
/*Using the Jackknife 1 formula: var(OR)=0.98*sum((OR_j-OR)^2);
/*For Odd Ratios, compute the confidence interval for log Odds first then
transfer the confidence interval to the original scale*/

/*exclude one replicate weights which didn't converge for model 3*/
data M3oddratio;
set M3oddratio;
OddsRatio58=OddsRatio;
run;

data Oddratio;
set M3oddratio (keep=Index OddsRatio OddsRatio1-OddsRatio100);
array ORREP[100] OddsRatio1-OddsRatio100;
array diff2_logOR[100] diff2_logOR1-diff2_logOR100;
do i=1 to 100;
diff2_logOR[i]=(log(ORREP[i])-Log(OddsRatio))**2;
end;
keep index diff2_logOR1-diff2_logOR100;
run;

proc transpose data=Oddratio out=Oddratio_transposed;
run;

data Oddratio_transposed;
set Oddratio_transposed;
if _N_ ^=1;
run;

/*for each Odd ratio, sum the square difference over the 100 replicates, the
column is for the predictor categories*/

proc summary data=Oddratio_transposed;
```

```
var Col1-Col24;
output out=OR_sum sum=Col1-Col24;
run;

/*Transpose the 24 column categories back to 24 rows*/

proc transpose data=OR_sum(drop=_TYPE_ _FREQ_) out=OR_sum_transpose;
run;

data OR_sum_transpose;
set OR_sum_transpose(drop=_NAME_ rename=(COL1=sum_diff2));
index=_N_;
run;

/*Merge the standard error to the original OR output*/

data fnl_M3Oddratio;
merge M3Oddratio (in=in1) OR_sum_transpose(in=in2);
by index;
if in1 and in2;
run;

data out.fnl_M3Oddratio;
set fnl_M3Oddratio;
LP_LogOR=log(ODDsRatio)-tinv(0.975, 49)*sqrt(0.98*sum_diff2);
UP_LogOR=log(ODDsRatio)+tinv(0.975, 49)*sqrt(0.98*sum_diff2);

LP_OR=exp(LP_LogOR);
UP_OR=exp(UP_LogOR);
run;

/*Compute the standard error for the random effect variance*/
/*exclude one replicate weights which didn't converge for model 3*/

data M3covparm;
set M3covparm;
covariance58=covariance;
run;

data covparm;
set M3covparm(keep=Covariance Covariance1-Covariance100);
array VARREP[100] Covariance1-Covariance100;
array diff2_VARREP[100] diff2_VARREP1-diff2_VARREP100;
do i=1 to 100;
diff2_VARREP[i]=(VARREP[i]-Covariance)**2;
end;
keep diff2_VARREP1-diff2_VARREP100;
run;

proc transpose data=covparm out=covparm_transposed;
run;

proc summary data=covparm_transposed;
var Col1;
output out=COV_sum sum=SUM_diff2;
run;
```

```
data COV_sum;
set COV_sum(drop=_TYPE_ _FREQ_);
run;

/*Merge the standard error to the original OR output*/
data fnl_M3covparm;
merge M3covparm COV_sum;
run;

data out.fnl_M3covparm;
set fnl_m3covparm;
SE_COV=sqrt(0.98*sum_diff2);
run;
```

Appendix D
Model-Based State Level Estimates for Cancer Related Knowledge Variables Using HINTS Data

1. WinBUGS Code for Model 5.1-5.2

```
model {
      # m states with samples

      for ( i in 1:m)  {
         y[i] ~ dnorm(theta[i], inv.D[i])
         theta[i]<-inprod(beta[], X[i, ])+v[i]
         v[i]~dnorm(0, tau)
         prop[i]<-sin(theta[i])*sin(theta[i])
                          }

  for ( i in 1:k)  {
             beta[i]~dflat()
             }
       tau<-1/A
       A~dunif(0, 100)
          }
```

Appendix E
Using Imputation to Enhance Multiple Iterations of HINTS Data

1. Single Imputation for Missing Income Data in HINTS 4 (Cycle 1) using Hot deck imputation method

```
libname dir 'your folder path';

%include 'your folder path\hints4cycle1_formats.sas';
data HINTS4CYCLE1; *temporal HINTS4 (cycle 1) data name;
      set h41.hints4cycle1_08152012; *HINTS4 (cycle 1) data name;
run;

***Set the Unknown and NA categories on variables involved to missing***;
data HINTS4CYCLE1;
      set HINTS4CYCLE1;
      COPY_Education = Education;
      if COPY_Education in (-9) then
      COPY_Education = .;
      COPY_RaceEthn = RaceEthn;
      if COPY_RaceEthn in (-9) then
      COPY_RaceEthn = .;
      COPY_RentOrOwn = RentOrOwn;
      if COPY_RentOrOwn in (-5, -9) then
      COPY_RentOrOwn = .;
      COPY_ComfortableEnglish = ComfortableEnglish;
      if COPY_ComfortableEnglish in (-5, -9) then
      COPY_ComfortableEnglish = .;
      COPY_BornInUSA = BornInUSA;
      if COPY_BornInUSA in (-9) then
      COPY_BornInUSA = .;
      COPY_IncomeRanges = IncomeRanges;
      if COPY_IncomeRanges in (-5, -9) then

      format COPY_Education Educati. COPY_RaceEthn RaceEthn. COPY_RentOrOwn
RentOrO.
      COPY_ComfortableEnglish Comfort. COPY_BornInUSA BornInU.;
run;


data HINTS4CYCLE1;
      set HINTS4CYCLE1;
      ID = _N_;
run;

proc sort data=HINTS4CYCLE1;
      by COPY_Education COPY_RaceEthn COPY_RentOrOwn COPY_ComfortableEnglish
COPY_BornInUSA;
run;
```

```
***Invoke Hotdeck imputation procedure in SUDAAN, Single Imputation***;
proc hotdeck data=HINTS4CYCLE1;
      weight person_finwt0; *weights;
      impvar COPY_IncomeRanges; *outcome;
      impby COPY_Education COPY_RaceEthn COPY_RentOrOwn
COPY_ComfortableEnglish COPY_BornInUSA; *covariates;
      impname COPY_IncomeRanges="IncomeRanges_IMP"; *name of imputed
outcome;
      impid ID;
      output IMPID IMPBY IMPUTEVAL / filename=impute1 replace; *output the
imputed data to a SAS dataset named impute1;
run;
```

2. Multiple Imputation (M=5) for Missing Income Data in HINTS 4 (Cycle 1) using Hot deck imputation method

```
***Invoke Hotdeck imputation procedure in SUDAAN, Multiple Imputation=5***;
proc hotdeck data=HINTS4CYCLE1;
      weight person_finwt0; *weights;
      impvar COPY_IncomeRanges / multimp=5; *outcome with M=5;
      impby COPY_Education COPY_RaceEthn COPY_RentOrOwn
COPY_ComfortableEnglish COPY_BornInUSA; *covariates;
      impname COPY_IncomeRanges="IncomeRanges_IMP"; *name of imputed
outcome;
      impid ID;
      output IMPID IMPBY IMPUTEVAL / filename=impute2 replace; *output the
imputed data to a SAS dataset named impute2;
run;
```

3. Multiple Imputation for HealthInfoSelf in 2008 Using IVEware

Note: The followings are example codes to invoke and execute IVEWare. Please consult with a statistician before trying to use them.

```
***Concatenate the data collected from all four times ***
data dir.youroutputdataname;
     set hints2003 hints2005 hints2008 hints4;
run;
***Invoke IVEWare in SAS***
%impute(setup=new, name=mysetup, dir="your path\");
datain dir.yourinputdataname; *input original data;
dataout dir.youroutputdataname all; *output data with imputed data;
categorical age gender educ ehealthdoc ehealthgrp ehealthmed everhadcancer
             familyeverhadcancer healthinfoother
            healthinfoself healthinsurance healthstatus hhinc maritalstatus
             occupationstatus raceethn rural
            seekcancerinfo seeprovider spaneng useinternet; *variables that
             are categorical and by default all variables are
             continuous;
```

```
restrict    ehealthmed(useinternet=1) ehealthgrp(useinternet=1)
             ehealthdoc(useinternet=1) healthinfoself(useinternet=1)
             healthinfoother(useinternet=1);
interact    year*useinternet year*healthinfoother year*healthinfosel
             year*trustdoc year*trustfam year*trustrad;
bounds      bmi (>=10.8,<=77.7);
iterations  5;
multiples   5;
run;
```


DEPARTMENT OF HEALTH & HUMAN SERVICES · USA
NIH
NATIONAL®
CANCER
INSTITUTE


T 2 3 0