# ASSESSING THE VALIDITY OF MATCHED ADDRESS DATA FOR UNLISTED TELEPHONE NUMBERS IN REACH 2010

**Joe Murphy, RTI International; Ann Burke, NORC; Whitney Murphy, NORC**

**KEY WORDS:** RDD, telephone, geocoding, screening

## Summary

Geographic pre-screening of survey households with listed telephone numbers is a time-saving, cost-effective measure. The listed numbers (which are usually included in traditional RDD frames) can be matched to an address and classified as inside or outside a target area using geocoding software. Those outside the geography can be designated as geographically ineligible without the cost of obtaining this information from household members. However, for unlisted numbers, conventional address matching provides no specific geographic information for pre-screening. Several vendors now offer tracing information on unlisted numbers from various sources, including warranty cards, web surveys, and other seco ndary listings. This paper aims to assess the quality of these data by comparing them with results captured in completed screeners from the second round of REACH 2010. Cost and quality data are compiled under the assumption that the cases were passively pre-screened alone or actively screened alone. Recommendations for future rounds and a discussion of the general utility of these data are also presented.

## Introduction

In household samples selected by random digit dialing (RDD), the exact location of each household is usually unknown at first. The samples can be drawn to target rough geographic areas, but numbers cannot be mapped to specific blocks, counties, or ZIP codes. Typically, when survey eligibility is based on such strict boundaries, cases are actively screened. That is, each telephone number in the sample is dialed, and if a household member answers, he or she is asked whether the household falls in the target geography. This method can be costly, however, particularly when the target geography is very small.

A method commonly used when the sample contains directory-listed telephone numbers is geographic pre-screening. A reverse-directory match can link each listed number to a specific address, which can then be categorized as geographically eligible or ineligible depending on its location. This process is often called "geocoding" and effectively eliminates the need to actively screen telephone numbers outside the area of interest. However, samples for telephone surveys often include unlisted telephone numbers as well as listed. By definition, directory-based address matching cannot be performed on unlisted numbers, which presents a challenge since unlisted numbers constitute the majority of an RDD sample frame.

To realize the cost savings of geographic-prescreening for unlisted numbers, geographic information must be obtained in another way. Recently, sample vendors have begun to offer address data for unlisted numbers, compiled from multiple non-directory sources. However, research regarding their quality is scant. This paper aims to bridge this gap and determine if geographic pre-screening of unlisted telephone numbers is a viable option.

## Data

The second round of the Racial and Ethnic Approaches to Community Health (REACH) 2010 survey afforded the opportunity to analyze non-directory addresses for a sample of unlisted RDD numbers. The purpose of REACH 2010 is to increase rates of protective behaviors and reduce rates of risk behaviors associated with disease, disability, and premature death that disproportionately affect certain minority racial and ethnic groups. The survey collects data on health risk behaviors among minorities in 27 communities by telephone for the purpose of evaluating coalition-driven health interventions in each community.

The REACH 2010 community samples are designed to capture information about each population of interest. Every community has a set of geographic boundaries and racial/ethnic targets. Most samples have a dual-frame design in which a portion is drawn from an RDD frame that excludes listed numbers, and the remainder is drawn from a listed telephone frame. The inclusion of unlisted numbers is important because a significant percentage of numbers in the United States are not included in telephone directories (Collins and Sykes, 1987). All numbers are drawn using a list-assisted method, which increases sampling efficiency while introducing a negligible amount of coverage bias (Brick, et al., 1995). Since the two frames are mutually exclusive, each number has only one chance of being selected.

For the first round of REACH 2010, directory-based addresses for listed numbers were purchased from three sample vendors. Traugott, Groves and Lepkowski (1987) and Brick, Collins, and Chandler (1997) showed that such addresses could effectively be used for the mailing of advance letters to reduce nonresponse. For this reason, letters were sent to geographically eligible households. The address data was not used for geographic pre-screening. Instead, the addresses were compared to those collected through active screening by telephone interviewers. In this way, it could be determined whether savings would have been realized had pre-screening been employed. The vendor that provided the most addresses on the listed numbers also turned out to have the highest level of congruence with the active screening data. That is, when the household was coded as inside the target geography from the address provided by that particular vendor, it was almost always coded as inside based on active screening as well. When the vendor address indicated the household was outside of the geography, the active screening almost always coded it as outside as well. The difference in data quality between vendors was attributed to their list sources and the frequency at which they were updated. The accurate and complete address matching offered by the selected vendor ensured cost savings and no loss of quality due to false negatives (vendor coded household outside the target geography when the household was actually inside). For these reasons, pre-screening was used for listed numbers in the second round of REACH 2010.

Based on the success with listed numbers, the experiment was extended to the unlisted numbers. For the unlisted RDD portion, samples are ordered by specifying the particular geographic boundaries in terms of counties, ZIP codes, or census tracts. The target geographies are translated into a set of telephone exchanges and analyzed as to how well they line up with the community's geographic boundaries. As a result of this analysis, the set of exchanges that represent the coverage level and geographic incidence most appropriate for the study are selected.

Two of the aforementioned sample vendors offer address data for unlisted numbers. They will be referred to as Vendor A and Vendor B. Addresses were purchased for the unlisted RDD telephone numbers from Vendors A and B and geocoded to indicate potential eligibility. Each unlisted RDD number was also actively screened by telephone interviewers who attempted to contact the sampled telephone numbers, administer a geographic screener, and record a disposition at the conclusion of each call. The passive (matched address) and active (telephone screener) data were then compared to determine their congruence.

The unlisted RDD samples, consisting of 9,969 unlisted telephone numbers in five communities, were sent to Vendors A and B for passive address matching. Because these numbers are unlisted, the address information was pulled from non-traditional non-directory sources as varied as warranty cards, web surveys, and even pizza delivery records. Table 1 presents the data sources reported by Vendors A and B. Each vendor returned partial and complete address information when available and a variable indicating whether the telephone number linked to a business or residence. The returned addresses were then matched to the community geography and categorized as inside or outside the community target geographies.

**Table 1. Vendor Data Sources**

| Vendor | Data Sources |
|---|---|
| A | • Call centers<br>• Companies in telecommuniciations and consumer goods<br>• Insurance and credit industries<br>• "Others" |
| B | • Recent new movers whose numbers have not yet been published, compiled from "various sources"<br>• Consumer databases<br>• "Others" |

For the same five communities, telephone interviewers attempted to contact each of the 9,969 unlisted telephone numbers. Responses to the geography screening questions were coded to indicate whether the respondent said the household was inside or outside of the geography, or if they did not know or refused to answer the question. Table 2 presents the eligibility criteria for each community. The specific screener questions are masked to protect the anonymity of the communities.

Vendor A provided full or partial address information for 9,965 of the 9,969 cases in the sample (99.96%). About 2,500 of these were also successfully screened by telephone interviewers. Vendor B provided address information for 2,103 cases (21.10%), 859 of which were also successfully screened by telephone interviewers.

**Table 2. Geographic Eligibility Criteria by Community**

| Community | Geographic Eligibility |
|---|---|
| 1 | Household is in one of 19 ZIP codes |
| 2 | Household is between 4 streets (bordering on north, south, east, and west) |
| 3 | Household is in one of 3 ZIP codes |
| 4 | Household is in one of 2 counties |
| 5 | Household is in one of 2 counties or 8 cities |

## Methods

To compare passive and active screening results for both vendors, cases were classified using the following match terminology: an *exact match* occurred when the vendor and screener geography codes were identical (both inside the geography or both outside the geography); a *false negative* occurred when the vendor address indicated the household is located outside the geography, but the screener respondent identified it as inside; a *false positive* occurred when the vendor address indicated the household is located inside the geography, but the screener respondent identified it as outside; *unknown* cases were those where the vendor address could not be geocoded definitely and/or the screener respondent was unable or unwilling to answer the geography questions.

As shown in Table 3, Vendor A's exact match rate was 83.7% (2,146 cases). This is comparable to results found in REACH with traditional address data for listed numbers, suggesting that some sources of address data from "non-traditional" sources may be as reliable for unlisted numbers as directory assistance data are for listed telephone numbers. Vendor A's false negative rate was 3.0% and the false positive rate was higher at 11.2%. Though it has not been fully investigated, this may be due to a tendency of certain respondents to "opt out" of the survey by falsely claiming geographic ineligibility. As suggested by Camburn, et al. (1995), if this occurs, the potential respondent is actually refusing to participate and will be incorrectly coded as ineligible.

Across communities, the results are fairly consistent, except in Community 2, where the false positive rate is very high. This could be due to the complicated nature of the geography questions for that community.

**Table 3. Vendor A Address Matches**

| Match with Screener | Overall | Community | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Exact Match | 2,146<br>83.7% | 585<br>85.5% | 184<br>54.8% | 70<br>79.6% | 511<br>89.3% | 796<br>90.2% |
| False Negative | 77<br>3.0% | 26<br>3.8% | 6<br>1.8% | 2<br>2.3% | 13<br>2.3% | 30<br>3.4% |
| False Positive | 288<br>11.2% | 55<br>8.0% | 132<br>39.3% | 15<br>17.1% | 43<br>7.5% | 43<br>4.9% |
| Unknown | 52<br>2.0% | 18<br>2.6% | 14<br>4.2% | 1<br>1.1% | 5<br>0.9% | 14<br>1.6% |
| Total | 2,563 | 684 | 336 | 88 | 572 | 883 |

As Table 4 shows, Vendor B's exact match rate was 85.9%. Although this rate is comparable to Vendor A, the absolute number of 738 exact matches for Vendor B was much lower than the 2,146 for Vendor A. Similar to Vendor A, Vendor B's false negative rate (2.4%) was lower than its false positive rate (10.4%). In Community 2, the false positive rate was also very high (53.5%) compared to the false negative rate (0.0%).

**Table 4. Vendor B Address Matches**

| Match with Screener | Overall | Community | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Exact Match | 738<br>85.9% | 110<br>83.3% | 32<br>45.1% | 28<br>87.5% | 193<br>89.8% | 375<br>91.7% |
| False Negative | 21<br>2.4% | 3<br>2.3% | 0<br>0% | 0<br>0% | 2<br>0.9% | 16<br>3.9% |
| False Positive | 89<br>10.4% | 15<br>11.4% | 38<br>53.5% | 3<br>9.4% | 17<br>7.9% | 16<br>3.9% |
| Unknown | 11<br>1.3% | 4<br>3.0% | 1<br>1.4% | 1<br>3.1% | 3<br>1.4% | 2<br>0.5% |
| Total | 859 | 132 | 71 | 32 | 215 | 409 |

Comparing Vendors A and B, the match rates are similar; Vendor A matched more cases than Vendor B; each had a high rate of false positives in Community 2; and each had more false positives than false negatives. Because false negatives represent the cases that could potentially introduce the most bias into the sample, they were compared to other cases in terms of screener demographic data to determine if they were systematically different. While these comparisons are not presented here, it should be noted that significant differences were not detected. The accidental exclusion of false negatives does not appear to introduce bias in this respect.

Geography was not the only consideration in the vendor comparison. Even if it is determined that a telephone number lies inside or outside the geography of interest, this does not insure that the number links to an actual residential address. Because of this concern, the results of the business or residential telephone number indicator provided by each vendor were examined. A total of 5,801 indicators were returned from Vendor A (58.2%) and 2,103 from Vendor B (21.1%). These indicators were compared to the CATI call dispositions presented in Table 5.

**Table 5.CATI Disposition Mappings**

| Screener Type Flag | Outcome Disposition |
|---|---|
| Business | Business |
| Household (Consumer) | Completed screener |
| | Soft refusal |
| | Hostile refusal |
| | HH line |
| | HH, eligibility unknown |
| | HH members away for field period |
| | Language barrier |
| | Mentally/physically incapacitated |
| | Privacy blocker |
| | Answering machine |
| Neither Business nor Household | Cell phone/pager |
| | Disconnected number |
| | Computer/fax line |
| Unknown | Ring-no-answer |
| | HH status unknown |

Vendor A provided exact matches with the CATI disposition for 3,606 of the cases (62.2%). The false negative rate was 4.4%. Vendor B provided exact matches for only 1,550 cases, but had a greater percentage of exact matches (73.7%). Vendor B's false negative rate of 1.3% was lower than Vendor A's (4.4%). The results of the business and residential number matching are presented in Table 6.

**Table 6. Business/Residential Matches**

| Vendor Code | Screener Code | Vendor A | Vendor B |
|---|---|---|---|
| Business | Business | 601<br>10.4% | 156<br>7.4% |
| | Household | 256<br>4.4% | 28<br>1.3% |
| | Neither | 415<br>7.2% | 23<br>1.1% |
| | Unknown | 255<br>4.4% | 14<br>0.7% |
| Household (Consumer) | Completed screener | 153<br>2.6% | 51<br>2.4% |
| | Soft refusal | 3,005<br>51.8% | 1,394<br>66.3% |
| | Hostile refusal | 904<br>9.1% | 368<br>17.5% |
| | HH line | 215<br>2.2% | 69<br>3.3% |
| TOTAL | | 5,801<br>100% | 2,103<br>100% |
| Exact Matches | | 3,606<br>62.2% | 1,550<br>73.7% |

**Results**

In terms of cost and quality, using Vendor A for geographic pre-screening of telephone numbers would have saved approximately 15% of the budget for screening. This figure was determined by starting with the costs of screening actively alone, subtracting the costs that were incurred by actively screening ineligible cases, subtracting the value of the time saved by eliminating the geography questions in the screener, and adding the cost of geocoding unlisted numbers. Using Vendor B would have resulted in a savings of approximately 5%. In terms of the introduction of potential bias, both vendors would not introduce so many false negatives as to make this method unreasonable. Being more conservative and using both vendors would be much more expensive and would not significantly improve the quality of the match results.

For the analysis of business and residential telephone number indicators, it was found that Vendor A provided three times as many exact matches, but Vendor B provided a greater percentage of exact matches. Using this method with Vendor A would have provided a cost savings of about 8% on screening, but less than 1% with Vendor B. In terms of potential bias, Vendor A's low match rate was deemed insufficient, and the gain in quality offered by Vendor B did not offset the low cost savings they offered. Using a hybrid approach that would incorporate data from Vendors A and B would not result in cost savings and would not greatly reduce potential bias.

## Conclusions

Geographic pre-screening on unlisted numbers is a viable cost savings option for REACH 2010 and possibly for other surveys. There were many differences in the results provided by the two vendors, so it is important to evaluate several alternative vendors to determine what can be offered in terms of cost savings and data quality. Vendors should be forthcoming with information and documentation regarding their data sources to assure quality results. While the sources of some vendor data may be proprietary information, this information is needed to understand the implications of a process that may begin with an individual ordering pizza and end with a survey research organization determining his or her eligibility for an important health survey.

The business and residential indicators provided by the vendors were of questionable quality. It was not found that these data could be used alone to code telephone numbers as eligible or ineligible without also determining the geographic eligibility of the number. Based on this experiment, if there is evidence that a telephone number links to any structure in the area of interest (whether it is a business or residence), it is advisable to call that number and actively determine whether it is actually a household or something else.

Looking to the future, a follow-up experiment is being considered which would compare vendor- supplied data against a database of unlisted telephone numbers with known addresses to further investigate the accuracy of vendor data. This may be conducted using the unlisted portion of an employee or professional organization phone and address database. Other vendors besides "A" and "B" should be evaluated, and interview responses, rather than screener demographics alone, should be examined to determine whether potential bias may be introduced using this method. The impact of this method on response rates should also be examined. Lastly, this experiment should be replicated in other communities and in a national study to determine whether it is a viable method across all types of households and surveys.

## References

Brick, M., Collins, M., and Chandler, K. (1997). U.S. Department of Education. National Center for Education Statistics. An Experiment in Random-Digit-Dial Screening, NCES 98-255,. Washington, DC: 1997.

Brick, J., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in List-Assisited Telephone Samples. *Public Opinion Quarterly*, 59:218-235.

Camburn, D., Lavrakas, P., Battaglia, M., Massey, J., and Wright, R. (1995). Using Advance Respondent Letters in Random-Digit-Dialing Telephone Surveys. *Proceedings of the American Statistical Association.*

Collins, M., and Sykes, W. (1987). The Problems of Non-coverage and Unlisted Numbers in Telephone Surveys in Britain. *Journal of the Royal Statistical Society*, 150, Part 3, pp. 241-253.

Traugott, M.R., Groves, R.M., and Lepkowski, J. (1987). Using Dual Frame Designs to Reduce Nonresponse in Telephone Surveys. *Public Opinion Quarterly*, 51:522-539.