Exploring the Impact of Various Control Total Sources for Adjusting for the Cell-only Population in the California Health Interview Survey

Ismael Flores Cervantes[1], Matt Jans[2], Trent D. Buskirk[3]

[1]Westat, 1600 Research Boulevard, Rockville, MD 20850-3129
[2] UCLA Center for Health Policy Research, California Health Interview Survey, 10960 Wilshire Blvd., Suite 1550, Los Angeles, CA 90034
[3] Marketing Systems Group, 755 Business Center Drive, Suite 200, Horsham, PA 19044

Abstract

Dual-frame (cell phone and landline) random digit dial (RDD) surveys greatly reduce coverage errors in a variety of survey estimates compared to landline-only RDD surveys. However, for state and local surveys, the choice of control total sources for post-stratification of the cell phone sample has been limited. Marketing Systems Group (MSG) produces quarterly county level estimates for proportion of cell phone-only households in a county based on phone number assignments and administrative databases. Since these estimates are produced for smaller geographical areas, their use can potentially improve estimates in dual-frame RDD surveys. We study the impact of using MSG county-level estimates as control totals for post-stratification weighting on estimates in California Health Interview Survey 2001-2012. Changes in the weighting process and qualities of the weights and estimates correlated with cell-only phone (CPO) use are addressed. Results show that although there are differences in the weights and estimates of CPO use, estimates produced using the new weights are not significantly different from estimates produced using the current weighting methodology at the state level.

Key Words: Cell phone only population, weighting adjustment, health survey, California Health Interview Survey, Dual-frame RDD Sampling

1. Motivation and Overview of the RDD Control Total Landscape

The California Health Interview (CHIS) uses a dual-frame (landline and cell) random digit dial (RDD) sample and post-stratifies its nonresponse-adjusted person-level weights to population totals of three telephone service domains as part of the weighting process (California Health Interview Survey, 2014a): total persons in households with only landline service, persons in households with only cell phones and persons in households with both landline and cell service (i.e., dual-user). The National Health Interview Survey provides the source of control totals for this adjustment. The CDC (Centers for Disease Control and Prevention) reports cell phone use in the U.S. through the NHIS Early Release Program, (NHIS Wireless Substitution series reports, Blumberg & Luke., 2013). These reports include estimates of households and persons by telephone use status for selected demographic groups and geographic areas in the US. While they were ground-breaking in their ability to track the cell-phone-only (CPO) population in the United States, they have some limitations when used as a control total source, especially for surveys of smaller and local geographies like CHIS. The standard NHIS Wireless

Substitution reports do not typically provide separate estimates for California or any other state for that matter, but include estimates for each Census Region. Census Region is the smallest geography available in the public reports. To obtain the estimate breakdown within that region that CHIS requires for weighting, a request is made to CDC for the special tabulation of the West region. While these estimates have served CHIS well in the absence of any other sources, the estimates are produced using the two latest quarters of NHIS data. As a result, there is a time difference between the period covered by NHIS estimates and when the CHIS data is weighted. The justification of using NHIS estimates for the West region for CHIS is that California comprises 52 percent of the West region.

Beginning in 2009, the CDC published separated reports with estimates by telephone use at the state level and for the largest counties in the state (Blumberg, Ganesh, Luke, & Gonzales, 2013). These reports are part of the NHIS wireless substitution state-level series which are published yearly since 2011. The report includes estimates for California, for six single counties and two groups of combined counties in the state. However, since these reports are not part of the NIHS early release program, they are not published as close to the time the data was collected as the national estimates. The sample used to compute the state level estimates is collected during the course of one full calendar year and estimates are computed using a combination of design based and synthetic estimates using small area methods. Consequently the estimates have a time lag of over a year from the time the data were collected to the time they are released. For example, the latest report published in December 2013 includes state level estimates for year 2011-2012. Although these estimates are a better geographic match for CHIS control totals, they may not be accurate enough as control totals for a current survey (i.e. conducted in 2013) due to rapid changes in cell phone use in the population.

In 2013, Marketing System Group (MSG) began releasing an alternative estimate of the cell phone only (CPO, or cell-only) households at the state and county levels. These estimates are based on a combination of administrative records for telephone assignments along with postal records on dwelling units along with state and 5-year county-level estimates of household telephone prevalence from the American Community Survey. Wherever possible, each of the components of these estimates is based on the most recent quarterly update. Estimates of the proportion of telephone households that are CPO are then computed for every county within the United States. While these estimates are based on multiple and primarily administrative sources, they do offer a few advantages when compared to the NHIS estimates. First, the control totals are developed at the county level and provide direct estimates of lower-level geography than is not available from NHIS (NHIS only provides modeled estimates at the county level, and then not for every county in the U.S.). Second, these estimates are released quarterly, making them more useful for short-term or one-time surveys over a specific time period. If for example, a sample survey is planned to field in June, then the CPO rates for that quarter can be used and are preferable to those that might have been released in say January as there are local fluctuations in CPO rates, especially for smaller counties. The frequency of release also likely means that the totals are a more accurate reflection of the telephone user populations at the time of sampling and represent more local fluctuations in the CPO landscape of a given area. While the general trend for CPO rates is in an upward direction (Blumberg & Luke., 2013), not every county exhibits a positive CPO growth quarter over quarter. This fluctuation is especially apparent in smaller counties that are sensitive to small changes in the landline population as well as seasonal counties that service recreational or educational populations. In those counties the underlying number of telephone households exhibits seasonal patterns and fluctuations that in turn affect the

overall CPO rates. Third, they may be more accurate for local areas simply due to their input sources (i.e., telecommunication active records databases that provide nearly complete coverage of every county in the U.S.)[1]. Overall, the frequency of their release and detailed geographical level make these MSG figures appealing as a potential source of control totals for the telephone service poststratification in CHIS and other dual-frame RDD surveys.

In this paper, we evaluate the impact of using the MSG telephone use domains as control totals for CHIS weighting as compared to the current weighting methodology. We assess the impact on the weighting process, final weights, and substantive estimates produced with CHIS data. In this analysis, we use data from CHIS 2011-2012 that was post-stratified to control totals derived from the first two quarters of the 2012 NHIS for the West region (California Health Interview Survey, 2014a). For the comparison, we developed an alternative set of weights where the weights are post-stratified to control totals derived for the MSG proportion of cell-only households by county in California for the first quarter of 2013. Section 2 describes the CHIS weighting process focusing on steps for telephone use poststratification. Section 3 highlights the changes to the CHIS weighting process required to produce weights based on the MSG CPO Control Total Estimates (CTEs). Section 4 compares the weights created with NHIS and MSG control totals and the results of analyses comparing the differences of state and county level estimates on key substantive survey variables. The final section discusses our findings, recommendations and future research.

2. Weighting adjustments in the California Health Interview Survey

The CHIS weighting approach is a standard design-based dual-frame method that is consistent with the sampling methods used during sample selection. In CHIS, adjustments are applied to the weights to compensate for both the telephone number and person probability of selection and other factors resulting from the design and administration of the survey. The specific details of the weighting methodology are reported elsewhere (California Health Interview Survey, 2014a), but we summarize the three distinct weighting passes in Figure 1A.

The first phase adjusts base sampling weights for nonresponse and subsampling at the screener and extended interview separately by source of sample: landline or cell phone frame. In Phase 2 the landline sample is further partitioned into those respondents who report living in landline-only households and those who report living in households with both landline and cell phones (i.e. dual users). The cell sample is partitioned similarly into those respondents who report living in cell-only households and those who are living in households with both types of phones (i.e. dual users). In each of these 4 subsets, a poststratification adjustment is performed using the appropriate CTEs from the NHIS. After poststratification, a compositing factor is applied to dual users from both the landline and cell phone samples so that the final weights sum to the population of California Adults (and children). In phase 3, the combined sample (i.e. from both landline and cell samples) is raked to population total for different demographic and socio-economic characteristics. The weights produced using the NHIS control totals are called NHIS weights thereafter. In this paper we focus our attention on the ADULT sample and derive estimates using responses and weights computed from the adult respondents.

[1]Wireless only household estimates are available at http://www.m-s-g.com/Web/genesys/wireless-estimates.aspx

To better understand the effect of these adjustments, we examined the estimates of the proportion of CPO adults in the state throughout the three phases. At the end of Phase 1, the proportion of CPO adults was 0.3469. In phase 2 the proportion of CPO adults in the NHS control total was 0.3472. After poststratification, the adult CPO proportion was 0.3483. After raking in Phase 3, the proportion of CPO adult estimate after raking to the demographic variables was 0.3535. The difference in the state CPO estimate computed the NHIS weights and the control total was 0.0052.

3. Modified CHIS weighting process with MSG county-level totals

The main difference between the NHIS and the MSG adjustments is the geographic level of the adjustment. While the NHIS estimates are used to poststratify the weights to the population in the state, the MSG estimates allows a CPO use adjustment at the CHIS reporting areas (i.e., the CHIS reporting areas correspond to single counties or group of counties). We expect this adjustment to improve the estimates of variables correlated to CPO use in the smaller reporting areas.

The CHIS-MSG weighting process modifies the CHIS standard weighting process by using MSG CTEs for CPO households available for each county in California. The CHIS-MSG process also has three phases as depicted in Figure 1B. Phase 1 and 3 remain unchanged and the differences between the weighting processes are found in Phase 2.

As in the CHIS weighting, the landline sample is partitioned into those landline-only respondents and dual users in Phase 2. The cell sample is partitioned similarly into cell-only respondents and dual users. The dual users from both samples are combined using a compositing factor in the same way as in the CHIS weighting process. The cell-only respondents and the grouped landline-only/dual users are then poststratified to MSG control totals at the county level separately. The landline users and dual users need be combined before they are poststratified because unlike the NHIS CTEs that are available for all groups defined by telephone service, the MSG CTEs are only available for cell-only respondents. When landline-only and dual-users respondents are combined and adjusted as a single group, it is implicitly assumed is that there are no differences in response rates between these groups. This assumption may not hold in practice but the landline-only population is becoming smaller as more people abandon their landlines.

Similar to the NHIS telephone use estimates, the MSG CPO estimates cannot be used directly. In the MSG estimates are percentages of CPO households by county. Assuming that CPO households have in average the same number of persons as all households in the population, we computed the percentage of CPO persons in the county. The percentage of CPO persons in the county is then applied to the county CHIS population estimates to produce the county control total by CPO status at the county level. In the MSG CPO poststratification, there are 44 reporting areas that yield 88 poststratification cells. The weights produced using the modified weighting process with the MSG totals are called MSG weights thereafter.

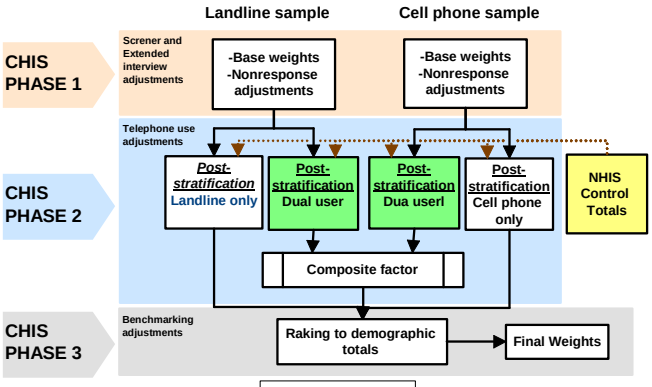


Figure 1A

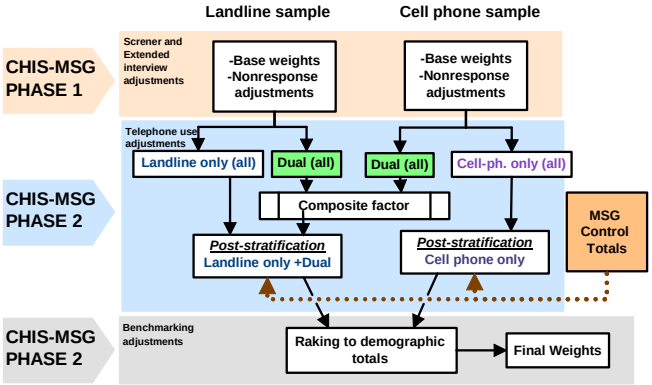


Figure 1B

Figure 1: A (top) Current CHIS weighting process and **B** (bottom) the modified CHIS weighting process using MSG county-level control totals.

As in the NHIS weights, we examine how the proportion of CPO adults changes throughout the three phases of weighting. As in the CHIS weighting process, the proportion of CPO adults was 0.3469 at the end of Phase 1. In the Phase 2, the proportion of CPO adults in California derived from the MSG control totals was 0.3640. After poststratification, the proportion of CPO adults in California was 0.3348. This proportion did not match the MSG numbers exactly because the MSG CPO estimates include children. In Phase 3, the proportion of CPO adults after raking was 0.3457. This was 0.0078 smaller than the adult CPO estimate computed using the NHIS weights. We were surprised at how close the CPO estimates were before and after the adjustments at the state level despite the counties were adjusted separately.

4. Comparison of results of NHIS and MSG weighting methods

Table 1 shows the summary statistics and distributions of the two sets of weights. The MSG weights have a smaller coefficient of variation, as a result; they have less variability than the NHIS weights.

Overall, an examination of the resulting final weights for the adults in the CHIS sample reveals a strong correlation between the CHIS and CHIS-MSG final weights (the regression line has a slope of 0.95 and the correlation is 0.97). These results are not

surprising because the weights share most of the weighting adjustments were raked to the same raking cells.

Table 1: Summary statistics for the NHIS and MSE weights

Statistics	Weights		Distribution	Weights	
	NHIS	MSG		NHIS	MSG
N	42,935	42,935	Minimum	3.3	3.7
Sum	27,796,484	27,796,484	Lower Quartile	147.9	159.7
Mean	647.4	647.4	Median	327.5	344.4
Standard Deviation	908.0	893.8	Upper Quartile	755.5	737.6
Coefficient of Variation	140.3	138.1	Maximum	15,315.2	13,583.9

There are also differences in the coefficient of variation (CV) for each set of final weights. On average, the MSG CVs are 4 percent smaller than the NHIS CVs but there is a large variation among the reporting areas. The MSG CVs can be 30 percent higher or lower than the CVs of the NHIS weights. These differences are mainly found in reporting areas for rural counties which also have smaller sample sizes.

4.1. Comparison of CPO estimates

In order to evaluate the impact on the estimates at the county level, we first analyzed how the estimated proportion of CPO adults changes after each adjustment separately for each set of weights and reporting area. We only consider changes in CPO estimates computed within each of the three distinct phases of each weighting method: Phase 1: pre-postratified weights; Phase 2- poststratified and composite weights and Phase 3- final/raked weights. These estimates for the 44 reporting areas (i.e., counties) for these two sets of weights are shown on the right side of Figure 2. The figure shows that the largest changes in the CPO estimates in the NHIS weights are the result of the raking adjustment. This is expected because the NHIS telephone use poststratification does not affect the weights within a county (e.g., most CPO estimates stay the same after this adjustment). There are also some reporting areas that stay constant throughout the adjustments. In contrast, the left side of Figure 2 shows the opposite behavior in CPO estimates computed using the MSG weights. The largest changes are the result of the CPO poststratification adjustment. In contrast, the weights after raking tend to maintain the same CPO estimates in most counties. These observations suggest that for a large number of reporting areas, the raking adjustment in the NHIS weights and the MSG CPO poststratification adjustments are similar. This is not entirely surprising because the raking dimensions include proxy socio-economic variables in addition to the demographic variables. In this situation, CPO use can be modeled using the variables used in raking.

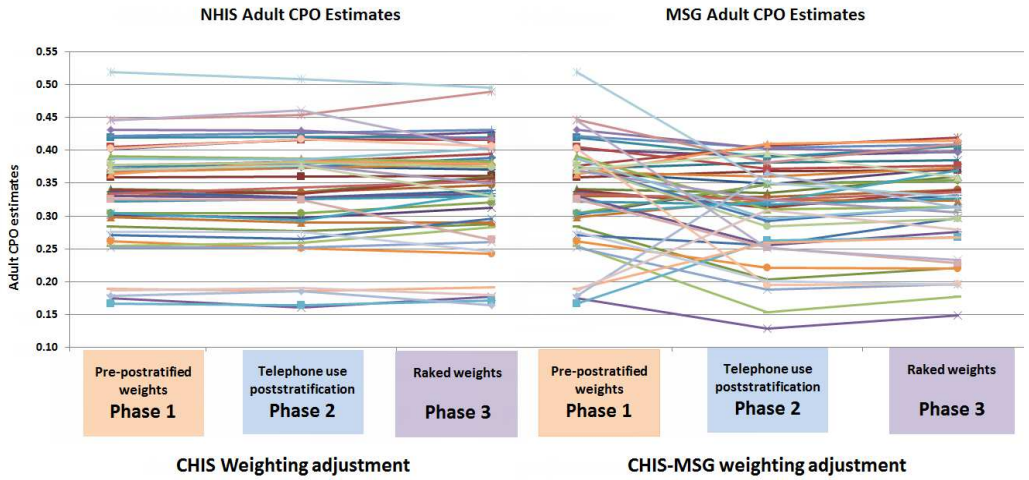


Figure 2: Changes in county-level CPO estimates from three steps in CHIS and CHIS-MSG weighting process; (1) after weight but before post-stratification, (2) after post-stratification, and (3) after raking. NHIS (a) and MSG (b) weights are compared.

Figure 3 directly compares the CPO adult proportions produced with the NHIS and MSG weights by county (i.e. reporting area). The blue series indicates the estimated proportion of CPO adult in the reporting area computed using the NHIS weights (i.e., p_{CPO}^{NHIS}) while the red series indicates the same proportions but computed using the MSG weights (i.e., p_{CPO}^{MSG}). The proportions are ordered from right to left by increasing values of the difference $d = p_{CPO}^{MSG} - p_{CPO}^{NHIS}$ so reporting areas with the similar estimates fall in the middle of the plot. The shaded rectangle indicates reporting areas where the relative difference $r\% = \left|100d/p_{CPO}^{MSG}\right|$ is less than 10 percent. The horizontal in the plot indicates the NHIS CPO state level estimate. In general, the proportions of CPO adults from the NHIS trace the MSG proportions. However, they have noticeably-different values for those areas at the both sides of the plot. For example, the MSG CPO adult proportion is almost half of the NHIS CPO adult proportion for the reporting area that includes Colusa County and Glenn County. The reverse is true for El Dorado County on the right side of the plot. Some of these differences can be the result of sampling variability because areas outside the shaded area have an average sample size of 650 adults while the average sample size of the reporting area within the shaded area is 1,023 adults (excluding Los Angeles County with a sample size of 9,009 adults). In the CHIS sample design, reporting areas that represent rural counties are allocated smaller samples.

4.2 Comparison of estimates correlated with CPO use

The primary goal of CHIS is to provide statewide and county-level estimates, so we need to assess whether the estimates computed using the MSG weights are systematically different from those estimates computed using the NHIS weights at the state and county levels.

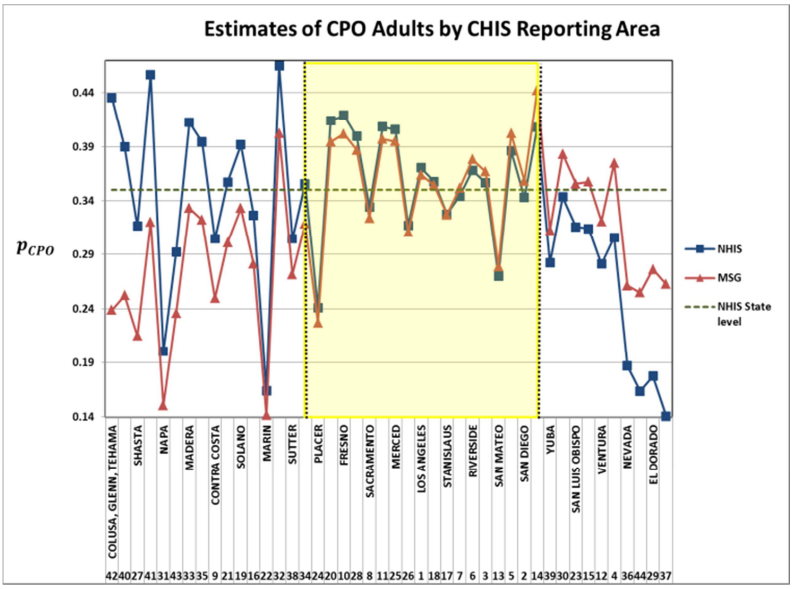


Figure 3: NHIS and MSG CPO adult estimates by reporting areas

Since there are no figures that can be used as gold standard, we cannot determine the set of weights that produces estimates with the smallest mean squared error. Instead we carried out an indirect analysis that examines the differences between the estimates and their correlations between these variables and CPO status. We focused on estimates of variables that are highly correlated to CPO status previously reported in the literature listed in Table 2. The table shows the correlation between the proportion of the variable (p_v^{NHIS}) and the proportion of CPO use by reported area (p_{CPO}^{NHIS}) computed using the NHIS weights for all reporting areas. The high correlation between the variables and the CPO use is also observed in the CHIS data as shown in the table except for the variables ASTCUR (diagnosed with asthma) and BINGE12 (Binge drinking in past 12 months).

Table 2: CHIS variables correlated to CPO use

Variables	*Description*	*Correlation*	*p-value*
POVLL	Under 100 % poverty line	0.69	<0.0005
INSANY	Uninsured in past 12 months	0.61	<0.0005
SRTENR	Own house	-0.53	<0.0005
AC32	Had alcohol past 12 months	-0.61	<0.0005
AE30	Had flu shot in past 12 months	-0.54	<0.0005
AH16	Delayed getting prescription in past 12 months	0.31	0.041
ASTCUR	Diagnosed asthma	0.01	0.935
SMKCUR	Current smoker	0.25	0.102
INSMC	Covered by Medicare	-0.62	<0.0005
INSMD	Covered by Medical	0.58	<0.0005
USUAL	Have usual place to go to when sick or needing health advice	-0.5	<0.0005
BINGE12	Binge drinking in past 12 months	-0.09	0.566

Several of these variables show a geo-spatial relationship between their mean or proportion and CPO rates across counties – as depicted in Figure 4 for the percentage of families in a county that are under the poverty line. This increases the likelihood of

detecting differences in the estimates produced by the two sets of weights, especially among the CPO telephone usage domains. While we did not present telephone household sub-domain estimates (i.e. adults from CPO households compared to adults from landline-only households, for example) we would expect even more marked differences in final estimates derived using these two weighting methods within the adult CPO subdomains for many of these variables.

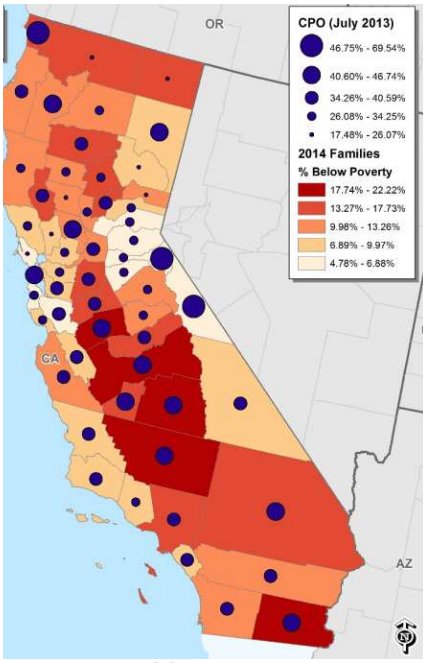


Figure 4: Map depicting the percentage of families within each county in California as well as the percentage of telephone households within the county that are Cell Phone Only (CPO). The map was generated by MSG using MSG CPO CTEs for each county.

In the next part of the analysis, we examine the state level differences between the estimates computed using the two sets of weights using the same variables. The differences between the state level estimates are listed in Table 3. Although there are differences, these are very small with an average difference of 0.002 and the maximum absolute difference of 0.0098 for variable AC32 (had alcohol past 12 months). In this case, the estimates are substantially similar and would lead to the same conclusions. At least for the variables highly correlated with telephone use, no significant changes in the state level estimates are observed. A more complete analysis includes other variables not correlated with CPO use to determine if the same pattern holds.

In the last part of the analysis, we compute the estimates and differences in estimates based on using the two sets of weights for the same set of variables for all reported areas in CHIS. There are a total of 1,056 estimates and 528 differences. The summary statistics of the differences is found in Figure 5a.

Table 3: State level differences of estimates computed NHIS and MSE weights

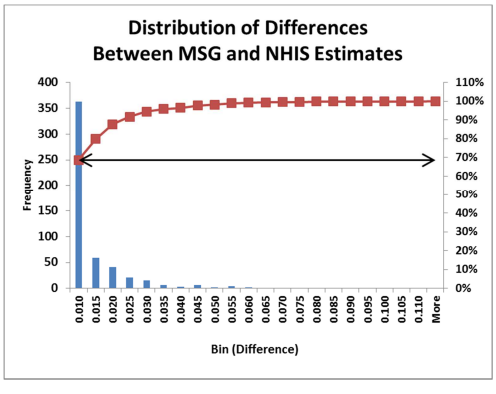
Variables	*Description*	*Difference*
POVLL	Under 100 % poverty line	+0.0062
INSANY	Uninsured in past 12 months	+0.0021
SRTENR	Own house	+0.0000
AC32	Had alcohol past 12 months	+0.0098
AE30	Had flu shot in past 12 months	+0.0002
AH16	Delayed getting prescription in past 12 months	-0.0012
ASTCUR	Diagnosed asthma	-0.0004
SMKCUR	Current smoker	-0.0002
INSMC	Covered by Medicare	-0.0015
INSMD	Covered by Medical	+0.0059
USUAL	Have usual place to go to when sick or needing health advice	+0.0026
BINGE12	Binge drinking in past 12 months	+0.0022
Summary Statistics of state level differences		
	Mean	0.0021
	Standard deviation	0.0035
	Minimum	-0.0015
	Maximum	0.0098

Figure 5a shows that for county level estimates, the differences between the estimates are very small with an average of 0.0017 (less than 0.2 percent). To examine the distribution of the differences, Figure 5b shows the plot of a Pareto analysis of the absolute value of the differences. The analysis shows that 69 percent of the differences (366 cases) are less than 0.01. Furthermore, 90 percent of the differences have an absolute value less than 0.023. Considering that in most cases the sampling error is larger than these magnitude of these differences, it is not possible to determine if these differences are statistically different from zero. Nevertheless, there are few areas where the differences are large. For example, the largest absolute differences are 0.068 for variable SRTENR (Own house) for El Dorado County and 0.065 for POVLL (Under 100 % poverty line) for Mendocino County. There may be some reporting areas where the estimates computed the MSG weights may have better properties, but the decision to replace the NHIS weights would depend on the particular area.

Summary statistics of differences of estimates based on NHIS and MSE weights by reporting area

Statistic	*Value*
Number of differences	528
Mean	0.0017
Standard deviation	0.0143
Minimum	-0.0534
Median	0.0016
Maximum	0.0682

(a)



(b)

Figure 5: Summary statistics of differences and Pareto analysis.

5. Conclusions and Extrapolations

We found no significant differences between estimates produced using the NHIS and MSG weights at either the county or state levels. We think that is because the raking adjustment in the NHIS weights has a similar effect as the county level CPO poststratification adjustment in the MSG weights. Although there are some small areas that could benefit from the change, for the most part, the estimates are same.

As a cautionary note, researchers should not jump to using county-level control totals, but neither should they be concerned about major changes in estimates. At the state level and for the largest counties, estimates were almost identical regardless of the control total source.

Small counties exhibited over-estimates, or under-estimates with the county-level source, relative to the region-level source, with no obvious hint as to what would cause a positive or negative difference. However, it is important to note that after controlling for sample size (e.g., testing the difference between county level estimates under each control total), none of the county level differences were statistically different. CHIS county-level samples can be small (n=400 in some areas), so different results might be seen with a different allocation or larger overall sample size per county. These caveats make it difficult to generalize our findings beyond CHIS to all state-level dual-frame RDD surveys with geographic areas.

The alternative control total source is not without its challenges. In addition to differences in its source, the specific control populations, specifically children and adolescents, are not available in the county-level source. The current comparison focused on adults; therefore the effect on these groups needs to be studied too. Although we have been comparing estimates from weights produced using two separate sources, there is no reason these cannot be combined. A combined approach may solve the problem of not adjusting children or persons in landline-only households and persons in households with both services.

Additional areas of research are the evaluation of estimates at the county level separately by telephone usage for children and alternative ways of combining both NHIS and MSG control totals.

References

Blumberg, S. J., Ganesh, N., Luke, J. V., & Gonzales, G. (2013). *Wireless Substitution: State-level Estimates From the National Health Interview Survey, 2012.* National Health Statistics Reports No. 70.

Blumberg, S., & Luke., J. (2013, December). *Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, January–June 2013.* Retrieved Mach 25, 2014, from http://www.cdc.gov/nchs/nhis/releases.htm#wireless

California Health Interview Survey. (2014a). *CHIS 2011-2012 Methodology Series: Report 5 – Weighting and Variance Estimation.* Los Angeles, CA: UCLA Center for Health Policy Research.

California Health Interview Survey. (2014b). *CHIS 2011-2012 Methodology Series: Report 1 - Sample Design.* UCLA Center for Health Policy Research, Los Angeles, CA.