Turning Knowledge into Practice

Research Triangle Park, North Carolina
Sample Frame Deduplication in the World Trade Center Health Registry

Joe Murphy, Paul Pulliam, and Randolph Lucas

RTI International

Joint Statistical Meetings
Toronto
August 12, 2004
The World Trade Center (WTC) Health Registry

- Database for following people exposed to the September 11, 2001 WTC disaster
- Purpose is to evaluate physical and mental effects of exposure
- Eligible population includes those who may have been present in lower Manhattan on the morning of 9/11 when the two towers collapsed or in the subsequent dust/debris cloud
  - Residents
  - Students and school staff
  - Building occupants, visitors, and passersby
  - Rescue, recovery, and cleanup workers
Two Priority Groups

- **Group 1: higher priority**
  - Residents south of Chambers Street
  - Students and school staff south of Canal Street
  - Occupants of 35 damaged or destroyed buildings
  - Rescue, recovery, and cleanup workers

- **Group 2: lower priority**
  - Residents between Canal and Chambers Streets
  - Building occupants, visitors and passersby south of Chambers Street (not in 35 buildings)
35 Damaged or Destroyed Buildings
Survey Design

- Sample frame is compiled from various sources
  - Purchased or acquired lists of likely eligible registrants ("preregistrants")
  - Self-identification through toll-free number
  - Self-identification through project web site
- Preregistrants are located and contacted primarily by phone to complete a 30 minute interview
- The Registry will represent a sample frame of individuals who may be contacted for future studies over the next 20 years
Sample Frame Duplication

- Hundreds of overlapping sample sources
  - Lists of residents, students, employees, rescue/recovery workers
  - Self-identification
- Sample frame is “perfect if every element appears on the list separately, once, and only once, and nothing else appears on the list.” (Kish, 1965)
- To avoid coverage error due to duplication, a systematic deduplication system was employed
Sample List Overlap

List 1

List 2

Self-ID
Deduplication Process

- Record entered into sample database
  - Nightly automated ChoiceMaker deduplication
    - Grey-area match
      - Human review w/ ReviewMaker
        - Record matches another
          - Record deactivated
        - Record is unique person
          - Tracing, contacting, and interviewing take place
### Active Clues

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Direction</th>
<th>Type</th>
<th>Matcher</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>imEditDistance2C mFirstLastName</td>
<td>match</td>
<td>clue</td>
<td></td>
<td>15.61</td>
</tr>
<tr>
<td>8</td>
<td>dCmFirstLastName</td>
<td>diff</td>
<td>clue</td>
<td></td>
<td>20.63</td>
</tr>
<tr>
<td>33</td>
<td>mJanCounter/Frequency/CMFmFirstLastName[4]</td>
<td>match</td>
<td>clue</td>
<td></td>
<td>8.34</td>
</tr>
<tr>
<td>24</td>
<td>mCmLastName</td>
<td>match</td>
<td>clue</td>
<td></td>
<td>1.90</td>
</tr>
<tr>
<td>30</td>
<td>mCmFrequency/CMFmLastName[4]</td>
<td>match</td>
<td>clue</td>
<td></td>
<td>0.48</td>
</tr>
<tr>
<td>55</td>
<td>cmCommonFirstNames</td>
<td>diff</td>
<td>clue</td>
<td></td>
<td>15.23</td>
</tr>
<tr>
<td>70</td>
<td>dGenderCode</td>
<td>match</td>
<td>clue</td>
<td></td>
<td>128.77</td>
</tr>
<tr>
<td>74</td>
<td>mHouseAndStreetAndApartment</td>
<td>match</td>
<td>clue</td>
<td></td>
<td>4.53</td>
</tr>
<tr>
<td>78</td>
<td>mZipCode</td>
<td>match</td>
<td>clue</td>
<td></td>
<td>6.80</td>
</tr>
<tr>
<td>79</td>
<td>rStateCode.NY</td>
<td>none</td>
<td>nul</td>
<td></td>
<td>2.12</td>
</tr>
<tr>
<td>94</td>
<td>Invalid/Ob/End/No</td>
<td>diff</td>
<td>clue</td>
<td></td>
<td>2.12</td>
</tr>
</tbody>
</table>

### Decisions

- **Probability**: 4.2
- **ChoiceMaker**: diff
- **Human**: match

### Navigation

- **Selection size**: 126
- **Next in selection**: ◄ ►
- **Current record**: 11
- **All records**: 0 / 1012

### Status Messages

- INFO: Clue evaluation complete, elapsed time: 13,823 milliseconds.
- INFO: Clue evaluation complete, elapsed time: 1,375 milliseconds.
- INFO: Clue evaluation complete, elapsed time: 13,358 milliseconds.
- INFO: Clue evaluation complete, elapsed time: 12,059 milliseconds.
Impact of Deduplication on Quality, Cost, and Respondent Burden

• Does deduplication reduce the potential for coverage error?

• Does it reduce data collection costs?

• Does it reduce respondent burden?
Data Quality

• Before deduplication, 207,660 records in sample database (as of 7/19/04)

• After deduplication, 187,211 records (9.8% of records eliminated)

• 20,449 deduplicated records represent 17,751 persons
Effect of Deduplication – Group 1

- Residents
- Students / School Staff
- Building Occupants
- Rescue / Recovery Workers
- Unknown (Mostly Ineligible)

Before and After comparison.
Effect of Deduplication – Group 2

- Residents
- Students / School Staff
- Building Occupants
- Rescue / Recovery Workers
- Unknown (Mostly Ineligible)

Comparison of numbers before and after deduplication.
Impact on Sample Composition

- Residents
- Students / School Staff
- Building Occupants
- Rescue / Recovery Workers
- Unknown (Mostly Ineligible)

Before and After comparison.
Impact on Cost

- $D =$ cost of deduplication per record
- $C =$ cost of contacting and tracing record
- For WTC Health Registry, $C$ is 16% greater than $D$
- *Without deduplication*: $207,660 \text{ cases} \times C = 240,886.$
- *With deduplication*: $(187,211 \text{ cases} \times C) + (20,449 \text{ duplicates} \times D) = 231,132$
- Actual costs for tracing and locating were 96% what they would have been without deduplication.
Impact on Respondent Burden

- Over 48,000 interviews to date * 28.6 minutes/interview = more than 22,000 hours of interview time
- Over 10,000 ineligible or noncompliant sample members * 3 minutes/call = 500 hours of call time
- Estimated 6,339 records that would have been contacted w/o deduplication * 30 seconds/call = 53 hours of unnecessary contact
- Respondent burden in time reduced by less than 1%
- But this does not factor in the annoyance and loss of perceived integrity that may have resulted from unnecessary contacts
Conclusions

• Use of deduplication across multiple list sources was successful for the WTC Health Registry across three dimensions:
  • data quality
  • cost
  • respondent burden
• Environmental exposure registries face the challenge of compiling a cohort exposed to substance across years or decades. Because these cohorts disperse across time, multiple data sources are often used to identify potential registrants.
• Deduplication may help maximize the analytic utility of existing databases.
For More Information

- WTC Health Registry website: www.wtcregistry.org

- Questions/comments on this paper, email: jmurphy@rti.org