Comparison of Quarterly and Yearly Calibration Data for Propensity Score Adjusted Web Survey Estimates

Survey Methods: Insights from the Field, Special Issue: Advancements in Online and Mobile Survey Methods

Katherine E. Irimata | National Center for Health Statistics

How to cite this article : Irimata, K.E., He, Y., Cai, B., Shin, H-C., Parsons, V.L. & Parker, J.D. (2020), Comparison of Quarterly and Yearly Data for Propensity Score Adjusted Web Survey Estimates in Survey Methods: Insights from the Field, Special issue: 'Advancements in Online and Mobile Survey Methods'. Retrieved from https://surveyinsights.org/?p=13426

DOI : 10.13094/SMIF-2020-00018

Abstract : While web surveys have become increasingly popular as a method of data collection, there is concern that estimates obtained from web surveys may not reflect the target population of interest. Web survey estimates can be calibrated to existing national surveys using a propensity score adjustment, although requirements for the size and collection timeline of the reference data set have not been investigated. We evaluate health outcomes estimates from the National Center for Health Statistics' Research and Development web survey. In our study, the 2016 National Health Interview Survey as well as its quarterly subsets are considered as reference datasets for the web data. It is demonstrated that the calibrated health estimates overall vary little when using the quarterly or yearly data, suggesting that there is flexibility in selecting the reference dataset. This finding has many practical implications for constructing reference data, including the reduced cost and burden of a smaller sample size and a more flexible timeline.

Introduction

Surveys are important tools for collecting information, particularly at the national level. Probability surveys allow one to estimate outcomes for a specified population of interest which is an important function for federal agencies. For example, the National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics (NCHS) has been used since 1957 to gauge the overall health of the United States. The NHIS is a multipurpose survey covering a range of health topics through in person interviews (https://www.cdc.gov/nchs/nhis/). While national surveys such as the NHIS are often conducted using face-to-face interviews, telephone interviews or mail surveys, more recently web collection has become an increasingly popular mode for surveys. Web surveys, which utilize questionnaires through an online interface, have potential advantages over traditional survey collection methods including lower cost and expedited collection and processing (Callegaro et al. 2015). Although there has been an increasing interest in web-based applications, web surveys have some limitations. For example, since web surveys require Internet access, there may be coverage error (Groves 1989) as non-Internet users are not represented among all survey participants. Although household Internet access has been steadily increasing over the past few decades, 19 percent of households reported not having a broadband Internet

subscription (cable, fiber optic, DSL, cellular data plan, satellite, or fixed wireless) in 2016. Lower income households and households led by an individual 65 years and older are less likely to have broadband Internet access and would be underrepresented in online surveys (Ryan 2017). In addition, web surveys administered via opt-in panels do not utilize a statistical sampling method and cannot produce estimates representative of the target population (Cornesse et al. 2020).

In order to study the properties of web survey estimates for health outcomes, the National Center for Health Statistics (NCHS) has been conducting a series of panel survey studies in the U.S. referred to as the Research and Development Survey (RANDS). The RANDS have been administered by external contractors, including Gallup and the National Opinion Research Center (NORC), using probability-sampled commercial panels. Three rounds of web surveys have been completed. The first two studies were administered by Gallup in quarter 4 of 2015 and in quarter 2 of 2016. The RANDS questionnaires in the first two rounds included a subset of questions from the NHIS as well as probe questions for cognitive evaluation. The third web survey was conducted by NORC in quarter 2 of 2018 and included split sample experiments in addition to select NHIS questions and probe questions. The surveys were conducted as probability surveys using the Gallup Panel and the AmeriSpeak Panel, proprietary sampling frames for Gallup and NORC, respectively. While Gallup and NORC calibrated the data using post-stratification weighting to maintain proportionality of demographic groups in the population, the panels and sample weighting methods differ between the external contractors. In this paper, we provide comparisons from the second round of RANDS collected by Gallup.

To adjust for potential selection bias in web survey estimates, methods such as propensity score weighting can be implemented (Taylor 2000, Terhanian and Bremer 2000). This technique utilizes a reference dataset, often a high-quality probability survey, to calibrate the web survey covariates to resemble the covariate distribution in the reference sample. Demographic variables are often specified in the propensity score model, although other variables can also be used (Rubin and Thomas 1996). There are many benefits to using propensity score weighting techniques including flexibility in model formation and small bias under model misspecification (Drake 1993). Previous studies have indicated that propensity score adjustments can reduce or remove biases due to noncoverage, nonresponse, or nonprobability sampling of web surveys (Lee 2006, Lee and Valliant 2009). But while calibration techniques have frequently been utilized to reweight web survey data, limited research has been performed to evaluate the impact of the properties of a reference dataset on the calibrated web survey estimates. In particular, there is interest in understanding how the size and collection timeframe of the reference dataset impact the calibrated web survey estimates. We consider situations where a limited subset of a reference dataset is available, such as one quarter of an annual survey, and compare the web survey estimates across the varying subsets to identify any differences. The usability of smaller reference datasets or reference datasets collected in different time periods would allow for flexibility in selecting a calibration dataset for propensity score weighting. This is particularly important for timely analyses as previous years of data or subsets of national health surveys could be utilized to produce calibrated web estimates without waiting for upcoming data releases.

This paper implements the propensity score adjustment method for estimation of health outcomes from probability-based survey panel data and evaluates the impact of implementing calibration datasets of varying sizes. We posit that reference surveys with well-matched covariate distributions will produce similar propensity score weighted estimates, even if the reference surveys vary in size or are collected over different time periods. In this paper, this is evaluated through the numerical investigation of the RANDS data. We focus on comparisons using the second round of RANDS which was collected in the second quarter of 2016. To compare health outcomes estimates across different reference datasets, we consider the NHIS 2016 public use file and evaluate five calibration datasets: the full year of NHIS 2016

data (NHIS 2016) as well as the four quarterly subsets (NHIS 2016 quarter 1 data, NHIS 2016 quarter 2 data, NHIS 2016 quarter 3 data, and NHIS 2016 quarter 4 data). This study assesses differences in the estimates based on the size of the reference dataset (full year versus quarterly) as well as on the collection period (quarter 2 versus other time periods).

In this paper, the RANDS data collected by Gallup in 2016 and the NHIS data from 2016 are introduced. Methods used to compare the RANDS and NHIS data, including comparisons of demographic features of respondents and major health outcomes are discussed. To assess similarities and differences between the propensity score calibrated health outcome estimates across the five reference datasets, the propensity score adjustment factors are evaluated. The results of these evaluations are reported followed by a summary and discussion of the findings of this study.

Data

RANDS

The second round of RANDS was conducted in the U.S. using the Gallup Panel. Gallup started the Panel in 2004 as a probability-based survey panel that is representative of the U.S. population. Potential panel members are selected using random-digit-dialing or address-based sampling. The panel is multimodal, and members of the panel are contacted using telephone, web, or mail about specific surveys. Approximately 80 percent of members can be reached by email to complete a web survey (How Does the Gallup Panel Work? 2019).

The RANDS questionnaire was developed by selecting a set of 88 questions from the NHIS as well as a series of 21 probe questions to conduct response pattern assessments (Scanlon 2017). The NHIS questions selected included questions on the topics of general health, food security, health insurance, working status, chronic conditions, smoking, physical activity, alcohol consumption, and mental health. Gallup invited 8,231 panel members to complete the RANDS web questionnaire. Only panel members who could complete the questionnaire by web were included; panel members who typically responded using telephone or mail modes were not sampled. Sampling strata were assigned by race, ethnicity, age group, and education level. The data collection began on March 16, 2016 and ended on April 13, 2016. Of the members contacted, 2,480 completed the survey, resulting in a 30.1 percent response rate. Gallup provided post-stratified weights which were raked to match characteristics of the U.S. population aged 18 and older from the Current Population Survey by age, race, ethnicity, sex, education level, and geographic region. The post-stratification steps were performed iteratively to converge to the target population proportions. Extreme weights were trimmed and provided as final weights for estimation (National Center for Health Statistics 2020).

NHIS

The NHIS is a cross-sectional health survey used to monitor health trends in the United States for the civilian noninstitutionalized population and to track progress toward achieving national health objectives. The survey has been conducted continuously since 1957 and collects data on a broad range of health topics annually including health status, health care access, and health care utilization. From 1997 to 2018, the NHIS had four main components including the Household Composition, Family, Sample Adult, and Sample Child. The survey content includes core questions as well as supplemental sections sponsored by other agencies. The 2016 NHIS contained supplementary questions on topics including health care access and utilization, functioning and disability, food security, mental health, balance, immunization,

vision, blood donation, chronic pain, Crohn's Disease, diabetes, e-cigarettes and use of tobacco products, hepatitis B/C screening, internet access and email usage, and heart disease and stroke prevention (National Center for Health Statistics 2017). We focus on a subset of questions that overlap with round 2 of RANDS.

Survey administration is conducted by the U.S. Census Bureau under contract with NCHS through in-person household interviews. Face-to-face interviews are conducted in the respondent's home, although follow up interviews may be conducted over the phone. The household response rate for the 2016 NHIS was 67.9 percent, with an 80.9 percent response rate of adults in responding households (54.3 percent unconditional response rate of sample adults). There were 33,028 adult NHIS respondents in 2016 selected for comparison to RANDS. Of this sample, 8,227 responded in quarter one, 8,256 responded in quarter two, 8,351 responded in quarter three, and 8,194 responded in quarter four.

The NHIS sample weights are derived from the probability of selection at each sampling stage. The final weights were calibrated for nonresponse and post-stratified by age, sex, race, and ethnicity classes using 2010 census-based population estimates (National Center for Health Statistics 2017). Quarterly weights were calculated as a function of the final annual weights.

Methods

Comparison of Demographics

Seven demographic variables and two general health variables were selected to evaluate differences between the characteristics of the RANDS and NHIS respondents. The variables considered included age group (18-34; 35-54; 55-64; 65-74; 75 and over), sex (male; female), race and ethnicity (non-Hispanic white; non-Hispanic black; non-Hispanic Asian; non-Hispanic other; Hispanic), education level (less than high school or GED; high school graduate; associate degree or some college; bachelor's or higher degree), family income (<$50,000$;50,000-$99,999$;>=100,000), geographic region (Northeast; Midwest, South; West), marital status (married/living with partner; single/never married; separated/divorced/widowed), self-rated health status (excellent; very good; good; fair; poor), and body mass index (BMI) category (underweight, BMI<18.5; normal, 18.5<= BMI <25; overweight, (25<= BMI <30; obese, BMI>=30) which was determined using reported height and reported weight from both surveys. The RANDS had item nonresponse for race and ethnicity, family income, geographic region, marital status, self-rated health status, and BMI category (due to missing values in reported height and weight). The NHIS had item nonresponse for education level, family income, marital status, self-rated health status, and BMI category (due to missing values in reported height and weight). Item nonresponse for family income was used to directly compare missingness in RANDS, however, National Health Interview Survey imputed income files are available online to address item nonresponse.

Estimates for RANDS and NHIS were obtained using SAS PROC SURVEYFREQ, which incorporates the survey weights and sample design into the estimation procedure. A sample design with stratification and clustering was specified. All estimates in this paper meet the NCHS standards of reliability (Parker et al. 2017). Differences in the observed demographics and general health covariates between RANDS and each of the NHIS datasets (2016 quarter 1, 2016 quarter 2, 2016 quarter 3, 2016 quarter 4, and the full year of 2016 data) were evaluated using the Rao-Scott chi-square test which is survey design adjusted (Rao and Scott 1981, 1984, 1987). Differences across the four NHIS quarters are also assessed using the Rao-Scott chi-square test, although the demographic distribution of the full year of NHIS 2016 data is not statistically compared to the quarterly subsets due to the overlapping sample.

Comparison of Health Outcomes

To evaluate the use of RANDS and web surveys to estimate major health measures, we consider six health outcomes related to smoking status, food security, health insurance, hypertension, asthma and diabetes. The six survey questions included on both RANDS and NHIS associated with these outcomes (with the possible categorical responses) are:

1. Have you smoked at least 100 cigarettes in your entire life? How often do you now smoke cigarettes? (Current smoker, former smoker, never smoker)
2. I worried whether my food would run out before I got money to buy more. (Often true, sometimes true, never true)
3. Do you have any of the following kinds of health insurance or health care coverage? Private health insurance, Medicare, Medi-Gap, Medicaid, SCHIP (CHIP/Children's Health Insurance Program), Military health care (TRICARE/VA/CHAMP-VA), Indian Health Service, State-sponsored health plan, Other government program, Single service plan (e.g., dental, vision, prescriptions)? (Yes, no)
4. Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure? (Yes, no)
5. Have you ever been told by a doctor or other health professional that you had asthma? (Yes, no)
6. Have you ever been told by a doctor or other health professional that you have diabetes or sugar diabetes? (Yes, no)

Differences between the RANDS estimates and the NHIS estimates were assessed using the Rao-Scott chi-square test. Propensity score adjustments were applied to the health outcomes estimates to calibrate the RANDS estimates to the NHIS. In this study, we consider various reference samples for the calibration, including the NHIS 2016 data and each of the subsets of quarterly data from the 2016 NHIS.

Propensity Score Adjustment

Propensity score weighting is a statistical method that has been used for calibrating survey weights to a reference survey. This approach is similar to post-stratification as it balances the covariates included in the propensity score model to the covariate distribution of a reference survey. For models that contain all potential confounders, the propensity score adjustment on the survey weights produces unbiased estimates of the treatment effect that are generalizable to the target population of the reference survey (Lee 2006, DuGoff, Schuler, and Stuart 2014). The probability of inclusion in the survey of interest (e.g., RANDS) is modeled using logistic regression on common covariates from the two surveys. This propensity score model is used for prediction of the estimated probability for adjusting the survey weights, although statistical tests on the model parameters can be used to evaluate significant associations between covariates and the odds of responding to the survey of interest. The inverse of propensity weighting method (Valliant and Dever 2011) is often utilized to adjust the survey weight to the target population represented by the benchmark survey through the adjustment factor $\frac{1-\hat{p}}{\hat{p}}$. Final propensity score adjusted weights are obtained by multiplying the original survey weight by the propensity adjustment factor.

For the RANDS data, the propensity score model was formed using the demographic covariates and general health variables (age group, sex, race and ethnicity, education level, family income, geographic region, marital status, self-rated health status, and BMI category) to adjust for any differences identified between RANDS and NHIS. Item nonresponse was treated as a separate category in the estimation procedure to account for differences in missing values. Prior to fitting the propensity score model, the NHIS weights were normalized to the sample size of the survey. A first order logistic model was formed

using the nine demographic and health covariates to estimate the propensity of participating in RANDS compared to NHIS. The reference category for each categorical variable was selected as the last nonmissing category shown in Table A1 (see Appendix). The RANDS weights were multiplied by the propensity adjustment factor $\frac{1-\hat{p}}{\hat{p}}$ to produce the pseudo sampling weights for calculating calibrated RANDS mean and variance estimates for each of the health outcomes. This process was repeated to calibrate RANDS to each of the NHIS datasets for comparison.

Comparison of Calibrated Estimates

The calibrated RANDS outcome estimates were not statistically compared to NHIS estimates as the calibrated dataset and the NHIS are not independent. However, differences between the propensity score calibrated RANDS estimates over the five reference surveys were statistically evaluated through an analysis of variance (ANOVA) of the propensity adjustment factors. Since the propensity score calibrated estimates are a function of the propensity adjustment factor, reference surveys that produce similar adjustment factors should be expected to result in similar calibrated estimates. Thus significant differences between the adjustment factors derived from the propensity score models indicate differences in the calibrated RANDS estimates due to the various reference surveys (NHIS 2016 quarter 1, NHIS 2016 quarter 2, NHIS 2016 quarter 3, NHIS 2016 quarter 4, and the full year of NHIS 2016 data) while a lack of statistically significant differences indicate that the reference surveys produced consistent calibrated RANDS estimates for the six health outcomes within the detection limits of the test.

Results

Comparison of Demographics

Table A1 reports the weighted estimates for demographic and general health covariates in RANDS and NHIS. For NHIS, the estimates are reported for surveys collected during each quarter of 2016 as well as an overall estimate from all surveys collected during 2016. The count of item nonresponse is also recorded in Table A1. While family income and marital status were missing in both surveys, the two demographic variables were missing more often in RANDS.

The distribution of demographic and health variables in the NHIS full year and quarterly data are very similar. The weighted estimates among the NHIS quarterly subsets are consistent with the estimates from the full year of NHIS data for most of the demographic variables, including age group, sex, race and ethnicity, and geographic region. More differences were seen between the NHIS full year data and the quarterly data for education level, family income, marital status, self-rated health status, and BMI category. Demographic variables across the four NHIS quarters were compared using the Rao-Scott chi-square test. This test confirmed that the weighted estimates for the demographic variables were consistent across the NHIS quarters, except for education level which differed significantly at the 5 percent significance level. Post hoc Rao-Scott chi-square tests with a Bonferroni correction did not identify statistically significant differences in education level between specific quarters as the Bonferroni correction is conservative, although we observe that quarters 1 and 2 had larger percent estimates for persons with less than high school education while quarter 4 had a larger percent estimate for persons with an associate degree or some college.

Rao-Scott chi-square tests were used to compare the demographic and general health variables from RANDS to each of the five NHIS datasets (four quarters and the full year data). Significant differences in the observed samples were identified between RANDS and each NHIS dataset for all variables except sex

and geographic region. We observe that RANDS reported a weighted estimate of 26.9 percent for the 18-34 age group and 5.0 percent for the 75 and over age group, while NHIS reported estimates ranging from 29.9 percent to 30.0 percent for the 18-34 age group and 7.9 percent to 8.0 percent for the 75 and over age group. In addition, RANDS reported an estimate of 73.2 percent non-Hispanic white adults while the NHIS datasets reported estimates ranging from 64.7 to 65.3 percent. For education level, RANDS reported a weighted estimate of 2.2 percent for less than high school or GED, while NHIS estimates ranged from 13.9 percent to 16.3 percent. The RANDS estimate for "Excellent" self-rated health status was 12.8 percent while the NHIS estimates ranged from 26.9 percent to 28.1 percent and the RANDS estimate for obesity (BMI category=obese) was 37.2 percent while NHIS estimates ranged from 29.7 percent to 30.6 percent.

Comparison of Health Outcomes

The unadjusted health outcome estimates in RANDS and NHIS are reported in Table A2 (see Appendix). The estimates of the observed health outcomes significantly differ between the two surveys for smoking status, food security, hypertension, and asthma. RANDS tends to underestimate the proportion of the population who have never smoked and overestimates the proportion that has had concerns about their food running out before they had money to buy more (food security responses "often true" or "sometimes true") compared to NHIS. Moreover, the estimates produced from RANDS tended to overestimate hypertension prevalence and asthma prevalence in the population compared to NHIS. Health insurance estimates and diabetes prevalence estimates were not found to significantly differ between RANDS and any of the five NHIS datasets.

Both RANDS and NHIS are missing reported health outcomes, although RANDS is missing a higher percentage of responses. The percent of missing health outcomes in RANDS ranged from 0.69 percent (health insurance and diabetes) to 1.41 percent (smoking status). Missing data patterns were fairly consistent between the full year of NHIS data and the quarterly subsets. Health insurance had the highest percent of missing responses for all five NHIS datasets (0.41 percent, 0.47 percent, 0.38 percent, 0.56 percent, and 0.46 percent for NHIS Q1, Q2, Q3, Q4, and full year, respectively). NHIS quarter 2, quarter 3, and the full year data had the lowest percent missingness in the outcome food security (0.01 percent, 0.06 percent, and 0.05 percent, respectively) while quarter 4 was missing asthma for only 0.02 percent of records. NHIS quarter 1 had the lowest percent of missing responses for the outcomes food security and asthma (0.06 percent for each).

Propensity Score Adjustment

Most demographic and health variables included in the propensity score models were identified as being significantly associated with the probability of inclusion in RANDS except sex which was not significant in any of the five propensity score models and BMI category which was not significant in four of the five propensity score models (mildly significant in the model using NHIS 2016 Q1 as a reference dataset). The parameter estimates for missing values for some of the covariates were relatively larger than other estimates due to the small number of missing values. Fit statistics indicated covariates improved model fit compared to the intercept model (not shown). The fit of each propensity score model varied slightly by reference dataset, although the fit for the propensity score models for the quarterly reference datasets were comparable (Table A3 in the Appendix).

Comparison of Calibrated Estimates

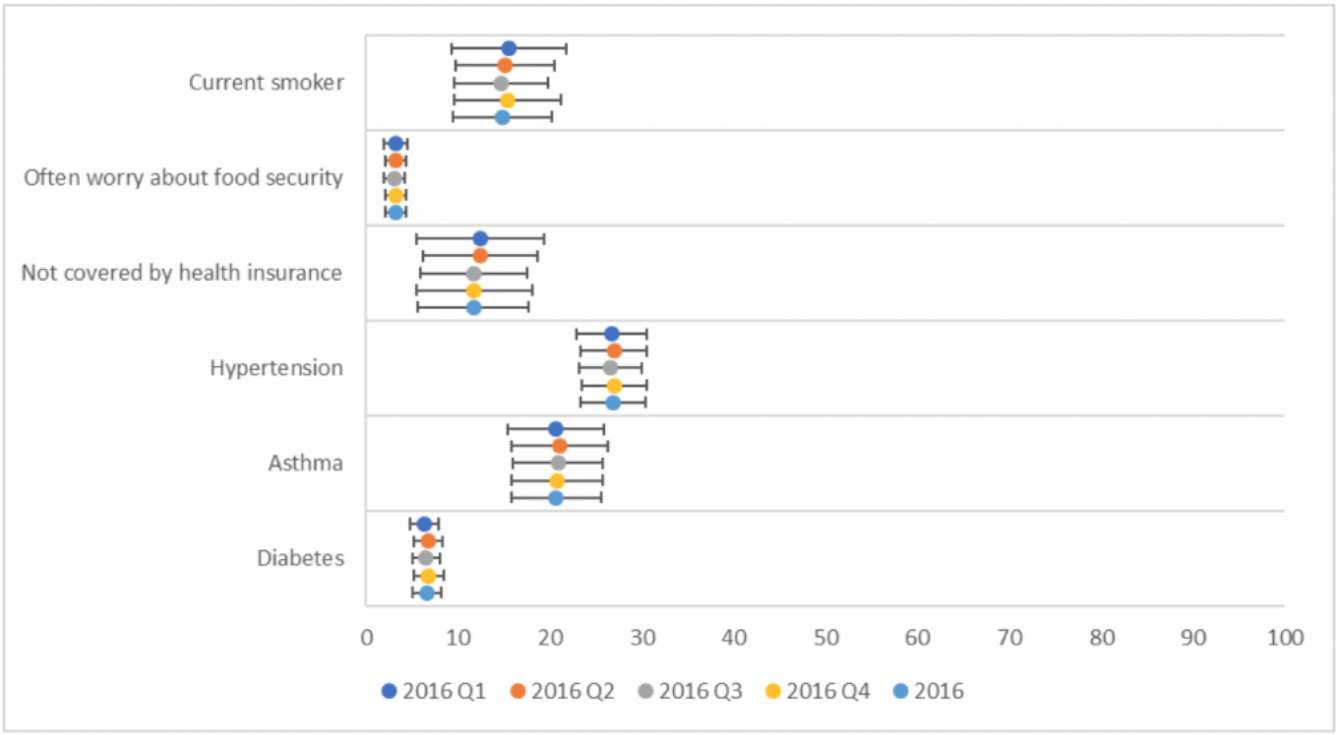
The calibrated RANDS estimates which use the full year of 2016 NHIS data as well as the quarter 1, quarter 2, quarter 3, and quarter 4 data from the 2016 NHIS as reference datasets are reported in Table 1. The extent of the calibration of the RANDS estimates compared to the NHIS health outcome estimates (reported in Table A2) varied. The propensity score weighting resulted in improved web estimates for smoking status that more closely reflected the NHIS estimates. Although the propensity weighting decreased the estimates for health insurance coverage, hypertension prevalence, and diabetes prevalence, the adjusted estimates were further from the NHIS estimates than the unadjusted estimates. While the RANDS estimates differed from the NHIS, this decrease due to weighting reflects the impact of calibrating to the NHIS as these outcomes were previously overestimated in RANDS. In the case of measures for food security and asthma, calibration to the reference datasets did not improve the RANDS estimates relative to the NHIS. The varied performance of the calibrated estimates suggests that additional research could be performed to improve propensity score weighting methods for web survey calibration of some outcomes.

Table 1: Propensity score adjusted RANDS health outcome estimates

	Unadjusted		Reference Dataset for Adjusted Estimates									
			NHIS 2016 Q1		NHIS 2016 Q2		NHIS 2016 Q3		NHIS 2016 Q4		NHIS 2016	
Variable	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
Smoking Status												
Current	15.2	1.12	15.5	3.19	15.1	2.76	14.7	2.61	15.4	2.97	14.9	2.75
Former	30.2	1.43	26.3	2.06	26.3	1.96	26.1	1.86	26.3	1.90	26.3	1.90
Never	54.6	1.55	58.1	3.05	58.6	2.82	59.2	2.67	58.3	2.84	58.8	2.75
Food Security												
Often true	4.3	0.63	3.2	0.62	3.2	0.59	3.1	0.56	3.2	0.58	3.2	0.59
Sometimes true	14.6	1.10	16.0	3.30	15.7	2.94	15.2	2.76	15.6	2.99	15.5	2.87
Never true	81.1	1.19	80.8	3.26	81.1	2.92	81.7	2.74	81.2	2.96	81.3	2.85
Health Insurance												
Yes	92.5	0.85	87.6	3.52	87.6	3.18	88.3	2.96	88.3	3.20	88.3	3.06
No	7.5	0.85	12.4	3.52	12.4	3.18	11.7	2.96	11.7	3.20	11.7	3.06
Hypertension												
Yes	35.2	1.42	26.7	1.92	27.0	1.84	26.5	1.74	27.0	1.79	26.9	1.79
No	64.8	1.42	73.3	1.92	73.0	1.84	73.5	1.74	73.0	1.79	73.1	1.79
Asthma												
Yes	19.2	1.28	20.7	2.67	21.1	2.68	20.9	2.47	20.8	2.53	20.7	2.48
No	80.8	1.28	79.3	2.67	78.9	2.68	79.1	2.47	79.2	2.53	79.3	2.48
Diabetes												
Yes	10.3	0.95	6.3	0.77	6.7	0.79	6.5	0.76	6.8	0.84	6.6	0.79
No	89.7	0.95	93.7	0.77	93.3	0.79	93.5	0.76	93.2	0.84	93.4	0.79
Note: Est stands for estimate, SE stands for standard error, estimates and standard errors are presented as percentages (%)												

However, while the calibrated health outcome estimates in RANDS differed from the NHIS estimates, the estimates across the five reference datasets were similar. Figure 1 displays the calibrated estimates for the health outcomes using each calibration dataset and the corresponding 95 percent confidence intervals. Estimates for all six health outcomes were consistent across the five NHIS datasets. Using the full year versus a single quarter of NHIS data did not greatly impact the calibrated RANDS estimates. In addition, estimates using reference datasets collected over different time periods than RANDS (i.e. quarters 1, 3, and 4) were similar. The standard errors were larger for the adjusted estimates compared to the unadjusted RANDS data, although the standard errors tended to be slightly lower when using the full year of NHIS 2016 data as the calibration dataset.

Figure 1: Calibrated RANDS estimates and 95% confidence intervals by reference dataset



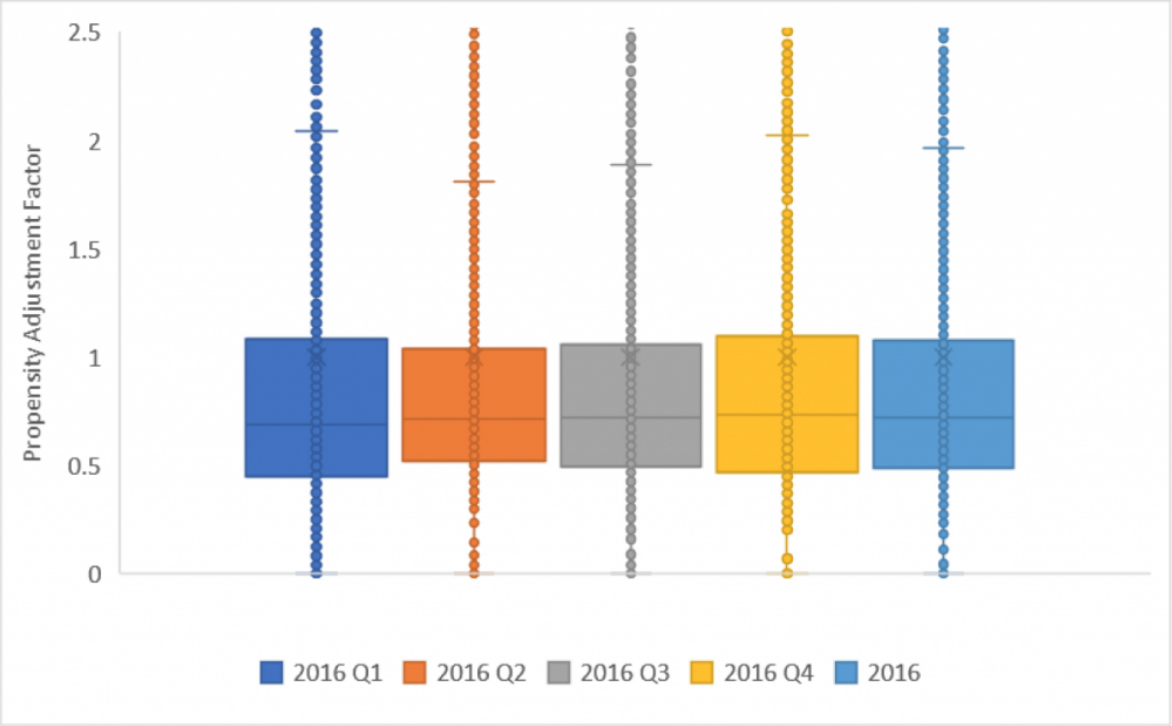
To further evaluate the effect of the reference dataset on propensity score weighting, we compare the propensity adjustment factors $\frac{1-p}{p}$ across the five reference datasets. The adjustment factors were normalized to the RANDS sample size (n=2,480) for comparison. The mean, standard deviation, minimum, maximum, and quartiles for each of the normalized propensity adjustment factors are reported in Table 2.

Table 2 : Descriptive statistics of normalized propensity adjustment factors ((1-p)/p) by reference dataset

Reference Data	Mean	Standard Deviation	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
NHIS 2016 Q1	1.00	1.75	5.57E-10	0.45	0.69	1.09	44.45
NHIS 2016 Q2	1.00	1.65	9.73E-10	0.52	0.72	1.04	41.37
NHIS 2016 Q3	1.00	1.54	1.53E-10	0.50	0.72	1.05	34.39
NHIS 2016 Q4	1.00	1.51	1.05E-10	0.47	0.73	1.10	35.75
NHIS 2016	1.00	1.54	1.12E-12	0.48	0.72	1.08	34.76

Statistical testing of the propensity adjustment factors using ANOVA indicate that the adjustment factors from each of the reference datasets do not significantly differ despite differences in the size and collection time period of the reference datasets. Figure 2 displays box plots for the adjustment factors ranging between 0 and 2.5 by reference dataset (full range of propensity adjustment factors is described in Table 2). The plot shows that the overall spread of the propensity adjustment factors is similar across all five reference datasets, which similarly suggests that the propensity score calibrated estimates are consistent across the reference datasets.

Figure 2: Boxplot of normalized propensity adjustment factors ((1-p)/p) by reference dataset



Discussion

While calibration techniques have been developed and used to improve survey estimates, research related to the selection of a calibration dataset has been limited. In this paper we consider propensity score adjusted estimates from the 2016 RANDS, a probability-based panel survey conducted by web, for five calibration datasets, including the NHIS data collected over the full year of 2016 and quarterly subsets of the 2016 NHIS data. Through a comparison of the calibrated estimates and statistical testing of the propensity adjustment factors $\frac{1-p}{p}$ using ANOVA, it is demonstrated that the estimates among the five reference datasets do not vary significantly. Although the Rao-Scott chi-square test identified statistically significant differences between estimates of education level between the four NHIS quarters, most demographic and health variable estimates were consistent across the reference datasets. Since the propensity adjustment factor is a function of the weighted marginal totals of the reference survey, the similar covariate distributions may have resulted in similar propensity adjustment factors and hence comparable calibrated estimates. This finding suggests flexibility in the selection of the reference dataset, under the condition that the selected reference dataset reflects the target population of interest. In this case the covariate distributions for the full year of NHIS as well as NHIS subsets by quarter were well-matched and comparisons demonstrated no significant differences between the propensity weighted estimates. Although the calibrated estimates of the health outcomes were similar, the standard errors of the calibrated estimates using the full year of NHIS 2016 as the reference dataset were smaller than the standard error estimates using the quarterly reference datasets which indicates that the reference dataset may impact the calibrated standard error estimates.

It is important to note that the NHIS is designed to have a similar demographic distribution in each quarter and to be a representative sample by quarter. However, while the results of this study may be dependent on the survey design of the NHIS, the findings may have implications for other surveys. This comparison indicates that a smaller reference dataset may be used for calibration, such as quarterly data rather than annual data from a national health survey, depending on the survey design. Based on the results from the RANDS study, panel survey estimates could be calibrated with data from NHIS as it became available rather than waiting for the entire year of data to be collected. The study results also

suggest that for the health outcomes investigated, the reference data did not necessarily need to be collected over the same time period as the web survey. For the RANDS comparison, estimates using NHIS data from quarters 1, 3, 4, and the full year of data were consistent with the health estimates produced using the quarter 2 subset of NHIS. Although one may prefer to select a reference dataset that overlaps the collection time frame of the panel survey, this finding is important for situations in which the collection time frames do not overlap or in situations where it is beneficial to use data from shorter (such as quarter vs. year) or prior time periods to expedite the production of panel survey estimates without waiting for upcoming data releases. While future studies should compare estimates using alternative calibration datasets to identify the importance of the survey design, this finding has many practical implications for the use of web surveys to produce national level estimates.

There are limitations in this study which should be evaluated before selecting a reference dataset for calibration to web surveys. As indicated previously, the estimates for the demographic and general health covariates were similar between the full year and quarterly NHIS subsets and statistical testing indicated that the quarterly estimates were consistent for all variables except education level. This is a feature of the survey design of the NHIS, and alternative calibration datasets which are not representative samples or that represent very different populations than the web survey may not result in similar estimates. In addition, calibration datasets that represent populations that change over time could impact the calibrated estimates. In this analysis, the demographic and health variables used in the calibration were time invariant and thus calibrated outcomes did not reflect variation in the estimates over time. Moreover, the NHIS datasets ranged in size from approximately 8,200 respondents per quarter to more than 33,000 respondents over the full year. These reference datasets are much larger than the RANDS survey, with a sample size in quarter 4 that was 3.3 times larger than the sample size of RANDS. Studies investigating the impact of small sample sizes in reference datasets may find that smaller reference surveys do not have the same level of estimation accuracy.

Appendix

Table A1: Weighted estimates of demographic and general health covariates
Table A2: Unadjusted RANDS and NHIS health outcome estimates
Table A3: Propensity score model estimates

References

1. Callegaro, M., Manfreda, K.L., and Vehovar, V. (2015). Web Survey Methodology. London: Sage.
2. Cornesse, C., Blom, A.G., Dutwin, D., et al. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, 8(1), 4-36.
3. Drake, C. (1993). Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics*, *49*, 1231-1236.

4. DuGoff, E.H., Schuler, M., Stuart, E.A. (2014). Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys. *Health Services Research* 49(1), 284-303.
5. Groves, RM. (1989). *Survey errors and survey costs*. New York: Wiley.
6. How Does the Gallup Panel Work? (2019). https://www.gallup.com/174158/gallup-panel-methodology.aspx. Accessed 22 October 2019.
7. Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics*, *22(2)*, 329-349.
8. Lee, S. & Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. Sociological *Methods & Research*, *37(3)*, 319-343.
9. National Center for Health Statistics. (2017). Survey Description, National Health Interview Survey, 2016. Hyattsville, Maryland.
10. National Center for Health Statistics. (2020). RANDS 2 Technical Documentation. Hyattsville, Maryland.
11. Parker, J.D., Talih, M., Malec, D.J., et al. (2017). National Center for Health Statistics Data Presentation Standards for Proportions. National Center for Health Statistics. *Vital Health Stat* 2(175).
12. Rao, J.N.K. & Scott, A.J. (1981). The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association*, *76*, 221–230.
13. Rao, J.N.K. & Scott, A.J. (1984). On Chi-Squared Tests for Multiway Contingency Tables with Cell Properties Estimated from Survey Data. *Annals of Statistics*, *12*, 46–60.
14. Rao, J.N.K. & Scott, A.J. (1987). On Simple Adjustments to Chi-Square Tests with Survey Data. *Annals of Statistics*, *15*, 385–397.
15. Rubin, D.B. & Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*, *52*, 254-268.
16. Ryan, C. (2017). Computer and Internet Use in the United States: 2016. *American Community Survey Reports*, ACS-39, US Census Bureau, Washington, DC.
17. Scanlon, P. (2017). Cognitive Evaluation of the 2015-2016 National Center for Health Statistics' Research and Development Survey, National Center for Health Statistics, Hyattsville, MD, https://wwwn.cdc.gov/qbank/report/Scanlon_2017_NCHS_RANDS.pdf.
18. Taylor, H. (2000). Does Internet Research Work? Comparing Online Survey Result with Telephone Survey. *International Journal of Market Research*, *42*, 58-63.
19. Terhanian, G. & Bremer, J. (2000). Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research. White paper: Harris Interactive.
20. Valliant, R. & Dever, J.A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research*, *40*, 105-137.