Application of State Health Status In the Design of the 2011-2014 National Health and Nutrition Examination Survey

Jay Clark[1], Leyla Mohadjer[1], Te-Ching Chen[2]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850
[2]National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782

Abstract
Traditionally in national surveys of the United States, census region is used to stratify the frame and ensure that regions are covered in the sample. In this paper, we focus on a health survey that applies a state-level healthiness index instead of regions (or other geographic identifiers) for initial stratification purposes. The National Health and Nutrition Examination Survey (NHANES) is an ongoing multistage annual survey with first-stage primary sampling units (PSUs) selected for four years at a time. NHANES collects thousands of health-related variables for each sampled person from an interview, physical examination, and laboratory tests. In the 2011-2014 four-year sample design, NHANES initially stratified the PSU frame by a state-level healthiness index, resulting in PSUs within the same strata (with the same general levels of healthiness) coming from different regions. This paper reviews this stratification process and examines the improvement in the precision of the estimates.

Keywords: stratification; health survey; sample design

1. Introduction to NHANES

The National Health and Nutrition Examination Survey (NHANES) is an ongoing survey of the noninstitutionalized civilian population of the United States conducted by the National Center for Health Statistics (NCHS). A unique feature of this survey is the collection of medical examination data for a nationally representative sample of the U.S. population. The information collected in the NHANES surveys is essential for estimating the prevalence of various diseases and conditions, explaining the mechanisms of disease development, and planning for health policy.

A primary purpose of NHANES is to produce a broad range of descriptive health and nutrition statistics for sex, race, ethnic, and age subdomains of the population. These data can then be used to measure and monitor the health and nutritional status of the noninstitutionalized population. The analytic goals of NHANES are as follows (Johnson, et al., 2014):

- Estimate the number and percentage of persons in the U.S. population and in designated subgroups with selected diseases and risk factors;
- Monitor trends in the prevalence, awareness, treatment, and control of selected diseases;
- Monitor trends in risk behaviors and environmental exposures;
- Study the relationship among diet, nutrition, and health;

- Explore emerging public health issues and new technologies; and
- Provide baseline health characteristics that can be linked to mortality data from the National Death Index or other administrative records (e.g., enrollment and claims data from the Centers for Medicare and Medicaid Services).

The sample design for NHANES involves four stages (Johnson, et al., 2014). The first stage is the selection of primary sampling units (PSUs). The PSUs are mostly single U.S. counties; in a few cases, adjacent counties are combined to keep PSUs above a minimum size. The second stage is the selection of segments (typically groups of adjacent census blocks) within sampled PSUs. Both PSUs and segments are sampled with probabilities proportional to size (PPS), where the measure of size (MOS) is the noninstitutionalized civilian population, weighted by race and Hispanic origin and other factors to ensure analytic domain targets are attained. At the third stage of selection, dwelling units (DUs) within each sampled segment are sampled. The final stage of selection is the sampling of eligible persons (called sampled persons, or SPs) within sampled DUs.

SPs are interviewed at home and appointed to a physical examination at a mobile exam center, which travels the country from PSU to PSU. The comprehensive medical exams can take up to four hours. Due to cost and operational constraints for the study, the four-year sample design is limited to 20,000 total examined persons in 60 PSUs (about 5,000 examined persons in each annual sample of 15 PSUs).

To stratify the PSUs for the four-year sample, after self-representing (SR) PSUs are removed, the frame of non-self-representing (NSR) PSUs is first divided into a number of major strata of approximately equal size, and within each major stratum, four minor strata are created, one for each year in the sample. The PSU stratification scheme should allow for efficient annual estimates without compromising the efficiency of the four-year estimates. The four-year sample has a one-PSU-per-minor-stratum design; each annual sample has a one-PSU-per-major-stratum design (Johnson, et al., 2014). With this method, annual samples contain only one PSU from each major stratum, and multiyear samples contain only one PSU per sampled minor stratum. When forming the PSU strata, both major and minor strata are constructed in such a way that they are as equal in total MOS as possible. Equalizing strata size can help to equalize the sample sizes across PSUs to improve operational efficiency and minimize weight variation.

In the 2011-2014 NHANES sample design, the initial stratification of the NSR PSU frame has used a state-level healthiness index, resulting in PSUs within the same strata (with the same general levels of healthiness) coming from different regions. This paper reviews this stratification process and examines whether sampling variability was reduced in 2011-2014 datasets by the new stratification method.

2. Stratification

2.1 NHANES Sample Designs 1999-2010

From 1999 to 2010, NHANES sample designs allowed for stratification to ensure that the PSUs making up the annual and multiyear samples were distributed evenly in terms of geography, urban-rural distribution, and certain population characteristics. This was especially important for this survey since only 15 PSUs are able to be fielded in each annual sample.

The NHANES 2007-2010 sample design resulted in 8 SR PSUs and required 52 NSR PSUs to be sampled. The NSR PSUs were stratified initially by census region, as in Figure 1 (Curtin, et al., 2013). Then, thirteen major strata were created using region and whether the PSU was part of a Metropolitan Statistical Area (MSA). One stratum included non-MSA PSUs from three different census regions (the Northeast, Midwest, and West regions), but the other twelve strata were all within one region. Section 2.4 compares this sample design with later NHANES sample designs.

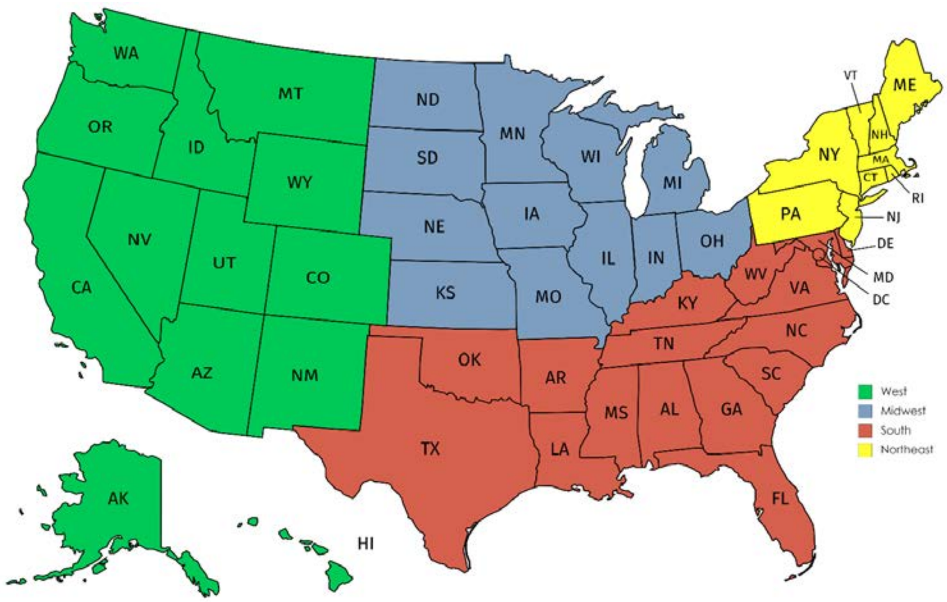

Figure 1: Four major "regions" according to the U.S. Census Bureau.

2.2 NHANES Sample Design 2011-2014

Beginning in 2011, NHANES sample designs added general health level as a component to the PSU stratification. General stratification theory suggests that stratifying by an outcome variable, or a variable highly correlated with the outcome variable, provides the greatest increase in precision (Cochran, 1977, p. 127). For NHANES, however, a compromise must be used since the survey has thousands of outcome variables with varying amounts of correlation, and other features of the design (distribution by geography and by size, for example) must be maintained.

To try to improve the precision for health-related estimates, the strata in the 2011-2014 design were initially formed based on state groups that were relatively homogeneous in terms of a derived health status (Hao, et al., 2011). To determine this status, a state-level data set was created for the 50 states and the District of Columbia using public data available from the National Vital Statistics System (NVSS) and the Behavioral Risk Factor Surveillance System (BRFSS), both conducted by the Centers for Disease Control and Prevention (CDC). For each state, six health-related variables were obtained. The infant mortality rate per 1,000 live births (2002-2004) and the age-adjusted death rate per 100,000 population (2003-2005) were obtained from NVSS. The death rate was adjusted for age so that states with older populations would not inherently be considered less healthy due to deaths from natural causes. The percentage of adults with self-reported high blood pressure (2001), the percentage of adults with self-reported overweight status or obesity (2002), the percentage of adults who reported eating fewer than five servings of fruits and vegetables per day (2002), and the percentage of adults who reported smoking (2002) were obtained from BRFSS.

Groups of states were formed using a cluster analysis, a method of identifying homogeneous groups of cases, in this case states, based on selected characteristics. A hierarchical cluster analysis was performed, with the within-group linkage used as the clustering method and the squared Euclidean distance used as the distance measure (Krenzke and Haung, 2009). To eliminate the effect of different units of measurement of variables, standardized measures were used. The sum of the standardized measures across the six health-related variables was considered an overall health score for the state.

The overall health scores and the individual values for the six health-related variables were reviewed to check that the clustering worked appropriately. For example, with five clusters, the District of Columbia was put into a cluster by itself, and with six clusters, Utah and the District of Columbia each were put into separate clusters. Thus, four clusters, with several states each, was considered optimal. Each of the 51 state-level units was assigned to a state group, where the first group consisted of states with the highest health level, and the fourth group included those with the lowest health level.

As with the 2007-2010 sample design, the 2011-2014 sample design required 52 NSR PSUs to be sampled, which needed to be split into thirteen major strata, and those strata needed to be approximately equal in MOS. To make the strata more similar in size, a few states were moved from one state group to the next state group. The goal was to create groups of states that had more similar levels of health than states within each census region, without accounting for healthiness variability at the county level within states. After the final state health groups were determined, the weighted averages of the health-related variables by state group were calculated and are shown in Table 1. The state groups show different health levels with clear trends. For all six variables, lower numbers indicate better health, and the healthiest state group was on average the healthiest for each one.

Table 1: Selected Health Indicators* by State Group

State Group	*Infant mortality rate per 1,000 live births*	*Age-adjusted death rate per 100,000 population*	*Percent adults having high blood pressure*	*Percent adults overweight or obese*	*Percent adults with poor nutrition*	*Percent adults smoking*
Healthiest	5.5	735.3	24.3	56.9	73.4	19.3
2nd Healthiest	6.9	786.6	24.9	57.7	75.7	22.6
3rd Healthiest	7.4	840.6	26.6	60.8	76.4	24.4
4th Healthiest	8.5	943.7	28.4	60.5	78.2	26.2

* Weighted by civilian noninstitutional population estimates.

As part of the 2011-2014 design, California, one of the healthiest states, was separated, since there was interest in sub-national, multi-year estimates for large states (Johnson, et al., 2014). Out of thirteen major strata, four strata were assigned to the healthiest states, and of those, 1.5 strata (that is, six minor strata) covered California and the remaining 2.5

strata (or ten minor strata) covered the remaining states in the group. Similarly, three strata covered each of the other healthy state groups. Within the major strata, PSUs were further subdivided by region and urban-rural population distribution. Four minor strata were formed within each major stratum by race/ethnicity density and poverty levels.

The final state groups by general health for the 2011-2014 design are shown in Table 2 and Figure 2.

Table 2: State Grouping for NHANES 2011-2014 PSU Sample

State Group	*State*
Healthiest	CA, CT, HI, IA, MA, MN, ND, NH, NJ, NY, RI, UT, VT, WA
2nd Healthiest	AK, AZ, CO, FL, ID, IL, KS, ME, MT, NE, NM, OR, SD, VA, WI, WY
3rd Healthiest	DE, IN, MD, MI, OH, PA, TX
4th Healthiest	AL, AR, DC, GA, KY, LA, MO, MS, NC, NV, OK, SC, TN, WV

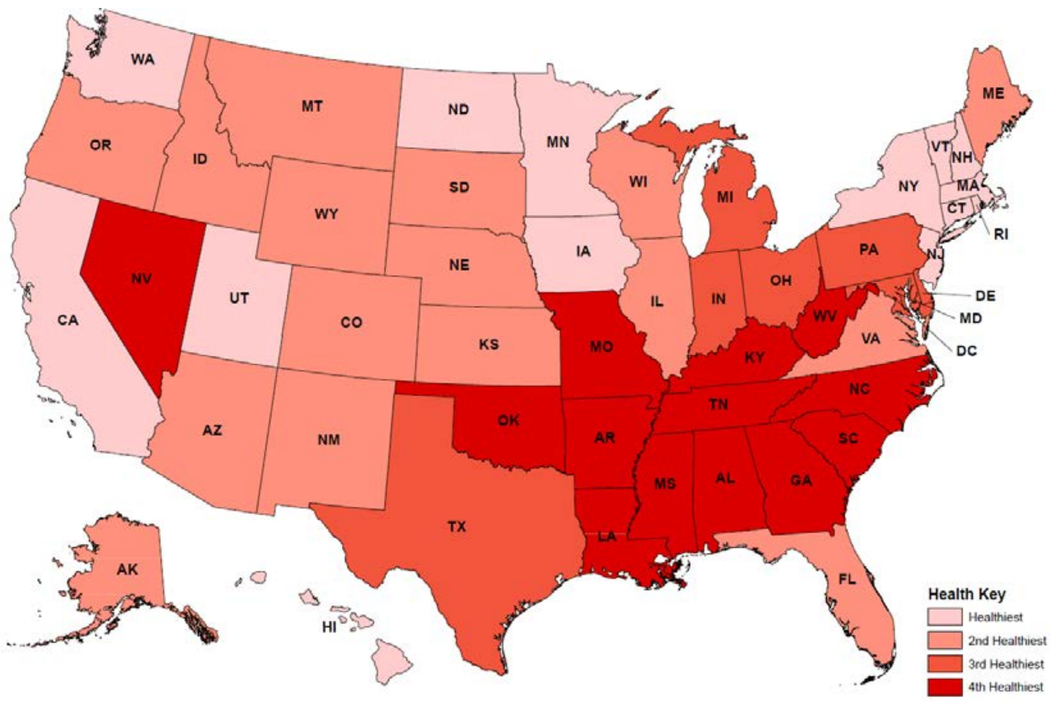

Figure 2: State grouping for NHANES 2011-2014 PSU sample.

Although some apparent correlation exists between these healthy state groups and census region (e.g., many states in the South region are also in the fourth-healthiest state group), quite a few differences are also easily found. The second-healthiest state group contains states from all four census regions, for example. Again, the intent for this grouping is not to assess the quality of health in any individual state, but to group states that have similar health characteristics.

2.3 Summary of NHANES Sample Designs 2007-2014

For comparison, Table 3 summarizes the strata definitions described previously for each of the NHANES sample designs in 2007-2014.

Table 3: Strata Definitions for Sample Designs 2007-2014

Sample Design	*SR PSUs*	*NSR PSUs*	*Major Strata*	*Minor Strata*
2007-2010	8	52	<u>13 total major strata:</u> 2 NE region, MSA 2 MW region, MSA 4 S region, MSA 3 W region, MSA 1 S region, non-MSA 1 NE/MW/W region, non-MSA	4 minor strata within each major stratum defined by: • race, Hispanic origin, and income level of PSUs
2011-2014	8	52	<u>13 total major strata:</u> 4 healthiest group 3 second healthiest group 3 third healthiest group 3 fourth healthiest group	4 minor strata within each major stratum defined by: • census region, • rural percentage, • race, Hispanic origin, and income level of PSUs

3. Analysis Results

NHANES interview and exam data from 2007-2008, 2009-2010, 2011-2012, and 2013-2014 were compared. The goal was to examine whether sampling variability was reduced in the later two datasets by the new stratification.

An unweighted analysis of variance (ANOVA), weighted ANOVA, and design effect approach were used to evaluate the new stratification design. Thirty-eight health-related outcome variables were reviewed in this analysis (see Appendix).

The relevant notation for the equations in this section is defined below.

N	=	Population Size (N is the sum of the PSU population for PSUs in the NSR frame);
M	=	Number of PSUs in NSR frame;
N_k	=	Population size in PSU *k*;
S_1	=	First Stage Sample: sample of PSUs;
S_2	=	Final Stage Sample: sample of people (includes all stages after PSU: segment, DU, person);
h	=	First Stage Stratum Index (h=1,2…H);
m_h	=	Number of PSUs in stratum *h*;
S_{1h}	=	Set of sampled PSUs in stratum *h* (2 PSUs in each S_{1h});
i	=	person index;
w_{1k}	=	First Stage Sampling Weights of PSU *k* (= 1/(PSU probability of selection));

w_{2i}	=	Final Stage Conditional Weights, which incorporate weight adjustments such as nonresponse adjustment, calibration and trimming (computed for example by division of final person weight by PSU weight);
$w_i = w_{1k} \times w_{2i}$	=	Unconditional person level sampling weights (final weight);
y_i	=	Study variable measurement for person i;
I_{ik}	=	Membership indicator for PSU k such that $I_{ik} = 1$ if person i belongs to PSU k and $I_{ik} = 0$ if not;
$\theta_k = N_k^{-1} \sum_{i=1}^{N_k} y_{ik}$	=	PSU k population average for variable y;
$\theta = M^{-1} \sum_{k=1}^{M} \theta_k$	=	NSR population average for variable y.

3.1 Unweighted ANOVA

The first comparison consists of an ANOVA for each study variable comparing the between-stratum variance to the within-stratum variance. Ideally, to have strata with smaller overall variance, the variance between PSUs within a stratum is small and the variance between strata is higher. Conducting the ANOVA without using PSU weights is equivalent to assuming simple random sampling within each stratum.

We computed the following overall F statistic $F = \frac{SSB/12}{SSW/13}$. With 13 major strata in the sample design, the statistic F will have an F distribution with 12 and 13 degrees of freedom under the null hypothesis that the underlying values of θ_k are equal for all k. The sum of squares between strata (SSB) and the sum of squares within strata (SSW) are calculated in the usual way, that is,

$$SSB = \sum_{h=1}^{H} m_h \left(\hat{\hat{\theta}}_h - \hat{\hat{\theta}}\right)^2 \text{ and}$$
$$SSW = \sum_{h=1}^{H} \sum_{k \in S_{1h}} (\hat{\theta}_k - \hat{\hat{\theta}}_h)^2 , \tag{1}$$

with

$$\hat{\hat{\theta}}_h = m_h^{-1} \sum_{k \in S_{1h}} \hat{\theta}_k,$$
$$\hat{\hat{\theta}} = m^{-1} \sum_{k \in S_1} \hat{\theta}_k,$$
$$\hat{\theta}_k = \hat{N}_k^{-1} \sum_{i \in S_2} w_i I_{ik} y_i, \text{ and} \tag{2}$$
$$\hat{N}_k = \sum_{i \in S_2} w_i I_{ik}.$$

For example, consider the outcome variable that is the indicator of diabetes. In 2007-2008, the F statistic of 1.181 has a p-value of 0.384, which is not significant. Similarly, the F statistics in 2009-2010 and 2011-2012 were not significant. The F statistic in 2013-2014 was 4.417, with a p-value of 0.006. So in only one of the four sets of data did the F statistic show that the stratification improved the precision. This result is affected by the small number of degrees of freedom (because of the small number of PSUs in this sample design), but provides no evidence that the precision improved in the later years. Figure 4 shows the p-values for the outcome variables indicating diabetes and total cholesterol.

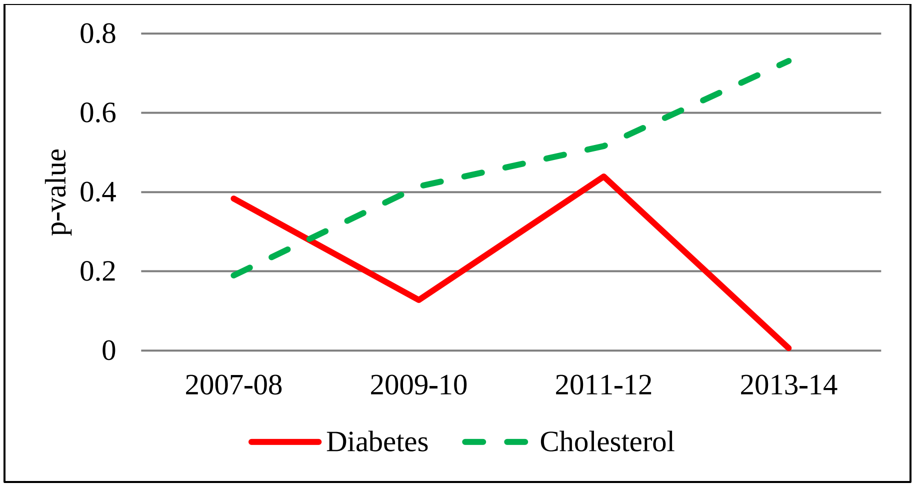

Figure 4: Unweighted ANOVA p-values for diabetes indicator and total cholesterol.

For the 2007-2008 sample, seven of the 38 tested variables show a p-value of 0.10 or less; for the 2009-2010 sample, nine variables do; for the 2011-2012 sample, only three variables do; and the 2013-2014 sample has 14 of these. The larger the F statistic is, the more the means vary among the strata. The smaller the p-value is, the more significant the differences. Again, a low number of degrees of freedom limited this analysis, but the result shows that for most of the variables, there was little evidence of improvement with the new stratification.

3.2 Weighted ANOVA

These ANOVAs were run using the PSU weights, w_{1k}, defined above. Using PSU weights in the ANOVA calculation includes the effect of the unequal sampling probabilities within each stratum. We compute the F statistic as before, $F = \frac{SSB/12}{SSW/13}$, with

$$\begin{aligned} SSW &= \sum_{h=1}^{H} \sum_{k \in S_{1h}} w_{1k} (\hat{\theta}_k - \hat{\hat{\theta}}_h)^2, \\ SSB &= \text{SST} - \text{SSW}, \text{ and} \\ SST &= \sum_{k \in S_1} w_{1k} (\hat{\theta}_k - \hat{\hat{\theta}})^2, \end{aligned} \tag{3}$$

with

$$\begin{aligned} \hat{\hat{\theta}}_h &= \widehat{M}_h^{-1} \sum_{k \in S_{1h}} w_{1k} \hat{\theta}_k, \\ \hat{\hat{\theta}} &= \widehat{M}^{-1} \sum_{k \in S_1} w_{1k} \hat{\theta}_k, \\ \hat{\theta}_k &= \widehat{N}_k^{-1} \sum_{i \in S_2} w_i I_{ik} y_i, \\ \widehat{N}_k &= \sum_{i \in S_2} w_i I_{ik}, \\ \widehat{M}_h &= \sum_{k \in S_{1h}} w_{1k}, \text{ and} \\ \widehat{M} &= \sum_{k \in S_1} w_{1k}. \end{aligned} \tag{4}$$

Again, as an example, consider the outcome variable that is the indicator of diabetes. In 2007-2008, the F statistic of 0.455 has a p-value of 0.909, which is not significant. However, the F statistics in 2009-2010, 2011-2012, and 2013-2014 were all significant, with p-values less than 0.03. As before, this result is affected by the small number of degrees of freedom in each dataset, but it provides no clear evidence that stratification using state health status improves the precision of the estimates. Figure 5 shows the p-values for the outcome variables indicating diabetes and total cholesterol. Total cholesterol had significant F statistics in 2009-2010 and 2011-2012, but not in the other two datasets.

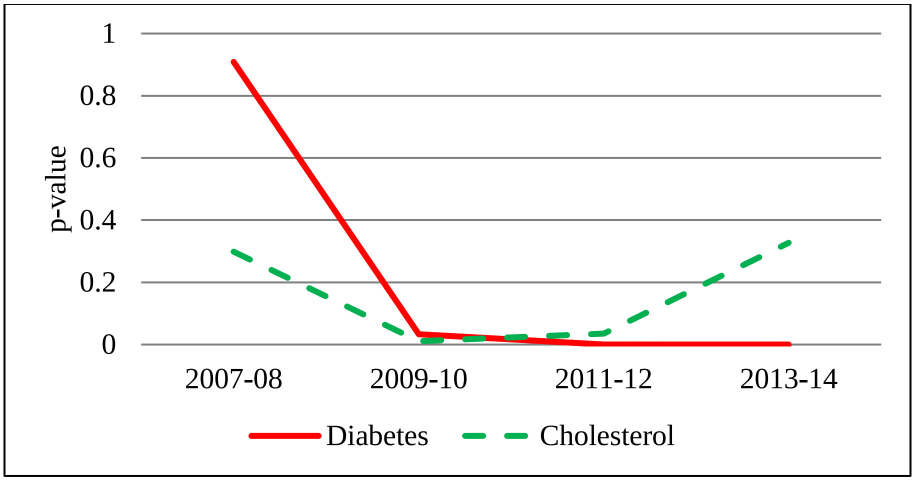

Figure 5: Weighted ANOVA p-values for diabetes indicator and total cholesterol.

For the 2007-2008 sample, 19 of the 38 tested variables show a p-value of 0.10 or less; for the 2009-2010 sample, 35 variables do; for the 2011-2012 sample, 22 variables do; and the 2013-2014 sample has 23 of these. Besides a small number of degrees of freedom, the limited number of PSUs also means that to some extent the F-value depends on how homogeneous the selected PSUs are. The result shows that although stratification in general improves the precision of the estimates, there was not enough evidence to show that the new stratification using healthiness had a large impact.

3.3 Design Effect Due to First Stage Stratification

The design effect due to the first stage stratification ($\widehat{DEFF}_{ST}$) is calculated as follows:

$$\widehat{DEFF}_{ST} = \frac{\hat{V}_{ST}(\hat{\theta})}{\hat{V}_{NST}(\hat{\theta})}, \text{ where} \quad (5)$$
$$\hat{\theta} = \sum_{i \in s_2} w_i y_i \,/\, \sum_{i \in s_2} w_i.$$

Here $\hat{V}_{ST}(\hat{\theta})$ denotes the estimated variance due to stratification, clustering, nonresponse adjustment, and calibration; and $\hat{V}_{NST}(\hat{\theta})$ denotes the estimated variance due to clustering, nonresponse adjustment, and calibration, but without consideration of stratification. Taylor series linearization was used to calculate both the numerator and denominator.

If the stratification had little effect on the variance, the first-stage design effect would be close to one. If the stratification decreased the variance, as hoped, then the first-stage design effect would be less than one. As before, the indicator for diabetes and the total cholesterol level are shown in Figure 6 by the first-stage design effect over the four tested datasets. For diabetes, the 2013-2014 design effect supports the newer stratification, but for cholesterol, the 2013-2014 design effect shows no evidence that the newer stratification has improved the precision.

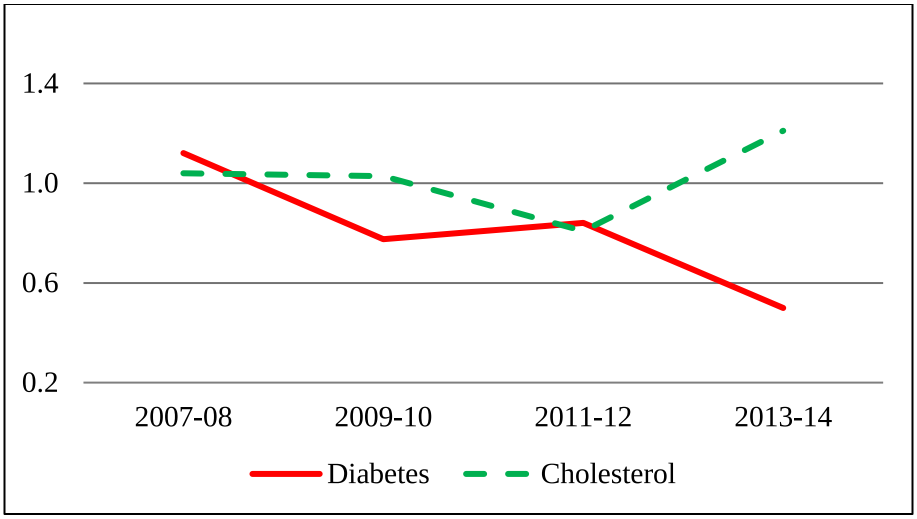

Figure 6: First-stage design effects for diabetes indicator and total cholesterol.

Figure 7 shows a boxplot for the first-stage design effect for all 38 tested variables over the four tested datasets. On average, the stratification does lower the variance, but no obvious improvement is shown in the later datasets with the newer health-based stratification.

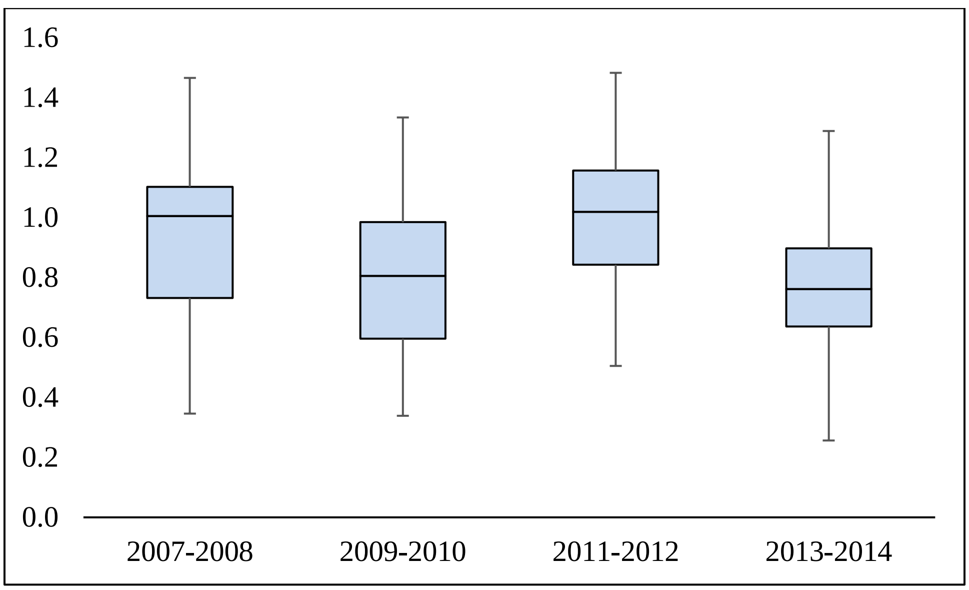

Figure 7: First-stage design effects for 38 outcome variables by 2-year data release.

4. Summary

In general, we did not find evidence that the newer stratification based on a healthiness index showed improvement in estimate precision, in part due to the limited number of NSR PSUs for which data are available in each 2-year data release. However, the newer stratification did not lead to a decrease in precision, either.

Figure 8 shows box plots for percent obesity among adults age 20 and older for all counties in the U.S., by region and by the 2011-2014 healthy state groups, from public data available from the Division of Diabetes Translation (DDT), provided by the CDC. While the healthy state groups show a monotonically increasing mean in obesity, the range shows there are many healthy (by this measure) counties within unhealthy states, and vice versa.

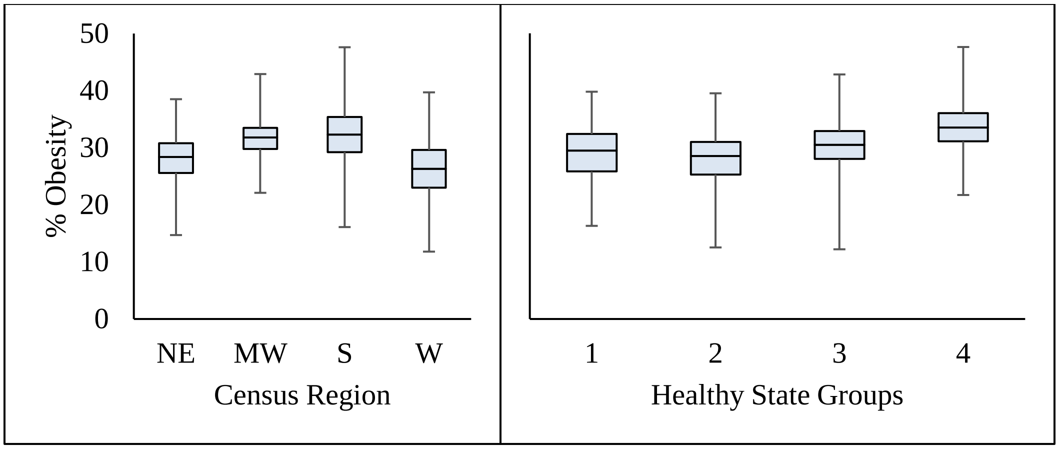

Figure 8: Percentage obesity at the county level for all U.S. counties, by region and healthy state groups.

For over 50 years, and continuously since 1999, data from NHANES have been used to measure and monitor the health and nutritional status of the civilian noninstitutionalized population of the United States. Improvements to the sample design, such as considering the overall health-level of the state in the stratification of the PSUs, are still being tested to try to increase the precision of the estimates from this important survey.

References

Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System (BRFSS). https://www.cdc.gov/brfss/data_documentation/index.htm.

Centers for Disease Control and Prevention (CDC). National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation (DDT). https://www.cdc.gov/diabetes/data/countydata/countydataindicators.html

Centers for Disease Control and Prevention (CDC). National Vital Statistics System (NVSS). https://www.cdc.gov/nchs/nvss/deaths.htm.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

Curtin L.R., Mohadjer L.K., Dohrmann S.M., et al. (2013). National Health and Nutrition Examination Survey: Sample design, 2007-2010. National Center for Health Statistics. *Vital Health Stat*, 2(160).

Hao, H., S. Dohrmann, L. Mohadjer, L. Li, W. Van de Kerckhove, and L. Curtin. 2011. Sample Design Changes for the NHANES 2011-2014 Annual Surveys. JSM Presentation, Section on Survey Research Methods.

Johnson C.L., Dohrmann S.M., Burt V.L., and Mohadjer L.K. (2014). National Health and Nutrition Examination Survey: Sample design, 2011-2014. National Center for Health Statistics. *Vital Health Stat*, 2(162).

Krenzke, T., and Haung, W. 2009. Revisiting Nested Stratification of Primary Sampling Units. *Proceedings of the Federal Committee of Statistical Methodology*.

Appendix

Table A: Outcome variables of interest

	Type	*Label*
1	Interview	Doctor ever said you had high blood pressure
2	Interview	Doctor ever said you had asthma
3	Interview	Doctor ever said you had arthritis
4	Interview	Doctor ever said you had coronary heart disease
5	Interview	Doctor ever said you had angina/pectoris
6	Interview	Doctor ever said you had heart attack
7	Interview	Doctor ever said you had weak/failing kidneys
8	Interview	Doctor ever said you had kidney stones
9	Interview	Doctor ever said you have diabetes
10	Interview	Taking insulin now
11	Exam	Average number of alcoholic drinks per day in past 12 months
12	Exam	Body mass index
13	Exam	Standing height
14	Exam	Recumbent length
15	Exam	Weight
16	Exam	Diastolic blood pressure
17	Exam	Systolic blood pressure
18	Exam	Ever used marijuana
19	Exam	Ever tried any form of cocaine
20	Exam	Cadmium level
21	Exam	Lead level
22	Exam	Hematocrit
23	Exam	Herpes I
24	Exam	Hemoglobin
25	Exam	Platelet count
26	Exam	Blood urea nitrogen
27	Exam	Creatinine level
28	Exam	Total cholesterol
29	Exam	Mercury level
30	Exam	Percentage obese
31	Exam	Percentage overweight
32	Exam	Percentage participating in vigorous sports or fitness activities
33	Exam	Any moderate sports/fitness activity
34	Exam	How many times have you been pregnant
35	Exam	Ever used female hormones
36	Exam	Used tobacco or nicotine in the last 5 days
37	Exam	Number of cigarettes smoked per day in the last 5 days
38	Exam	Presence of chlamydia