Consumer Cellular Database: More Efficient, but at What Cost?

Caroline Scruggs[1], Marcus Berzofsky[1], Thomas Duffy[1]
Bo Lu[2], Timothy Sahr[3]
[1]RTI International, 3040 E Cornwallis Road, Research Triangle Park, NC 27709
[2]Division of Biostatistics, The Ohio State University, 1841 Neil Avenue, Columbus, OH 43210
[3]Ohio Colleges of Medicine Government Resource Center, 150 Pressey Hall, 1070 Carmack Road, Columbus, OH 43210

Abstract
As telephone surveys increase the proportion of respondents who come from the cellphone frame, the question of how to accurately and efficiently target respondents in small geographic areas (e.g., counties) has become a major issue for survey methodologists. In this paper, we evaluate Marketing System Group's (MSG) Consumer Cellular Database to determine: (1) the accuracy level of the address information on the database (2) if those obtained through the database are different in terms of demographic characteristics and key outcome variables compared to those obtained through traditional random cellphone number selection and (3) if the database yields respondents who live in their expected county at a lower cost (e.g., better yield rate). The Consumer Cellular Database uses multiple publicly available sources to construct household and person-level files. These files cover about half the U.S. population. For example, in Ohio the database includes 2.5 million out of the 5 million households. Using the 2017 Ohio Medicaid Assessment Survey, our evaluation selected a stratified random sample of cellphone numbers using the Rate Center Plus method (Berzofsky, et al. 2017) to minimize the classification error associated with using rate centers as proxies for counties. After selecting the sample, the sample was further stratified by whether the number was contained in the database. We assess efficiency through yield rates and number of required call attempts. We evaluate differences in the random cellphone sample and Consumer Cellular Database by comparing demographic characteristics such as gender, age, race/ethnicity, income level, and marital status. We also assess survey outcomes including the percentage of insured persons, the percentage of persons on Medicaid, the percentage of persons with a chronic condition, and the percentage of persons with unmet medical needs to contrast sample sources on key outcome variables.
Key Words: Cellphone, telephone survey, classification error

1. Introduction

In the past several years telephone surveys are having a steady increase in the proportion of respondents who come from the cellphone frame. With this, the question of how to accurately and efficiently target respondents in small geographic areas (e.g., counties) has become a major issue for survey methodologists. State-based studies often want county or sub-state level estimates requiring accurate geographic identifiers on the cellphone frame.

Currently, sample vendors, such as Marketing System Group (MSG), identify a cellphone number's rate center and determine the county from which the rate center most likely resides. A rate center is the location in which a cellphone number was activated. Rate centers can then be clustered into rate center counties based on the county in which the majority of a rate center falls. While the rate center county assignment can predict where the cellphone user resides, there is classification error in where a person actually lives and their assigned rate center county (Berzofsky, Scruggs, Speizer, Peterson, Lu, and Sahr 2017). Because of this further research is necessary for methods that can increase the likelihood that the sub-state allocation is achieved.

More recently, MSG, and other sample vendors, have developed databases such as the Consumer Cellular Database which uses multiple publicly available sources to construct household and person-level files. These files cover about half the U.S. population. For example, in Ohio the database includes 2.5 million out of the 5 million households. The files contain identifying information, including address. MSG has access to three sources which have linked cellphone and address information:

Source 1: Has a rich set of additional characteristics beyond address (e.g., race, age of person).

Source 2: Has some additional characteristics beyond address but not as many as Source 1.

Source 3: Only has address information beyond phone number.

1.1 Motivation

The Ohio Medicaid Assessment Survey (OMAS) is a periodic general population survey of residents in Ohio. The survey has been conducted approximately every two years since 2004. The survey collects information on and produces estimates related to access to health care and health care status. The survey is a dual frame RDD survey. In 2017, OMAS collected 36,000 interviews – 70% of which were collected through the RDD cellphone frame. Because the outcomes of interest are highly correlated to where a person lives, county level estimates are critical to understanding the populations at greatest risk.

Using the 2017 OMAS, our evaluation selected a stratified random sample of cellphone numbers using the Rate Center Plus method (Berzofsky, et al. 2017) to minimize the classification error associated with using rate centers as proxies for counties. After selecting the sample, cases were flagged if they were in the Consumer Cellular Database. Therefore, while we did not incorporate the Consumer Cellular Database in the stratification, we did account for it when producing and releasing replicates into the field. The 2017 OMAS collected a total of 29,899 cellphone interviews from the RDD frame. Of those, 6,987 (23.4%) were flagged as also being in the Consumer Cellular Database.

1.2 Research Questions

To evaluate MSG's Consumer Cellular Database, we developed three research objectives:

1. Determine the accuracy level of the address information on the database.
2. Determine if those obtained through the database are different in terms of demographic characteristics and key outcome variables compared to those obtained through traditional random cellphone number selection.
3. Determine if the database yields respondents who live in their expected county at a lower cost (e.g., better yield rate).

2. Methods

2.1 Consumer Cellular Database Accuracy

When initially drawing the sample, we were only aware of the availability of source 1 from the Consumer Cellular Database. Therefore, for purposes of data collection, source 1 was used to assign cases to replicates by county. However, after data collection, in discussions with MSG, we determined source 2 and source 3 could be used to link to address information, but these sources were only available post selection of a sample. In other words, source 1 is the only source which is available for stratification prior to selection the sample. Because of our knowledge at the time of sampling, we were only able to append source 1 on the full sample when we initially drew the sample. Then, after collection ended, we had the Consumer Cellular Database sources 2 and 3 appended to the completed cases from the survey for this evaluation. For our evaluation, we examine the accuracy of the address information on all three sources. However, for comparisons of the respondent differences and cost efficiency, we will compare source 1 to the cellphone RDD samples since it was appended to the full sample, not just completed cases. Table 1 shows the number of completed cases that were contained on each data source.

Table 1: Completed Cases on Each Database Source

	Cases	Percent	Unique Cases
Cellphone RDD Completes	29,899	--	--
DB – Source 1	6,987	23.40%	56.10%
DB – Source 2	2,320	7.80%	20.10%
DB – Source 3	4,729	15.80%	70.10%

In terms of how the three sources overlapped with each other, there are differences across the sources. Source 2 had the fewest completed cases as well as unique cases, it also showed a big overlap with source 1 having 76% of its cases shared. Comparatively, source 3 only had 28% of cases overlap with source 1. Also, to note, source 1 was on 20.5% of the working cellphone RDD cases that were released during sampling.

The final evaluation of the sources related to the accuracy of their address information. To do this we looked at the county match rate between the county of residence identified in the database and the self-reported value provided by respondents.

2.2 Respondent Characteristics

We utilized the 6,987 completed cases from the Consumer Cellular Database (source 1) to compare to 22,912 completed cases from the Cellphone RDD only sample. We first compared the demographic characteristics of persons among the two sample types as well as the population, using the 2016 American Community Survey 1-year estimates for Ohio. The demographic characteristics considered were

- Age group (< 45, >=45)
- Gender
- Income (Below 138% of FPL, above 138% of FPL)
- Race/ethnicity (White, non-White)
- Marital status (married, non-married)
- Employer offers health insurance (yes, no)

Next, we conducted bivariate analyses comparing the Consumer Cellular Database and the Cellphone RDD sample for the following outcomes:

- Percentage uninsured
- Percentage on Medicaid
- Percentage on employer insurance
- Percentage having problems getting healthcare
- Percentage with high blood pressure or hypertension
- Percentage who have ever had a heart attack
- Percentage who have ever had coronary heart disease
- Percentage who have ever head congestive heart failure
- Percentage who have ever had diabetes

Finally, to determine if any difference in the demographics where causing bias, for each outcome, we fit a logistic model containing the demographic variables previously considered.

2.3 Sample Performance

To assess the efficiency of the Consumer Cellular Database (source 1) and the Cellphone RDD sample we compared yield rates (the ratio of completed interviews over sampled telephone numbers), the number of required call attempts, and response rates (AAPOR RR #4) (American Association for Public Opinion Research, 2016).

3. Results

3.1 Consumer Cellular Database Accuracy

As shown in Table 2, the Consumer Cellular Database source 3 has the highest match rate between the frame and the respondent's reported county (87.4%). Source 3 only contains address information beyond the phone number, but this shows that the information is most accurate. Meanwhile, source 1 has a lower match rate (54.0%) than the Cellphone RDD sample using rate center county. Therefore, even though there is a size disparity between the two sources, if county-level accuracy is paramount to a design, source 3 results in more accurate cases than source 1 – 4,135 cases in source 3 compared to 3,638 for source 1.

Table 2: County Match Rate

	% County Match	SE
Cellphone RDD	66.2%	0.00
DB – Source 1	54.0%	0.01
DB – Source 2	81.6%	0.01
DB – Source 3	87.4%	0.00

Furthermore, when comparing the three Consumer Cellular Database sources at the individual county match rate, we see in Figure 1, that source 3 clearly out performs sources 1 and 2.

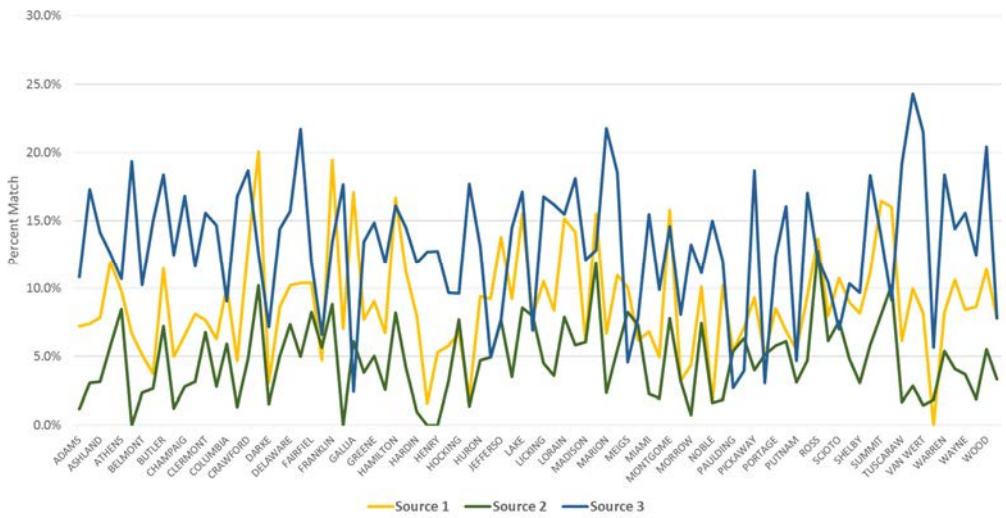

Figure 1: County Match Rate by County and Consumer Cellular Database Source

3.2 Respondent Characteristics

3.2.1 Demographics

Figure 2 presents a comparison of respondents on the Consumer Cellular Database (source 1) and the Cellphone RDD only sample. All demographic characteristics are statistically different. Specifically, those on the Consumer Cellular Database are more likely to be older, female, White, married, on employer insurance and have income over 138% of the federal poverty level (FPL). These differences range from 1.8% (gender) to 12.5% (FPL). This is expected since a person on the Consumer Cellular Database would be more established where they live.

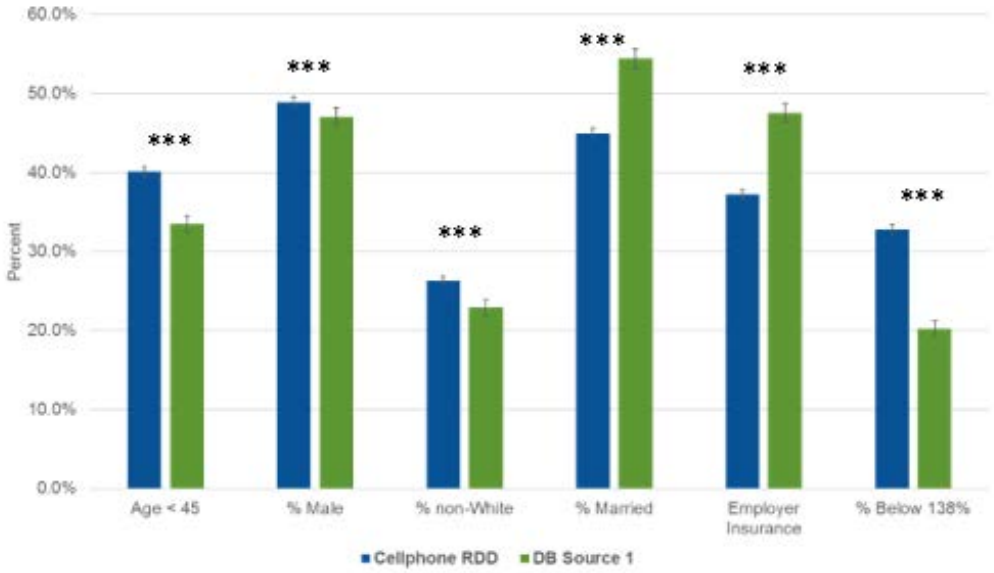

Figure 2: Percentage of respondents in the Cellphone RDD sample and Consumer Cellular Database by demographic characteristics.

When examining age, gender and race characteristics we also compared them to the population in Ohio, shown in Figure 3. The Consumer Cellular Database trends older than Cellphone RDD sample but even more so than the population. The male percentage still underperforms on the Consumer Cellular Database, but by a 1.3% difference when compared to the population. While we do have more white respondents on the Consumer Cellular Database than the Cellphone RDD sample, this does trend closer to the make-up of the population.

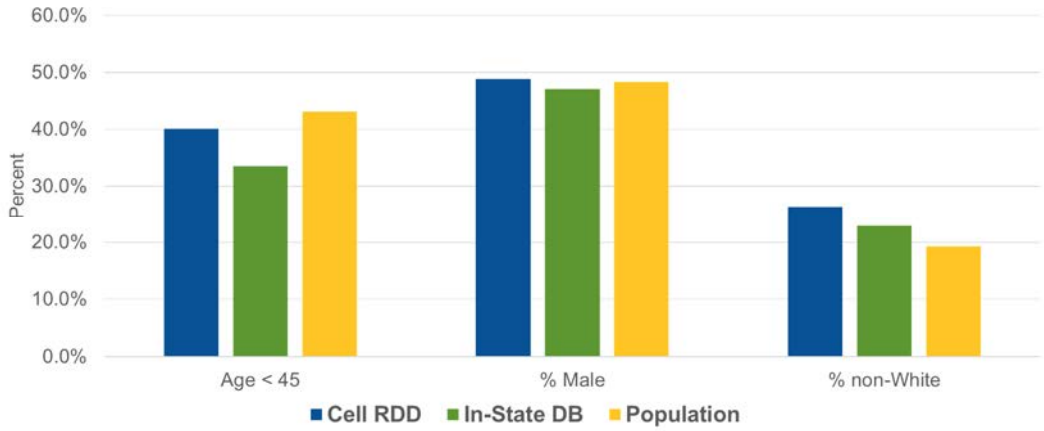

Figure 3: Percentage of respondents in the Cellphone RDD sample and Consumer Cellular Database compared to the population by select demographic characteristics

3.2.2 Key Outcomes

Figure 4 presents the bivariate comparison of outcomes between respondents in the Consumer Cellular Database and those in the Cellphone RDD sample. For outcomes related to a respondent's health insurance status/coverage the differences are significant. Also, we see those on the Consumer Cellular Database were statistically less likely to have congestive heart failure.

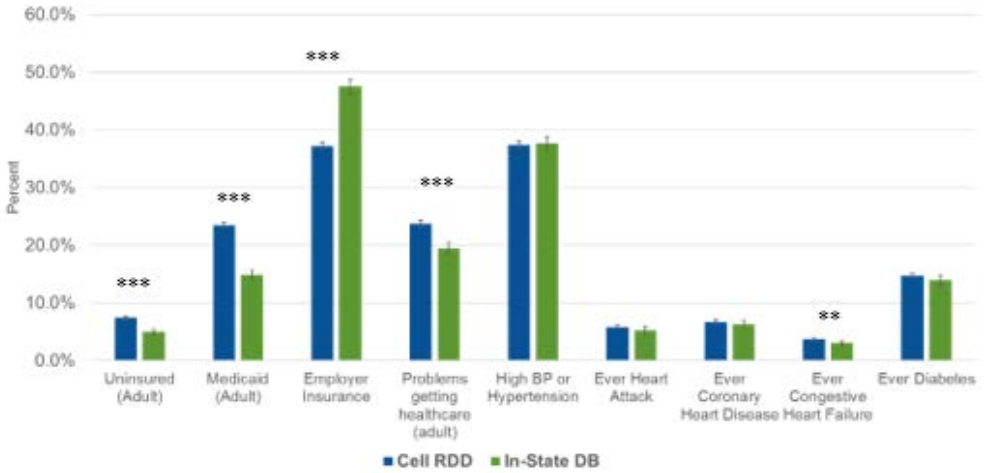

Figure 4: Percentage of respondents in the Cellphone RDD sample and Consumer Cellular Database by health insurance status and health outcomes

Figure 5 presents the marginal probability of a respondent's health insurance status and health outcome status controlling for respondent characteristics. Uninsured, on Medicaid, and problems getting healthcare remain to be statistically different; however, the percent differences are smaller than before controlling for respondent characteristic difference. Meanwhile, all health outcomes show no statistical differences.

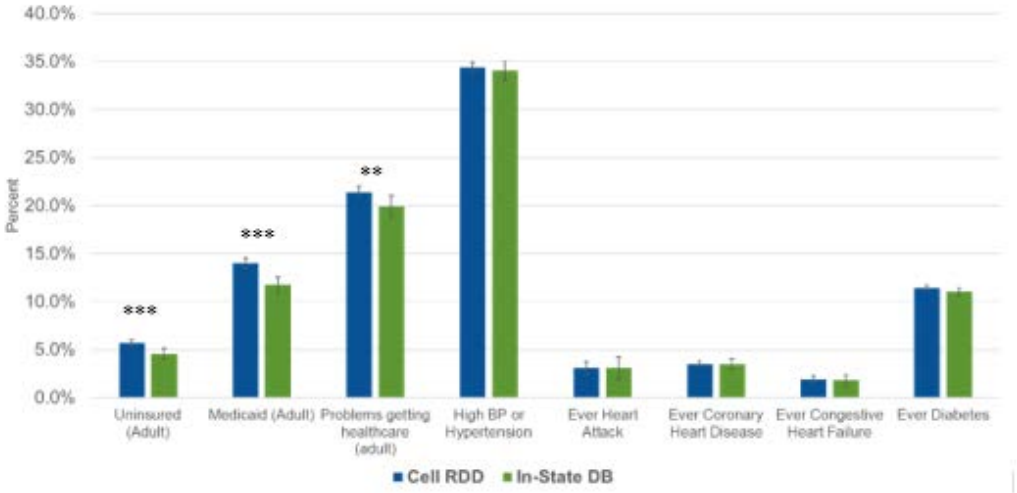

Figure 5: Marginal probabilities of respondents in the Cellphone RDD sample and Consumer Cellular Database controlling for respondent characteristics by health insurance status and health outcomes

3.3 Sample Performance

To examine the sample performance of the Consumer Cellular Database and if it can help reduce the cost burden of cellphone sampling, we first obtained the number calls per complete, which showed the Consumer Cellular Database resulting in 3.2 fewer calls per complete. However, when pulling the average number of calls necessary on completed cases only, we see the Consumer Cellular Database is slightly higher than the Cellphone RDD cases as shown in Table 3.

Table 3: Call Statistics

	Average Calls of Completed Cases	SE	Calls per Complete
Cellphone RDD	2.58	0.01	100.14
DB Source 1	2.71	0.02	96.94
Difference	0.13***	0.02	-3.20

Table 4 shows the yield rates, the number of phone numbers worked per completed case, obtained from the two sample frames. The Consumer Cellular Database required 9 fewer numbers than the Cellphone RDD sample.

Table 4: Yield and Response Rates

	Yield Rate	e-Factor	AAPOR RR4
Cellphone RDD	48.1	0.14	22.4%
DB Source 1	39.2	0.20	19.2%

Surprisingly, the Consumer Cellular Database had a lower response rate than the Cellphone RDD sample, as shown in Table 4. This result can be seen as being driven by the higher e-factor which stems from the higher proportion of unknown cases on the Consumer Cellular Database, seen in Table 5. While the overall response rate is lower for the Consumer Cellular Database, it does have a higher proportion of completed cases, which is significantly different than the Cellphone RDD completes, as is the lower proportion of ineligibles.

Table 5: Final Call Disposition Percentage

	Cellphone RDD	DB Source 1
Completes ***	2.90%	3.50%
Partial	0.30%	0.30%
Noncontact *	0.10%	0.10%
Refusal	4.20%	4.20%
Ineligible ***	45.40%	33.20%
Unknown ***	47.10%	58.80%

4. Conclusion

We found the Consumer Cellular Database, specifically source 3, produces a high county accuracy rate. There are some differences in the demographics and key outcomes between the Consumer Cellular Database and the Cellphone RDD sample. The database trends older and towards higher income respondents. Health outcomes are not significantly different for outcomes associated with older persons. The database resulted in three fewer calls per complete, a lower yield rate, and a lower response rate but less ineligibles and more unknown cases.

While the sample on the Consumer Cellular Database does perform better, the increased cost of purchasing the sample may not be offset by the better call counts and yield rate it produces. Among the different sources for address information, source 3 of the Consumer Cellular Database is clearly the best indicator. While not available prior to drawing an initial RDD sample, a two-stage design could be implored by which samples could be stratified by the phone number being in source 3 or not, with an oversample of numbers in source 3. Because of the significant differences between persons in and not in the Consumer Cellular Database, even after controlling for demographics, this could impose bias on a study. To mitigate this potential bias, additional weighting steps to control for demographic differences could be implemented. If a study does need to target smaller geographic levels, the Rate Center Plus method could be altered to use the address information obtained from the Consumer Cellular Database.

References

Marcus E Berzofsky, Caroline B Scruggs, Howard Speizer, Kimberly C Peterson, Bo Lu, Timothy Sahr. (2017) A Method for Accounting for Classification Error in a Stratified Cellphone Sample. *Journal of Survey Statistics and Methodology*,,smx033, https://doi.org/10.1093/jssam/smx033

American Association for Public Opinion Research. “Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys” 2016. Available at https://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx.

U.S. Census Bureau. “American Community Survey 1-Year Estimates” 2016. Available at https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_1YR_S0101&prodType=table.

U.S. Census Bureau. “American Community Survey 1-year Public Use Microdata Samples” 2016. Available https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_pums_csv_2016&prodType=document