Implementation of Adaptive Design on the Medicare Current Beneficiary Survey

Christopher Ward, Felicia LeClere, Kari Carris, Stephen Cohen, Dean Resnick, Micah Sjoblom, Jennifer Vanicek, Ying Li

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

Abstract

In light of the steady decline in response rates across a range of surveys, interest has grown among survey researchers to establish methods of measuring representativeness of the collected sample. On the Medicare Current Beneficiary Survey (MCBS), we have implemented an adaptive survey design to ensure that the survey respondents are representative of the sample population. The MCBS is a continuous, multipurpose survey of a nationally representative sample of the Medicare population, conducted by the Centers for Medicare & Medicaid Services (CMS) through a contract with NORC. In this presentation, we review key measures used in adaptive design and describe our findings based on data collected during Fall 2017. We also compare this period to the same period in 2016, provide displays that help analysts interpret R-indicators, and discuss interventions proposed to respond to these measures. We discuss the use of R-indicators that measure both the representativeness to the sample population as a whole and also R-indicators for specific variables that are important. We propose practical data collection interventions as a means of implementing real-time adaptive design.

Key Words: Medicare, survey, adaptive design, R-indicators, bootstrapping, variance estimation, representativeness, sample, responsive design, CAPI, longitudinal, field, interview

1. Background

1.1 Introduction to the MCBS

The Medicare Current Beneficiary Survey (MCBS) is a continuous, in-person, multipurpose survey of a nationally representative sample of the Medicare population, conducted by the Centers for Medicare & Medicaid Services (CMS) through a contract with NORC at the University of Chicago. The survey covers a variety of topics, including health care utilization and expenditures, all sources of health insurance coverage, and health status and functioning.

The MCBS employs a rotating panel sample design and represents the population of beneficiaries in the 50 states and District of Columbia. Each sampled beneficiary is scientifically selected as part of an annual panel and is interviewed up to three times (Fall, Winter, or Summer round) per year for four consecutive years to form a continuous profile of their health care experiences. One panel is retired each summer, and a new panel is selected to replace it each fall. Panels in their first round of interviewing are called "incoming" panels; panels in their second through twelfth round are called "continuing" panels[1]. Sampled beneficiaries may be living in the community (e.g., their homes) or a facility (e.g., nursing homes).

1.2 Analysis Background and Definitions

[1] Beginning in 2018, the number of rounds of participation was reduced from 12 to 11 and the last round of interviewing is the winter round, not the summer round.

This paper presents the implementation of an adaptive survey design on the MCBS during the Fall 2017 round of data collection. In contrast to uniform survey design, an adaptive design may adjust the data collection strategy in response to performance measures or metrics. In this paper, we demonstrate the process of monitoring three key performance measures during Fall 2017 data collection of the MCBS: the sample-level R-indicator, the variable-level R-indicator, and the unconditional partial category R-indicator. Each R-indicator is calculated from a model, whose specification is outlined in the next section. We compare these measures against the benchmark R-indicators from Fall 2016 and discuss examples where changes in data collection protocol may be warranted. We then discuss intervention options, which are tailored to the performance measures. Finally, we discuss a few avenues for future research on the use of R-indicators in adaptive design.

2. Methods

2.1 Overview of Adaptive Design

Uniform survey design has long been the standard approach to large-scale surveys. Uniform survey design assumes that all phases of a survey, including sampling, data collection, and post-processing methods, should remain consistent across all phases of data collection and for all sampled persons. This approach allows for standardization and for straightforward measurement of bias and variance of estimates (Schouten, Peytchev, & Wagner, 2017).

This approach to survey design carries some limitations. Groves and Heeringa (2006) proposed responsive design to address underlying problems with the uniform methodology. For example, under the uniform methodology, it is not possible to know the optimal protocol for data collection a priori, as the optimal approach may vary by different segments of sampled persons. Given these considerations, it may be wise to tailor protocols to groups of sampled persons. In contrast to the uniform approach, adaptive design requires a multi-phase structure whereby early phases gather requisite information and later phases integrate that information into improved, targeted protocols whose goals are strong sample representativeness and, likewise, high data quality.

2.2 Why Adaptive Design?

The recent decline of response rates across surveys has been well documented (Galea & Tracy, 2007; Biener et al., 2004; Curtin, Presser, & Singer, 2005). As response declines, the risk of bias to key survey estimates increases (Groves et al., 2011). Further, as response rates decline, their usefulness as a performance metric diminishes. One limitation of response rate is that, in isolation, it reveals little about the quality of the data collected unless response is nearly universal among sampled persons.

Looking beyond the response rate may yield valuable insights into the representativeness of the collected sample, especially when response is not close to 100 percent. By understanding whether the yielded sample is representative, we can implement data collection interventions to target underrepresented persons to help bring the collected sample back into balance. This practice assumes there exists a relationship between data quality and representativeness of the collected sample, and a robust research literature has flourished in recent years to examine this relationship (Wagner, 2012; Schouten, et al., 2012; Wagner, 2010).

A key limitation of uniform survey design is that it assumes an equal propensity to respond among sampled persons. In practice, one's response propensity may vary on a number of dimensions. A sampled person's likelihood to respond may vary by interviewer effects (Davis et al., 2009), mode of data collection offered (Newman et al., 2002), or persistence of interviewer contact attempts (Schouten, Cobben, & Bethlehem, 2009). Adaptive design – in contrast to uniform survey design – acknowledges that flexible data collection methods can respond to these differential propensities to help improve data quality, reduce costs, or both.

Adaptive design, as it is discussed in this paper, is therefore focused on yielding a representative sample (as measured by R-indicators) as a complement to achieving a high response rate. If we can minimize the risk that response propensity covaries with key survey estimates for any one group, then we can minimize measurement bias in those estimates.

2.3 R-Indicator Computation and Interpretation

The R-indicator, or Representativeness Indicator, is one mechanism for understanding whether the collected sample is at risk of introducing bias into key survey estimates. Where there exists substantial variation in response – for example, if women respond at a much higher rate than men – the R-indicator implies that the collected sample is not representative of the sample population (in this case, women are overrepresented among collected sample in comparison to men).

R-indicators are calculated from a model of response propensity (the model used for MCBS is outlined in the "Model Specification" section). The R-indicators are hierarchical and computed at three levels: sample, variable, and variable category. Formulas for computing the various R-indicators are provided in the appendix.

These values provide complementary but distinct perspectives on sample representativeness. The sample R-indicator gives a high-level overview of representativeness and is calculated from the covariates included in the response propensity model. It measures representativeness of the responding sample compared to the sample population. Values range from 0 to 1, where 1 indicates that the collected sample is fully representative of the sample population. Values less than 0.75 indicate that one or more model parameters is unbalanced, and that a data collection intervention may be warranted to return the collected sample back to balance.

The unconditional partial variable-level R-indicator measures the representativeness of the responding sample associated with each variable in the response propensity model (e.g., sex or age). It does not convey information about the overall sample. It is used to ascertain the relative strength of each covariate or variable in the model (e.g., race, age) to predict response propensity. Values range from 0 to 0.5, where a value of 0 indicates that variable is balanced; higher values signify that the variable causes more variation in response propensity relative to other variables in the model. Values greater than 0.125 indicate that an intervention may be warranted.

The unconditional partial category-level R-indicator shows whether a given category or value within a variable (e.g., non-Hispanic Black or age under 65 years) is balanced with other categories within the same variable. In other words, the category-level R-indicator measures which categories of a variable are over- or under-represented in the responding sample. Values range from -0.5 to +0.5. A value of 0 indicates that the category is balanced, while values that deviate from 0 indicate that the variable category causes more variation in response propensity than the other variable categories. Values greater than +0.125 or less than -0.125 indicate that an intervention may be warranted.

2.4 Model Specification

In the response propensity model we specified for MCBS, there were 47 total R-indicators: one sample-level, six variable-level, and 40 category-level. These R-indicators were derived from a model of response propensity, which is described in this section. The model was first developed on the Fall 2016 Incoming Panel (the benchmark round) and then applied to the Fall 2017 Incoming Panel for comparison. The model predicted whether an interview would be completed given the set of sample-frame covariates. Comparisons between these two rounds of data collection appear throughout the report.

2.5 Criteria for Model Variable Selection

The calculation of R-indicators requires a model of response propensity. To select covariates for the model, we considered a number of parameters on which we sought balance (representativeness) in the collected sample. To be eligible for inclusion in the model, a variable must have been available at the outset of data collection. Given that we were interested in assessing the representativeness of the sample collected at the baseline interview (that is, the Incoming Panel), the set of available variables was restricted to those appearing on the sample frame. The final set of sample frame variables included in the model were race, sex, age, Hispanic ethnicity, stratum, and region (Exhibit 1).

Exhibit 1: R-Indicator Model Variables

Variable	Categories
Hispanic Ethnicity	Hispanic Non-Hispanic
Race	White non-Hispanic Black non-Hispanic Hispanic Other non-Hispanic Unknown
Age	45 or younger 45 to 64 65 to 69 70 to 74 75 to 79 80 to 84 85 or older
Sex	Male Female
Region	1. CT, MA, ME, NH, RI, VT 2. NJ, NY, PR 3. DC, DE, MD, PA, VA, WV 4. AL, FL, GA, KY, MS, NC, SC, TN 5. IL, IN, MI, MN, OH, WI 6. AR, LA, NM, OK, TX 7. IA, KS, MO, NE 8. CO, MT, ND, SD, UT, WY 9. AZ, CA, HI, NV 10. AK, ID, OR, WA

2.6 Bootstrap Method for Confidence Interval Estimation

To understand the extent to which random variation may cause R-indicator measures to fluctuate from week to week, it was necessary to estimate variances on the R-indicators. Estimating variances also allowed for the comparison of R-indicators at the same point of data collection (e.g., week 8) across multiple rounds of data collection.

2.6.1 Motivation for Using Bootstrapping Method

A closed-form solution to variance estimation has yet to be published in the research literature. In its place, we developed a bootstrapping procedure to calculate 95% confidence interval estimates for each R-indicator. These calculations were performed on both Fall 2016 and Fall 2017 of data collection to allow for comparison between rounds. The following procedure was conducted to compute each bootstrapped confidence interval:

1. 10,000 samples with replacement were obtained from the panel. The size of each sample was equal to the total size of that panel.
2. From each of the 10,000 simulated samples, R-indicators were calculated using the full-sample base weight.
3. The 2.5th and 97.5th percentiles were computed for each R-indicator among the 10,000 simulated samples. These values bounded an estimated 95% confidence interval for each R-indicator.

2.6.2 Use of Confidence Interval Estimates for Comparisons

In general, we recommend the following interpretation of bootstrapped confidence intervals for R-indicators: in the case that an R-indicator value falls outside a previous value's confidence interval, then that change should be interpreted as significantly different. Otherwise, the change should not be considered to be significantly different. Overlapping confidence intervals imply that the difference is likely due to normal fluctuations. Non-overlapping confidence intervals suggest a substantive difference.

3. Results

3.1 Model Fit Statistics

The Fall 2016 model exhibited acceptable fit statistics. The Hosmer-Lemeshow goodness-of-fit test yielded a p-value of 0.168 with eight degrees of freedom. The percent-concordant rate was 55.5%, with 43.1% discordant and 1.4% tied. Wald and likelihood-ratio rests of the global null hypothesis were all significant at the $p < 0.0001$ level. Response propensities for Fall 2016 were calculated from this model, and those propensities were used in the calculation of the R-indicators. A Fall 2017 model was also established from the same covariates as those in the Fall 2016 model, and the former model was used to compute R-indicators for the Fall 2017 round of data collection.

3.2 R-indicator Results

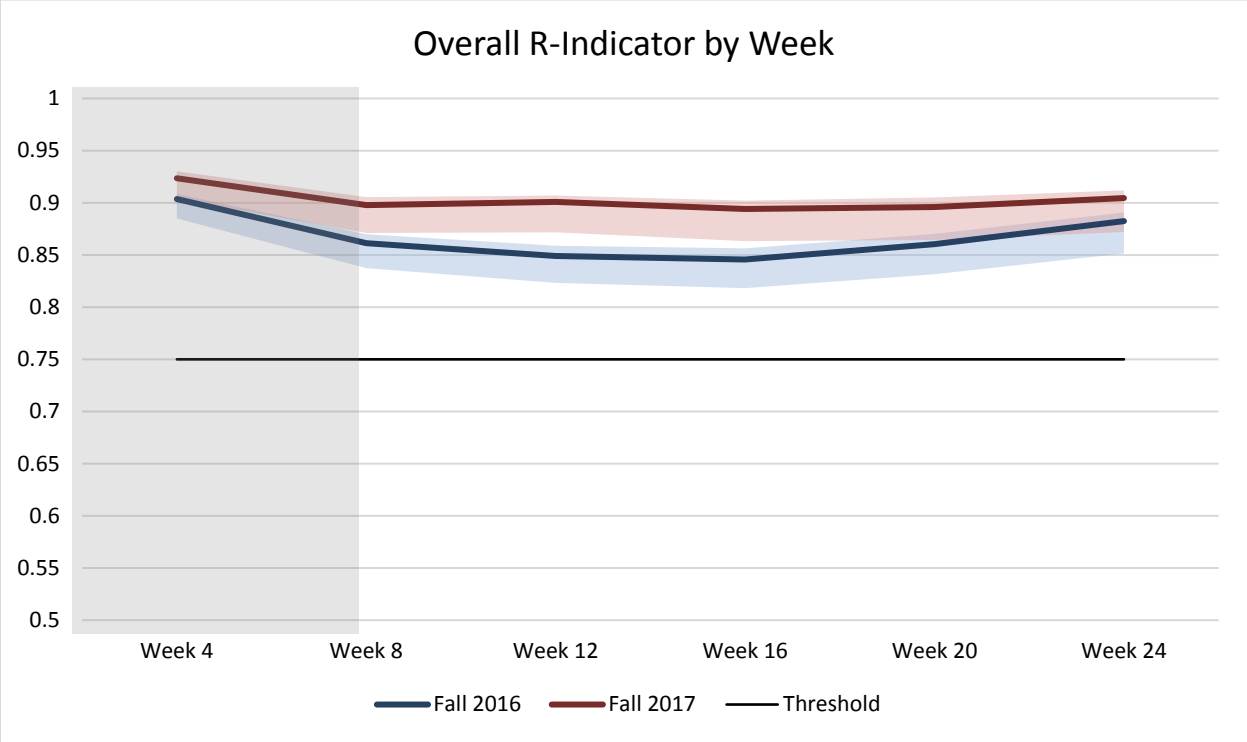
We tracked R-indicators weekly and evaluated them against their respective benchmarks. In the sections that follow, we present a series of graphs which depict the various R-indicator values observed for the Incoming Panels throughout the Fall 2016 and Fall 2017 data collection periods. In each graph, the dark lines trace the R-indicator values, and the lighter-colored regions around them indicate the corresponding 95% confidence interval estimates. The solid black line indicates the threshold against which the R-indicator value was compared.

R-indicator values tended to stabilize over the course of data collection. We advise that R-indicators should be interpreted with caution until approximately halfway through the data collection period. In the case of the MCBS Baseline Interview, approximately half of the targeted number of interviews are collected by the eighth week of data collection (out of a total of 24 weeks). It is for this reason that we have provided the R-indicator values for the weeks leading up to Week 8 of data collection (the point at which approximately half of the targeted number of interviews is collected), but we have shaded in gray those regions of exhibits 2 through 8 as a reminder that R-indicator values prior to Week 8 may be unstable and should be interpreted with caution.

3.2.1 Sample R-indicators for Fall 2016 Incoming Panel and Fall 2017 Incoming Panel

Exhibit 2 presents the overall sample R-indicator values over time in Fall 2016 and Fall 2017 of data collection. The dark lines trace the R-indicator values, and the lighter-colored regions around them indicate the corresponding 95% confidence interval estimates. The black line indicates the threshold against which the R-indicator values are compared; for the overall sample R-indicator, a value of at least 0.75 indicates a minimum threshold of representativeness.

Exhibit 2. Overall Sample R-Indicator by Week of Data Collection

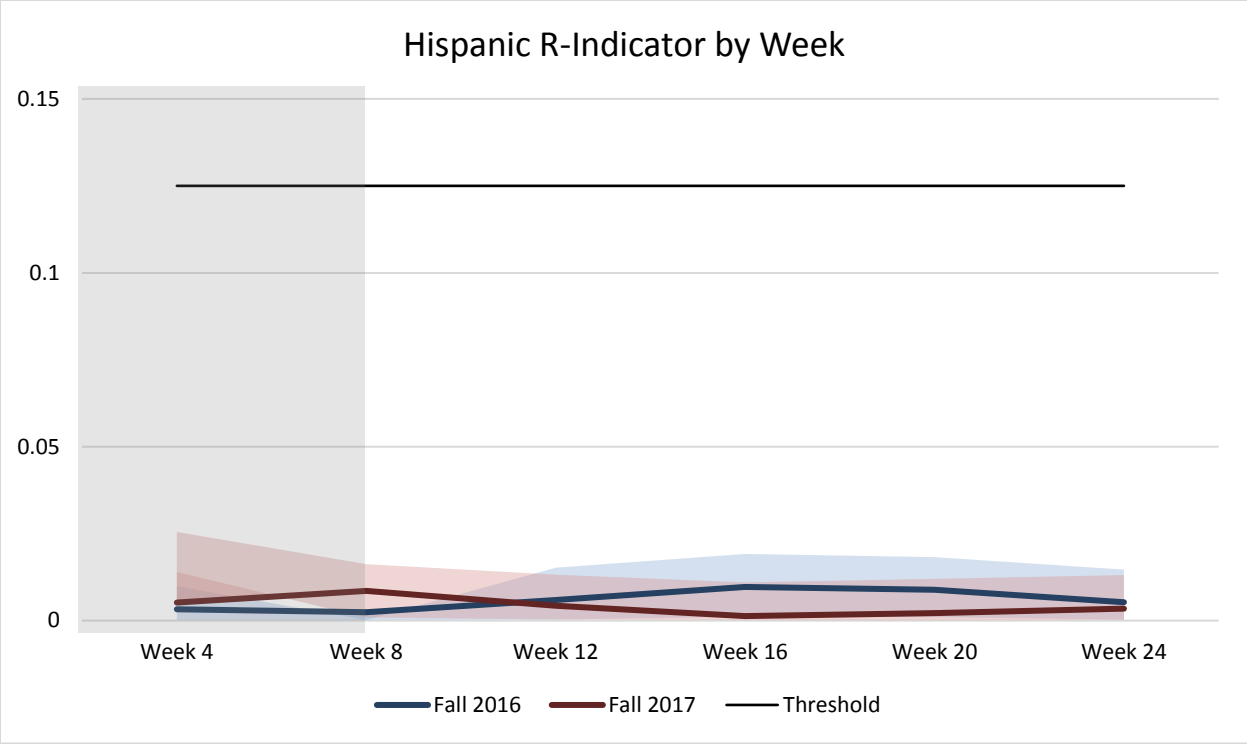


In all observed weeks of data collection for Fall 2016 and Fall 2017, the overall R-indicator exceeded the target benchmark of 0.75. In Fall 2017 of data collection, the overall R-indicator was higher at each interval of data collection than in Fall 2016 of data collection. By week 16 of data collection, the confidence intervals of the Fall 2016 and Fall 2017 R-indicators did not overlap, which suggests that the sample collected in Fall 2017 was more representative with respect to the Fall 2017 model parameters compared to that collected in Fall 2016 during those weeks of data collection. Because they were bootstrapped, confidence intervals were asymmetrical for all observed periods of data collection.

3.2.2 Unconditional Partial Variable-Level R-indicators for Fall 2016 Incoming Panel and Fall 2017 Incoming Panel

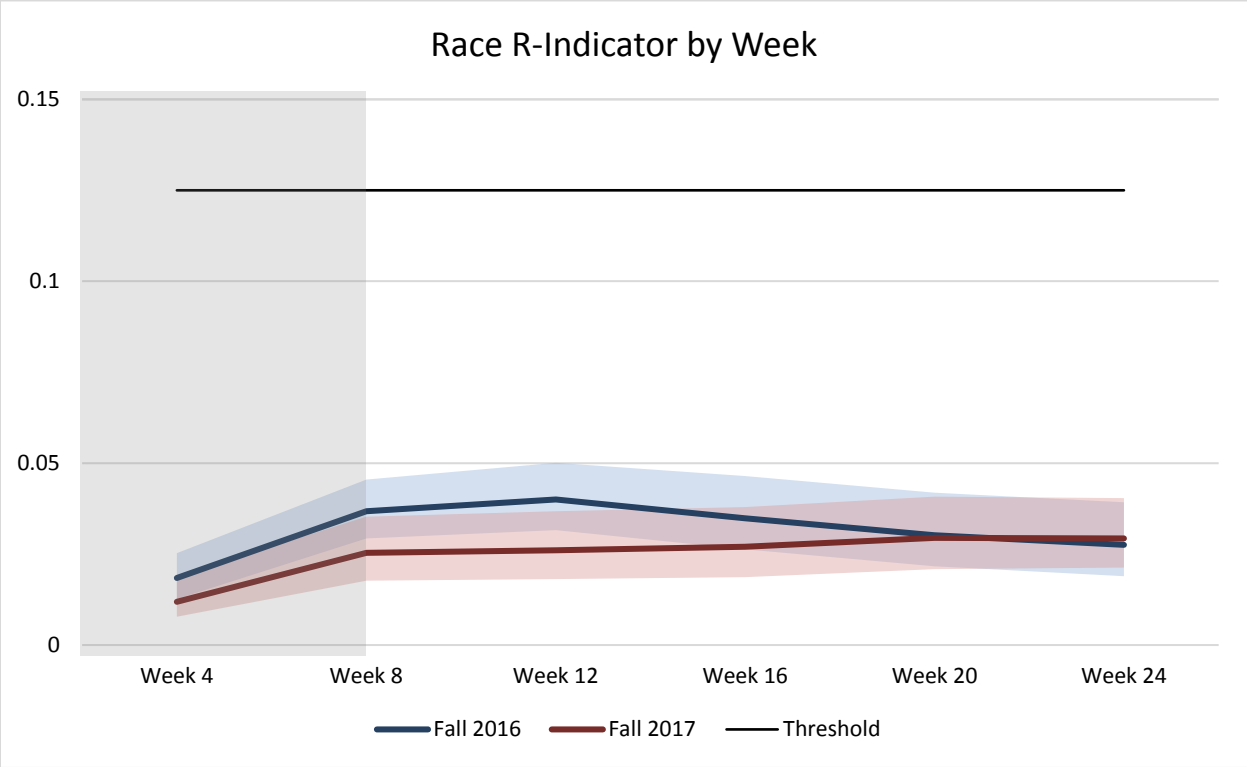
Exhibits 3 through 8 present the unconditional partial variable-level R-indicators for each model variable over time in Fall 2016 and Fall 2017. The dark lines trace the R-indicator values, and the lighter-colored regions around them indicate the corresponding 95% confidence interval estimates. The black line indicates the threshold against which the R-indicator values are compared. For the unconditional partial R-indicator, a value less than 0.125 indicates a desirable level of representativeness on that model parameter.

Exhibit 3. Hispanic Variable-level R-indicator by Week of Data Collection



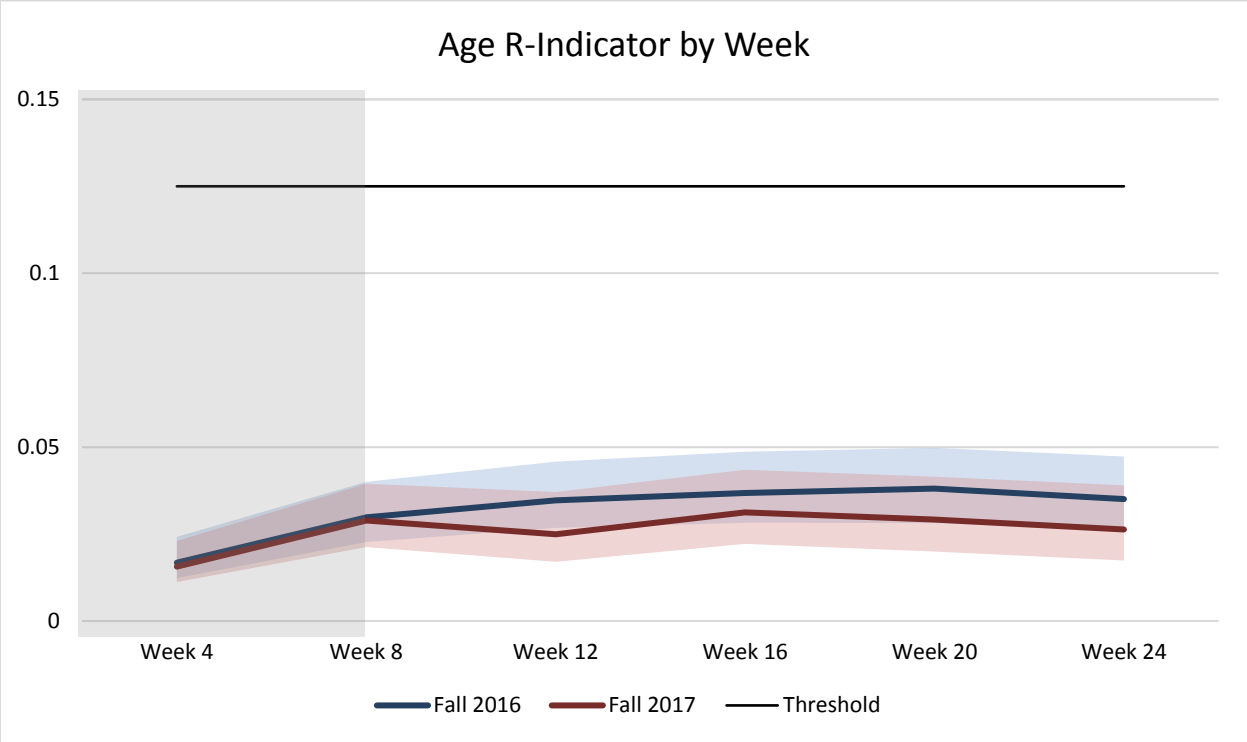
The Hispanic variable-level R-indicator was well below the benchmark of 0.125 for all observed weeks of data collection in both Fall 2016 and Fall 2017 of data collection. The level of representativeness with each round's model was comparable through the remaining weeks of data collection. In contrast to the overall R-indicator, confidence intervals for the partial variable R-indicators are generally symmetrical like a standard confidence interval.

Exhibit 4. Race Variable-level R-indicator by Week of Data Collection



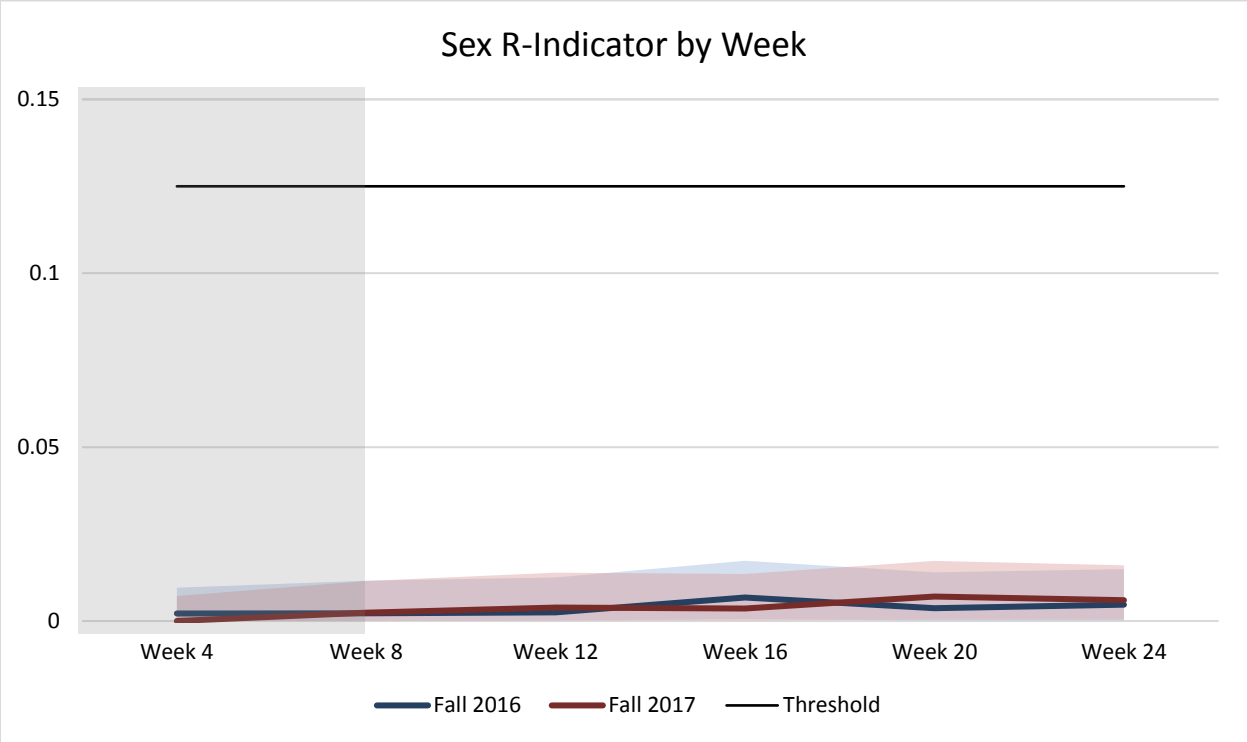
Representativeness with respect to race was very good in Fall 2016 and Fall 2017 of data collection. Values for variable-level R-indicators for race were modestly, but not significantly, lower through week 16 of data collection in Fall 2017 compared to the same period in Fall 2016 of data collection. The observed values at weeks 20 and 24 were nearly identical. All observed values fell below the benchmark value of 0.125.

Exhibit 5. Age Variable-level R-indicator by Week of Data Collection



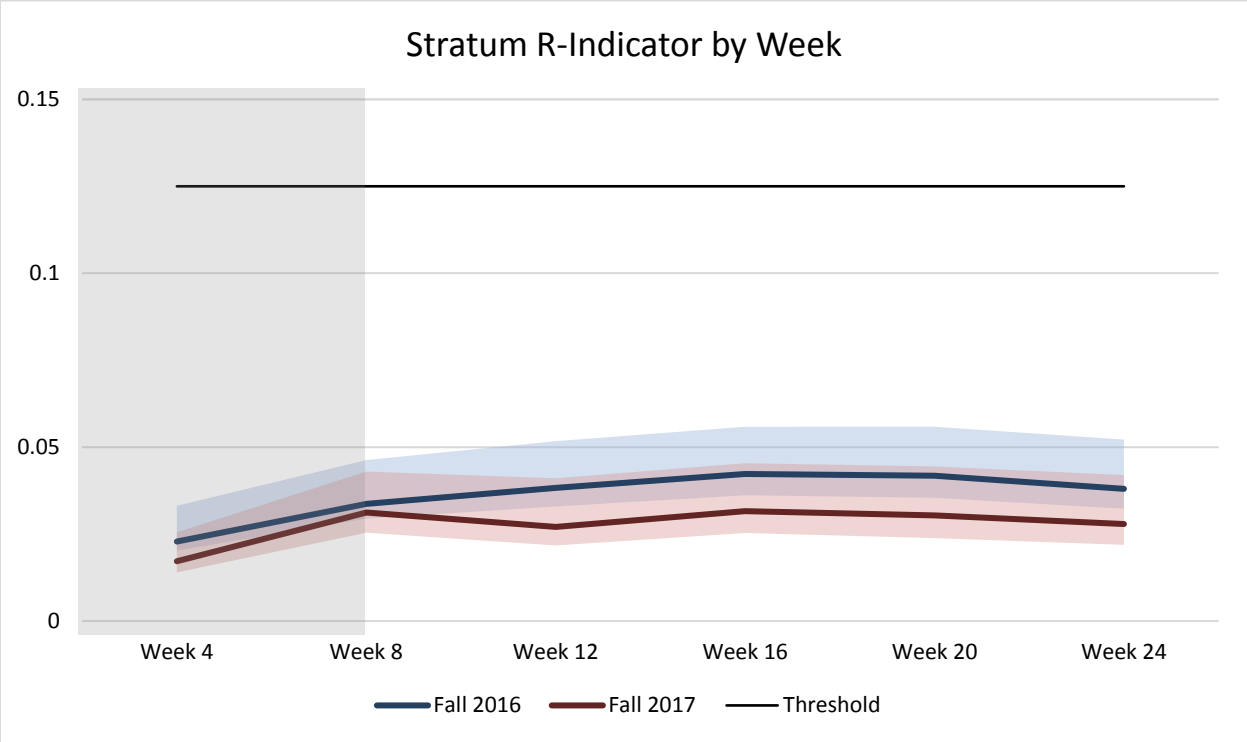
The age variable-level R-indicator was well below the threshold benchmark of 0.125 for all observed weeks of data collection in Fall 2016 and Fall 2017. The level of representativeness observed in Fall 2017 was nearly identical to that of Fall 2016 through week 8 and improved modestly into week 12 (i.e., the R-indicator value is lower for Fall 2017 at week 12). Fall 2016 and Fall 2017 of data collection exhibited comparable levels of representativeness with respect to age throughout data collection.

Exhibit 6. Sex Variable-level R-indicator by Week of Data Collection



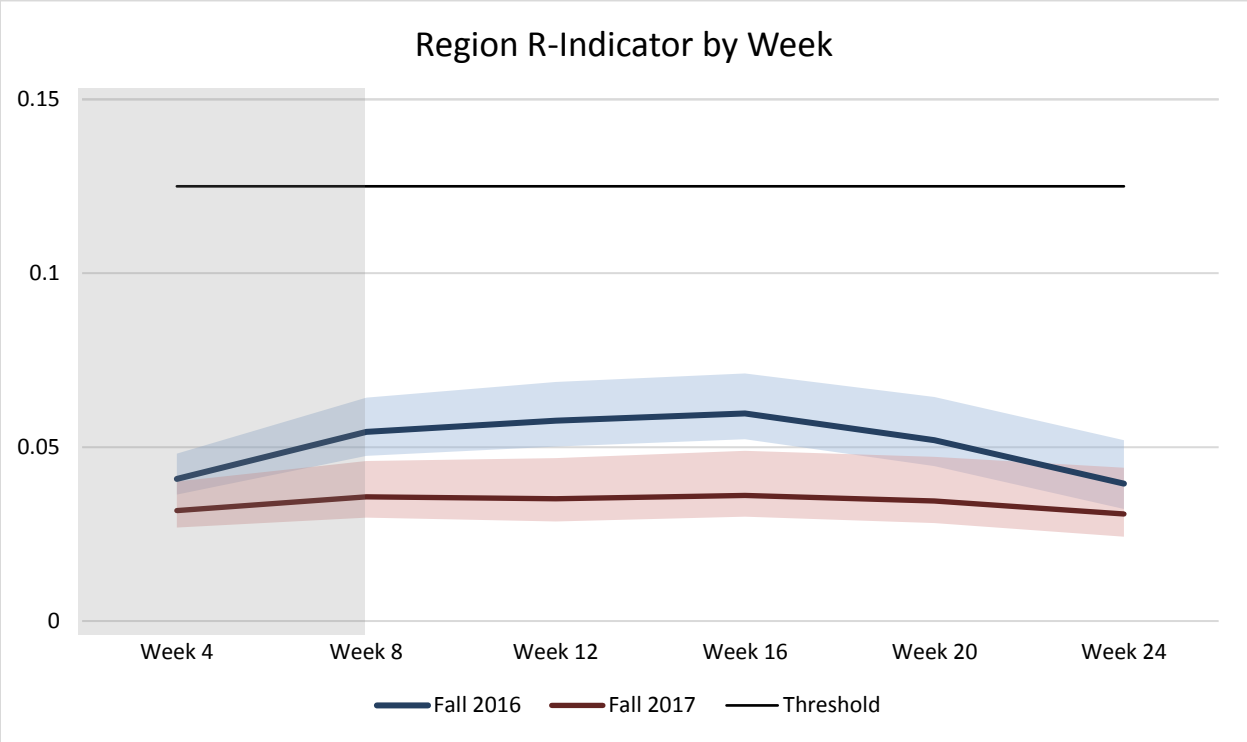
Representativeness of the sample based on sex of the beneficiary was excellent with respect to the propensity models used in Fall 2016 and Fall 2017 of data collection. Through week 16 of Fall 2017 and across all weeks of Fall 2016, the variable-level R-indicator for sex was nearly 0, which indicated a good balance between men and women in the collected sample.

Exhibit 7. Stratum Variable-level R-indicator by Week of Data Collection



The stratum variable-level R-indicator was well below the threshold benchmark of 0.125 for all observed weeks of data collection in Fall 2016 and Fall 2017. The level of representativeness observed in Fall 2017 was slightly, but not significantly, better than that observed in Fall 2016 through all weeks of data collection.

Exhibit 8. Region Variable-level R-indicator by Week of Data Collection



This variable-level R-indicator suggests that collected sample was more representative with respect to region in Fall 2017 as of week 16 compared to the same point in data collection in Fall 2016. By week 20, the confidence intervals of the R-indicators overlapped, which indicates that Fall 2016 and Fall 2017 exhibited similar levels of performance on this indicator by the end of data collection. All observed R-indicator values in both rounds fell well below the target benchmark of 0.125, which indicated the collected sample was acceptably balanced with respect to region.

4. Discussion and Adaptive Intervention Techniques

4.1 Criteria to Activate Adaptive Intervention

For each type of R-indicator, there were different thresholds against which we evaluated the representativeness of the collected sample. For the overall sample R-indicator, a value greater than 0.75 suggested that the overall composition of beneficiaries who completed interviews is reasonably representative of the target sample. For the unconditional partial variable-level R-indicator, a value less than 0.125 is preferable. For the unconditional partial category-level R-indicator – which may take on positive or negative values – a value between -0.125 and +0.125 was desirable. Were observed R-indicators to fall outside these ranges, we may have sought data collection interventions to yield a more representative sample.

In Fall 2016 and Fall 2017, we did not observe R-indicators outside these thresholds; consequently, no data collection interventions were implemented to improve the representativeness of the achieved sample. In the discussion that follows, we present examples of paradata elements that are monitored throughout data collection and the possible interventions that could be implemented in response to R-indicator values deviating from the established thresholds.

4.2 Paradata Monitoring

Weekly monitoring of R-indicators allows for early identification of any problems with sample representativeness. In the event that one or more R-indicator values violates a threshold criterion, we could intervene in an attempt to correct the imbalance. Paradata analysis could help pinpoint which interventions might be most effective. We have listed below a number of paradata elements that might be monitored throughout data collection. They were selected for their relevance to data collection costs, data collection quality, and their ability to indicate aberrations in data collection. These paradata were refreshed on a daily basis, which means that any problems with data collection could be diagnosed in the event that an R-indicator threshold criterion were violated. These include:

1. Average number of contact attempts
2. Rate of field manager intervention
3. Rate of refusal
4. Number of days sampled beneficiary has spent in data collection
5. Rate of appointment-making

Once a data collection aberration is identified, we could devise interventions to correct a sample imbalance. These strategies should be identified prior to the fielding period, be easily implemented by the field, and reflect interviewer or respondent behavior that is modifiable (e.g., directing interviewers to increase their contacts to a particular segment of the sample is a behavior that they can easily change).
Example strategies or interventions that could be implemented to correct a sample imbalance on MCBS include:

1. Redirecting efforts to segments of the population that are out of balance
2. Adjusting case assignments to other interviewers in the same sample area
3. Modifying the frequency or timing of contacts
4. Altering the mode of contact (e.g., in-person visits, phone calls, Express mail delivery of letters describing study)

Other interventions or strategies, such as providing a respondent incentive or fielding a “critical items” questionnaire rather than the entire questionnaire to a segment of the population are strategies that might be implemented on MCBS too, but those types of intervention would require alterations to the project protocol and require Institutional Review Board and Office of Management and Budget approval; they would also have budget implications and require advance project planning to implement.

4.3 Paradata Monitoring to Inform Data Collection Intervention

The set of interventions available is closely related to the paradata being monitored. This is where the adaptive aspect of adaptive survey design comes into the fore. For example, if non-Hispanic Black beneficiaries were found to be underrepresented in the collected sample, we would evaluate paradata trends for those beneficiaries. If, for example, the number of call attempts for those beneficiaries is low, then we may ask interviewers to make additional call attempts.

In this section, we describe examples of paradata monitoring from Fall 2016 data collection and how they could inform methods of intervention. Should an R-indicator suggest that an intervention is necessary, it would be important to assess the paradata for “pending” sample (i.e., not yet completed or otherwise finalized) because any data collection intervention could be implemented only on pending cases. Exhibits 9 and 10 illustrate how the number of contact attempts by race could be monitored for pending cases using data from the Fall 2016 Incoming Panel.

Exhibit 9 shows the percentage of pending sample that falls into various categories of race over the weeks of data collection for Fall 2016. Exhibit 10 shows the mean number of contact attempts for pending sample by week of data collection for each category of race.

Taken together, these two graphs show the percentage of cases in each category of race that are pending at a given week of data collection and the average number of contact attempts. If the R-indicators had suggested that the Fall 2016 sample was imbalanced by race, these graphs could have been used to suggest which types of cases might need additional contact attempts to improve the representativeness of the collected sample, and to assess the volume of pending sample that could receive those additional contact attempts. As noted above, however, the Fall 2016 sample was not imbalanced and there was no need to implement an intervention.

Exhibit 9. Percentage of Pending Sample by Race

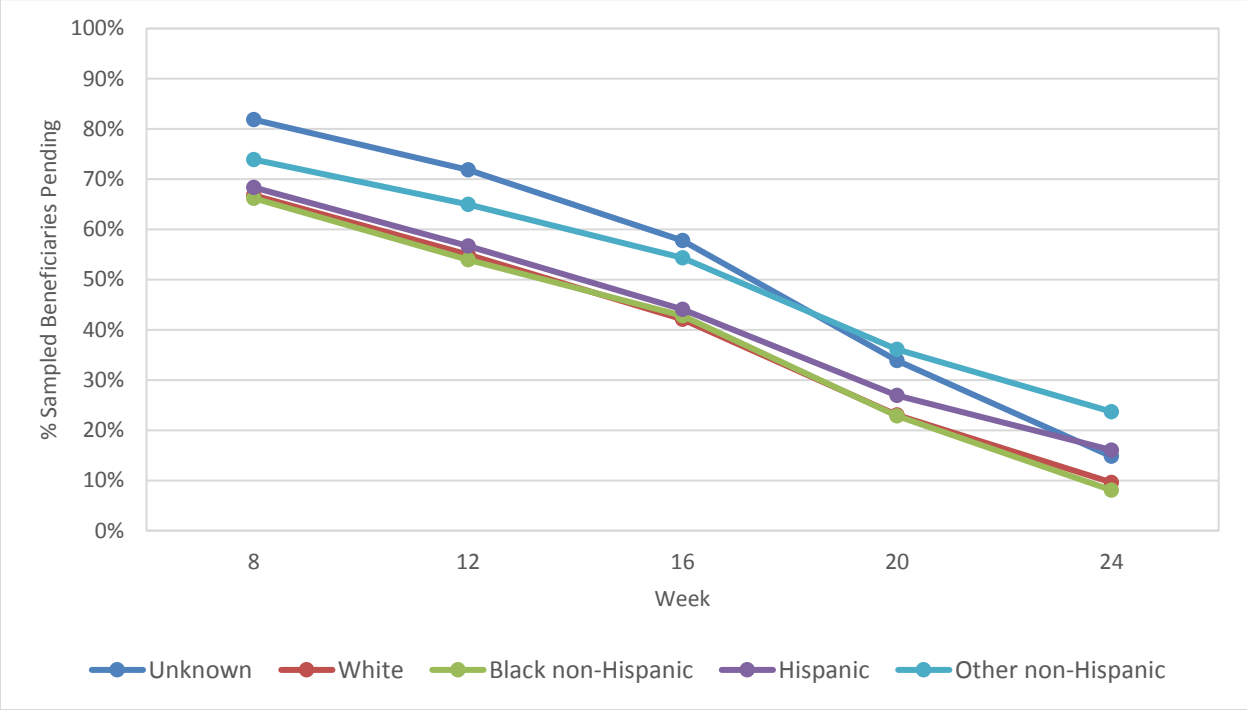
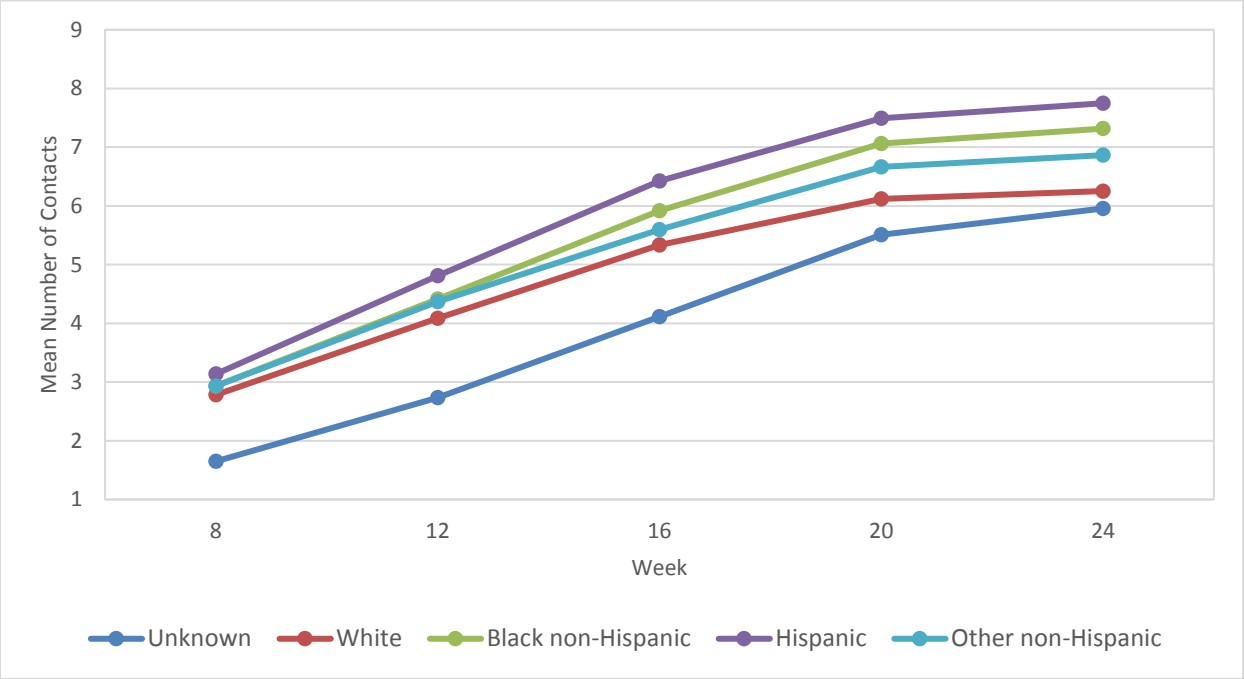


Exhibit 10. Mean Number of Contacts among Pending Sample by Race



5. Areas for Future Research

R-indicators are a promising metric of sample quality, and we expect to explore further their use in evaluating sample quality in longitudinal surveys such as MCBS. For the MCBS in 2017, we focused our efforts on applying adaptive design techniques to and calculating R-indicators for the Incoming Panel. This section outlines promising avenues for new research in 2018 related to adaptive design. We discuss methods for computing R-indicators on continuing sample and a technique for computing variance estimates with the replicate weights.

5.1 Avenues for Future Research: R-indicators on Continuing Sample

Our current R-indicators work has been limited to the Incoming Panels fielded in Round 76 and Round 79. The R-indicators afford an additional lens to evaluate the representativeness of the collected sample when evaluated in conjunction with response rate. The composition of the baseline sample of beneficiaries provides a natural benchmark against which the representativeness of collected sample can be readily evaluated. For continuing sample, however, the benchmark is less well defined. One area for future research is to develop a method of calculating R-indicators on continuing sample. A first step in that line of research is to devise a definition of the benchmark of representativeness against which R-indicators can be calculated.

5.2 Closed-form Variance Estimation of R-indicators

The bootstrapping procedure we proposed for variance estimation allows for estimates of a 95% confidence interval, but it does not take advantage of the replicate weights. In future work, we propose the following alternate approach for variance estimation using the replicate weights:
Obtain the 2016 panel full-sample and replicate base weights. These would not normally be created at the outset of when data collection begins, but in future years it will be possible to derive them at that time.

For each R-indicator, 100 R-indicators are calculated using the full-sample weight and each of the 100 replicate base weights. This yields 100 total derivations of each R-indicator of interest.
Finally, the following formula is used to calculate the estimated standard error of each R-indicator using Fay's balanced repeated replication method (Judkins, 1990):

$$var(\hat{y}) = \frac{1}{K(1-0.3)^2}\sum_{k=1}^{K}(\hat{y} - \hat{y}_k)^2$$

where K = 100 (for the 100 replicates), $\hat{y}$ is the calculated R-indicator using the full-sample base weight, and $\hat{y}_k$ is the calculated R-indicator for each of the 100 replicates, and 0.3 is the Fay coefficient. The square root of this estimated variance is the estimated standard error, which can be used to conduct statistical tests or construct confidence intervals as usual.

Acknowledgements

The research in this article was supported by the Centers for Medicare & Medicaid Services, Office of Enterprise Data and Analytics under Contract No. HHSM-500-2014-00035I, Task Order No. HHSM-500-T0002 with NORC at the University of Chicago. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of NORC at the University of Chicago or the Centers for Medicare & Medicaid Services. Research from the analysis was presented at the 2018 Joint Statistical Meetings in Vancouver, BC, Canada.

References

Biener, L., Garrett, C. A., Gilpin, E. A., Roman, A. M., & Currivan, D. B. (2004). Consequences of declining survey response rates for smoking prevalence estimates. American journal of preventive medicine, 27(3), 254-257.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. Public opinion quarterly, 69(1), 87-98.

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H., & Resnicow, K. (2009). Interviewer effects in public health surveys. Health education research, 25(1), 14-26.

Galea, S., & Tracy, M. (2007). Participation rates in epidemiologic studies. Annals of epidemiology, 17(9), 643-653.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). Survey methodology (Vol. 561). John Wiley & Sons.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(3), 439-457.

Judkins, D. R. (1990). Fay's method for variance estimation. Journal of Official Statistics, 6(3), 223.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. Survey Methodology, 35(1), 101-113.

Schouten, B., Peytchev, A., & Wagner, J. (2017). Adaptive Survey Design. CRC Press.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Schlomo, N., & Skinner, C. (2012). Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response through R-Indicators and Partial R-Indicators. International Statistical Review, 80(3), 382-399.

Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. Public Opinion Quarterly, 74(2), 223-243.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. Public Opinion Quarterly, 76(3), 555-575.

Appendix: Formulas for Calculating R-indicators

1. **Overall Sample R-indicator**

$$R(\hat{\rho}) = 1 - 2\left(\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\hat{\rho}_i - \hat{\bar{\rho}})^2}\right)$$

where s_i indicates whether the case is in the sample; π_i indicates probability of selection; $\hat{\rho}_i$ is the estimated response propensity; and $\hat{\bar{\rho}}$ is the average response propensity over the entire sample.

2. **Unconditional Partial Variable-Level R-indicator**

This value is obtained by computing the variation in the propensities across the categories or domains, $k = 1,..,K$ defined by the variable x in question following equation:

$$R_u(x,\hat{\rho}) = \sqrt{\sum_{k=1}^{K}\frac{N_k}{N}(\hat{\bar{\rho}}_{x,k} - \hat{\bar{\rho}}_x)^2}$$

In this formula, k indexes the categories of variable x; $\hat{\bar{\rho}}_{x,k}$ is the estimated response propensity of category k; $\hat{\bar{\rho}}_x$ is the average response propensity across the categories.

3. **Unconditional Partial Category-level R-indicator**

$$R_u(x,k,\rho) = \sqrt{\frac{N_k}{N}}(\bar{\rho}_{x,k} - \bar{\rho}_x)$$

In this formula, N_k refers to the population size for the stratum for category k.