Survey Research Methods (2008)
Vol.2 , No.2 , pp. 47-62
ISSN 1864-3361
http://www.surveymethods.org

Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?

Jill A. Dever
University of Maryland

Ann Rafferty
Michigan Department of Community Health

Richard Valliant
University of Maryland/University of Michigan

The Internet is an attractive mode of data collection to survey researchers due to cost savings and timeliness in comparison with other modes. However, survey estimates are subject to coverage bias if sampled persons with Internet access are systematically different from those without Internet access who were excluded from the survey. Statistical adjustments, either through weighting or modeling methods, can minimize or even eliminate bias due to non-coverage. In the current paper, we examine the coverage bias associated with conducting a hypothetical Internet survey on a frame of persons obtained through a random-digit-dial sample. We compare estimates collected during telephone interviews from households with and without Internet access using data from the 2003 Michigan Behavioral Risk Factor Surveillance System in the United States. A total of 25 binary variables (e.g., the percent of adults who have asthma or who are classified as being obese) and four count variables (e.g., the number of alcoholic drinks consumed per month) were analyzed for this study in addition to eight demographic characteristics. Weights based on the general regression estimator are computed such that the coverage bias is reduced to undetectable levels for most of the health outcomes analyzed from the Michigan survey. Though not definitive, the analysis results suggest that statistical adjustments can reduce, if not eliminate, coverage bias in the situation we study.
Keywords: Internet penetration, undercoverage, calibration estimation, poststratification, US Behavioral Risk Factor Surveillance Survey (BRFSS)

Introduction*

Internet surveys have been used for several years to obtain data in the social sciences (e.g., Schonlau et al. 2002; Ballard and Prine 2002; Suh and Han 2003), health research (e.g., Alexander and Trissel 1996; Braithwaite et al. 2003), and other disciplines. Internet surveys offer a less expensive option and a shorter data collection period in comparison to telephone and in-person household surveys (Couper 2000). Additionally, administration through the Internet can enhance the survey experience through the use of sound and video (Couper et al. 2004). The European WebSM project (www.websm.org), the 2006 special issue of the *Journal of Official Statistics* on Web surveys, and the Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences (MESS)[1] in the Netherlands are all testimony to the fact that this remains an area of active research.

One of the greatest disadvantages of the Internet as a mode of data collection is the limited access of some in the general population. For example, even though a November 2007 Net-Ratings report by Nielsen[2] names the United States (US) as having the fifth highest Internet penetration rate in the World (71.4 percent), a 35.4 point increase over figures calculated in 2000, this rate is substantially less than 100 percent. In most other countries, Internet penetration is lower. For the European Union, InternetWorldStats reports the Internet penetration was 55.7 percent in November 2007.[3] Penetrations for individual countries range from 30 percent for Bulgaria to 87.8 percent for Netherlands.

Surveys that require Internet access from a specified location such as the home will have an even more restrictive coverage rate. Harwood and Rainie (2004), using data from the Pew Internet and American Life Project, report that approximately 64 percent out of the 128 million American adults (18 years or older) in 2002 used the Internet from any number of locations. However, only roughly 88 percent of those same adults had access to the Internet from home resulting in a potential undercoverage rate of over 43 (=100x(1 - 0.64x0.88)) percent.

Internet surveys, by design, exclude the entire non-Internet population. What is meant by 'Internet' and 'non-Internet' naturally varies depending on how access locations (home, work, or elsewhere) are counted. In this study, we consider the Internet population to be those persons who have access at home and study those properties of samples

Contact information: Jill A. Dever, Joint Program in Survey Methodology, 1218 LeFrak Hall, University of Maryland, College Park, MD 20742 USA, Email: jdever@survey.umd.edu

* This second version of the article includes the tables 4 and 5 which were missing in the first published version.

[1] http://www.uvt.nl/centerdata/en/mess/

[2] http://www.internetworldstats.com/stats14.htm#north

[3] http://www.internetworldstats.com/stats9.htm#eu

selected from this home-Internet population. Some authors (e.g., Lee 2006) distinguish between Web and Internet surveys. Web surveys are presented via browsers while Internet surveys can be done through browsers or email. This distinction is not important for the investigation in this paper.

If estimates are desired for the complete household population and persons without access to the Internet are systematically different from the survey participants, the estimates are subject to bias due to coverage error (Groves 1989). In the US, lower Internet penetration has been observed for older, unemployed, less educated, rural, and disabled populations in comparison to their complements (NTIA 2002). Though men and women are equally likely to access the Internet from work, men are slightly more likely to access the Internet from home (Fallows 2005). Still, the preference for Internet surveys will likely increase in the future based on economic advantages and timeliness, especially with ever increasing Internet penetration rates (Beniger 1998; Couper 2000; Couper et al. 2001; Dillman 2002).

The purpose of this article is twofold. First, we describe detectable coverage biases by examining differences in health outcomes for adults with and without home Internet access. Second, we investigate whether additional model covariates can be used to successfully eliminate the detectable differences and the corresponding coverage bias. Estimates are derived from telephone interviews conducted for the 2003 US Behavioral Risk Factor Surveillance System (BRFSS) within the US state of Michigan. Section 2 (*Background*) gives some background on surveys that are conducted via the Web and the types of coverage errors that may occur. Section 3 (*Michigan Behavioral Risk Factor Surveillance System*) describes the BRFSS data used in our analysis. In section 4 (*Models for Health-Related Characteristics*), we present the results for models fit to various health characteristics and, in particular, examine whether having access to the Internet at home is an important predictor. In the fifth section (*Survey Weights for the Internet Cases*), we examine whether the general regression estimator can be used to calculate survey weights that will reduce coverage errors. Section 6 (*Conclusion*) contains our concluding remarks.

Background

The Internet, as a data collection medium, offers several advantages over other methods. Internet surveys, like mail surveys, offer a less expensive data collection option in comparison to telephone and in-person household surveys (Couper 2000). For example, Internet survey budgets do not include costs associated with interviewer training, travel, address-listing procedures, and the professional time involved in designing and selecting multistage, area probability samples. Internet surveys can also require less ramp-up time than other surveys. For telephone surveys, months may be needed to recruit and train interviewers. When a survey sponsor does not have access to a complete address list, area household surveys traditionally employ counting and listing procedures to develop sampling frames from which households are selected. Depending on the number of sampled areas, this process can take several months to conduct.

A shorter data collection period is another benefit listed for Internet surveys. For example, the Navy Personnel Research, Studies, and Technology (NPRST) laboratory, a research and development unit within the Department of the US Navy, conducted a seven-day quick poll in April 2005 to determine the prevalence of sexual assault victims in the active-duty Navy (Newell et al. 2005). Schonlau et al. (2004) completed a Web survey in 3.5 weeks with twice as many households as selected for a parallel telephone survey completed in three months. Knowledge Networks provides "eight-day turnaround studies" known as KN/Quick View in which interviews are obtained from "1,000 adults (residing in the U.S) from a nationally representative sample".[4]

Whether a survey done by Internet provides useful information depends, in part, on the population for which inferences are desired. Couper (2000) lists eight types of Web surveys that range from non-probability volunteer samples to probability samples from all or part of a target population. At one extreme is a volunteer panel recruited through advertisements on various Web sites. A list of these volunteers may be accumulated and a sample of the volunteers selected for a particular survey. Such a sample may represent the set of persons who originally volunteered but not the general population. At the other extreme is a target population that is a well-defined group for which an email address list frame is available and from which a representative sample can be selected. This is different from a situation in which the general household population is the target. In that case, heroic assumptions, sketched below, are needed to say that estimates from an Internet sample, that covers only a subset of all households, apply to the entire population. As observed by Couper (2000) and Best et al. (2001), coverage error is one of the biggest threats to inference when a Web survey has the household population as its target.

Examples of populations in which list frames may be used are students at a university, employees of a particular company, active-duty members of a branch of the military, or residents of one of the Scandinavian countries (Denmark, Finland, Iceland, Norway, and Sweden) which maintain total population registers. For these populations, a complete, or nearly complete, list of all population members along with contact information is available from administrative records. A sample can be selected in which (nearly) every member of the population has a prescribed, positive probability of inclusion, and sample persons can be directed to a Web site to complete the survey. Person-specific identification numbers and passwords may be used to ensure linkage between the sample member and their responses and to additionally minimize that likelihood of someone other than the intended participant filling in the responses.

Our emphasis here is on the more difficult case of a household population where no complete list is available of either all or a subset of households (or persons) with home Internet access. Internet surveys that hope to represent the

[4] http://www.knowledgenetworks.com/ganp/quickview/knqv-specs.htm

entire household population may be selected from various types of frames (including volunteer panels), subsampled from large, initial telephone samples, or subsampled from area probability samples. If the sample persons are expected to complete the survey from their homes (rather than at work or another location), then anyone without home Internet access is ineligible for the study.

The fact that an Internet sample covers only persons with Internet access means that over 40 percent of the US household population would have been excluded from a general population Web survey conducted in 2003 (Harwood and Rainie 2004). With such severe undercoverage, heavy reliance on statistical adjustments is needed to make estimates for the full household population. Efforts are often made to correct for poor sample coverage by calculating weights using poststratification, raking, or more elaborate regression methods (e.g., Kalton and Flores-Cervantes 2003; Kott 2006). Control totals for the complete target population are used even though the sample itself may be selected from a subset of that population. These weights are applied to both non-probability and probability samples and will produce estimators that are unbiased in a model-based sense if the sample data follow the same model as the larger population. This type of coverage correction through weighting is common practice even in large, well-established surveys like the US Current Population Survey (CPS) (Kostanich and Dippo 2000). However, because Internet surveys cover a much smaller proportion of the household population, their dependence on weighting adjustments is much greater than for a survey like the CPS (Vehovar et al. 1999).

One approach to selecting an Internet sample would be to recruit a panel of persons through a telephone survey and then select a subsample from the panel that has Internet access. This method raises two questions: (i) how different is the telephone sample from the general population, and (ii) how different from the telephone sample is the subset of persons that has home Internet access? We study these issues by comparing characteristics of adults living in Michigan as estimated from the CPS, an area probability sample, with those estimated from the BRFSS, a random digit dialing (RDD) telephone sample. We additionally compare BRFSS estimates for those respondents with and without home Internet access. A third issue that we are unable to address with the data is whether the respondents to an Internet survey are different from the nonrespondents.

In our analyses, the effect of coverage error is not confounded with nonresponse error. We consider this an advantage since it permits us to see whether statistical adjustments can correct for the first of these types of errors. If coverage error could not be corrected by weight adjustments (or similar means), then there is little hope of correcting for the compound effect of both coverage and nonresponse errors.

Michigan Behavioral Risk Factor Surveillance System

The Centers for Disease Control and Prevention (CDC) established the BRFSS as a mechanism to collect US state-level data on "preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases in the adult population" (CDC 2003). The BRFSS is composed of annual state-level telephone surveys conducted by state health departments. One randomly chosen adult 18 years of age or older is selected for the survey from (in most states) a list-assisted RDD sample of households. The BRFSS telephone questionnaire contains three parts: 1) a core set of questions administered by all states; 2) a set of optional modules; and 3) state-added questions (CDC 2002).

In 2003, the Michigan Department of Community Health created an instrument that included an Internet-usage module along with the core questions. This module was included in all four quarters of the 2003 Michigan BRFSS (MI BRFSS). Additional questions specific to the MI BRFSS instrument are located in *Appendix A*. The MI BRFSS produces weighted estimates that are intended to apply to all persons living in the state who are 18 years of age or older. However, some coverage bias is introduced into the estimators from the survey because only those adults living in households with a residential telephone line are interviewed. Nonresponse biases may also exist in the estimators due to less than a 100 percent response rate - the MI BRFSS achieved an unweighted response rate of 49.8 percent (National Center for Chronic Disease Prevention and Health Promotion 2004) using AAPOR definition RR4 (AAPOR 2004). Adjustments to account for these and other biases are made to the weights; further details are provided below.

As with other surveys, the final BRFSS analysis weights were calculated by poststratifying the base weights (inverse of the inclusion probabilities) to external control totals. The MI BRFSS base weights were adjusted using the 2002 "bridged-race post-censal" population estimates by single-year of age, race, Hispanic origin, and gender.[5] The final weights account for differences in inclusion probabilities, nonresponse, and noncoverage. These weights were used in the subsequent analysis tables to produce population based estimates of health risks. The estimated standard errors (SEs) were computed using Taylor series linearization in SUDAAN®, a software package created by RTI International to analyze correlated data (Research Triangle Institute 2004).

Comparisons of BRFSS with CPS

The US Bureau of Labor Statistics (BLS) conducts the CPS, a monthly survey of over 50,000 randomly chosen US households. The design is best described as a large stratified, multi-stage random sample with an in-person administration of the survey instrument. The primary focus of the CPS is to collect national and state-level labor force characteristics,

[5] 'Bridging' refers to procedures used by the US Census Bureau to make data collected using one set of race categories consistent with data collected using a different set of race categories. Details are provided at http://www.cdc.gov/nchs/about/major/dvs/popbridge/popbridge.htm

such as the number of hours worked for pay and any unemployment earnings, for the civilian, non-institutionalized US population of persons at least 16 years of age. Supplemental topics, such as those related to home Internet access addressed in this paper, are added to the base instrument (BLS 1997).

Estimates from the CPS, in addition to counts from the US Decennial Census, are regularly used to poststratify weights from other surveys due to the high quality of the data collected (see, e.g., Nadimpalli, Judkins and Chu 2004). For example, the overall response rate for the October 2003 CPS exceeded 92 percent (US Census Bureau 2004). The CPS, as noted earlier, is an area probability sample that, in principle, covers all households in the US. The BRFSS covers only the households that have landline telephones. A comparison of the percent distribution across various domains for the MI BRFSS with the Michigan CPS (MI CPS) may be used to examine any differences that exist between the populations covered by CPS and BRFSS. We chose the October 2003 MI CPS survey for comparison with the MI BRFSS due to the inclusion of a CPS Internet usage supplement and the comparable time periods.

The estimated percent distribution in the target population (adults ages 18 years and older residing in Michigan) for the MI BRFSS is provided in Table 1 by demographic group. Figure 1 shows the estimated difference in percentage distributions between MI BRFSS and the MI CPS for persons living in all households (All HHs) and in telephone households (Phone HHs). Limits of 95 percent confidence intervals are shown as red lines. In general, the percentages were comparable with few exceptions. The distributions for age group, race/ethnicity, gender, employment status, and marital status are all comparable. Minor differences between the 2003 BRFSS estimates (adjusted using 2002 Census data) and the 2003 CPS estimates are attributed at least in part to the variable growth rates within the state. The MI BRFSS estimated a lower percent of persons in families with incomes less than $20,000 (15.2 vs. 18.4 and 16.8) or greater than $75,000 (23.2 vs. 29.0 and 30.4) and estimated a higher percent in families with children less than 18 years of age (42.4 vs. 35.5 and 35.7). Additionally, the MI BRFSS estimated a slightly higher percent of persons with at least a four-year college degree (28.9 vs. 22.1 and 22.9).

Table 2 provides a comparison of the estimated home-Internet penetration rates by demographic group for the MI BRFSS and the MI CPS. Overall, the MI BRFSS estimates a rate 9.2 and 6.7 percentage points higher than the "All HHs" and "Phone HHs" MI CPS estimates, respectively, as predicted above. Higher rates are also seen for each of the various demographic groups. For example, the MI BRFSS estimates that 41.0 percent of persons with a family income less than $20,000 have access to Internet in the home, approximately 16.4 (and 13.3) percentage points higher than the MI CPS. Also, the MI BRFSS Internet-penetration estimate for persons with less than a high school education is 12.5 (and 9.7) percentage points higher than the CPS (40.4 vs. 27.9 and 30.7). The smallest difference between the MI BRFSS and CPS was estimated for persons with a family income that exceeds $75,000. For many of the demographic groups, the BRFSS estimates (taken from a telephone sample) are closer in value to the CPS telephone household estimates.

There are some patterns within the set of MI BRFSS estimates that are worth noting. The Internet penetration rate increases with family income. For persons with family income of less than $20,000, 59.0 percent do not have home Internet access. This implies that a person's family income is correlated with home Internet access and should be included as a predictor in model-based estimation. A similar increasing trend is visible with education. Non-Hispanic (NH) Blacks are less likely to have Internet access at home compared with NH Whites and other race/ethnicity groups. The penetration rate increases with age through the 35-44 group; the oldest age group (65+) has the lowest penetration rate at 36.3 percent.

The estimated demographic distribution of the MI BRFSS population is quite similar to that of the MI CPS for characteristics we examined. On dimensions used to poststratify the MI BRFSS, the estimates will naturally be exactly the same as the 2002 post-censal estimates produced by the US Census Bureau. However, the fact that the estimated home-Internet penetration rates in the MI BRFSS are consistently higher than those estimated from the October 2003 MI CPS does raise questions. The CPS asked *"Does anyone in this household connect to the Internet from home?"* while the BRFSS wording was *"Do you have access to the Internet at home?"* Since one can have 'access' without 'connecting', this may have contributed to a lower CPS penetration estimate. In addition, if nonrespondents to a telephone survey have a lower Internet penetration rate than respondents, this would lead to the higher BRFSS estimates in Table 2. In summary, the population represented by the MI BRFSS does appear to use the Internet at somewhat higher rates than the MI CPS telephone HHs.

Discussion of Health Characteristics from the MI BRFSS

Data for 29 MI BRFSS health questions were identified for the subsequent analyses. A binary variable was created for each of 25 categorical variables; four continuous variables were used as collected in the BRFSS. For example, the five-level question on general health, labeled as "V1_1" below, was recoded to a binary variable using the specifications below. Every person providing either a "Don't Know" or "Refused" response was excluded from the analyses for all variables.

V1_1: Would you say in general your health is:

1 *Excellent*
2 *Very good*
3 *Good*
} → 1 Good to Excellent

4 *Fair*
5 *Poor*
} → 2 Poor to Fair

Table 1: Weighted Demographic Characteristics for the 2003 MI BRFSS, Sample Sizes of Persons (*n*), and Estimated Standard Errors (*se*).

Household Characteristics	n	% (se)
Family Income		
< $ 20,000	553	15.2 (0.7)
$20,000-$34,999	756	24.4 (0.9)
$35,000-$49,999	555	17.9 (0.8)
$50,000-$74,999	580	19.3 (0.8)
$75,000+	682	23.2 (0.9)
Children in HH		
Yes	1,230	42.4 (1.0)
No	2,299	57.6 (1.0)

Person Characteristics	n	% (se)
Age		
18 - 24	237	12.6 (0.8)
25 - 34	483	17.8 (0.8)
35 - 44	655	21.2 (0.8)
45 - 54	735	19.3 (0.7)
55 - 64	597	12.7 (0.6)
65 +	825	16.5 (0.6)
Race/Ethnicity		
NH White	3,024	81.5 (0.8)
NH Black	309	12.8 (0.7)
Other	163	5.7 (0.5)

Person Characteristics	n	% (se)
Gender		
Male	1,398	48.1 (1.0)
Female	2,134	51.9 (1.0)
Education		
Less than HS	361	11.2 (0.7)
HS Grad/GED	1,105	31.3 (0.9)
<4yrs College	994	28.5 (0.9)
College Grad+	1,062	28.9 (0.9)
Employment		
Employed	1,897	58.2 (1.0)
Unemployed	169	5.9 (0.5)
NILF*	1,459	35.9 (0.9)
Marital Status		
Married	1,972	59.7 (1.0)
Not Married	1,558	40.3 (1.0)
Total persons	3,532	

* NILF = not in labor force

Figure 1. Differences of Weighted Demographic Characteristics for the 2003 MI BRFSS and the October 2003 MI CPS. Point estimates are dots; 95% confidence intervals are shown as red lines.

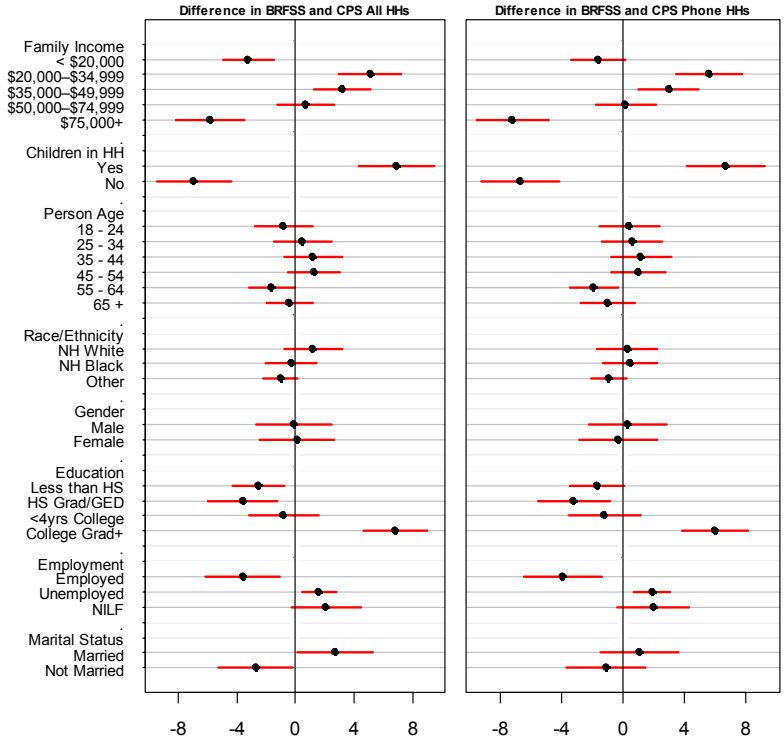


Table 2: Comparison of Estimated Home-Internet Penetration Rate by Demographic Characteristic for the 2003 MI BRFSS and the October 2003 MI CPS. Estimates are for percentages of persons.

			October 2003 MI CPS					
Characteristics	MI BRFSS		All HHs			Phone HHs		
Household	n	% (se)	n	% (se)	Diff (se)	n	% (se)	Diff (se)
Family Income								
< $20,000	540	41.0 (2.6)	409	24.6 (0.9)	16.4 (2.8)	357	27.7 (0.9)	13.3 (2.9)
$20,000-$34,999	739	52.9 (2.2)	438	42.4 (1.0)	10.5 (2.4)	407	44.4 (1.0)	8.5 (2.6)
$35,000-$49,999	545	69.7 (2.2)	333	63.1 (1.0)	6.6 (2.4)	321	65.1 (1.0)	4.6 (2.6)
$50,000-$74,999	574	82.3 (1.7)	430	70.9 (0.9)	11.4 (1.9)	420	70.4 (1.0)	11.9 (2.2)
$75,000+	669	91.7 (1.1)	690	89.9 (0.6)	1.8 (1.3)	687	90.1 (0.6)	1.6 (1.4)
Children in HH								
Yes	1,194	74.6 (1.5)	1,063	69.6 (1.0)	5.0 (1.8)	1,012	72.1 (0.9)	2.5 (2.0)
No	2,248	61.1 (1.2)	1,913	51.0 (1.0)	10.1 (1.6)	1,808	53.5 (1.0)	7.6 (1.9)
Person								
Age								
18 - 24	230	67.4 (3.3)	366	60.9 (1.0)	6.5 (3.5)	319	68.9 (1.0)	-1.5 (3.6)
25 - 34	469	73.7 (2.3)	492	59.3 (1.0)	14.4 (2.5)	463	61.9 (1.0)	11.8 (2.7)
35 - 44	637	76.0 (1.9)	623	68.5 (1.0)	7.5 (2.1)	589	71.5 (0.9)	4.5 (2.3)
45 - 54	720	75.2 (1.8)	557	70.6 (0.9)	4.6 (2.0)	537	72.7 (0.9)	2.5 (2.2)
55 - 64	588	67.7 (2.1)	429	52.7 (1.0)	15.0 (2.3)	414	54.3 (1.0)	13.4 (2.6)
65 +	801	36.3 (1.9)	509	30.8 (1.0)	5.5 (2.1)	498	31.1 (1.0)	5.2 (2.3)
Race/Ethnicity								
NH White	2,959	69.8 (0.9)	2,402	60.9 (1.0)	8.9 (1.4)	2,300	62.7 (1.0)	7.1 (1.7)
NH Black	298	47.3 (3.3)	381	35.7 (1.0)	11.6 (3.4)	341	40.3 (1.0)	7.0 (3.6)
Other	153	66.8 (4.3)	193	61.4 (1.0)	5.4 (4.4)	179	64.9 (1.0)	1.9 (4.5)
Gender								
Male	1,362	68.7 (1.4)	1,407	58.4 (1.0)	10.3 (1.7)	1,325	61.1 (1.0)	7.6 (2.0)
Female	2,083	65.0 (1.2)	1,569	56.9 (1.0)	8.1 (1.6)	1,495	59.2 (1.0)	5.8 (1.9)
Education								
Less than HS	350	40.4 (3.2)	410	27.9 (0.9)	12.5 (3.3)	371	30.7 (1.0)	9.7 (3.5)
HS Grad/GED	1,080	54.4 (1.8)	1,020	45.9 (1.0)	8.5 (2.1)	953	48.3 (1.0)	6.1 (2.3)
<4yrs College	962	71.8 (1.6)	863	67.9 (1.0)	3.9 (1.9)	828	70.0 (1.0)	1.8 (2.1)
College Grad+	1,043	85.5 (1.2)	683	81.0 (0.8)	4.5 (1.5)	668	81.8 (0.8)	3.7 (1.7)
Employment								
Employed	1,855	75.5 (1.1)	1,827	65.6 (1.0)	9.9 (1.5)	1,738	68.3 (1.0)	7.2 (1.8)
Unemployed	165	56.9 (4.5)	123	54.7 (1.0)	2.2 (1.2)	108	62.6 (1.0)	-5.7 (1.5)
NILF*	1,420	54.4 (1.5)	1,025	43.4 (1.0)	11.0 (1.4)	973	44.8 (1.0)	9.6 (1.7)
Marital Status								
Married	1,934	75.3 (1.1)	1,722	65.9 (1.0)	9.4 (1.5)	1,673	67.1 (1.0)	8.2 (1.8)
Not Married	1,509	54.0 (1.6)	1,254	46.6 (1.0)	7.4 (1.9)	1,147	50.3 (1.0)	3.7 (2.2)
Total	3,445	66.8 (0.9)	2,976	57.6 (1.0)	9.2 (1.4)	2,820	60.1 (1.0)	6.7 (1.7)

* NILF = not in labor force

Table 3: Estimated Percent of MI Adults with a (BRFSS) Health Characteristic by Presence of Home Internet

	Health Characteristics		Internet at Home?		
		Response	YES	NO	Difference
Variable	Description	Category	Pct (se)	Pct (se)	Pct (se)
V1_1	Would you say that in general your good, health is excellent, Very good, fair, or poor?	Good to Excellent	89.8 (0.7)	75.6 (1.4)	14.2*** (1.6)
V2_1	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare?	Yes	90.7 (0.8)	86.3 (1.3)	4.4*** (1.5)
V2_2	Do you have one person you think of as your personal doctor or health care provider?	One or more	84.0 (1.0)	83.1 (1.4)	1.0 (1.7)
V2_3	Was there a time in the past 12 months when you needed to see a doctor but could not because of the cost?	Yes	9.4 (0.7)	13.5 (1.2)	-4.1*** (1.4)
V3_1	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?	Yes	82.6 (0.9)	69.5 (1.5)	13.1*** (1.8)
V4_1	Have you ever been told by a doctor that you have diabetes?	Yes[a]	5.5 (0.5)	12.7 (1.0)	-7.2*** (1.1)
V5_1	Have you ever been told by a doctor, nurse, or other health professional that you have high blood pressure?	Yes[a]	22.6 (1.0)	35.5 (1.5)	-12.9*** (1.8)
V5_2	Are you currently taking medicine for your high blood pressure (among those ever diagnosed)?	Yes	74.1 (2.2)	75.1 (2.5)	-1.0 (3.3)
V6_3	Have you ever been told by a doctor, nurse, or other health professional that your blood cholesterol is high (among those ever tested)?	Yes	36.6 (1.3)	40.5 (1.7)	-3.9 (2.2)
V8_1	Are you now trying to lose weight?	Yes	46.7 (1.2)	41.6 (1.6)	5.1** (2.0)
OBESE	Calculated Variable: Obese	Obese	23.9 (1.1)	28.5 (1.5)	-4.6* (1.9)
V9_1	Have you ever been told by a doctor, nurse or other health professional that you had asthma?	Yes	13.8 (0.8)	13.5 (1.2)	0.2 (1.5)
V9_2	Do you still have asthma (among those ever diagnosed)?	Yes	65.8 (3.2)	78.0 (4.0)	-12.2** (5.1)
V10_1	During the past 12 months, have you had a flu shot?	Yes	27.9 (1.0)	36.5 (1.5)	-8.6*** (1.8)
V11_1	Have you smoked at least 100 cigarettes in your entire life?	Yes	48.2 (1.2)	57.8 (1.6)	-9.6*** (2.0)
CURRSMKR	Calculated Variable: Current Smoking Status	Current Smoker	23.0 (1.1)	31.8 (1.6)	-8.8* (2.0)
V14_18	Are you currently pregnant (among females 18-44)?	Yes	3.8 (0.8)	1.8 (0.9)	2.0 (1.3)
AR_STAT	Arthritis status (diagnosed, joint symptoms, neither)	Diagnosed or Joint symptoms	47.2 (1.2)	58.6 (1.7)	-11.4*** (2.1)
V15_5	Are you now limited in any way in any of your usual activities because of arthritis or joint symptoms?	Yes	23.1 (1.4)	33.6 (1.9)	-10.5*** (2.4)
V15_6	In this next question we are referring to work for pay. Do arthritis or joint symptoms now affect whether you work, the type of work you do, or the amount of work you do?	Yes	19.4 (1.4)	32.4 (2.7)	-13.0*** (3.0)

Note: [a] Excludes health conditions during pregnancy. ∼0 indicates that the estimate rounds to but is not equivalent to zero. Two-tailed p-Value significance: *(0.05,0.1]; **(0.01,0.05]; *** ≤ 0.01.

Table 3: Continued

Health Characteristics		Response	Internet at Home? YES	Internet at Home? NO	Difference
Variable	Description	Category	Pct (se)	Pct (se)	Pct (se)
V16_1	In the past 3 months, have you had a fall (among those aged 45 years or older)?	Yes	12.1 (1.1)	15.0 (1.3)	-2.8* (1.7)
V17_1	Are you limited in any way in any activities because of physical, mental, or emotional problems?	Yes	17.5 (0.9)	27.2 (1.4)	-9.6*** (1.7)
V17_2	Do you now have any health problem that requires you to use special equipment, such as a cane, a wheelchair, a special bed, or a special telephone?	Yes	3.6 (0.4)	8.9 (0.8)	-5.3*** (0.9)
V18_2	Now, thinking about the moderate activities you do [fill in (when you are not working,) if "employed" or "self-employed"] in a usual week, do you do moderate activities for at least 10 minutes at a time, such as brisk walking, bicycling, vacuuming, gardening, or anything else that causes small increases in breathing or heart rate?	Yes	88.6 (0.8)	77.2 (1.4)	11.5*** (1.6)
V18_5	Now, thinking about the vigorous activities you do [fill in (when you are not working) if "employed" or "self-employed"] in a usual week, do you do vigorous activities for at least 10 minutes at a time, such as running, aerobics, heavy yard work, or anything else that causes large increases in breathing or heart rate?	Yes	53.8 (1.2)	37.8 (1.7)	16.0*** (2.1)

Health Characteristics		Response	Internet at Home? YES	Internet at Home? NO	Difference
Variable	Description	Category	Mean (se)	Mean (se)	Mean (se)
TOTALFRU	Fruit & juice times/day	*na*	1.4 (∼0)	1.5 (∼0)	∼0 (0.1)
TOTALVEG	Vegetables times/day	*na*	2.1 (∼0)	2.1 (∼0)	0.1 (0.1)
FRVEG	Fruits & Vegetables times/day	*na*	3.6 (∼0)	3.5 (0.1)	∼0 (0.1)
NDRINKMO	Number of Alcoholic Drinks per Month	*na*	13.1 (0.7)	13.3 (1.3)	-0.3 (1.5)

Note: [a] Excludes health conditions during pregnancy. ∼0 indicates that the estimate rounds to but is not equivalent to zero. Two-tailed p-Value significance: *(0.05,0.1]; **(0.01,0.05]; *** ≤ 0.01.

A comparison of the health characteristics for those with and without home Internet access is provided in Table 3. If the subset with Internet access differs appreciably from those without and from the MI telephone population as a whole, this could signal that the scope of inferences from the Internet survey is limited. In fact, a statistically significant difference was detected in 15 of the 29 analytic variables at a significance level of 0.01 or lower, two variables were different at the 0.05 level, and three variables were different at the 0.10 level. Significant differences were not detected for nine of the 29 variables. The Michigan adult population with home Internet access generally has better health and is more health conscious than the non-Internet population. This is seen, for example, in the higher levels of physical exercise (V3_1: 82.6 vs. 69.5), better perceived health (V1_1: 89.8 vs. 75.6), and lower rates of health conditions such as diabetes (V4_1: 5.5 vs. 12.7) for the home Internet group. Additionally, they are less likely to have arthritis related limitations (V15_5: 23.1 vs. 33.6) and to have fallen in the past three months (V16_1: 12.1 vs. 15.0). As noted previously, home Internet access declines for groups beyond age 44 in the study population. Therefore, these findings are consistent with the previous re-

search findings that note an average age difference between users and non-users of the Internet.

Models for Health-Related Characteristics

As discussed previously, some differences in the health-related outcomes exist for certain domains between persons with Internet access at home and those without. If the detectable differences can be eliminated, or at least, substantially reduced by adjusting for covariates like those in Table 1, then it may be feasible to adjust data from an Internet sample to represent the adult target population. In this section we examine whether presence/absence of home Internet access is a significant predictor of the various health-related variables in Table 3 using models that include various, personal demographic characteristics. Due to the lack of a detectable difference in the continuous variables, we chose to exclude them from subsequent analyses and focus only on the binary variables. As a convenient short-hand, we will refer to persons with Internet access at home as the 'access' group and those without Internet access at home as 'non-access'.

In Table 4, we test for the significance between presence/absence of Internet access at home as a predictor for each of the 25 binary analysis variables in a logistic regression setting.[6] The model covariates include an indicator for Internet access at home and the eight demographic characteristics discussed in Tables 1 and 2. When controlling for the demographic characteristics, the significant difference between the access and non-access groups shown in Table 3 disappears for 16 of the 25 health outcomes; non-significance is maintained for four of the outcomes. In only three models shown in Table 4 (any physical activity, high cholesterol, and current smoking status) is Internet at home significant at the 0.05 level or lower, suggesting the need for additional covariates. A slight significance at the 0.1 level still exists for moderate physical activity (V18_2). The introduction of the model covariates for high cholesterol rates (V6_3) introduced a significant difference (0.05) between access and non-access that was not detected in previous tests (Table 3). Ninety-five percent confidence intervals for the ratio of the odds of having the particular health characteristic for access vs. non-access were also examined (Table 4). For most of the health characteristics, the confidence intervals include the value 1.0 indicating that the difference in the odds for access and non-access is small. However, differences were detectable for any physical activity (V3_1) and high cholesterol rates (V6_3) even after controlling for the other covariates.

Note that the significant difference between presence/absence of Internet access at home is eliminated for twelve of the health outcomes with a 'minimal' logistic model that includes only family income and age category. The significance was eliminated for only two variables with a model containing only age and for five variables for a model containing only household income. That is, 18 of 25 health variables had significant differences between the access and non-access means after adjustment for age group only; 15 of 25 variables had significant differences after adjusting only for income. Eight models required more explanatory variables to eliminate the significant difference. These results suggest that household income and age are the strongest correlates of (MI BRFSS) home Internet access but that other covariates are often needed to make the Internet access variable unnecessary when modeling health characteristics. Of course, a non-significant test on the Internet-at-home variable does not mean that there is no effect at all. A larger sample size would likely detect a non-zero coefficient. However, the practical question for survey estimation is whether the effect is small enough, after accounting for other covariates, that a sample of Internet-at-home persons can be used to make estimates for the entire population that are nearly unbiased. We address this issue more directly in section 5.

We also investigated whether an "intensity of Internet use" variable was related to the health characteristics in Table 4. The questions in Appendix A were used to create an intensity variable with categories: heavy (Q. 31.21=1), medium (Q. 31.21=2 or 3), light (Q. 31.21=4, 5, or 6), and no use (Q. 31.20=1). For each health variable we tested whether the proportion with the characteristic was the same across the four categories using a Wald statistic. Ten of 25 tests were significant at the 0.05 level. We also used the intensity variable as an independent variable in the models in Table 4 in place of the Internet-at-home variable. The coefficient of the intensity variable was significant at the 0.05 level in 5 of the 25 models. Thus, inclusion of the other covariates reduced the number of health variables for which intensity was a potentially useful predictor but did not eliminate it entirely.

Although we ran logistic models to predict the binary characteristics, linear models implicitly underlie weighted survey estimators of the form $\hat{T} = \sum w_i y_i$ Thus, we also fit linear models to predict health characteristics using the same covariates as in Table 4. These models produced the same general conclusions as the logistic models - accounting for the demographic characteristics led to non-significant Internet-at-home variables in 20 of 25 models.

Survey Weights for the Internet Cases

The problem of adjusting for nonresponse and coverage errors is common to many surveys and is usually addressed by weighting the survey sample up to the desired target population, even when the sample does not fully cover that population. For example, Part II of Lepkowski et al. (2007) discusses this approach extensively for telephone surveys. Kott (2006) and Särndal and Lundström (2005) describe the use of calibration weighting methods to adjust for nonresponse. To see whether survey weights could be computed that effectively adjust for coverage errors, we calculated general regression (GREG) weights, a specific type of calibration adjustment, using the MI data for the sample persons who

[6] Table 4 contains the results of many explicit and implicit significance tests. We have not adjusted the levels of the tests to account for multiple comparisons, but rather use the test results as an exploratory tool to suggest whether home Internet access is needed to model health characteristics.

Table 4: Ninety-five percent confidence intervals for the ratio of the odds of having a characteristic to the odds of not having the characteristic for 25 health-related variables.

Health Characteristics	Significance in Table 3	Internet at home in model	95% CI on odds ratio for Internet at Home	Health Characteristics	Significance in Table 3	Internet at home in model	95% CI on odds ratio for Internet at Home
V1_1: General Health	***	*	(0.96, 1.69)	V9_2: Still Have Asthma	**		(0.78, 1.23)
V2_1: Any Health Care Coverage[a]	***		(0.59, 1.26)	V10_1: Flu Shot (12 Mo) [a]	***		(0.73, 1.12)
V2_2: Personal Doctor			(0.65, 1.19)	V11_1: Smoked 100 Cigs[b]	***		(0.73, 6.80)
V2_3: Cost Prevented Dr Visit[a]	***		(0.83, 1.71)	CURRSMKR: Current Smoking Status[b]	*	**	(0.69, 1.08)
V3_1: Physical Activity	***	**	(1.05, 1.72)	V14_18: Now Pregnant (Age < 45)			(0.64, 1.18)
V4_1: Diabetes[b]	***		(0.61, 1.18)	AR_STAT: Diagnosed Arthritis/CJS[a]	***		(0.56, 1.20)
V5_1: High BP Ever[a]	***		(0.79, 1.27)	V15_5: Limited by, Diagnosed Arth/CJS[a]	***		(0.63, 1.31)
V5_2: Taking BP Meds			(0.85, 2.01)	V15_6: Diag Arth/CJS Affects Work[a]	***		(0.72, 1.18)
V6_3: Ever Told Cholesterol High		**	(1.07, 1.72)	V16_1: Fell in Past 3 Mo[a]	*		(0.64, 1.52)
V8_1: Trying to Lose Wt[a]	**		(0.91, 1.38)	V17_1: Now Limited in Any Way[a]	***		(0.99, 1.75)
OBESE: Obese (bmi=30) vs Not Obese[a]	*		(0.75, 1.35)	V17_2: Health Probs, Special Equip[a]	***		(0.91, 1.41)
V9_1: Ever Told Asthma[b]			(0.35, 1.32)	V18_2: Mod Physical Activity /Week Activity /Week	***	*	(0.93, 1.71)
				V18_5: Vig Physical Activity /Week[b]	***		(0.78, 1.64)

Note: The second column (Significance in Table 3) is the significance level of the test that the difference in proportions having a health characteristic is zero for persons with and without Internet access at home. P-Value significance: *(0.05,0.1]; **(0.01,0.05]; *** ≤ 0.01. [a] The "Internet at Home" variable is not significantly different from zero in a model accounting only for family income and age in a minimal model. [b] The minimal model for this set of variables contains fewerthan the full set of covariates but extends beyond a covariate set containing income and age.

reported having Internet access at home. As shown in Table 2, 66.8 percent (2,179 persons) of the MI BRFSS sample had access at home. GREG estimators are described in Särndal, Swensson, and Wretman (1992) and are motivated by linear relationships between an analysis variable y and a set of covariates, x_1, x_2, K, x_p. The form of a GREG estimator of a population total is $\hat{T} = \hat{T}_{0y} + \sum_{j=1}^{p} b_j (T_{xj} - \hat{T}_{0xj})$ where $\hat{T}_{0y} = \sum_{i \in s} w_{0i} y_i$ is an estimator of the population total of y based on an initial set of weights, w_{0i}; s is the set of sample units; $\hat{T}_{0xj} = \sum_{i \in s} w_{0i} x_{ji}$ is the estimator of the population total for the jth covariate; T_{xj} is the population total for that covariate calculated here using the MI CPS; and b_j is an estimator of a regression coefficient. The estimator of the slope vector $\mathbf{b} = (b_1, \ldots, b_p)$ is obtained via weighted least squares using the w_{0i} survey weights. The initial weights may be base weights, i.e., inverses of inclusion probabilities, or nonresponse-adjusted base weights. A GREG implies a weight, w_{Gi} , for sample unit i (see, e.g., Särndal, Swensson, and Wretman 1992:232) so that the usual procedure of computing survey estimates as weighted sums of data fields can be used. The estimated mean of y is then

$$\hat{\bar{y}} = \sum_s w_{Gi} y_i / \sum_s w_{Gi}.$$

Software for computing GREGs and more general calibration estimators is now freely available. The French Institut National de la Statistique et des Études Économiques has written a SAS® macro called CALMAR that can be downloaded from www.insee.fr (see Sautory 2003). The GREG and other calibration functions are also part of the R® survey package (Lumley 2004, 2005; R Development Core Team 2005).

A GREG estimator is motivated by the linear model, $E_M(y_i) = bfx_i'\beta$ where E_M denotes expectation with respect to a model, $\mathbf{x}_i'$ is the vector of p covariates for unit i, and β is a slope parameter. The GREG is model-unbiased in the sense that $E_M(\hat{T} - T) = 0$ where $T = \sum_{i \in U} y_i$ is the population total. This follows since $E_M(T) = \sum_U \mathbf{x}_i'\beta$ and $E_M(\hat{T}_{0y}) = \sum_{ji}^{p} \beta_j \hat{T}_{0xj}$ as long as $\mathbf{b}$ is a model-unbiased estimator of β. A key requirement is that the same model, $E_M(y_i) = \mathbf{x}_i'\beta$ hold for the entire population. If, for example, a separate model holds for the access and non-access groups, then samples are needed from both groups in order to estimate the model parameters.

A GREG can also be thought of as reducing coverage error by using the population covariate totals as part of the estimator. For example, if the estimated number of persons aged 65 and over is too small based on the initial w_{0i} weights, the GREG weights are calibrated in the sense that the estimate based on the GREG weights will equal the control count of 65+ year olds. More generally, for each covariate, the GREG reproduces the population totals, i.e., $\sum_{i \in s} w_{Gi} x_{ji} = T_{xj}$. Other calibration estimators, like raking, will also reproduce population covariate totals and are reasonable alternatives to the GREG. However, we focus here on the use of calibration to minimize undercoverage, and not on the particular calibration algorithm used to accomplish our goal.

To check the efficacy of this method, we computed three sets of GREG weights for the 2,301 cases with Internet at home based on the following sets of covariates: (1) seven covariates listed in Table 4 (age group, race/ethnicity, gender, education, presence of children in household, employment, and marital status); (2) the four covariates currently used in the BRFSS poststratification (age group, race, gender, and Hispanic origin); and (3) age group only. The population values were taken from the MI CPS. Although we found earlier for the BRFSS that household income was a useful predictor of health characteristics, we had to exclude the variable from the list of model covariates because of the incomplete CPS data (22.7 percent of the MI CPS records were missing income). Comparisons of the estimated percentages from the full MI BRFSS using the original weights and the Internet-at-home subset with the three sets of GREG weights are shown in Table 5. An insignificant difference between the estimates for the full MI BRFSS and the GREG-adjusted estimates suggests that the coverage errors have been reduced, although not necessarily eliminated, through the weight adjustment.[7]

Note that when more covariates are added to the GREG, there is a tendency for the GREG weights to become more variable in order to hit more control totals. This leads to estimated totals and proportions with slightly larger SEs. For example, the GREG SEs for V1_1 (General Health) under the age-group, 4-covariates, and 7-covariates models are 0.7, 0.8, and 1.0, respectively. But, the point estimates of the proportions of persons with access also change as more covariates are added to the GREG, typically making them closer to the full MI BRFSS estimates.

A significant difference (0.05 level and lower) between the estimates for the full MI BRFSS and GREG weights incorporating only age group exists for 21 of the 25 health characteristics. The significant difference was eliminated for four of these variables by additionally incorporating race, gender, and Hispanic origin into the weight adjustment (V5_1, V8_1, OBESE, and V10_1). A minimally significant difference at the 0.10 level remained for only four of the 25 health characteristics once the more complete list of seven covariates was used to calculate the GREG weights. The significant difference in the estimates for V5_2 (taking blood pressure medication) for the full MI BRFSS and the 7-covariate GREG remains even after adjustment.

The reduction in coverage errors for 25 health characteristics with the 7-covariate GREG in comparison with the other GREG weights is shown graphically in Figure 2. Here, we examine the percent relative difference (PRD) of the GREG estimates ($\hat{p}_{GREG}$) from the full MI BRFSS estimates ($\hat{p}_{full}$) using the formula $100 \times (\hat{p}_{GREG} - \hat{p}_{full}) / \hat{p}_{full}$. Variables are sorted by PRD to make patterns more evident. As Figure 2 makes clear, the effectiveness of statistical adjust-

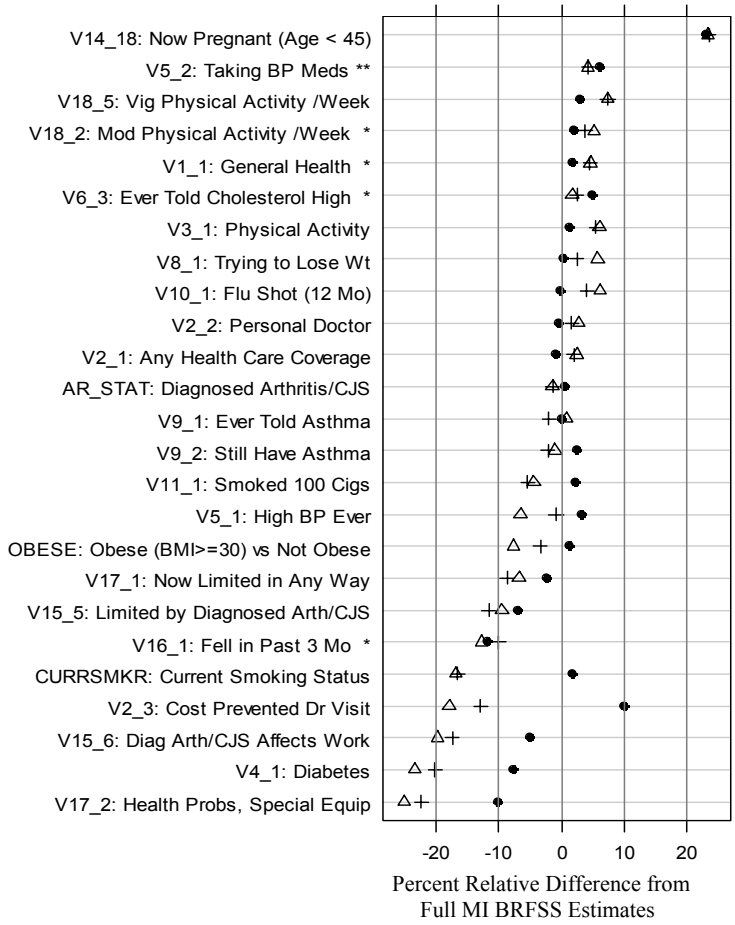
[7] Variances of differences were estimated using SUDAAN in a way that accounted for the correlation between estimates. The fact that the MI CPS covariate totals are subject to sampling error was not accounted for, implying that estimated variances of differences are likely to be too small.

Table 5: Estimated Percent of MI Adults with a Health Characteristic for Full BRFSS and Home-Internet Subset by Original and GREG Weights

			Internet-at-Home					
		Full MI	Age Group		4 Covariates		7 Covariates	
Health Characteristics	Response	BRFSS Pct (se)	GREG Pct (se)	Diff Pct (se)	GREG Pct (se)	Diff Pct (se)	GREG Pct (se)	Diff Pct (se)
V1_1: General Health	Good to Excellent	84.9 (0.7)	88.8 (0.7)	-3.9*** (0.6)	88.7 (0.8)	-3.8*** (0.6)	86.4 (1.0)	-1.5* (0.8)
V2_1: Any Health Care Coverage	Yes	89.3 (0.7)	91.5 (0.7)	-2.2*** (0.5)	91.0 (0.8)	-1.7*** (0.6)	88.4 (1.1)	0.9 (0.8)
V2_2: Personal Doctor	One or more	83.6 (0.8)	85.9 (0.9)	-2.3*** (0.6)	84.9 (1.0)	-1.3** (0.7)	83.2 (1.2)	0.4 (0.9)
V2_3: Cost Prevented Dr Visit	Yes	10.8 (0.6)	8.9 (0.7)	1.9*** (0.5)	9.4 (0.8)	1.4** (0.6)	11.9 (1.1)	-1.1 (0.8)
V3_1: Physical Activity	Yes	78.2 (0.8)	82.9 (0.9)	-4.7*** (0.6)	82.5 (1.0)	-4.2*** (0.7)	79.2 (1.2)	-1.0 (0.9)
V4_1: Diabetes	Yes[a]	7.9 (0.5)	6.1 (0.5)	1.8*** (0.4)	6.3 (0.6)	1.6*** (0.5)	7.3 (0.8)	0.6 (0.6)
V5_1: High BP Ever	Yes[a]	26.8 (0.8)	25.0 (1.0)	1.8** (0.7)	26.5 (1.2)	0.2 (0.8)	27.6 (1.4)	-0.8 (1.1)
V5_2: Taking BP Meds	Yes	74.6 (1.6)	77.7 (1.9)	-3.1** (1.4)	77.7 (2.2)	-3.2*(1.7)	79.2 (2.4)	-4.6** (2.0)
V6_3: Ever Told Cholesterol High	Yes	37.5 (1.0)	38.2 (1.3)	-0.6 (0.8)	38.4 (1.3)	-0.9 (0.9)	39.4 (1.5)	-1.9* (1.1)
V8_1: Trying to Lose Wt	Yes	45.0 (1.0)	47.6 (1.2)	-2.6*** (0.8)	46.1 (1.3)	-1.1 (0.8)	45.1 (1.4)	-0.1 (1.0)
OBESE: Obese (BMI$\geq$30) vs Not Obese	Obese	25.4 (0.9)	23.4 (1.0)	1.9*** (0.7)	24.5 (1.2)	0.9 (0.8)	25.7 (1.3)	-0.3 (1.0)
V9_1: Ever Told Asthma	Yes	13.6 (0.7)	13.7 (0.8)	-0.1 (0.5)	13.3 (0.8)	0.3 (0.5)	13.6 (0.9)	0 (0.7)
V9_2: Still Have Asthma	Yes	69.1 (2.5)	68.4 (3.1)	0.7 (1.9)	67.6 (3.2)	1.5 (1.9)	70.7 (3.2)	-1.7 (2.2)
V10_1: Flu Shot (12 Mo)	Yes	30.6 (0.8)	32.4 (1.1)	-1.9** (0.7)	31.8 (1.2)	-1.2 (0.8)	30.5 (1.2)	0.1 (0.9)
V11_1: Smoked 100 Cigs	Yes	51.4 (1.0)	49.1 (1.2)	2.3*** (0.8)	48.6 (1.3)	2.8*** (0.8)	52.6 (1.4)	-1.2 (1.0)
CURRSMKR: Current Smoking Status	Current Smoker	25.8 (0.9)	21.5 (1.0)	4.4*** (0.7)	21.5 (1.1)	4.3*** (0.7)	26.3 (1.3)	-0.5 (1.0)
V14_18: Now Pregnant (Age $<$ 45)	Yes	3.2 (0.6)	3.9 (0.8)	-0.7** (0.4)	3.9 (0.9)	-0.7* (0.4)	3.9 (1.4)	-0.7 (1.0)
AR_STAT: Diagnosed Arthritis/CJS	Yes	50.7 (1.0)	50 (1.2)	0.7 (0.8)	50.0 (1.3)	0.7 (0.8)	51.0 (1.4)	-0.3 (1.0)
V15_5: Limited by Diagnosed Arth/CJS	Yes	27.2 (1.1)	24.6 (1.4)	2.6** (1.1)	24.1 (1.5)	3.2*** (1.1)	25.3 (1.7)	1.9 (1.3)
V15_6: Diag Arth/CJS Affects Work	Yes	23.8 (1.3)	19.1 (1.4)	4.7*** (1.0)	19.7 (1.5)	4.1*** (1.0)	22.7 (1.8)	1.2 (1.3)
V16_1: Fell in Past 3 Mo	Yes	13.3 (0.8)	11.6 (1.0)	1.7** (0.8)	12.0 (1.1)	1.3* (0.8)	11.8 (1.1)	1.6* (0.9)
V17_1: Now Limited in Any Way	Yes	20.7 (0.8)	19.3 (0.9)	1.4** (0.7)	18.9 (1.0)	1.8*** (0.7)	20.2 (1.2)	0.5 (0.9)
V17_2: Health Probs, Special Equip	Yes	5.6 (0.4)	4.2 (0.5)	1.4*** (0.4)	4.3 (0.6)	1.2** (0.5)	5.0 (0.8)	0.6 (0.7)
V18_2: Mod Physical Activity /Week	Yes	84.7 (0.7)	89.1 (0.7)	-4.3*** (0.6)	87.9 (0.9)	-3.2*** (0.6)	86.3 (1.1)	-1.6* (0.8)
V18_5: Vig Physical Activity /Week	Yes	48.3 (1.0)	51.9 (1.2)	-3.5*** (0.8)	51.9 (1.3)	-3.5*** (0.8)	49.7 (1.4)	-1.4 (1.0)

Note: The "Diff" column contains the percentage difference between the original MI BRFSS and GREG estimates and the associated standard error. The covariates in the "4 Covariates" GREG weight adjustment includes the four the BRFSS poststratification variables (age group, race, gender, and Hispanic origin); the "7 Covariates" adjustment includes age group, race/ethnicity, gender, education, presence of children in household, employment, and marital status. P-Value significance: *(0.05,0.1]; **(0.01,0.05]; *** $\leq$0.01. Standard errors of differences between the full MI BRFSS estimates and the three sets of GREG estimates were computed with SUDAAN, accounting for the correlation between estimates.

Figure 2. The Percent Relative Difference from the Full MI BRFSS Estimate for 25 BRFSS Health Estimates calculated with Three GREG Weights.



Note: "•"= Seven auxiliary variables (Age group, Race/Ethnicity, Gender, Education, Presence of Children in Household, Employment status, and Marital status); "+"= Four BRFSS poststratification variables (Age group, Race, Gender, and Hispanic Origin); "△"= Age group. P-Value significance ($^{*}(0.05, 0.1]$; $^{**}(0.01, 0.05]$; $^{***} \leq 0.01$) denotes the cases from Table 5 where the 7-covariate GREG estimate was significantly different from the full MI BRFSS estimate.

ment is somewhat mixed. Assuming that the full MI BRFSS estimates are closest to the truth for the target population, there are a number of estimates where the percentage difference between them and the GREG estimates is relatively large. This is especially true for the age-only and 4-covariate GREGs. However, most of the larger PRDs (e.g., V2_3, V15_6, V4_1, and V17_2) are for variables where the full MI BRFSS and the 7-covariate GREG estimates were not significantly different.

Estimates for 20 of the 25 variables showed the lowest PRD with the 7-covariate GREG weight adjustment (symbol •). In the five cases where the 7-covariate GREG did not have the smallest PRD, it was competitive with the best choice. Using only age in the weight adjustment resulted in the largest PRD in absolute value for 17 of the 25 health characteristics suggesting that an insufficient amount of coverage error has been eliminated with this technique.

Conclusion

Using the Internet to survey household populations is extremely appealing because of both timeliness and cost. However, Internet surveys are obviously restricted to persons who can access the Internet. Whether estimates from this restricted group can be used to make inferences about a larger population depends on whether households that have Internet access are different from the general population of households. The standard randomization justification of weighting the random sample to represent the target population does not apply to Internet survey estimates because the sample itself is not selected from the correct population. Therefore, weighting estimators only by inverse inclusion probabilities will not result in design-unbiased estimators. Instead, we must rely on statistical models to attempt to create unbiased estimators for the complete household population as in Valliant et al. (2000).

The question of representativeness of an Internet sample

is complicated by the type of frame used for selecting the initial sample. Two of the *better* choices for frames would be a list-assisted telephone sample and an area probability sample. Landline telephone samples can have coverage problems related to the exclusion of persons with Internet access living in either non-telephone or cell-phone only households. In theory, all households are available for selection from area samples; however, these samples can suffer from some undercoverage in certain race/ethnicity and age groups. The level of undercoverage in area samples is typically less than experienced with telephone surveys. To combat such coverage problems, surveys use poststratification or more elaborate calibration estimation to form weighted sample distributions that match those of the target population. Even if an area sample is the starting point for an Internet sample, the problem remains that households without Internet access are not covered by the sample. There may also be problems in getting persons to participate within households that have access to the Internet. Gaining cooperation from older persons and others who do not often use the Internet may be a particular challenge.

We examined the coverage and estimation issues by comparing demographic distributions in data collected in 2003 from the US state of Michigan in the Current Population Survey (CPS) and the Behavioral Risk Factor Surveillance System (BRFSS). We also compared the health characteristics of persons with home access to the Internet and those without based on the BRFSS.

Distributions do differ between CPS and BRFSS within some categories that were not used in poststratifying the BRFSS, such as family income and education. Internet penetration rates are also significantly different between CPS and BRFSS within many demographic categories. Using BRFSS data, we also found significant differences in health-related characteristics between persons in Internet and non-Internet households. For example, persons with Internet access reported having better health in general, were more likely to have health care coverage, were more likely to exercise, and were less likely to have high blood pressure or diabetes. Thus, based on these marginal differences, it appears that a sample of persons in Internet households cannot be used to represent all households.

However, when models were fitted to predict the probability of having certain health characteristics, like insurance, diabetes, and a number of others, we found that an indicator for having the Internet at home typically was not a useful predictor after a sufficient number of demographics like family income, gender, education, and age group were included in the model. In other words, the predicted value for a person is essentially the same regardless of whether the person has Internet access or not after controlling for other demographic characteristics.

To study whether statistical adjustments can reduce or eliminate coverage biases in actual survey estimators, we weighted the MI BRFSS sample of persons with Internet access at home using general regression (GREG) techniques. GREG estimation is a flexible way of accounting for regression relationships like the ones described above. We found that, with a rich enough set of covariates, GREG-weighted estimates were quite close to estimates from the full MI BRFSS sample. However, adjusting by age-group only or by age, race, gender, and Hispanic origin (which are the variables normally used in raking for BRFSS) produced estimates that were statistically significantly different from the full sample MI BRFSS estimates for most of the variables we studied. Only when we incorporated education, presence of children in the household, employment status, and marital status, were we able to produce estimates that were statistically close to the full BRFSS estimates.

A weakness of our analysis is that we could not compare adjusted estimates of health characteristics to ones from a survey with higher quality than the MI BRFSS, e.g., to estimates from the CPS or the US National Health Interview Survey[8], because such information was not available for Michigan. There may also be economic (e.g., employment, income, etc.) or other types of variables where the GREG-adjustment would be less effective. Additional covariates, like household size, may be needed for reducing coverage bias, depending on the subject of the survey.

Our analysis highlights one situation in which, for many characteristics, predictions for populations that include persons with and without Internet access at home can be legitimately made based only on a sample of Internet households. Thus, survey weights can be constructed using a method like the GREG based on explanatory variables similar to the ones we have studied here. This requires that population totals for the explanatory variables be known from some external source, such as projections based on the decennial census or estimates from a large well-executed survey like the CPS. Ideally, the weighted estimates will be model-unbiased for population quantities for the combined Internet and non-Internet population, even in cases where a repeated sampling justification does not exist. An alternative is to estimate population values using propensity weighting as in Schonlau et al. (2004) and Schonlau, van Soest, and Kapteyn (2007). However, using that method for coverage adjustments requires data on both covered and non-covered persons, making it considerably less flexible.

The external validity of the Internet-based estimates needs to be carefully examined and not assumed. Simple poststratification, which accounts for a limited number of variables and their interactions, is not likely to adequately adjust for coverage differences in estimates for persons with and without Internet access. General regression estimators or more elaborate calibration estimators are needed. The calibration estimators can flexibly incorporate income, education, race/ethnicity, and other variables as long as control totals are available from demographic projections or large surveys like the CPS. Of course, as home use of the Internet becomes more prevalent, the magnitude of the coverage problem will decrease but nonresponse to Internet surveys may still be substantial and may vary by demographic group. Calibration with many variables will still be essential in those cases.

[8] http://www.cdc.gov/nchs/nhis.htm

Based on the results discussed here, we conclude that there is some hope for using well-designed Internet surveys to make estimates for the general household population. The situation we addressed was one in which a well-controlled sample of persons with access to the Internet could be selected, e.g. from a larger telephone survey in our case, and in which any nonresponse in the Internet survey is ignorable. Our study also requires the implicit assumption that persons would report via Web the same data that they would report in a telephone survey. That is, there is no mode difference between telephone and Web.

Our results do not apply to the types of uncontrolled samples seen, for example, in volunteer Web surveys. Those surveys may entirely omit important demographic groups and have biases that cannot be eliminated. We acknowledge that the results discussed in this paper pertain to one telephone survey of one US state-based population. Nonetheless, the methodology used here shows promise and should be considered for other such analyses of the Internet population.

Acknowledgements

The supplemental BRFSS questions were developed as a cooperative effort among Richard Curtin and James Lepkowski (University of Michigan), Colm O'Muircheartaigh (University of Chicago), Larry Hembroff (Michigan State University), and Harry McGee (Michigan Department of Community Health). This research was funded in part through grant no. UR6/CCU517481-03 from the National Center for Health Statistics to the Michigan Center for Excellence in Health Statistics.

References

Alexander, R. B., & Trissel, D. (1996). Chronic Prostatitis: Results of an Internet Survey. *Urology*, *48*, 568-574.

Ballard, C., & Prine, R. (2002). Citizen Perceptions of Community Policing: Comparing Internet and Mail Survey Responses. *Social Science Computer Review*, *20*, 485-493.

Beniger, J. R. (1998). Presidential Address: Survey and Market Research Confront Their Futures on the World Wide Web. *Public Opinion Quarterly*, *62*, 442-452.

Best, S. J., Krueger, B., Hubbard, C., & Smith, A. (2001). An Assessment of the Generalizability of Internet Surveys. *Social Science Computer Review*, *19*, 131-145.

Braithwaite, D., Emery, J., de Lusignan, S., & Sutton, S. (2003). Using the Internet to Conduct Surveys of Health Professionals: a Valid Alternative? *Family Practice*, *20*, 545-551.

Bureau of Labor Statistics. (1997). *Methodology and Documentation for the Current Population Survey (CPS) - CPS Questionnaire for the Basic Monthly Survey.* (Washington, D.C.: U.S. Bureau of the Census)

Centers for Disease Control and Prevention.(2002). *Behavioral Risk Factor Surveillance System (BRFSS) - State Questionnaire.* (Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (CDC), December, 2002 (V1.5))

Centers for Disease Control and Prevention. (2003). *Technical Information and Data for the Behavioral Risk Factor Surveillance System (BRFSS) - 2003 BRFSS Overview.* (Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention)

Couper, M. P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, *64*, 464-494.

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture This!: Exploring Visual Effects in Web Surveys. *Public Opinion Quarterly*, *68*, 255-266.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web Survey Design and Administration. *Public Opinion Quarterly*, *65*, 230-253.

Dillman, D. A. (2002). *Navigating the Rapids of Change: Some Observations on Survey Methodology in the Early* 21^{st} *Century.* (Draft of Presidential Address to the American Association for Public Opinion Research Annual Meeting)

Fallows, D.(2005). *How Women and Men Use the Internet.* (Report from the Pew Internet and American Life Project)

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

Harwood, P., & Rainie, L. (2004). *People Who Use the Internet Away from Home and Work.* (Report from The Pew Internet and American Life Project)

Kalton, G., & Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, *19*, 81-97.

Kostanich, D., & Dippo, C. (2000). *Current Population Survey: Design and Methodology.* (Technical paper 63. Washington DC: Department of Commerce)

Kott, P. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology*, *32*(2), 133-142.

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics*, *22*(2), 329-349.

Lepkowski, J., Tucker, C., Brick, J. M., De Leeuw, E., Japec, L., Lavrakas, P., et al. (2007). *Advances in Telephone Survey Methodology*. New York: John Wiley.

Lumley, T.(2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*, 1-19.

Lumley, T.(2005). *Survey: Analysis of Complex Survey Samples. R package version 3.01*. Seattle: University of Washington.

Nadimpalli, V., Judkins, D., & Chu, A. (2004). *Survey Calibration to CPS Household Statistics.* (Proceedings of the Survey Research Methods Section, American Statistical Association, 4090-4094)

National Center for Chronic Disease Prevention and Health Promotion. (2004). *2003 Behavioral Risk Factor Surveillance System - Summary Data Quality Report.* (Centers for Disease Control and Prevention)

National Telecommunications and Information Administration. (2002). *A Nation Online: How Americans are Expanding their Use of the Internet.* (Washington, DC)

Newell, C., Whittam, K., & Uriel, Z.(2005). *2005 SAVI Quick Poll Executive Summary.*

R Development Core Team. (2005). *R: A language and environment for statistical computing.* (Vienna, Austria: R Foundation for Statistical Computing)

Research Triangle Institute. (2004). *SUDAAN User's Manual, Release 9.0.* (Research Triangle Park, NC: Research Triangle Institute)

Särndal, C.-E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Sautory, O. (2003). *CALMAR 2: A new version of the CALMAR calibration adjustment program.* (Proceedings of Statistics Canada's Symposium 2003: Challenges in Survey Taking for the Next Decade)

Schonlau, M., Fricker Jr., R. D., & Elliott, M. N. (2002). *Conducting Research Surveys via E-mail and the Web*. Arlington: VA: RAND Publications.

Schonlau, M., van Soest, A., & Kapteyn, A. (2007). Are "Webographic" or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring? *Survey Research Methods*, *1*(3), 155-163.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., et al. (2004). A Comparison Between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, *22*, 128-138.

Suh, B., & Han, I. (2003). The Impact of Customer Trust and Perception of Security Control on the Acceptance of Electronic Commerce. *International Journal of Electronic Commerce*, *7*, 135-162.

The American Association for Public Opinion Research. (2004). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.*

US Census Bureau. (2004). *CPS Basic Monthly Survey: Quality Measures.* (Report from the Joint Project Between the Bureau of Labor Statistics and the Bureau of the Census)

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley and Sons, Inc.

Vehovar, V., Manfreda, K. L., & Batagelj, Z. (1999). *Web Surveys: Can The Weighting Solve The Problem?* (Proceedings of the Section on Survey Methods Research, American Statistical Association, 962-967)

Appendix A: Internet Questions Specific to the 2003 MI BRFSS

31.20 Do you have access to the Internet at home?
< 1 > Yes
< 2 > No [Go to Closing Statement]
< 7 > Don't know [Go to Closing Statement]
< 9 > Refused [Go to Closing Statement]

31.21 How often do you use the Internet at home? Would you say, at least once a day, five to six times a week, two to four times a week, about once a week, less than once a week, or have you not used the Internet in the last month?
< 1 > At least once a day
< 2 > 5-6 times a week
< 3 > 2-4 times a week
< 4 > About once a week
< 5 > Less than once a week
< 6 > Not in the last month
< 7 > Don't know
< 9 > Refused