Investigating Internet Opt-in Panels for Behavioral Surveillance

Carol A. Gotway Crawford, PhD
Chief, Population Health Surveillance Branch

Catherine Okoro, PhD
Epidemiologist

Satvinder Dhingra, MPH
Epidemiologist
Northrop Grumman

Federal Committee on Statistical Methodology Research Conference
November 4, 2013
Washington, DC
What Are Internet Opt-in Panels?

- **Potential panelists are recruited via the Internet**
  - Banner ads, email lists, promotions, and offers
  - Double opt-in process to become a panel member

- **Panelists become the pool for sample selection**

- **Panel may or may not be representative of the population**
  - Coverage is limited to Internet users (~80% of the population)
  - Respondent selection and motivation
Why Use Internet Opt-in Panels?

- Lower cost than probability-based sampling
- Shorter collection and prep time for data release to the public than current methods (RDD, face-to-face)
- Expands the surveillance and study tool-kit
- Permits longitudinal and in-depth follow-up studies
- Increases administrative and design flexibility and efficiency
Pilot Study

- **4 States**
  - Cooperative agreements in GA, IL, NY, and TX

- **3 Vendors**
  - Different sampling methodologies
  - Cooperating and collaborating
    - De-duplication of respondents
    - Nearly identical questionnaire format

- **3 Levels of Geography**
  - National
  - State
  - Metropolitan Statistical Area
Pilot Objectives

- Compare sampling methodologies
  - Sample matching, source blending, and quota
- Assess feasibility and accuracy for public health
- Compare estimates with those from other surveys
- Evaluate across a range of parameters:
  - Cost, geographic granularity, and timeliness
Sampling Methodologies

- **Sample Matching**
  - Different modes of recruitment are used to ensure representativeness for hard-to-reach populations
  - Potential respondents are selected by matching to a random sample from the American Community Survey
  - Final responses are weighted to known characteristics in the U.S. using propensity score weighting

- **Sample Blending**
  - Uses population segments designed to reflect behavioral differences but based on Census data
  - Apply the segmentation structure locally to balance, weight, and blend sample

- **Quota Sampling**
  - A non-probability sample in which respondents take the survey on a first-come, first-served basis according to a fixed quota
Questionnaire Development

- Survey consists of ~80 questions (20 minutes)

- Questions drawn from:
  - CDC: BRFSS, NHANES, & NHIS
  - NIH: PROMIS
  - SAMHSA: NSDUH
  - ONC: Consumer Survey of Attitudes Toward the Privacy and Security Aspects of EHR and HIE
  - NPWF (National Partnership for Women and Families)
  - NSF supported Cooperative Congressional Election Study
National: Demographics (Unweighted)

- **NHIS 2012 HH CAPI**
- **BRFSS 2012 DF-RDD CATI**
- **YouGov 2013 IPS Matched**
- **CPS 2012**

### Demographics

- **Gender**
  - Male
  - Female

- **Age Groups**
  - 18 - 29
  - 30 - 44
  - 45 - 64
  - 65+

- **Education**
  - <HS
  - =HS
  - >HS

### Comparison of Data Sources

- **YouGov 2013 IPS Matched**
- **BRFSS 2012 DF-RDD CATI**
- **NHIS 2012 HH CAPI**
- **CPS 2012**
MSA: Race/Ethnicity

Atlanta

Chicago

New York

Houston

- Mktg Inc. 2013
- uSamp 2013
- IPS Quota
- YouGov 2013
- IPS Matched
- SMART BRFSS 2012
- DF-RDD CATI
State: Education

![Graphs showing education levels by state](image)

- **GA**: Bars for <HS, =HS, >HS categories with data from various sources like Mktg Inc. 2013, uSamp 2013, YouGov 2013.
- **IL**: Similar bars for <HS, =HS, >HS with data from NHIS 2011, HH CAPI.
- **NY**: Similar bars for <HS, =HS, >HS with data from BRFSS 2011, DF-RDD CATI.
- **TX**: Similar bars for <HS, =HS, >HS with data from IPS Matched, uSamp 2013.
National: Outcomes

Obesity (BMI ≥30)

Diabetes

Fair/Poor

Ex/VG/GD

YouGov 2013

BRFSS 2012

NHIS 2012

IPS Matched

DF-RDD CATI

HH CAPI
MSA: Obesity (BMI ≥30)
State: Diabetes

[[Bar charts for GA, IL, NY, TX showing data for Diabetes]]
MSA: Diabetes

![Graphs showing diabetes prevalence in Atlanta, Chicago, New York, and Houston. Each graph compares different sampling methods: Mktg Inc. 2013, uSamp 2013, YouGov 2013, IPS Quota, IPS Matched, SMART BRFSS 2012, DF-RDD CATI.](image-url)
State: Disability

GA

IL

NY

TX

Mktg Inc. 2013
Blended

uSamp 2013
IPS Quota

YouGov 2013
IPS Matched

BRFSS 2013
Preliminary
DF-RDD CATI

NHIS 2011
HH CAPI
MSA: Disability

**Atlanta**

**Chicago**

**New York**

**Houston**

- **Mktg Inc. 2013 Blended**
- **uSamp 2013 IPS Quota**
- **YouGov 2013 IPS Matched**
- **SMART BRFSS 2012 DF-RDD CATI**
State: Cost Barrier

- **GA**
  - Mktg Inc. 2013: 30
  - uSamp 2013: 20
  - YouGov 2013: 20
  - NHIS 2011: 10
  - HH CAPI: 10
  - BRFSS 2011: 10
  - DF-RDD CATI: 10
  - IPS Matched: 10
  - IPS Quota: 10

- **IL**
  - Mktg Inc. 2013: 40
  - uSamp 2013: 30
  - YouGov 2013: 30
  - NHIS 2011: 20
  - HH CAPI: 20
  - BRFSS 2011: 20
  - DF-RDD CATI: 20
  - IPS Matched: 20
  - IPS Quota: 20

- **NY**
  - Mktg Inc. 2013: 20
  - uSamp 2013: 10
  - YouGov 2013: 10
  - NHIS 2011: 10
  - HH CAPI: 10
  - BRFSS 2011: 10
  - DF-RDD CATI: 10
  - IPS Matched: 10
  - IPS Quota: 10

- **TX**
  - Mktg Inc. 2013: 30
  - uSamp 2013: 20
  - YouGov 2013: 20
  - NHIS 2011: 10
  - HH CAPI: 10
  - BRFSS 2011: 10
  - DF-RDD CATI: 10
  - IPS Matched: 10
  - IPS Quota: 10
State: Current Smoker

**GA**

**IL**

**NY**

**TX**

![Bar graphs for different states showing current smoker rates.](image-url)
State: Heavy Drinker

GA

IL

NY

TX

Mktg Inc. 2013
uSamp 2013
YouGov 2013

IPS Quota
IPS Matched

NHIS 2011
HH CAPI

BRFSS 2011
DF-RDD CATI
Quantifying Uncertainty

- The use of Frequentist confidence intervals with data from a non-probability sample is theoretically inappropriate.

- Bayesian credible intervals are a more appropriate way to quantify uncertainty when analyzing data from a non-probability sample.

- In our pilot studies, however, both methods yielded highly similar, if not identical, results.
## Uncertainty Comparison

<table>
<thead>
<tr>
<th>Variable</th>
<th>Confidence Interval</th>
<th>Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>29.22</td>
<td>32.61</td>
</tr>
<tr>
<td>Diabetes</td>
<td>9.88</td>
<td>11.95</td>
</tr>
<tr>
<td>High BP</td>
<td>26.64</td>
<td>29.62</td>
</tr>
</tbody>
</table>
Major Benefits

• **Time** (samples constructed to be representative):
  < 15 days for a national survey ~ 4,000 interviews
  ~ 30 days for most states ~3,000 interviews
  ~ 30 days for large (5+ million) MSAs ~2,000 interviews

• **Cost:**
  – Internet opt-in panels: $5-$15 per completed interview
    • Costs include editing and weighting
  – Dual-frame RDD State direct costs average ~$70/CI
    • Considerable additional costs for editing and weighting
Preliminary Results

• **Great deal of similarity**
  – Results of sample matching comparable with BRFSS and NHIS
  – Variation among surveys consistent across states
  – Internet opt-in panels fairly accurate at lower levels of geography
  – Quota sampling not as accurate

• **Differences can be attributed to:**
  – Coverage effects (sample selection*outcome interaction)
  – Use of different control totals and weighting methods
  – Mode effects (face-to-face, telephone, Internet)
  – Question differences and order effects
  – Temporal changes (2013 vs. 2011)
  – Sample size differences
  – Cross-sectional differences
Acknowledgements

- Stephen Ansolabehere, Harvard University & CCES
- Steven Gittelman, Mktg. Inc.
- Douglas Rivers, Stanford University & YouGov
- Meena Khare, NCHS
- Georgia, Illinois, New York, and Texas Departments of Health, BRFSS
  - Rana Bayakly, Madhavi Vajani, Francis Annor, GA
  - Bruce Steiner, IL
  - Bethany Hawke, NY
  - Rebecca Wood, TX
- Haci Akcin, CDC
- Derek Ford, CDC
- Guixiang Zhao, CDC
- Soyoun Park, CDC