Cell Phone Methodology

PAPOR Mini Conference
June 28, 2013

Paul Johnson
SSI
PAPOR Mini Conference
June 28, 2013

Papers/slides obtained with author permission. Any misinterpretations are my own.
Recent Methodological Updates Adopted for the National Immunization Survey (NIS)

Vicki Pineau¹, Kirk Wolter¹, Robert Montgomery¹, Bess Welch¹, and Stacie Greby²
¹NORC at the University of Chicago
²Centers for Disease Control and Prevention

2013 AAPOR Annual Conference
May 17, 2013
Dual-Frame Results

- Minimized the perceived non-representative sample risk
- Maintained reasonable vaccination coverage estimates
  - Comparable to LL only data and NHIS PRC
- Response rates decreased
- Costs increased
  - Expansion was supported by cost savings associated with the following methodological changes
    - Expansion of the NIS-Child age-eligibility criteria
    - Shortening the household telephone questionnaire
    - Incorporating efficient sample weighting methods via shrinkage weighting
Dual-Frame* vs. Landline Only Vaccination Coverage Rate Estimates, NIS, United States, 2011

* Dual-frame vaccination coverage rate estimates computed using shrinkage weighting to minimize MSE
Comparison of Bias and Variance of Shrinkage Weighting to Full Weighting, NIS, United States, Q2-Q4 2011

No statistically-significant differences at the $\alpha=0.05$ level.

Reduction in width of 95% confidence intervals.
### Birth Cohorts Using Any Day of Quarter Design, NIS, United States, 2011

<table>
<thead>
<tr>
<th>Month of Birth</th>
<th>Q1 2011</th>
<th>Q2 2011</th>
<th>Q3 2011</th>
<th>Q4 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 09</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Jan 09</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Feb 09</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>Mar 09</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Apr 09</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>May 09</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Jun 09</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Jul 09</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Aug 09</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Sep 09</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Oct 09</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Nov 09</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Dec 09</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

**Color Key**
- Age Eligible Day of Screening
- Age Eligible Any Day of Quarter (Old)
- Age Eligible Any Day of Quarter (Young)
Sampling Cell Phones by Rate Center: Efficacy, Coverage and Incidence

David Dutwin, SSRS
David Malarek, MSG
Example

Map of Rate Centers

DC Metro
Data

- MSG’s Rate Center Dataset:
  - Rate center geographic data
  - Appended Census data for demographics
  - Square Miles
  - Population
  - Households
  - 18+ Population
  - Ethnicity
  - Vacancy
  - Tenure
  - CPO Households

- SSRS EXCEL omnibus survey:
  - 12,229 cell phone interviews
  - Full demographic battery

- Unified EXCEL/Rate Center Dataset
Significant Differences by Respondent Characteristics

Rate Center Mismatches

50 percent of cell owners live in their rate center.

- **Center City**
  - Match: 67%
  - Mismatch: 60%

- **Have Child**
  - Match: 56%
  - Mismatch: 59%

- **College Degree**
  - Match: 28%
  - Mismatch: 33%

- **$50,000 or More**
  - Match: 39%
  - Mismatch: 42%

- **Male**
  - Match: 56%
  - Mismatch: 59%

- **Under 35**
  - Match: 37%
  - Mismatch: 42%
Differences by Rate Center Characteristics

Percent Cell Owners Not Living in Home Rate Center by Distribution of Key Rate Center Parameters

- Square Miles
- Population
- Population Density
Significant Differences by Respondent Characteristics

CBSA Mismatches

**Center City**
- Match: 67%
- Mismatch: 60%

**Own Home**
- Match: 42%
- Mismatch: 51%

**College Degree**
- Match: 29%
- Mismatch: 38%

**3+ Persons**
- Match: 53%
- Mismatch: 46%

**Black/Hispanic**
- Match: 30%
- Mismatch: 22%

**Under 35**
- Match: 37%
- Mismatch: 51%
Differences by Rate Center Characteristics, CBSA Level

Percent Cell Owners Not Living in Rate Center CBSA by Distribution of Key Rate Center Parameters

- Square Miles
- Population
- Population Density
Rate Centers: Conclusions

1. Half of cell phone owners live in their rate center
2. Over three quarters live in the CBSA of their rate center
3. Mismatches more likely in mid-atlantic and northeast; in small rate centers; in rate centers high in population density
4. Mismatches more likely to be non-metro, educated, and young; CBSA mismatches more likely to be home owners, educated, white, young, and in single family households
5. CBSA on average attain 73 percent incidence; 75 percent coverage; logically, geographies smaller than a CBSA will be worse on both metrics
Alternative Sample Selection and Data Collection Strategies for Balancing Cell Phone Response Distribution Across County/Region Level Geographies in a Dual Frame Telephone Survey

AAPOR 2013 Methodological Briefs: Cell Phones

Howard Speizer, Marcus Berzofsky, Tom Duffy, Jamie Ridenhour (RTI)
Tim Sahr (Ohio State University)

May 17, 2013
Rate Center

- Rate Center holds some promise as a stratification variable because it is assigned to every cell number.
- Classification error however, is an issue:
  - Incorrect state=7.7%
  - Classification error statewide=33%
  - In Hamilton County the false-positive rate was 48%
# OMAS Results with Adjusted Rate Center

<table>
<thead>
<tr>
<th>Medicaid Region</th>
<th>Estimate of cell completes based on population totals</th>
<th>Estimate of cell completes based on %COP</th>
<th>Estimate of cell completes based on rate center</th>
<th>Estimate of cell completes based on adjusted rate center</th>
<th>Cell achieved - expected based on modified rate center</th>
<th>Non-response ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central</td>
<td>1039</td>
<td>1162</td>
<td>1056</td>
<td>1094</td>
<td>39</td>
<td>1.04</td>
</tr>
<tr>
<td>East Central</td>
<td>695</td>
<td>675</td>
<td>745</td>
<td>701</td>
<td>-21</td>
<td>0.97</td>
</tr>
<tr>
<td>Northeast</td>
<td>1048</td>
<td>891</td>
<td>721</td>
<td>678</td>
<td>-77</td>
<td>0.89</td>
</tr>
<tr>
<td>Northeast Central</td>
<td>257</td>
<td>222</td>
<td>322</td>
<td>309</td>
<td>-37</td>
<td>0.88</td>
</tr>
<tr>
<td>Northwest</td>
<td>573</td>
<td>569</td>
<td>626</td>
<td>633</td>
<td>8</td>
<td>1.01</td>
</tr>
<tr>
<td>Southeast</td>
<td>302</td>
<td>249</td>
<td>253</td>
<td>254</td>
<td>-31</td>
<td>0.88</td>
</tr>
<tr>
<td>Southwest</td>
<td>772</td>
<td>887</td>
<td>940</td>
<td>978</td>
<td>42</td>
<td>1.04</td>
</tr>
<tr>
<td>West Central</td>
<td>507</td>
<td>536</td>
<td>527</td>
<td>544</td>
<td>77</td>
<td>1.14</td>
</tr>
</tbody>
</table>
Sample Design for Next OMAS

- Using 2012 survey data we can estimate rate center error and non-response by county to improve the cell phone sample stratification.

- Allocate desired number of cell phone completes to counties based on corrected rate center distribution to get desired nominal sample size by county.
Steps to Determine Sample Size by County

- Adjust nominal sample size by following adjustments to get starting sample size of cell phone numbers

1. Frame error adjustment: ratio of initial rate center distribution and corrected rate center distribution
2. Response rate adjustment: expected aggregate response rate times the ratio of prior survey number of completes over prior survey expected number of completes
3. Out of state adjustment: ratio of number of contacted persons in rate center county indicating they lived out of state over number of contacted persons in rate center county
4. Ineligibility adjustment: aggregate ineligibility adjustment specific to study design other than out of state adjustment
Appending Billing ZIP Process

- SSI’s *Wireless Geo-ID Process*
  - Post-production process (run after sample is pulled)
  - Appends billing ZIP code of the cell phone number using information in proprietary data sources

- Three geographic data sources for comparison:
  - *Rate center data* (all sample records): based on the FIPS of the rate center in the cell sample
  - *Billing data* (matched cases only): based on data appended from Wireless Geo-ID process
  - *Respondent-reported data* (screened cases only): geographic residence information reported by respondent in survey
National Data: Accuracy of Billing ZIP Compared to Rate Center

**Billing vs. respondent-reported data**

<table>
<thead>
<tr>
<th>Level</th>
<th>Accuracy (Match) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>90.3%</td>
</tr>
<tr>
<td>County</td>
<td>72.5%</td>
</tr>
<tr>
<td>ZIP</td>
<td>53.9%</td>
</tr>
</tbody>
</table>

**Rate center vs. respondent-reported data**

<table>
<thead>
<tr>
<th>Level</th>
<th>Accuracy (Match) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>91.0%</td>
</tr>
<tr>
<td>County</td>
<td>58.3%</td>
</tr>
<tr>
<td>ZIP</td>
<td>N/A</td>
</tr>
</tbody>
</table>
How do matched and unmatched cases differ: *National vs. NYC Demographics*

<table>
<thead>
<tr>
<th>Unmatched are more likely to be:</th>
<th>National Sample (Unmatched vs. Matched)</th>
<th>NYC Sample (Unmatched vs. Matched)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less educated</td>
<td>HS or less (54.1% vs. 37.5%)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>HS or less (60.9% vs. 28.6%)&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Lower income</td>
<td>Less than $50K (64.4% vs. 48.3%)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Less than povertyX2 (68.1% vs. 35.0%)&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Hispanic, non-White</td>
<td>Hispanic (26.0% vs. 14.2%)&lt;sup&gt;2&lt;/sup&gt; White (47.8% vs. 66.3%)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Hispanic (42.4% vs. 24.0%)&lt;sup&gt;2&lt;/sup&gt; White (14.0% vs. 35.9%)&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Renters</td>
<td>Rent (48.4% vs. 39.9%)&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Not asked</td>
</tr>
<tr>
<td>Younger</td>
<td>16 to 34 (48.3% vs. 44.0%)&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Not significant</td>
</tr>
<tr>
<td>Cell-phone only</td>
<td>Not significant</td>
<td>CPO (72.5% vs. 62.6%)&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>1</sup>p<.05,  <sup>2</sup>p<.001
Conclusion & Recommendations

- Appended billing ZIP can be used to improve efficiency
  - Low match rate can limit use/benefit
  - Unmatched cases are different, should include at least some

- Lots of regional variability in accuracy of rate center & appended ZIP data (completeness/accuracy)
  - Use all available info when defining frame, but beware variation due to change over time or small sample sizes

- Develop initial sampling design/stratification but plan to evaluate after implementation and adjust if needed
  - Pilot test can be useful, but needs to be large enough
  - Test with limited replicates, but need time to evaluate/refine
Cell-Phone Sampling Frames: Effectiveness and Dependability of Recent-Usage Data

American Association for Public Opinion Research
Boston, Massachusetts
May 19, 2013

Presented by: Robert DeHaan and Robin Gentry
Acknowledgements: SSI, Neustar
Why Cell Phone RDD and Why Now?

» Dual Frame (Address Based v Cell Phone RDD)

» Increasing Incidence of Cell Phone Only/Mostly Households

» New Innovations
  • Usage Based Activity Indicators
  • Billing Zip Append
  • Pre-placement mailings
Grouped according to usage over last 10 months

- 16,050 “active” numbers hand-dialed
- 2 weeks across 6 diverse markets
- Pre-placement postcard (50/50)
- Placement phone calls made
- Cell Phone vs. Landline usage questions asked

Study Overview (Activity Indicator Split)

1. High Usage
2. Moderate Usage
3. No Usage

Contactability Study
Contact/Consent Rates

Consent Rate = \frac{\text{Agree}}{\text{Agree} + \text{Non-Agree}}

<table>
<thead>
<tr>
<th>Disposition</th>
<th>High Usage</th>
<th>Moderate Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Operational Consent Rate</strong></td>
<td>17.11%</td>
<td>16.39%</td>
</tr>
<tr>
<td>Agree</td>
<td>12.80%</td>
<td>9.75%</td>
</tr>
<tr>
<td>No Answer</td>
<td>19.66%</td>
<td>16.24%</td>
</tr>
<tr>
<td>Won’t Talk</td>
<td>38.36%</td>
<td>28.68%</td>
</tr>
<tr>
<td>Answering Machine</td>
<td>4.00%</td>
<td>4.27%</td>
</tr>
<tr>
<td>Disconnect</td>
<td>19.18%</td>
<td>35.86%</td>
</tr>
<tr>
<td>Other</td>
<td>6.00%</td>
<td>5.20%</td>
</tr>
</tbody>
</table>

**Usable: Agree**

**Usable: Non-Agree**

**Unusable**
Conclusions

» Activity Indicators are useful when screening out unused numbers (Indicator 3)

» Pre-placement postcard was not particularly effective, but better results could be expected with a letter and a dollar

» Improved proportionality amongst 12-24 year olds, offset by reduced representation amongst 25-34 year olds

» Cell Phone Only density in frame is nearly double what was found in a 2006 Arbitron study
Attempting to Boost RDD Cell Sample Productivity by Identifying Non-working Numbers Prior to Dialing

Michelle Mosher – SSI
Jonathan Best – Princeton Survey Research Associates International
What is the ‘Phone Activity Flag’?

- Phone activity = the frequency of calls generated from a specific wireless telephone number
  - Sourced from caller ID network information
  - Information from over 175 carriers
    - Majority of regional carriers
    - Majority of cable companies
    - Inbound calls from ‘out of network’ also captured
    - 95% of outgoing calls captured (remaining 5% are mostly activity from text only numbers with no voice capability)
### Omnibus Sample – Screening Results

<table>
<thead>
<tr>
<th>Sample/Comps</th>
<th>Active</th>
<th>Inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feb 2013 week 4 Sample</td>
<td>71.3%</td>
<td>28.7%</td>
</tr>
<tr>
<td>Feb 2013 week 4 Comps</td>
<td>95.6%</td>
<td>4.4%</td>
</tr>
<tr>
<td>Feb 2013 week 3 Sample</td>
<td>70.3%</td>
<td>29.7%</td>
</tr>
<tr>
<td>Feb 2013 week 3 Comps</td>
<td>96.0%</td>
<td>4.0%</td>
</tr>
<tr>
<td>Aug 2012 week 3 Sample</td>
<td>67.7%</td>
<td>32.3%</td>
</tr>
<tr>
<td>Aug 2012 week 3 Comps</td>
<td>95.0%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Aug 2012 week 2 Sample</td>
<td>66.9%</td>
<td>33.1%</td>
</tr>
<tr>
<td>Aug 2012 week 2 Comps</td>
<td>94.8%</td>
<td>5.3%</td>
</tr>
</tbody>
</table>
Disposition Comparison

Outcome Rates by Activity Status

<table>
<thead>
<tr>
<th></th>
<th>Inactive</th>
<th>Active</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working</td>
<td>21.7%</td>
<td>47.3%</td>
<td>62.1%</td>
</tr>
<tr>
<td>Contact</td>
<td>37.7%</td>
<td>46.2%</td>
<td></td>
</tr>
<tr>
<td>Cooperation</td>
<td>13.2%</td>
<td>19.0%</td>
<td>18.5%</td>
</tr>
<tr>
<td>Eligibility</td>
<td>47.9%</td>
<td>60.8%</td>
<td>60.0%</td>
</tr>
</tbody>
</table>
## Effects on Productivity

### Interviewer Hours by Activity Status

<table>
<thead>
<tr>
<th></th>
<th>Active</th>
<th>Inactive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviews (Ints)</td>
<td>480</td>
<td>20</td>
<td>500</td>
</tr>
<tr>
<td>Interviewer hours (Int hr)</td>
<td>123.8</td>
<td>31.7</td>
<td>155.5</td>
</tr>
<tr>
<td>Productivity (Ints/Int hr)</td>
<td>3.9</td>
<td>0.6</td>
<td>3.2</td>
</tr>
</tbody>
</table>

\[
\frac{(3.9 - 3.2)}{3.2} = 21\%
\]
Conclusions

- Cell phone screening does a good job at identifying unproductive numbers. Over 60% of all non-working numbers are flagged as inactive.
- Inactive numbers lead to few interviews – only about 4 to 5 percent of total completes.
- The screening increases the working rate of a Continental US cell sample from 62% to 80%.
- No evidence that excluding the inactive sample leads to bias in basic sample demographics.
- We estimate that excluding the inactive numbers would lead to increases in cell sample productivity of 17-21%.
The Mechanics of GPS Geo-Location for Mobile Devices:

Their Potential for Measurement Error and Some Illustrative Data

Dr. Max Kilger and TraShawna Boals
OVERVIEW

- Uses of geo-location in survey and market research
- Technologies for generating Geo-location data
- How GPS works
- Measurement Errors
- Illustrative Data from passive meter technology
Uses of geo-location in survey and market research

- What is geo-location research?

- Examples:
  - Active: When a participant is located near a point of interest to the research, a survey is triggered.
  - Passive: When a participant has a mobile app installed on device and it periodically captures GPS coordinates.
Technologies for generating Geo-location data in mobile devices

- Global Positioning Systems (GPS)
- Cell Tower Triangulation
- Wifi geo-location
Technologies for generating Geo-location data in mobile devices

- Global Positioning Systems (GPS)
  - U.S. system - 31 satellites in medium earth orbit
  - At least 4 satellites visible at any point on Earth
  - More satellites visible = more accurate geo-location fix
How GPS Geo-Location Works

- GPS signal has three basic components
  - **Date/Time/Satellite Status** – ultra-accurate time signal used in calculating satellite-GPS receiver trip time
  - **Ephemeris** – very accurate estimate of the satellite position at the time of signal
  - **Almanac** – reference data with satellite id’s, approximate orbits, satellite clock statuses

- GPS-equipped device position determined by combining time and ephemeris data (position) from multiple satellites

Accuracy: +/- 10 meters
How aGPS Geo-Location Works

- aGPS – assisted GPS is a way to speed up GPS-based geo-location and is found on many smartphones
  - Obtains a quicker TTFF – Time to first fix – time it takes to determine geo-location by acquiring and storing information about the location of satellites via cellular network.
  - aGPS utilizes GPS information from cell towers with known geo-locations to establish rough location and shorten iterative geo-location process for the GPS signal calculations
Technologies for generating Geo-location data in mobile devices

- Cell Tower Triangulation
  - Multiple cell towers with known locations
  - Cell tower signals overlap by purpose
  - More towers = better geo-location fix
How Cell Tower Triangulation Works

- Geo-location using multiple cell tower signals

- Multiple pieces of information utilized
  - Signal strength from each tower
  - Known geo-location of each tower
  - When available, directional information based on which antenna segment is active

Accuracy: +/- 1000 meters down to 50 meters
Technologies for generating Geo-location data in mobile devices

- Wifi geo-location
  - Typical 30 meter signal
  - Wifi signal detected by mobile device
  - Large databases of wifi points of presence used to ascertain fix
How Wifi Geo-location Works

- Wifi geo-location uses signals from one or more wifi sources
  - Typical sources include privately/publically owned wifi routers and access points
  - These devices use TCP/IP protocols – this means that each wifi device has a unique MAC address – no other device has the same address
  - Entities like Google used initiatives like StreetView to collect wifi MAC addresses and exact geo-location data
  - Large databases containing MAC address and geo-location allows mobile devices that encounter wifi devices to acquire fix

Accuracy: +/- 100 meters down to 5 meters
Illustrative Data from Passive Measurement Technology

- Sample data from mobile measurement panel utilizing passive application
- App collects GPS data from mobile device – 2 day path shown below
GPS Measurement Errors

- GPS satellite signal is line of sight – thus things like buildings are a problem

- Esoteric error sources
  - Ionosphere disturbances
  - Solar flares
  - Relativistic effects due to speed of satellite

- Common error sources
  - Multipath errors due to buildings
  - Faraday effects due to things like steel vehicle bodies
  - Clock drift on GPS-enabled devices
Example of GPS Error on Smartphone

- The GPS calculated path is ~150 meters off the roadway
- Likely to be Faraday cage effect due to steel body of vehicle
Cell Tower Triangulation Error

- Scene from “Breaking Bad” burying a body? No, cell tower triangulation error
- Person likely entered building 300 meters south and lost satellite signal
- Star Trek “Transporter Effect”
- Ft Lauderdale → Tampa → Ft Lauderdale in 90 seconds
Privacy Issues

- Significant privacy issues with geo-location data
  
  - Only 4 date/time/geo-location data points needed to identify single smartphone from 1.5 million smartphones\(^1\)
  
  - 34.8% of U.S. adults say that they are willing to provide some personal information to a company in order to get something they want\(^2\)

2. Experian Simmons Fall 12 Month National Consumer Study
Three current geo-location technologies – often used in concert

Each of these technologies has inherent measurement error sources

Geo-location technologies and data present non-trivial privacy issues
Contact

- Max Kilger – max.kilger@experian.com
- TraShawna Boals – trashawna.boals@experian.com
Cell Phone Revolution
Communication Revolution
New Challenges
New Opportunities