Improvements in the MCBS Sample Design

Kirk Wolter, Whitney Murphy, and Nicholas Davis

NORC at the University of Chicago

I. Introduction to the MCBS

The Medicare Current Beneficiary Survey (MCBS), sponsored by the Centers for Medicare and Medicaid Services (CMS) in partnership with the Center for Medicare and Medicaid Innovation (CMMI), is a continuous, in-person, longitudinal survey of a representative national sample of the Medicare population. Linked to Medicare claims data, the survey was designed to aid CMS in administering, monitoring, and evaluating the Medicare programs. It has been carried out continuously for more than 20 years, encompassing more than one million interviews. The MCBS is the leading source of information and analysis of Medicare and its impact on beneficiaries and plays an essential role in monitoring and evaluating key provisions of the Affordable Care Act (ACA). Two key sets of data result from the MCBS Survey: a person-level Access to Care (ATC) data file, and a heath care event-level Cost and Use (CAU) data file. Historically, MCBS samples have been drawn using a stratified multistage probability sample design, which covers the population of beneficiaries in the 50 states, District of Columbia, and Puerto Rico.

This paper reports on three recent improvements made to the MCBS sampling design. First, we discuss a redesign of the second stage of sampling implemented in 2014. Instead of the survey's historical practice of using ZIP-code areas as second stage units (SSUs), we selected a set of Census tracts or tract-based clusters to serve as SSUs. The use of Census tracts in place of ZIP-code areas reduces the burden of maintaining the SSUs over time and allows for easier merging of MCBS data with Census and other aggregate-level geographic or environmental data. Second, we report on an expansion of the MCBS sampling frame being implemented in 2015 to include persons who turn age 65 and enroll in Medicare on or prior to December 31 of the current year. Historically, enrollees would not be sampled and interviewed until the fall of the year following their year of enrollment, and their data would not be compiled until two years after the year of enrollment. This expansion will permit CMS to release MCBS data products up to a year earlier than in the past. Finally, we give methods of oversampling Hispanic and Asian Medicare beneficiaries being implemented in 2015 and 2016.

II. Historical Sampling Design

The MCBS has used a rotating panel design in which a new sample, or *panel*, is selected each year to replace an outgoing panel. Each panel is retained in the study for the full four years specified under the MCBS sample rotation scheme and is designed to: (a) replace approximately one-third of the respondents in the existing MCBS sample; and (b) extend coverage to persons added to the Medicare rolls during the previous year. Under the rotating panel design, each case is interviewed a total of 12 times, including three times per year spread over a four-year period. The historical procedure has been to determine the sampling frame for the current year panel to include all beneficiaries who enrolled by January 1 of the current year. This means the sampling frame includes beneficiaries who enrolled in the prior year or in earlier years, but does not include current year enrollees.

Sampling for the new annual panel is conducted in spring and initial interviews are conducted in the fall round. The overall sample interviewed at the fall round encompasses the newly selected annual panel and three continuing panels of beneficiaries selected in the previous three years. One continuing panel is retired in the summer round of each year.

Through 2013, panels were selected using a stratified multistage probability sample design. At the first stage of sampling, 107 large geographical areas, called primary sampling units (PSUs), were selected using probability proportional to size sampling procedures. The current set of PSUs was selected in 2001. At the second stage of sampling, ZIP code areas (referred to as "ZIP clusters") were selected within the 107 PSUs. The third and final stage of sampling for a panel consisted of a stratified random selection of beneficiaries within the sampled ZIP clusters. Seven age strata were used as follows: under 45, 45 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, and 85 and older.

While ZIP codes (and fragments and clusters thereof) have historically been used as SSUs in the MCBS, there are several reasons why they are particularly ill-suited to the task of design, conduct, and analysis of sample surveys. First, ZIP codes are not areal features; instead, they are a changing collection of mail delivery routes. Because they are constantly changing and new ZIP codes are being created, they require the MCBS to conduct expensive maintenance of SSUs. For this reason, SSUs can be defined differently from year to year. In fact, between 1992 and 1999, 356 ZIP clusters (including 993 ZIP fragments) were added to the sample to compensate for new or changing ZIP boundaries and to improve coverage (OMB Information Collection Request (ICR) 2010), and between 2001 and 2012, an additional 48 zip clusters (including 108 zip fragments) were added. Furthermore, it is difficult to merge census or other environmental data that uses Census geographies onto MCBS data files for analytical purposes. Census geographies do not overlay ZIP code geographies. And even while the Census Bureau has constructed "ZCTAS" (ZIP code tabulation), they are merely generalizations of ZIP codes, and not all ZIP codes are included. The fact that ZIP code boundaries change often means that an address in a particular ZIP code in one data collection round could be in a very different one demographically in a future round. And further, ZIP codes are relatively large areas and thus may encompass heterogeneous environments.

III. New Sampling Design

Beginning in 2014, we improved the sampling design by shifting from the old ZIP-based SSUs to new tract-based SSUs. Census tracts are generally smaller and relatively more permanent subdivisions of counties. The entire U.S. is divided into tracts. Tract boundaries rarely change, and when changes do occur, they are made only once a decade, following a decennial census. As such, SSUs based on census tracts will need little or no maintenance for the MCBS. Further, because the SSUs would be constructed from actual census geographies, the merging of census or environmental variables onto the MCBS data files for analytic purposes would be simple and direct. Sampling based on Census tracts rather than the previous approach that used fragments of ZIP codes will also ensure greater consistency and harmonization of geography over time. It will take four years of selecting the new fall panel for CMS to fully realize the efficiencies of this innovation. Three continuing panels based on ZIP SSUs will be retired, one per year, during this period.

Geocoding played a major role in the construction of the SSUs. In order to determine the measures of size and select SSUs, the Medicare Enrollment Database (EDB), a

comprehensive master list of all active beneficiaries and used as the basis for the sampling frame at the third stage of sampling, was geocoded to the tract level. We used a two-step geocoding process to accomplish this. First, a list of all ZIP codes intersecting the 107 PSUs was created, and all beneficiary records for which the address was classified in one of these ZIP codes were extracted from the EDB. Second, we used the GIS tool MapMarker Plus™ to standardize the address fields for all addresses in the extraction and geocode them to the tract level.

EDB counts at the tract level were used to construct SSUs and calculate SSU measures of size: the weighted sum across the seven age strata of the number of Medicare beneficiaries in each stratum. Tracts that were large enough to meet the minimum measure of size comprised single-tract SSUs. For tracts that were not large enough, each tract was combined with the adjacent tract having the smallest land area. Collapsing continued in tract number order, which approximates a serpentine sort, until all SSUs were large enough to meet the minimum measure of size. A total sample of 703 SSUs was selected from the 107 PSUs, consisting of a proportional allocation of 242 SSUs to the 29 certainty PSUs subject to a minimum of 6 SSUs per PSU, and an equal allocation of 6 SSUs to each of the 76 non-certainty PSUs. There were three noncertainty SSUs that contained 6 or fewer SSUs, in which case all of the SSUs in the PSU were selected.

IV. Sampling Current-Year Enrollees

The current MCBS design has historically resulted in delivery of data products containing information about the cost and use (CAU) of health care services in reference year t during the middle of the year two years later (year $t + 2$). Such late delivery arises because the year *t* cohort of beneficiaries, which contributes to the cost and use of health care services in reference year t, was not even sampled until year $t + 1$ and not initially interviewed until the fall round of year $t + 1$.

Beginning in 2015, we implemented a second sample design innovation, in which the year *t* cohort[1] of beneficiaries was included in the sampling frame of beneficiaries from which the year *t* panel[2] is selected. Under the historical MCBS system, this cohort is not sampled until a full year later, in year *t*+1. Members of this cohort become eligible for Medicare through age, disability, or other specific health issues. The largest of these groups is made up of beneficiaries who attain Medicare entitlement through age eligibility; these people will have reached their 64th birthdays by January 1st of year *t*. This group becomes eligible for Medicare benefits at some time during year *t*, and thus is not included in a traditional panel until year *t*+1, and is not represented in the traditional year *t* CAU files until year *t*+2. A minority of the year *t* cohort is made up of members who become eligible for Medicare through disability, and will become eligible for Medicare benefits at some time during year *t*. Thus, similar to the age-eligible portion, this group is not included in a traditional panel until year *t*+1, and is not represented in the traditional year *t* CAU files until year *t*+2.

In 2015, we included these groups in the year *t* sampling frame for panel selection for the first time. The set of beneficiaries from the current year cohort is similar in size to the

[1] An annual cohort is the set of beneficiaries that are enrolled in Medicare and appear on the Medicare Enrollment Data Base (EDB) within a given year.

[2] An annual panel is the set of beneficiaries sampled in a given year and initially interviewed in the fall round of that year.

corresponding set of current-year beneficiaries encountered in the historic MCBS design, around 300 to 400 cases, and is thus a relatively small part of the new panel. Because this new approach to sampling includes the year *t* cohort in the MCBS sample one year earlier, it will allow the delivery of CAU data up to one year earlier than previously feasible, which is expected to be of considerable value to CMS and other users of MCBS data.

Exhibit 1 compares the timing of sampling and data delivery for the old design to the new design, which samples current-year enrollees. A detailed description of the sampling of future beneficiaries follows.

Exhibit 1: Comparison of 2015 Cohort Sampling and Interviewing under the Old and New Designs

Event	2015 Cohort Hypothetically under the Old MCBS Design	2015 Cohort Actually under the New MCBS Design
First Sampled	2016	2015[1]
Access to Care Interview	Fall 2016 (R76)	Fall 2015 (R73)
Delivery of 2015 Cost and Use Files	Winter/Summer 2017	Winter/Summer 2016

[1] A very close approximation of the 2015 cohort would be sampled in 2015

Timing of the Interview. Members of the year *t* cohort of beneficiaries sampled under the new design will all be enrolled in Medicare sometime during sampling year *t*. Because we expect better cooperation after these individuals become eligible and have a connection to Medicare, and because the interview is geared toward those who are already enrolled, these sampled individuals are interviewed only after they are enrolled. The majority will become eligible and enroll before Fall interviewing begins; for those not enrolled until after September 1, we conduct an interview with the sampled person after he or she enrolls in Medicare (i.e., on or after their enrollment date in the EDB file).

Sampling Frame. The inclusion of current year enrollees in the sampling frame requires additional steps to be taken in the building and sampling of the frame. Under the old design, one extract of the EDB was required for sampling. In the spring of year *t*, an extract of the EDB containing all beneficiaries who had enrolled by January 1 of year *t* was used as a base for the construction of the sampling frame. From that extract, beneficiaries falling within the selected MCBS PSUs and SSUs were identified, and ultimately this subset of beneficiaries constituted the frame from which the year *t* panel was selected. Including current year enrollees introduced some new challenges, since not all year *t* enrollees are included in the EDB by the spring of year *t*, when sampling occurs. Instead, year *t* enrollees are added to the EDB in two distinct manners. First, beneficiaries who will be automatically enrolled in Medicare appear on the EDB up to four months prior to their automatic enrollment. These beneficiaries can be included in the frame (and ultimately the sample) up to four months before they are enrolled, but will not be interviewed until after their

enrollment date. Second, beneficiaries who self-enroll appear on the EDB within a month[3] after their enrollment in Medicare. Thus, someone enrolling in December of 2015 may not appear on the EDB until January of 2016.

In April, when the EDB extract is pulled to facilitate sampling for the fall round, only a portion of the current year enrollees will be included on the EDB. Beneficiaries who enrolled by April 1 of year *t* or who will be automatically enrolled within four months of April (i.e., by August 1 of year *t*) are included in the EDB extract. However, any beneficiary who self-enrolls after April 1 or is automatically enrolled after August 1 of year *t* will not yet appear on the EDB. Thus, multiple EDB extracts are required to facilitate sampling of the full year *t* cohort. Two additional EDB extracts are pulled each year and contribute to the year *t* sampling frame: (1) an extract in August, which contains additional self-enrollees through August 1 of year *t* and scheduled automatic enrollees through December 1 of year *t*; and (2) an extract in October, which contains additional self-enrollee's through October 1 of year *t* and scheduled automatic enrollees through January 1 of year *t+1*. The October extract is scheduled for the latest date possible to facilitate sampling and fielding in year *t*; however, it will still leave a slight undercoverage of any self-enrollees between October 2 of year *t* and January 1 of year *t+1*. A final exploratory extract is scheduled for mid-January of year *t+1* to identify this undercoverage and account for it in weighting adjustments.

V. Oversampling Race/Ethnic Minorities

Both the Office of Minority Health and the Office of Enterprise Data and Analytics at the CMS recognize that the current MCBS sample is not large enough for precise estimates of health disparities experienced by specific race/ethnicity subpopulations of interest. Efforts were undertaken beginning in 2015 to focus on two of these populations: beneficiaries of Hispanic, Latino/a, or Spanish[4] origin and beneficiaries of Asian origin. By oversampling these populations, additional analyses could be accomplished both by CMS and the many data users interested in health disparities among these unique populations.

The main goals of the oversampling are to increase the number of Hispanics and Asians in the MCBS enough to allow for precise estimates of health disparities experienced by these populations, and in the case of the Hispanic population, to reduce the proportion of MCBS Hispanic beneficiaries who reside in Puerto Rico and thereby increase the proportion of MCBS Hispanic beneficiaries from outside Puerto Rico. Oversampling of beneficiaries is facilitated by the availability of a race classification code for beneficiaries on the EDB. The race classification code, a variable used to predict race (including Hispanic origin) for Medicare beneficiaries on the EDB frame, is created and maintained by CMS using a combination of surname lists and several pieces of information from the EDB. A measure of the accuracy of this code is displayed in Exhibit 2. We matched the race classification code to the true reported race/ethnicity for members of the 2011 and 2012 panels. Approximately 90 percent of beneficiaries flagged as Hispanic report themselves as truly Hispanic, and 74 percent of beneficiaries flagged as Asian report themselves as truly Asian. Thus, this race code will result in some misclassification. Some cases coded and sampled

[3] This is our current understanding; however, from some ongoing analyses conducted by NORC, we know that some cases are not added immediately. Further investigation will be needed to determine the number of cases this affects, the magnitude of the delays, and the impact these delays may have on the new design.

[4] For the remainder of this paper, Hispanic, Latino/a, or Spanish origin beneficiaries will be referred to as Hispanic.

as Hispanic will turn out to be non-Hispanic, and some cases coded and sampled as Asian will turn out to be non-Asian.

Exhibit 2: Conditional Probabilities of Self-Reported Race Given Coded Race in 2012* ATC

True (Survey-Reported) Race	Race Classification Code			
	Hispanic	Asian	All Other	Total
Hispanic	0.901	0.020	0.020	0.081
Asian	0.011	0.740	0.003	0.016
All Other	0.087	0.240	0.977	0.903
Total	1.000	1.000	1.000	1.000

*Includes 2011 and 2012 panels only. Excludes beneficiaries residing in Puerto Rico.

We investigated five basic sampling designs for implementing the oversamples. Each has its own statistical and cost implications. For each of the methods, two sampling strata are established: one for beneficiaries classified as Hispanic (or Asian) by the race code, and one for all other beneficiaries, including those classified as non-Hispanic (or non-Asian) by the race code together with and those for whom the race code is missing. The five methods are as follows:

1. Select the oversample from within the existing sample of SSUs.
2. Select the oversample from two HISKEW[5] files. (This would approximately double the frame from which the oversample could be drawn.)
3. Select the oversample within an independent sample of SSUs selected with probability proportional to the size of the population of interest (Hispanic or Asian).
4. Select the oversample from within existing PSUs at large (ignoring the selected SSUs).
5. Select two additional PSUs with certainty.

For the Hispanic oversample, which was first implemented in 2015, we used an ultimate goal of 1,500 completed Hispanic interviews across all panels by 2018 as a driver for the oversample size each year. Based on this goal and the current size of the Hispanic population within the MCBS, 75 additional completed interviews with Hispanic beneficiaries are required to be sampled each year. In order to achieve this goal, Method 1 proved sufficient to fulfill our needs. The Hispanic population in the current set of MCBS core SSUs is large enough to support the selection of an additional sample large enough to achieve 75 additional Hispanic completed interviews per year. After four years, we expect to achieve the targeted goal of 1,500 completed Hispanic interviews across the four active panels. This overall sample size is expected to remain steady indefinitely, assuming an

[5] A HISKEW is a unique 5-percent subsample of the full Medicare EDB. A different HISKEW is identified and used for sampling for the MCBS each year.

additional 75 Hispanic completes is added to each new panel moving forward. Our approach to oversampling Hispanics features the following:

1. The oversample is drawn from our existing selected MCBS SSUs, which fall within the existing MCBS PSUs. Further, it is only drawn from the 685 (of 703 total) SSUs that fall in the 104 (of 107 total) PSUs outside of Puerto Rico (due to the desire to decrease the proportion of total MCBS Hispanic beneficiaries from Puerto Rico).
2. Cases are selected for oversampling based on the race classification code.
3. Screening is not used. Instead, the population of beneficiaries within the HISKEW, within the 685 SSUs, are classified into two primary sampling strata based on the race code: Hispanic and Other (including non-Hispanic and missing race/ethnicity).
4. The sampling fraction for the Hispanic stratum is determined to achieve the required sample size of true (self-reported) Hispanics.
5. The sampling fraction for non-Hispanic stratum is determined to achieve the required sample size of true (self-reported) non-Hispanics.
6. The sampling fractions in points 5 and 6 are calculated simultaneously to achieve the required sample sizes while taking into account of the probabilities of race misclassification inherent in the race classification code.
7. The Hispanic and non-Hispanic strata described above are crossed by the seven standard strata defined by age group. Thus, in conducting sampling operations, 14 (= 2×7) sampling strata are used.

For the Asian oversample, which will be implemented beginning in 2016, a similar goal of 1,500 completed Asian interviews across all panels within four years of implementation was used to drive the sample sizes for the annual Asian oversample. An additional 470 completed interviews with Asian beneficiaries each year will be required to meet this target by 2019. Because the Asian population in America, and in the current selected MCBS SSUs, is relatively small, achieving a large sample of Asians, at least large enough for separate analysis, is quite challenging. To confront and solve this challenge, we are testing the five methods presented above. Our current plan is to implement Method 2 for the Asian oversample; i.e., we will use two HISKEWs for selecting the Asian oversample. This method provides double the number of flagged Asians for oversampling than would be available if we sampled from our core SSUs across only one HISKEW.

When the Asian oversample is implemented, we will also be oversampling Hispanic beneficiaries. Thus, we expect our approach to oversampling Hispanics and Asian to be similar to the approach for Hispanics and to feature the following:

1. The oversamples will be drawn from our existing selected MCBS SSUs, which fall within the existing MCBS PSUs. Further, they will only be drawn from the 685 (of 703 total) SSUs that fall in the 104 (of 107 total) PSUs outside of Puerto Rico. (For Hispanics, this is purposeful, due to the desire to decrease the proportion of total MCBS Hispanic beneficiaries from Puerto Rico; for Asians, this is due to the fact that there are virtually no coded Asians in Puerto Rico.)
2. The Hispanic and Other strata will be defined in terms of one HISKEW file, while the Asian stratum will be defined in terms of two HISKEW files.
3. Cases will be selected for oversampling based on the race classification code.
4. Cases without a race code will be selected based on the EDB race/ethnicity code.
5. Screening will not be used. Instead, the population of beneficiaries within the HISKEW(s), within the 685 SSUs, will be classified into three primary sampling

strata based on the race code: Hispanic, Asian, and Other (i.e., non-Hispanic and non-Asian).

6. The sampling fraction for the Hispanic stratum will be determined to achieve the required sample size of true (self-reported) Hispanics.
7. The sampling fraction for Asians will be determined to achieve the required sample size of true (self-reported) Asians.
8. The sampling fraction for Others will be determined to achieve the required sample size of true (self-reported) Others.
9. The sampling fractions in points 5, 6, and 7 are calculated simultaneously to achieve the required sample sizes while taking into account of the probabilities of race misclassification inherent in the race classification code.
10. The Hispanic, Asian, and Other strata described above are crossed by the seven standard strata defined by age group. Thus, in conducting sampling operations, 21 (= 3×7) sampling strata will be used.

VI. Summary and Implications

NORC and CMS introduced three innovations to the MCBS sample design. First, we introduced tracts as second-stage sampling units beginning in 2014. Tracts are relatively stable and change far less often than ZIP codes, and it is easier to merge Census data and other geographic data by tract. This improvement provides both increased analytical utility and some cost savings. Second, we expanded the sampling frame to include the current-year cohort of enrollees beginning in 2015. As a result, data products will be available for analysis about one year earlier than ever before. Third, we introduced Hispanic oversampling in 2015 and will introduce Asian oversampling in 2016. These will allow for greater power in estimation of health disparities between and within these populations.