Coverage and Data Quality Association in Enhanced Address-Based Sample Frames

Ipek Bilgen, Ned English, and Lee Fiorio

NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603

Abstract

Racial and Ethnic Approaches to Community Health across the U.S. Risk Factor Survey (REACH U.S.) provides the Centers for Disease Control and Prevention (CDC) and the involved communities with quantitative data to track the progress and achievements of the community intervention programs to eliminate health disparities. For REACH U.S., NORC conducts multi-mode surveys using address-based sampling frames enhanced with race/ethnicity information. Specifically, REACH U.S. employs two sampling frames: 1- An address-based sampling (ABS) frame derived from the U.S. Postal Service (USPS) Delivery Sequence File (DSF), and 2- A race-targeted list. REACH U.S. collects data primarily via two modes of data collection (telephone and mail interviews). This paper examines the REACH U.S. Year 3 achieved sample and investigates the impact of increasing efficiency via list-based frames as opposed to ABS frames. We examine whether the two frames significantly differ on key health measures and investigate whether using an enhanced DSF affects key statistics. According to our results, we see some significant response differences among respondents who were covered by the targeted lists versus those who were not covered. Specifically, among the examined key estimates, two of the key health estimates significantly differ between DSF-only list and the enhanced DSF list. Overall, however, the source frame does not seem to have much impact on the relationship between the key respondent characteristics and the key health estimates.

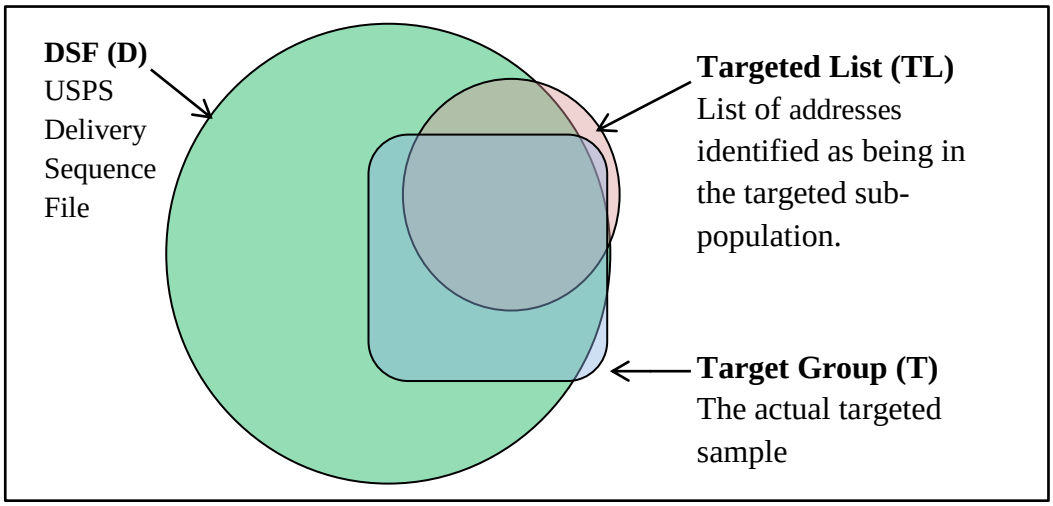
Key Words: Address-based Samples, Enhanced Frames, Targeted Lists, Sample Frame Construction, Coverage

1. Introduction

Survey research has recently undergone a transformation from surveys primarily based on random-digit-dialing (RDD) to multi-mode surveys using address-based sampling (ABS) strategies due to increased costs and decreasing response rates (Iannacchione, 2011; Link, et al. 2008; Link, et al. 2009). Consequently, different screening methodologies and strategies have emerged from studies surveying specific targeted subpopulations (including age or racial/ethnic groups). While some subpopulation surveys use address-based sampling during frame construction and combine multiple modes (e.g., phone and mail) during data collection, others exclusively use ABS and employ mail during both recruitment and data collection (i.e., two-phase mail ABS design). This study focuses on the former approach and specifically examines the use of targeted lists for sub-populations in the United States (such as certain age or racial groups) to enrich ABS designs for greater efficiency. The rationale for such a strategy is to decrease costs and increase efficiency when targeting rare populations.

ABS studies generally base their sampling frames on the United States Postal Service delivery-sequence file (CDSF or DSF) due to its near-universal address coverage. While research has shown that the DSF-derived ABS frame contains essentially all households receiving mail in the U.S., they do not generally contain information about household characteristics. Consequently, ABS-based design efficiency and associated costs can be challenged in surveys targeting hard-to-reach populations. In the case of REACH U.S., which targets several different races/ethnicities in a variety of communities across the United States, one method for combatting declining efficiency of the DSF is to make use of race/ethnicity targeted lists. Licensed by market research companies, targeted lists are compilations of households expected to contain members of a particular population based on surname information and other consumer behavior (Kennel & Mei, 2009). Taken alone, we would not expect targeted lists to be able to cover all of the targeted population in a particular area; however, one strategy to avoid coverage loss would be to enhance an ABS list using a targeted list (TL∩D in Figure 1), and use such a hybrid list in surveys which target a specific subpopulation (T). In Figure 1, the target group that is covered by the enhanced list is indicated via TL∩D∩T and will be referred to as the enhanced DSF throughout the paper. As illustrated in Figure 1, the enhanced DSF does not fully cover the target group (i.e., T not in TL∩D in Figure 1). Thus, a section of the targeted group is not covered via targeted lists. However, as the DSF list is more complete, the majority of the target group is covered by the DSF (D∩T). The sections the DSF covers also include the section that is not covered via targeted lists and will be referred to as the DSF-only section throughout the paper (D∩T not in TL∩D∩T).

Figure 1: The Use of Enhanced ABS Frame for Targeted Groups

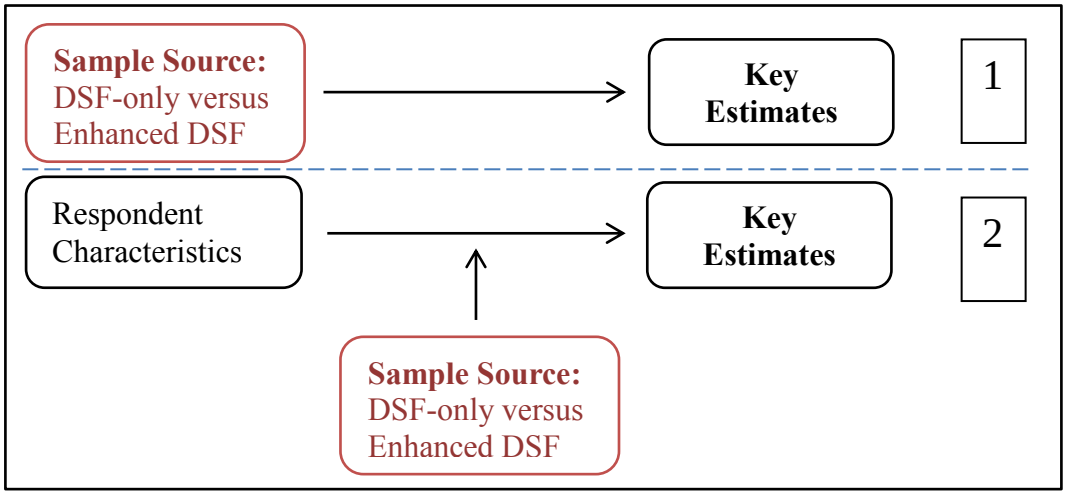


Research Questions

We examined the impact of the use of the enhanced ABS frame on key survey estimates. Specifically we investigated the following research questions (see Figure 2):

1) Do key survey estimates differ depending on the sample source (enhanced DSF versus DSF-only)? Specifically, do key estimates obtained from the households that are covered via the enhanced DSF (i.e., TL∩D∩T in Figure 1) differ from those that are DSF-only (i.e., D∩T not in TL∩D∩T in Figure 1)?
2) Does the sample source (enhanced DSF versus DSF-only) moderate the relationship between respondent characteristics and key estimates?

Figure 2: The Relationship between Sample Source and Key Estimates



Note that in our analyses we excluded the small section of the target group (T) that is covered by targeted list (TL) but is not a part of the DSF (TL∩T not in D∩T in Figure 1). We also excluded the section of the target group that is not covered by either DSF or targeted list (TL) as this section was not the scope of our study.

2. Data and Methods

Data: REACH U.S.

The Racial and Ethnic Approaches to Community Health across the U.S. (REACH U.S.) risk-factor survey is a Centers for Disease Control and Prevention (CDC) program designed to eliminate racial and ethnic health disparities. NORC at the University of Chicago conducts the REACH U.S. survey in 28 communities to track the progress of the community intervention programs and monitor health indicators in the defined geographical areas. REACH U.S. transitioned to multi-mode ABS from RDD in 2008 and currently uses telephone, mail, and face-to-face data collection modes.

REACH U.S. uses targeted lists to enhance its ABS design for greater efficiency. Hence, targeted lists in the context of REACH U.S. are designed to enumerate and identify households containing race/ethnicities of study focus at higher rates than using the USPS DSF alone. The USPS DSF provided by Valassis may be enhanced via household-level data containing demographic information (e.g., gender, age, and race-ethnicity) from a second vendor[1]. REACH U.S. first uses an enhanced ABS design to create the sampling frame, and then selects addresses at varying rates depending on their inclusion. Using the sampled addresses, households are screened to determine eligibility, after which eligible respondents complete the questionnaire. We used the data from these completed questionnaires in our analyses.

Since 2009, NORC has conducted the REACH U.S. survey in 28 communities to assess disparities in health outcomes and evaluate the effectiveness health intervention programs. REACH U.S. communities vary greatly in size, location, and the

[1] Examples of vendors include InfoUSA, Marketing Systems Group (MSG), Targus, Survey Sampling International, Valassis.

concentration of priority racial/ethnic groups. For the purposes of this analysis, we limit the number of REACH U.S. communities of interest to 20, excluding those that did not employ targeted lists. Because some REACH U.S. communities target multiple ethnicities, it is possible that within a given community we evaluated multiple flags. The REACH U.S. priority groups considered in this analysis are African American, Hispanic, and Asian. At question is what impact using enhanced DSF frames (via targeted lists) may have on substantive results. We first examine whether households covered by enhanced DSF list differ from those who are from the DSF-only list. Secondly, we evaluate the degree to which key statistics would be impacted through the sole use of enhanced DSF frames. Our analysis focuses on the 21,377 completed interviews achieved among the 20 REACH U.S. communities. Of these completed interviews, 81.7% originated from the intersection of the DSF and targeted list (enhanced DSF) and 18.3% came from the DSF-only portion of the frame.

Methods

In this paper, we examined responses to four key health estimates: *Body Mass Index (BMI)*, *Smoking Summary*, *Health Care Coverage*, and *Diabetes Summary*. These outcome variables were obtained from the substantive portion of the REACH U.S. completed questionnaires. REACH U.S. collected the respondent's self-reported height and weight which were then used to calculate the *Body Mass Index (BMI).* For purposes of the analysis, we recoded BMI as a binary variable in which respondents with a BMI of 30 or less (normal or overweight) were coded as 0, and respondents with a BMI greater than 30 (obese) were coded as 1. The *smoking summary* variable has two categories, one for those that smoke every day and the other for those that did not smoke every day. *Health care coverage* is dichotomized into a category for respondents with health care coverage and another category for respondents without. *Diabetes* is a binary variable with 1 meaning the respondent has been diagnosed with diabetes and 0 meaning he or she had not. In the models, we used the *sample source (enhanced DSF versus DSF-only)* as an independent variable for research question 1 and a moderator for research question 2 (see Figure 2). The *sample source* is a binary variable and dichotomized into two categories as enhanced DSF (TL∩D∩T in Figure 1) versus DSF-only (D∩T not in TL∩D∩T in Figure 1) which indicates whether the address was from the enhanced DSF or the DSF-only portion of the frame.

Independent variables in the models were obtained from the completed interview data as well. Of interest to us were the *age* and *gender* of the respondent as well as his or her *education*, *race/ethnicity*, and *household income*. *Education* was dichotomized into two groups, one containing people with less than a high school education and another containing people with a high school education or more. *Race/ethnicity* was included as dummy variables in which each race group has represented one race/ethnicity group (e.g. Hispanic versus Non-Hispanic). *Household income* was split into two groups, one comprising respondents who reported less than $20,000 a year and another group reporting $20,000 a year or more. We also included variables of interest which may relate to the key health estimates in the targeted groups such as *place of birth* (whether the respondent is foreign born or not), whether or not the respondent spoke English (non-English speaker), as well as whether they are worried or stressed about having enough money to pay rent. Lastly, we controlled for block group characteristics using data from the 2010 Census. Of interest was the *percentage of population made up by the REACH* U.S. *target group*, *the housing unit density per square mile*, and *urbanicity*. Urbanicity was approximated using the Census Type of Enumeration Area (TEA) code. If the census used mail-out/mail-back to enumerate all the blocks within a block group, it was coded as

urban; if not, it was deemed rural. We also controlled for the *mode* in which the survey was completed, either by telephone or mail. Face-to-face interviews are conducted in only a handful of the communities that are not a part of the 20 communities that we are investigating; therefore, they were excluded from our analyses.

Separate logistic models were conducted for four dependent variables to examine both research question 1 and research question 2. For the first research question, among the completed interviews we compare the responses of individuals from the enhanced DSF with the responses of individuals from the DSF-only portion. For the second research question we examine whether sample source changed the relationship between respondent characteristics and the key health estimates. Therefore, we include the frame source of each completed questionnaire as a moderator variable via interaction terms.

3. Results: Analyses of the Key Health Measures from the Completed Questionnaires

Table 1 illustrates the relationship between the sample source (enhanced DSF versus DSF-only) when controlling for respondent demographics such as age, race/ethnicity, sex, education status, place of birth, income, whether the respondent worried about paying rent, as well as block group characteristics such as urbanicity, target density, housing unit density, and mode. The focus of interest was to examine whether the address source related to key health estimates, as per research question 1. We included relevant respondent and block-level characteristics to control for the known coverage differences among DSF-only list and the enhanced DSF list. For instance, English, Bilgen, and Fiorio (2012) found that block groups in highly dense and low income communities experience lower levels of targeted list coverage despite relatively high concentrations of the target group. Also, target list coverage is more likely to be adequate in block groups in REACH U.S. communities with stable housing (i.e., areas with high occupancy and low percentage of households renter occupied).

Moreover, respondent characteristics also play a role in terms of being covered by the targeted lists. For instance, the English et al. (2012) findings also reveal that coverage tends to be higher in communities that target Asians/Pacific Islanders or Hispanics/Latinos, while communities that target African Americans had relatively lower coverage. Because these race/ethnicity flags are at least partially created using surname lists, it is not surprising that African Americans would be more difficult to identify. Also, we included age as older respondents are more likely to be in the targeted lists (and hence covered by the enhanced DSF). Lastly, we included education in the models as a control variable as well, because people with higher levels of as education are more likely to be covered by the enhanced DSF, as education is positively correlated with income.

Two of the four key health estimates significantly differed between the DSF-only list and the enhanced DSF list (Table 1). Specifically, the DSF-only potion of the list was more likely to cover people who smoke every day as well as respondents who are more likely to be diabetic. This indicates that the respondents who are covered by the enhanced DSF ($TL \cap D \cap T$) provide different responses to two of the four examined key health estimates than the people who are from the DSF-only portion of the frame ($D \cap T$ not in $TL \cap D \cap T$). While Table 1 indicates that the descriptive information for key health estimates may differ depending on their source frame, Table 2 suggests that overall the source frame may not influence the relationship between respondent characteristics and key health estimates. Table 2 specifically explores whether the sample source plays a moderator role between the key independent variables (i.e. respondent characteristics) and the explored key health indicators (i.e., Research Question 2).

Table 1. Binary Logistic Regression: Estimates for Key Health Statistics (Research Question 1)

	BMI		Smoking		No Health Care Coverage		Diabetes	
	B	**Odds Ratio**	**B**	**Odds Ratio**	**B**	**Odds Ratio**	**B**	**Odds Ratio**
Intercept	-1.42***	0.24	-1.80***	0.17	-0.24+	0.79	-4.39***	0.01
Sample Source (Enhanced DSF vs. DSF-only)	0.07	1.07	0.23***	1.25	0.07	1.07	0.11**	1.11
Age (in years)	0.00	1.00	-0.01**	0.99	-0.04***	0.96	0.05***	1.05
Hispanic	0.35***	1.43	-0.52***	0.60	-0.42***	0.66	0.20+	1.22
White non-Hispanic	0.01	1.01	0.04	1.04	0.15+	1.16	0.06	1.06
African-American non-Hispanic	0.24**	1.27	0.19	1.21	0.20+	1.22	0.17	1.18
Asian non-Hispanic	-0.33***	0.72	0.21	1.23	-0.61***	0.54	0.04	1.04
American Indian	0.11	1.11	0.15	1.16	0.01	1.01	0.13	1.14
Other Race	0.05	1.05	0.13	1.14	0.07	1.07	0.07	1.08
Sex (1=Female)	0.39***	1.48	-0.42***	0.66	-0.37***	0.69	0.07+	1.08
Education (1=Less than HS)	0.12**	1.13	0.46***	1.59	0.19***	1.21	0.25***	1.29
Non-English Speaker	-0.28***	0.75	-0.17**	0.84	0.41***	1.50	0.17**	1.19
Worried about paying rent	0.38***	1.47	0.42***	1.52	0.41***	1.50	0.28***	1.33
Place of birth (1=Foreign Born)	-0.56***	0.57	-0.82***	0.44	0.54***	1.71	-0.20***	0.82
Household (HH) Income (0 = >20K; 1 = <20K)	0.13***	1.14	0.64***	1.89	0.67***	1.96	0.22***	1.24
Mode (telephone and mail)	0.01	1.01	-0.12**	0.89	0.09**	1.09	-0.03	0.97
Urbanicity (1=Rural)	0.07	1.07	0.21+	1.23	0.02	1.02	0.07	1.07
Target Density	0.32***	1.37	0.29**	1.34	0.51***	1.66	0.04	1.04
Housing unit density	-0.09**	0.91	0.11**	1.11	-0.34***	0.71	-0.10**	0.91
AIC[2]	25299.262		14274.427		18547.998		18834.679	
Likelihood Ratio χ^2	1234.9386		1141.2327		2294.1000		1879.3268	

+ p<.10, **p<0.05, ***p<0.001

[2] The fit statistics are provided for the comparison of Table 1 and 2.

Table 2. Binary Logistic Regression: Estimates for Key Health Statistics (Research Question 2)

	BMI		Smoking		No Health Care Coverage		Diabetes	
	B	**Odds Ratio**	**B**	**Odds Ratio**	**B**	**Odds Ratio**	**B**	**Odds Ratio**
Intercept	-1.51***	0.22	-1.77***	0.17	-0.22	0.81	-4.40***	0.01
Sample Source (Enhanced DSF vs. DSF-only)	0.53**	1.71	0.11	1.12	-0.18	0.83	-0.01	0.99
Age (in years)	0.00	1.00	-0.01***	0.99	-0.04***	0.96	0.05***	1.05
Hispanic	0.43***	1.54	-0.55***	0.57	-0.41**	0.66	0.17	1.19
White non-Hispanic	0.06	1.06	0.02	1.02	0.17+	1.19	0.05	1.05
African-American non-Hispanic	0.36***	1.44	0.13	1.14	0.18	1.20	0.06	1.07
Asian non-Hispanic	-0.38***	0.69	0.14	1.15	-0.60***	0.55	-0.03	0.97
American Indian	0.13	1.13	0.14	1.15	0.02	1.02	0.14	1.15
Other Race	0.07	1.07	0.12	1.13	0.08	1.08	0.06	1.06
Sex (1=Female)	0.39***	1.47	-0.42***	0.66	-0.37***	0.69	0.07**	1.08
Education (1=Less than HS)	0.16***	1.18	0.45***	1.56	0.25***	1.28	0.26***	1.30
Non-English Speaker	-0.30***	0.74	-0.18**	0.84	0.40***	1.50	0.16**	1.17
Mode (telephone and mail)	0.01	1.01	-0.12**	0.89	0.09**	1.09	-0.03	0.97
Worried about paying rent	0.38***	1.46	0.42***	1.52	0.41***	1.50	0.28***	1.33
Place of birth (1=Foreign Born)	-0.48***	0.62	-0.75***	0.47	0.52***	1.69	-0.15**	0.86
Household (HH) Income (0 = >20K; 1 = <20K)	0.15***	1.16	0.65***	1.91	0.67***	1.95	0.25***	1.29
Urbanicity (1=Rural)	0.07	1.07	0.21+	1.24	0.01	1.01	0.08	1.09
Target Dens	0.32***	1.37	0.31**	1.36	0.54***	1.72	0.11	1.12
Housing unit density	-0.09**	0.92	0.11**	1.11	-0.35***	0.71	-0.11**	0.90
Age*Sample Source	0.00+	1.00	0.00	1.00	0.00	1.01	0.00	1.00
Hispanic*Sample Source	-0.45**	0.64	0.12	1.13	-0.03	0.97	0.08	1.08
White non-Hisp.*Sample Source	-0.28**	0.75	0.05	1.05	-0.07	0.93	-0.04	0.96
AfAm non-Hisp.*Sample Source	-0.59**	0.55	0.17	1.19	0.13	1.14	0.44+	1.55
Asian non-Hisp.*Sample Source	0.12	1.13	0.23	1.26	0.06	1.06	0.45+	1.57
Education*Sample Source	-0.20+	0.82	0.07	1.08	-0.33**	0.72	-0.03	0.97
Place of birth*Sample Source	-0.49***	0.61	-0.32+	0.73	0.09	1.10	-0.35**	0.70
HH Income*Sample Source	-0.08	0.92	-0.04	0.96	0.03	1.03	-0.21	0.81
AIC	25248.776		14285.203		18551.129		18817.241	
Likelihood Ratio χ^2	1301.4240		1170.5423		2306.9695		1912.7649	

+ p<.10, **p<0.05, ***p<0.001

In the models illustrated in Table 2, first we explored all the two-way interactions between the key independent variables (respondent characteristics) and sample source. In all of four models, the two-way interactions were consistently non-significant and did not improve the model fit. Therefore, these two-way interactions were excluded from the final analyses illustrated in Table 2. The comparison of the fit statistics between Tables 1 and 2 indicates that the addition of the interaction terms improved the model fit in two of the four models. Models for smoking and health coverage, however, do not significantly improve the model fit. This indicates that in two out of the four models, the sample source did not moderate the relationship between key health statistics and key independent variables. According to Table 2, overall the relationships between the key health statistics and the respondent characteristics (such as age, race, and annual income) did not vary by the sample source in most models. Nevertheless, the relationship between place of birth and being obese, smoking every day, and having diabetes is significantly different between respondents covered by the enhanced DSF and those covered by the DSF-only list. Regardless of whether the respondents are covered by the enhanced DSF or DSF-only list, U.S.-born respondents were more likely to report being obese, being an every day smoker, and having diabetes. However, the estimated differences between the U.S. born and foreign born respondents were more pronounced among those who were covered by the enhanced DSF than who were not covered by the enhanced DSF.

Similarly, the relationship between education and being obese and not having health coverage is significantly different between the people who are covered by and the enhanced DSF and those who are from the DSF-only portion of the frame. Specifically, among respondents who are from the DSF-only list, those with less than a high school education are significantly more likely to report being obese and not having health coverage than the respondents who have at least a high school education. The difference between respondents with disparate levels of education, however, is among respondents recruited from the enhanced DSF. These results may indicate that overall the respondents from the enhanced DSF may be less educationally diverse than the respondents who are from the DSF-only portion of the frame.

Lastly, the relationship between being African-American, having diabetes, and being obese is significantly different between those who are covered by the different frame sources. Specifically, among respondents recruited from the DSF-only portion of the frame, those identifying as African-American were more likely to report being obese than those who did not identify as African-American. Interestingly, this relationship is reversed among respondents recruited from the enhanced DSF: those identifying as African-American were more likely to report being less obese than those not identifying as African-American. Such differences may be partially explained by income, with more affluent respondents being found on the enhanced DSF across races/ethnicities.

4. Conclusion and Discussion

English, Bilgen, and Fiorio (2012) found that the people who are from the DSF-only portion of the frame are demographically different than the respondents who are from the enhanced DSF. The current paper examines the overall impact of these differences among respondents who are covered by enhanced DSF versus those who are covered by the DSF-only portion of the frame. According to our results, we see some significant differences in key health estimates among respondents who are covered by the enhanced DSF versus those who are not. Specifically, among the examined key estimates, two of the key health estimates significantly differ between DSF-only list and the enhanced DSF list. Therefore, the results indicate that the descriptive information for the key health estimates may differ depending on the source frame. However, overall the source frame

does not seem to have much impact on the relationship between the key respondent characteristics and the key health estimates.

The few differences between key respondent characteristics and the key health estimates based on sample source may indicate that the respondents from the enhanced DSF are less diverse than the respondents who are from the DSF-only portion of the frame. Our results are not sufficient enough to indicate the existence of coverage bias due to the differences among sample members who are covered by enhanced DSF versus those who are from the DSF-only portion of the frame. They may imply that the survey estimates that are obtained from respondents covered by these two types of frames are somewhat different. However, the reader should note that one limitation of our analysis is due to the large portion of the frame, 81.7%, composed of enhanced DSF addresses. More research is necessary to support the idea of using only enhanced frames. One could use randomized experiments which compare the key survey estimates from respondents who are sampled among different lists (enhanced DSF frame versus DSF frame) and investigate how using enhanced frame alone) would potentially influence key estimates. Future studies could also employ validation data to examine the differences in the quality of reports among frames.

5. References

English, Ned, Ipek Bilgen, and Lee Fiorio, L. (2012). "Coverage Implications of Targeted Lists for Rare Populations. *Proceedings of the Joint Statistical Meetings.*

Iannacchione, Vincent G, Jennifer M Staab, and David T Redden. (2003) "Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey." *Public Opinion Quarterly* 67:2: 202-210.

Kennel, Timothy L., and Mei Li. (2009) "Content and Coverage Quality of a Commercial Address List as a National Sampling Frame for Household Surveys." *Proceeding of the Joint Statistical Meetings.*

Link, Michael W., Gail Daily, Charles D. Shuttles, Tracie L. Yancey, and H. Christine Bourquin. (2009) "Building a New Foundation: Transitioning to Address-Based Sampling After Nearly 30 Years of RDD". *Proceedings of the American Statistical Association*, AAPOR [CD ROM], Alexandria, VA: American Statistical Association.

Link, Michael W., Michael P. Battaglia, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad. (2008). "A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) For General Population Surveys." *Public Opinion Quarterly* 72(1), 6-27.