Benchmarking Mobile App Geofenced Samples: Adjusting for National Coverage and Selection Bias

Davia Moyse[1], Yangyang Deng[1], Matt Jans[1], Ronaldo Iachan[1], Lee Harding[1], Kristie Healey[2], James Dayton[2], Scott Worthge[3], Laura O'Campo[3]
[1]ICF, 530 Gaither Rd., Suite 500, Rockville, MD 20850
[2]ICF, 126 College St., Suite 2, Burlington, VT 05401
[3]MFour Mobile Market Research, 19800 MacArthur Blvd., Suite 700, Irvine, CA 92612

Abstract
Mobile phone technology and geolocation advances have made it simple to locate survey respondents in locations outside of their home, such as while shopping or passing a store. As this nonprobability sampling method relies on smartphone ownership and specific store presence, demographic and geographic coverage or selection biases, common to probability surveys of the general population, may be exaggerated. To explore this, we sampled mobile panel members in the U.S. when they entered geofenced areas around grocery, convenience, and home improvement stores, and asked them health-related questions. This paper discusses the demographic and geographic attributes of the nonprobability geofenced respondents relative to population totals and a gold-standard probability sample health survey. We compare four raking approaches to account for the potential biases: along basic demographics; along expanded demographics; along demographic and geographic dimensions as independent margins; and along controlled, cross-classified margins of geographic by demographic characteristics. The four methods are evaluated based on the distribution of the raked weights and by benchmarking weighted estimates of survey responses to a comparable probability survey.

Key Words: nonprobability surveys, mobile phones, innovative data collections methods, nonprobability benchmarking, coverage bias, selection bias, raking

1. Introduction

Nonprobability panels continue to increase in popularity and sophistication, but remain largely untested as replacements for or complements to probability samples. One of the most promising nonprobability panels on the market is MFour's geofenced *Surveys on the Go®* panel, which uses the geolocation technology on panel members' smart phones to sample them from specific locations defined by points around which a "geofence" is drawn.

Geofences are virtual geographic boundaries that are set around real-world locations, and enable mobile phone applications to trigger an action when the device enters or leaves the area. For example, a common market research application is to pick a point of interest (e.g., a shopping center or store that wants to sample its patrons), and place a geofence around the entrance to that shopping center or store. Then, patrons who are also members of the research company who set the geofence will be invited to complete a questionnaire when they trip the geofence on entry or exit. While geofenced surveys are usually used for intercept market research like this, this innovative technology can be used to capture a sample of the general population and invite them to complete a survey on any topic.

This sampling approach has several potential benefits. Logistically, it provides the opportunity to access potential respondents outside of their home and without the use field interviewers. It is also more cost- and time-efficient than probability samples or on-the-ground intercept surveys that can be used for general population surveys, recreational or environmental surveys, and surveys targeting rare or hard-to-reach populations. For example, under traditional approaches, constructing a sampling frame and obtaining a respondent pool to represent "current tobacco users who have also visited a doctor in the past month" would be very challenging, expensive, and likely result in a small analytical group. Using mobile nonprobability sampling to reach the same group of people allows access to a large potential respondent pool at a lower cost per eligible and per complete. In addition to sampling efficiencies, mobile panel methods offer measurement opportunities not feasible in traditional household surveys. For example, it is possible to capture details about events and behaviors while they are happening, which mitigates recall error. In a traditional survey, respondents would be asked to recall whether they had medical lab tests completed within the past year, but would likely have difficult remembering all lab tests conducted, and certainly would have trouble remembering their exact cholesterol levels from a given test. A geofenced sampling approach could sample participants during a doctor's visit while they are receiving cholesterol test results. There is also the option to collect "bonus" data elements, such as capturing images of test results or videos of interactions with doctors or the doctor's office via the mobile phone's camera. Such options are simply included as response tasks within the questionnaire.

Given these potential benefits, ICF and MFour have been aiming to fully understand the extent to which sampling panelists at geofenced locations can be a feasible alternative or complement to traditional probability sampling. As with any survey methodology, this sampling method invites potential biases, depending on the level of population representativeness obtained through survey respondents. To justify the use of geofenced sampling, we must answer three questions:

1. Are there certain populations that are unreached by the geofences?
2. Are there certain populations that have a lower chance of being included via the geofences?
3. Can we correct for these potential biases—coverage and selection—that may be exaggerated by this method?

To do so, we assessed whether a geofenced sample of grocery, convenience, and home improvement stores can produce useful population estimates of public health outcomes and health risk factors, and the extent to which this method yields population representativeness.

2. Methods

2.1 Geofenced Sampling and Data Collection

The target population for this study was noninstitutionalized adults age 18 and older living in the United States. Geofences and survey data collection were provided by MFour's *Surveys on the Go*® mobile opt-in panel, which includes approximately two million active users. MFour traditionally specializes in dairy studies; in-home measurement; advertisement, entertainment, and behavior trackers; and, more generally, in geo-targeting measurement to engage respondents in the middle of or just after completing an activity. Their panel is single-source (i.e., not combined with other Web or smartphone panels), which limits overlap with other online opt-in panels.

For this proof of concept study, over 47,000 geofences with a fifty-meter radius were drawn around entrances. Twenty-eight large and well-known national chains were included. Figure 1 provides a visual representation of geofences in Tennessee for two store chains.

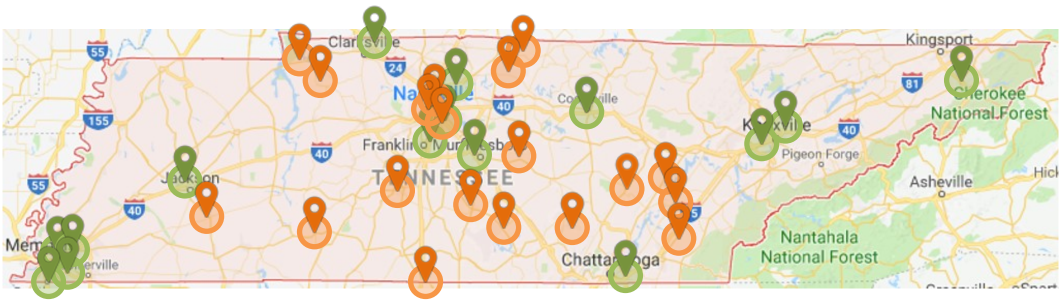


Figure 1: Hypothetical representation of two geofenced store chains in Tennessee (Geofences are not presented to scale.)

Panelists received push notifications from the *Surveys on the Go*® app to complete a brief survey immediately upon entering a geofence. The app produces a visual notification and a cash register “cha-ching” sound. Figure 2 shows how panel members see what surveys they have been invited to (left screenshot) and an example demographics question (right screenshot).

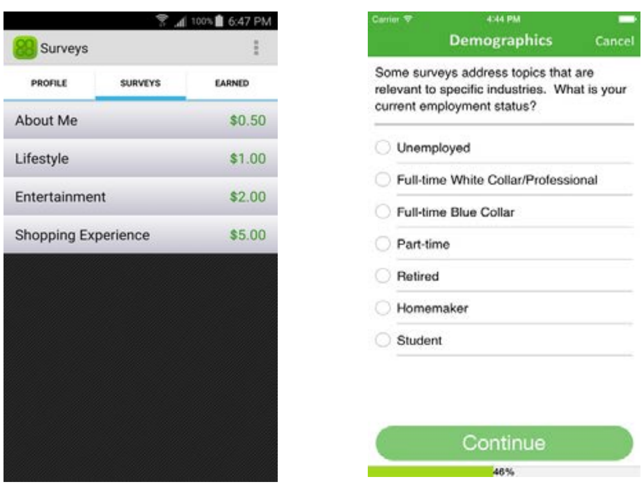


Figure 2: MFour *Surveys on the Go*® smartphone app interface and example survey question

For this study, the survey remained available to the panelist for 48 hours from the push notification and could be completed after they left the geofenced area. Reminders were sent via the app to nonrespondents at one, twenty-four, and thirty hours after the initial invitation. The survey remained in the field until the quota of 1,000 completed questionnaires was obtained. Analyses in this report are based on 998 respondents.

2.2 Questionnaire Topics

The brief questionnaire, estimated to take five to six minutes to complete, included two components: basic demographics and health topics. Demographic information was used for eligibility determination and to compare the composition of respondents to the composition of established population surveys and the target population overall. These demographics included state and zip code of residence, place of residence (e.g., private, college housing), age, gender, sexual orientation, ethnicity, race, marital status, education, employment status, and the number of adults in their household by gender.

Health topic data were used to benchmark geofenced survey respondents to known, well-accepted estimates. The health topic questions are presented in Table 2.2-1. They are borrowed from the Behavioral Risk Factor Surveillance System (BRFSS) core section and collect data on tobacco, alcohol, and sugar-sweetened beverage behaviors, in addition to other key health behaviors.

Table 2.2-1: BRFSS Health Topic Questions

Topic	*Question*	*Response Options*
Tobacco Behavior	Have you smoked at least 100 cigarettes in your entire life? Do not include electronic cigarettes (e-cigarettes, NJOY, Bluetip), herbal cigarettes, cigars, cigarillos, little cigars, pipes, bidis, kreteks, water pipes (hookahs), or marijuana. Please note that 100 cigarettes is equal to 5 packs of cigarettes.	1. Yes 2. No
Tobacco Behavior	Do you now smoke cigarettes every day, some days, or not at all?	1. Every day 2. Some days 3. Not at all
Tobacco Behavior	Do you currently use chewing tobacco, snuff, or snus every day, some days, or not at all?	1. Every day 2. Some days 3. Not at all
Alcohol Behavior	During the past 30 days, how many days per week or per month did you have at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor?	_____ days per: 1. Week 2. Month Don’t know/Not sure
Alcohol Behavior	During the past 30 days, on the days when you drank, about how many drinks did you drink on the average? Please note: One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. A 40-ounce beer would count as 3 drinks, or a cocktail drink with 2 shots would count as 2 drinks.	_____ Number of drinks Don’t know/Not sure
Alcohol Behavior	Considering all types of alcoholic beverages, how many times during the past 30 days did you have [IF MALE, INSERT “5 or more”, ELSE IF FEMALE, INSERT “4 or more”] drinks on an occasion?	_____ Number of times None Don’t know/Not sure
Alcohol	During the past 30 days, what is the largest	_____ Number of drinks

Behavior	number of drinks you had on any occasion?	None Don't know/Not sure
Sugar-sweetened Beverage Behavior	Not including fruit-flavored drinks or fruit juices with added sugar, how often in the past 30 days did you drink 100% fruit juice such as apple or orange juice? Enter '0' if you did not drink 100% fruit juice in the last 30 days.	_____ times per: 1. Day 2. Week 3. Month Don't know/Not sure
Sugar-sweetened Beverage Behavior	Now, thinking about sugar-sweetened beverages including regular soda, sports drinks, energy drinks, coffee, tea, and juices that have added sugar, how often in the past 30 days did you drink sugar-sweetened beverages? Enter '0' if you did not drink any sugar-sweetened beverages in the last 30 days.	_____ times per: 1. Day 2. Week 3. Month Don't know/Not sure
Other Health Behavior	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, government plans such as Medicare, or Indian Health Service?	1. Yes 2. No
Other Health Behavior	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	1. Yes 2. No

2.3 Weighting

Geofenced results were weighted using a step-wise poststratification raking approach, resulting in four independent sets of survey weights. The first method used only four basic demographics. The second method added three additional demographic dimensions. The third method then included geographic dimensions. Finally, method four was a fine-tuning step where key cross-classifications of demographic and geographic dimensions were used. Single margins and key cross-classifications were selected to either mirror those used in the comparable BRFSS probability survey or to address coverage or selection biases, while also accounting for small cell size issues. Table 2.3-1 provides more details on the specific margins used for each method. All control totals were determined using population statistics from the 2017 American Community Survey (ACS).

Table 2.3-1: Four Weighting Methods

Method	*Description*	*Poststratification Raking Dimensions*	*CV*	*Design Effect*
1	Basic Demographics	• Age: 18-24, 25-34, 35-49, 50+ • Sex at Birth: Male, Female • Race/Ethnicity: Non-Hispanic White, Non-Hispanic Black, Hispanic, Other/Multirace • Education: High School Degree or Less, Some College, College Degree or More	151.5	3.30
2	Expanded Demographics	• Age, Sex at Birth, Race/Ethnicity, and Education from Method 1 • Tenure: Own, Rent • Household Income: < \$35,000, \$35,000 to < \$50,000, \$50,000 to < \$75,000, ≥ \$75,000	170.5	3.91

		• Employment Status: Employed for Wages, Self-employed, Unemployed		
3	Expanded Demographics + Geographic	• Age, Sex at Birth, Race/Ethnicity, Education, Tenure, Household Income, and Employment Status from Method 2 • U.S. Region: Northeast, South, Midwest, West • Metro Status: Metro, Non-Metro • Urbanicity: Urban, Rural	176.8	4.13
4	Cross-classifications	• U.S. Region by Age • U.S. Region by Race/Ethnicity • Metro Status by Sex at Birth • Metro Status by Education • Tenure • Household Income • Employment Status • Urbanicity	185.1	4.43

We weighted the BRFSS data to be representative of the US using standard BRFSS margins plus state in collapsed categories (Iachan, et. al., 2016). BRFSS margins include sex by age, race/ethnicity, education, marital status, home ownership, sex by race/ethnicity, race/ethnicity by age, and type of phone in the household (cell only, landline only, or both).

Weighted estimates of key health behaviors from each of the four methods were computed and compared to national BRFSS estimates using t-tests.

3. Results

3.1 Geographic Comparison of Sampled Geofences to U.S. Adult Population

We calculated the distribution rates of U.S. adults to the states, to regions, and to divisions, as well as to metro and non-metro counties within division. We determined the same distribution rates for the 47,000 plus geofences used for sampling. Figure 3 shows a heat map of the geofence rate minus the U.S. adult rate by state. Although these are different units of analysis, low coverage bias can be assumed in states where the geofence rate and U.S. adult rate are reasonably similar. California, Texas, and New York may have had too few geofences relative to the target population to mitigate coverage bias, whereas Florida, North Carolina, Illinois, and Virginia may have had too many geofences.

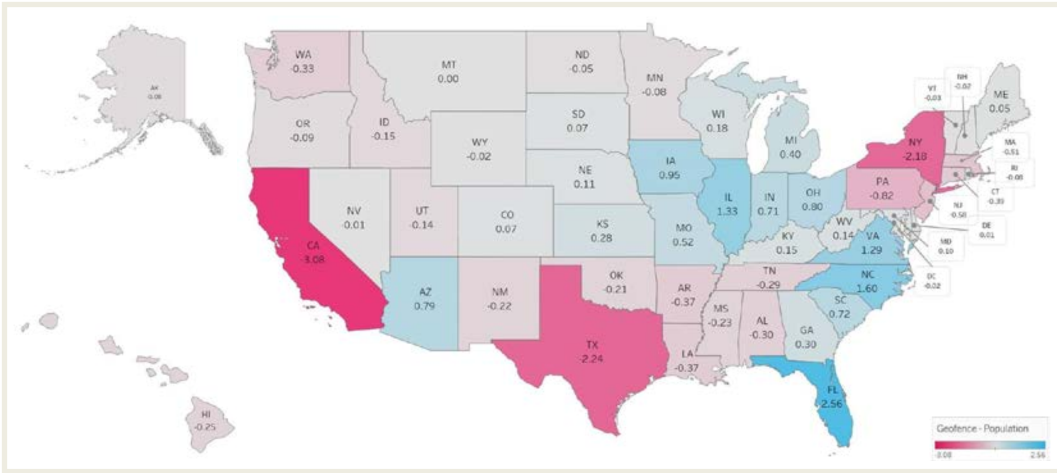


Figure 3. Distribution of Geofences vs. Distribution of U.S. Adults, by State

Table 3.1-1 provides similar information at the U.S. region, U.S. division, and U.S. division by metro status levels. West and Northeast regions may have had too few geofences relative to the target population, which appeared to be largely driven by differential rates in California (Pacific division) and New York (Middle Atlantic division). In the more rural areas of the country, such as West North Central, where national grocery, convenience, and home improvement chains may not be as prevalent, the geofences were drawn at a reasonably similar rate to the U.S. adult population and had over coverage in non-metro counties.

Table 3.1-1: Four Weighting Methods

U.S. Region	*U.S. Division*	*Geofence Rate – U.S. Adult Rate*		*Non-Metro Coverage**
Midwest	East North Central	5.25	3.44	---
	West North Central		1.81	Over
West	Mountain	-3.53	0.30	---
	Pacific		-3.83	---
Northeast	New England	-4.56	-0.97	Over
	Middle Atlantic		-3.59	---
South	South Atlantic	2.83	6.70	---
	East South Central		-0.68	Under
	West South Central		-3.18	---

* More than a 5 percentage point absolute difference between geofence and U.S. adult rates

The distribution of survey respondents to U.S. regions and divisions was similar to the distribution of the target population. Fewer respondents were obtained from the Northeast compared to the population (11% and 18%, respectively). We also did not have respondents from Alaska, South Dakota, or Wyoming.

3.2 Demographic Comparison of Panel Members and Survey Respondents to U.S. Adult Population

Figure 4 shows the distribution of MFour's *Surveys on the Go* ® active adult (18+) panel members, the survey's respondents, and the target U.S. adult population along key demographic lines. The adult panel members and the respondent pool skewed younger and more female than the target population. Although the panel members reasonably represent the target population in terms of education and employment status, the respondents to our survey tended to be more educated and employed than the overall U.S.

adult population.

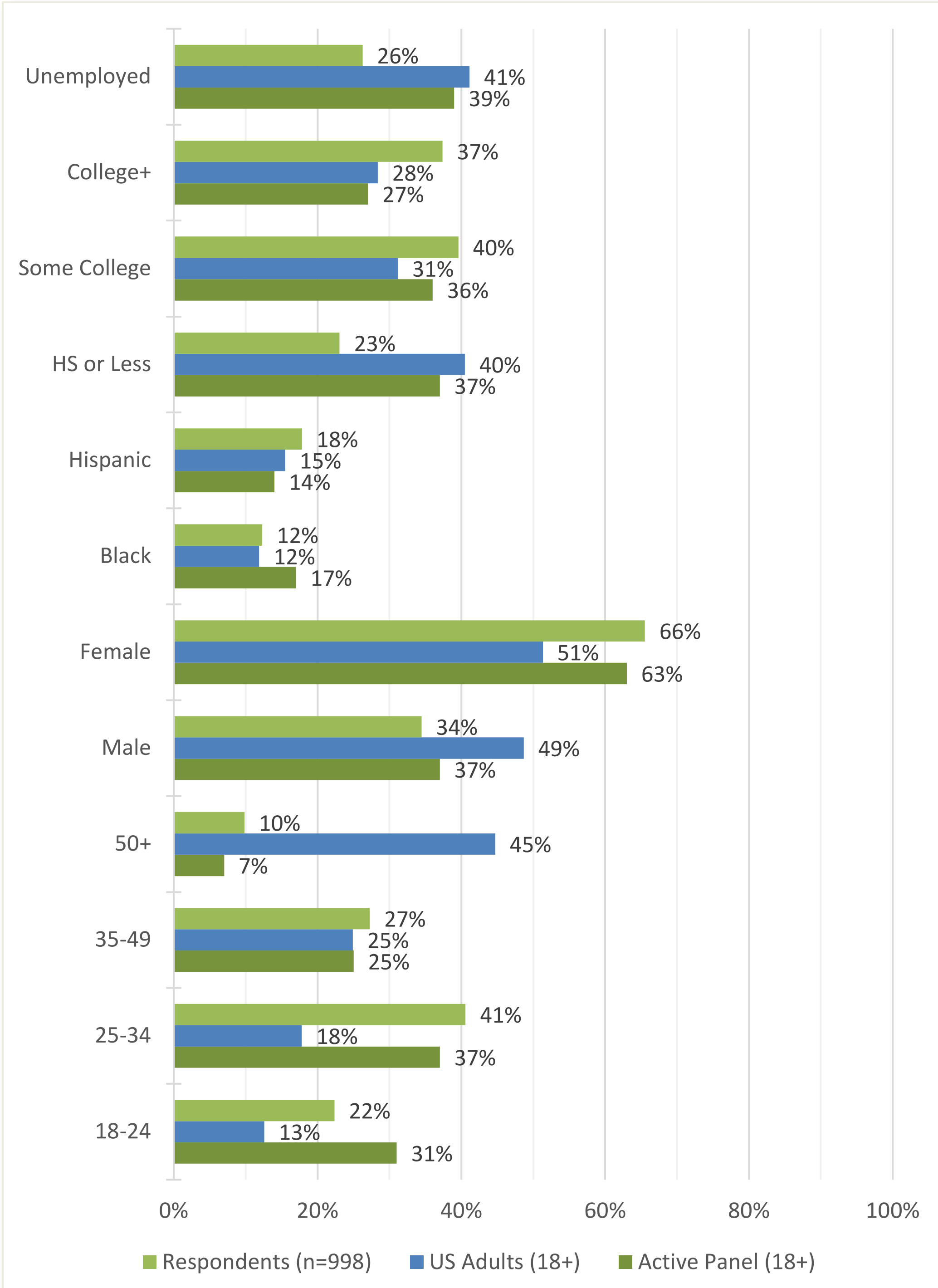
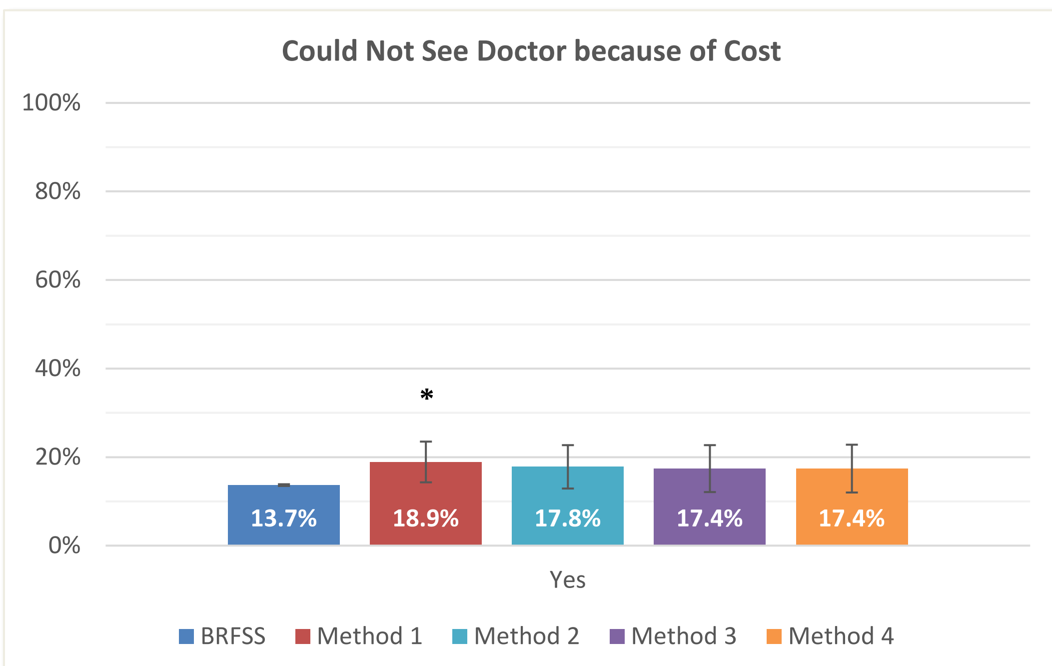


Figure 4. Demographic Comparison of Active Panel Members & Respondents vs. Distribution of U.S. Adults

3.3 Key Health Behavior Estimates Benchmarked to BRFSS

Figures 5 through 8 show key health behavior estimates for the gold-standard probability survey (BRFSS) vs. estimates obtained via the geofenced sample under the four weighting methods described earlier. Estimates of being unable to see a doctor in the past 12 months because of cost under Method 1 weighting (i.e., raking by four basic demographics) were significantly different than estimates from BRFSS ($p < 0.05$; Figure 5). Estimates of this health behavior under weighting Methods 2, 3, and 4, in which employment status and household income were included as raking dimensions, were not significantly different than BRFSS. Having any health care coverage (Figure 6) was not significantly different for any weighting method when compared to BRFSS. Both of these health behaviors are highly correlated with employment status and it appears that the inclusion of expanded demographics as raking dimensions was able to account for biases for these constructs.

Conversely, estimates of both binge drinking in the past 30 days and being a current every day smoker was significantly different than BRFSS regardless of weighting method ($p < 0.05$; Figures 7 and 8). Although not fully depicted here, we found it was common for the geofence sample under any weighting method to be significantly different from BRSS and to overestimate measures of extreme tobacco use (i.e., everyday smokers, and people who use smokeless tobacco), two of the three measures of binge drinking assessed, and consumption of sugar-sweetened beverages. Interestingly, the geofence sample produced an underestimate of fruit juice consumption. While 100% fruit juice is often a high-calorie drink, it is a healthier drink than sugar-sweetened beverages from a nutrition standpoint. Thus, it appears that when the geofenced sample isn't representative and biases cannot be accounted for via weighting, it tends to overestimate unhealthy behaviors and underestimate healthy ones.



* $p < 0.05$; significant difference found between geofenced nonprobability sample estimate and BRFSS benchmark using t-test

Figure 5. Could Not See Doctor because of Cost: BRFSS vs. Geofenced Weighting Methods

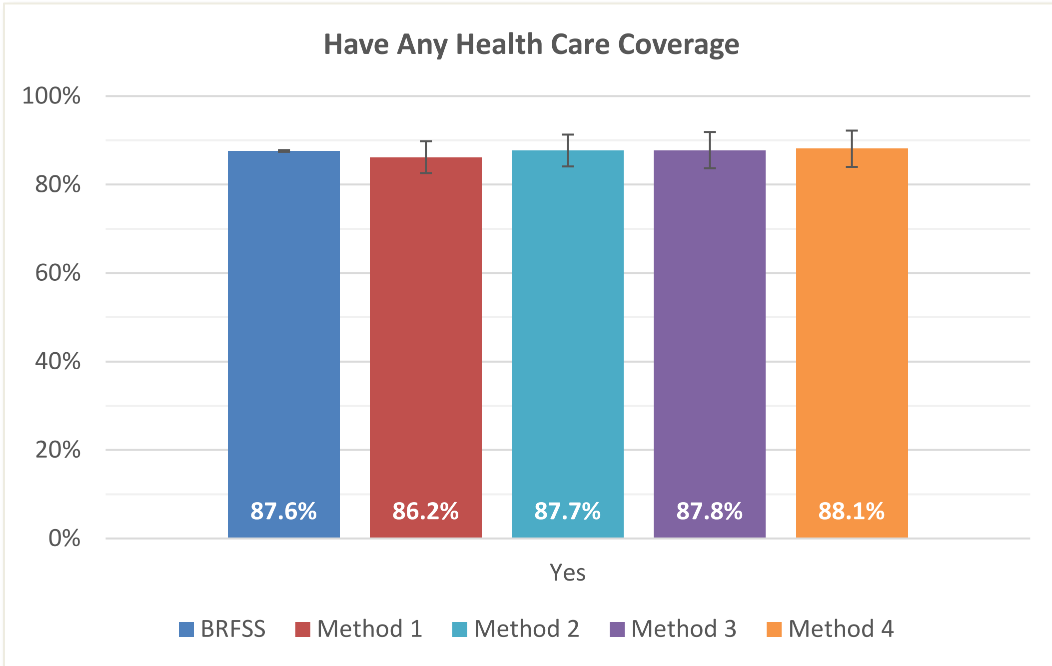
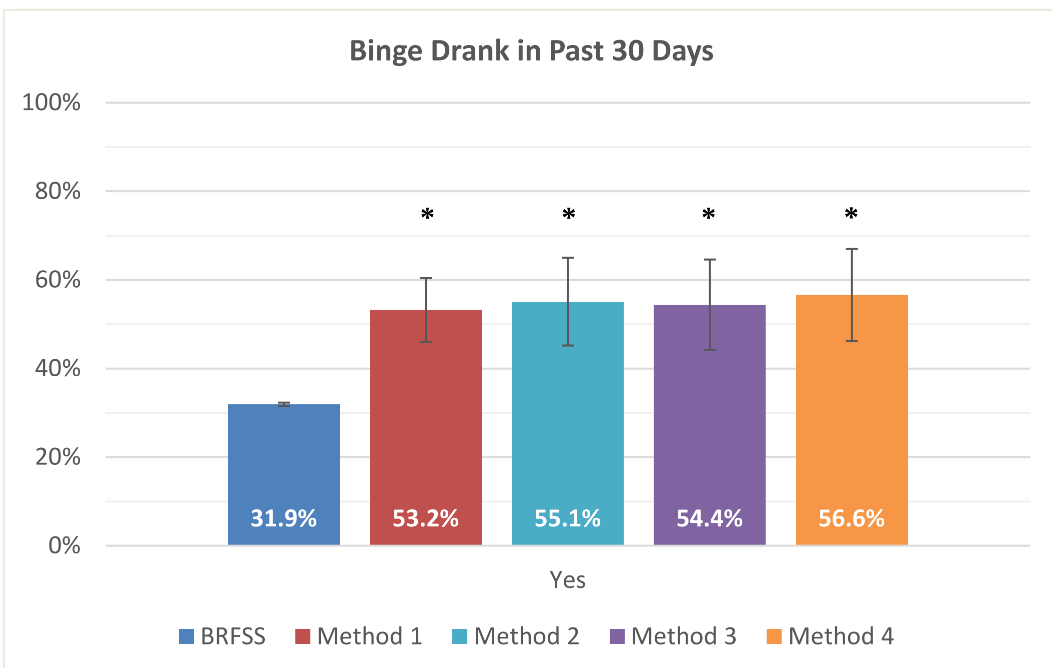
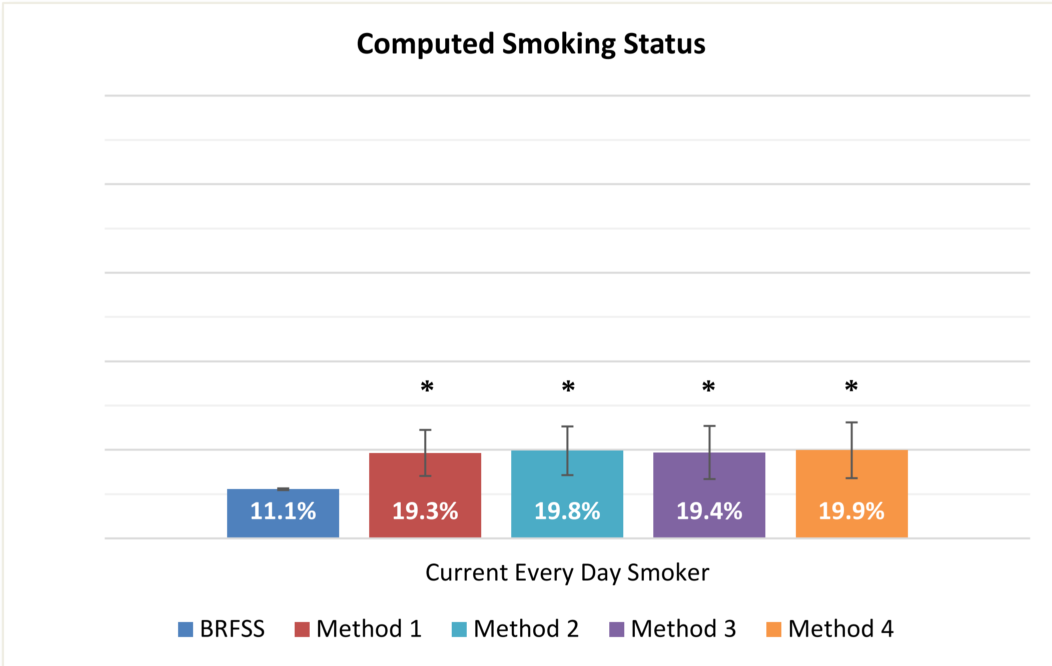


Figure 6. Have Any Health Care Coverage: BRFSS vs. Geofenced Weighting Methods



* $p < 0.05$; significant difference found between geofenced nonprobability sample estimate and BRFSS benchmark using t-test

Figure 7. Binge Drank in Past 30 Days: BRFSS vs. Geofenced Weighting Methods



* $p < 0.05$; significant difference found between geofenced nonprobability sample estimate and BRFSS benchmark using t-test

Figure 8. Computed Smoking Status: BRFSS vs. Geofenced Weighting Methods

4. Conclusions

There are several logistical and measurement benefits to using a nonprobability sample derived from geofences around grocery, convenience, and home improvement stores. They are time- and cost-efficient and may reduce measurement error. As a complement to or alternative for general population surveys, specifically, the geolocation technology on panel members' smart phones can aid in sampling potential respondents outside of their home, which has historically been the easiest place to locate individuals. As a new and innovative approach to nonprobability sampling, geofenced sampling requires rigorous testing. Our proof of concept design was developed with the need for this rigorous testing in mind. This paper evaluated the extent of coverage and selection bias and whether these biases could be mitigated during the production of weights and key estimates.

Overall, certain biases existed at both the frame and respondent level for this geofenced nonprobability sampling approach along both demographic and geographic lines. We determined that the geofenced sample yielded good geographic coverage in the vast majority of U.S. states, but additional geofences should be drawn in future studies in California, New York, and Texas. The geofences also provided over-coverage of rural counties at the U.S. division level in areas where rurality is more prevalent, which helped to ensure that this key population was represented. On the other hand, no responses were obtained from Alaska, which is a fairly unique population. Demographically, we found under representation of males, those aged 50+, those who were unemployed, and those with less education, when compared to the target population of U.S. adults.

Our weighting methods to account for these observed biases had mixed results. As such, geofenced samples may be better used as a complement to, as opposed to a replacement for, traditional sampling and only for particular constructs. Health cost barriers and care coverage, which are more directly related to the raking dimensions of income and employment, produced estimates in line with the gold standard estimates. Despite these successes, health behaviors with more complex and indirect relationships with the dimensions used for weighting may not be accurately measured by a geofenced sample. Based on this study, the estimates to be most concerned about include: extreme tobacco use, two definitions of binge drinking in the past month, and consuming high-calorie drinks. Overall, the geofence sample seems to overestimate unhealthy behaviors and health risks.

References

Iachan, R., Pierannunzi, C., Healey, K., Greenlund, K. J., & Town, M. (2016). National weighting of data from the Behavioral Risk Factor Surveillance System (BRFSS). *BMC Medical Research Methodology*, 16(1), 155. https://doi.org/10.1186/s12874-016-0255-7