An Internet-based Method of Selecting Control Populations for Epidemiologic Studies

Mary Bishop Stone, Joseph L. Lyon, Sara Ellis Simonsen, George L. White, Jr., and Stephen C. Alder

From the Public Health Program, Department of Family and Preventive Medicine, University of Utah, Salt Lake City, UT.

Received for publication August 30, 2005; accepted for publication May 24, 2006.

Identifying control subjects for epidemiologic studies continues to increase in difficulty because of changes in telephone technology such as answering services and machines, caller identification, and cell phones. An Internet-based method for obtaining study subjects that may increase response rates has been developed and is described. This method uses information from two websites that, when combined, provide accurate and complete lists of names, addresses, and listed phone numbers. This method was developed by use of randomly selected streets in a suburb of Salt Lake City, Utah, in June 2005.

control groups; epidemiologic methods; patient participation; patient selection; research design

Abbreviation: RDD, random digit dialing.

Epidemiologists have long recognized the importance of selecting appropriate controls in case-control studies (1–3). Over the last two decades, random digit dialing (RDD) (4, 5) has been used to identify and recruit controls, but advances in telephone technology have reduced response rates and increased selection bias when obtaining controls via RDD (5–7). These advances include answering machines, caller identification, and “do not call” listings. The number of households with caller identification increased nearly 500 percent between 1996 and 2000, with nearly half using caller identification in 2000 (8). During the 1980s and early 1990s, the response rates obtained via RDD ranged from 60 to 75 percent (9, 10). However, RDD response rates have declined in recent years. For example, the Centers for Disease Control and Prevention’s Behavioral Risk Factor Surveillance System survey, which has been conducted annually via RDD since 1994, observed a decline in response rate from 70 percent to less than 50 percent between 1994 and 2000 (5, 11). The declining effectiveness of RDD is a concern to epidemiologists (5, 6). Techniques such as advance mailings have been proposed to improve response rates (12, 13), but the names and addresses of potential participants are necessary to utilize such methods. We propose an Internet-based method for the recruitment of controls. This method involves the use of inexpensive, online, commercial databases for the identification of controls, and it allows researchers to obtain contact information on all potential participants within a geographic area that can be used for techniques such as advance mailings (14).

MATERIALS AND METHODS

The alternate method of identifying study subjects described here was developed from the Thyroid Cohort Study that was initiated in 1965 when 4,818 schoolchildren were identified and examined yearly from 1965 to 1968 (15). In 1985 and 2003, the cohort was relocated and reexamined. In 2003, using Internet database resources and previous examination data, we located 96.5 percent of the cohort, including 451 (80.5 percent) who had not been examined since the 1960s. This was accomplished over 3 months by 2.75
full-time equivalent employees. Our success compelled us to ask whether these Internet resources could be used to identify controls for other types of epidemiologic studies; we investigate this question in the present study.

Accurint (16) is a commercially available database containing information from a variety of data sources; most are publicly available. Accurint can be searched by name or by street to find every address and occupant on the street. An Accurint query costs $0.25, which covers one of the following: a single name search, a single address search, or a search for all names and addresses on a particular street. We conducted a pilot study to explore the feasibility of using this database to identify subject subjects within a geographic area. The target community for this pilot study is a suburb of Salt Lake City, Utah (Magna). Magna has a population of 22,300 and is defined by a single ZIP Code. The community includes single-family dwellings (the majority of residences), some duplexes, and a few large apartment complexes.

To test the completeness of capture, we chose a random sample of six streets in the community, recorded the addresses on the streets by visual inspection, and compared these addresses with those obtained from a street search in the Accurint database. Six staff members visited the selected streets and recorded every address and name that were visible from the sidewalk or the street. They were blind to the information available from Accurint (16). They made an effort to scrutinize each property to identify dwellings not visible from the street. This visual inspection was the standard against which we compared address information from the Accurint database.

In addition to identifying addresses during visual inspection, we also recorded the names appearing on the property. Using occupant information from the Accurint (16) database, we compared the names obtained by visual inspection with those available in Accurint. Each set of names was assigned a code indicating the degree of matching. If the same first and last name or last name only were identified, the pair was considered a match. Otherwise, the pair was considered to be a mismatch.

The information from the visual inspection of addresses was entered into a spreadsheet and compared with the information obtained from Accurint (16). The Accurint (16) database contained substantially more addresses on each street than were found on visual inspection, and only one address was missing from Accurint (16). In order to investigate these discrepancies, two staff members revisited the six streets with a list of the addresses obtained from Accurint (16) to verify the additional addresses. The team was unable to locate any of the additional addresses provided by Accurint (16).

To determine whether visual inspection or Accurint was more accurate, a second Internet database, MelissaData (17), was searched. This is also a commercially available data source that will provide lists of addresses by street name. One of the data sources used in MelissaData is the US Postal Service Delivery Sequence File. Information from MelissaData is available for $9.50 for 1,000 addresses, with a $25.00 minimum charge per use.

### RESULTS

A summary of the results of the comparison among the Accurint (16) database, MelissaData (17) information, and the field visual inspection is displayed in table 1. Of the 387 addresses identified, 167 addresses were identified by Accurint (16) but not found by visual inspection.

In order to determine whether these 167 addresses were actual residences, we compared the MelissaData (17) database information with the results of visual inspection. We found a 98.2 percent agreement between MelissaData (17) and visual inspection.

We concluded that the extra addresses found by Accurint (16) were not actual dwellings. Using other Internet data sources such as InfoSpace (18), QuestDex (19), or the US version of Infobel (20), we confirmed that 119 (71.3 percent) of the 167 addresses were included because of clerical errors. This determination was made by locating the current occupant in the Internet databases listed above and comparing the address information with that found in Accurint. The potential for inaccuracy due to clerical errors is acknowledged on the Accurint (16) website. Additionally, we investigated the apparently erroneous 167 addresses by interviewing US Postal Service personnel. They stated that some of the addresses could have been assigned to vacant lots in anticipation of future dwellings, or the addresses could have been lots where dwellings may have once existed but were razed.

Comparing current occupant names from Accurint (16) with the names of the 39 occupants whose names were visible on the properties yielded 79.5 percent agreement, as shown in table 2.

Most epidemiologic researchers continue to use telephones to identify controls. To assess the availability of telephone numbers, we evaluated the 217 addresses identified in MelissaData (17) and confirmed by visual inspection, and we found that 175 (79.5 percent) had a telephone number listed in the Accurint (16) database.

### TABLE 1. Addresses found in the Accurint* and MelissaData* databases compared with those found by visual inspection, Magna, Utah, 2005

<table>
<thead>
<tr>
<th>Addresses found by</th>
<th>No.</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual inspection or in Accurint</td>
<td>387</td>
<td>100.0</td>
</tr>
<tr>
<td>Visual inspection and in Accurint</td>
<td>219</td>
<td>56.6</td>
</tr>
<tr>
<td>Visual inspection but not in Accurint</td>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>MelissaData but not by visual inspection</td>
<td>167</td>
<td>43.1</td>
</tr>
<tr>
<td>Visual inspection and in MelissaData</td>
<td>221</td>
<td>100.0</td>
</tr>
<tr>
<td>Visual inspection but not in MelissaData</td>
<td>217</td>
<td>98.2</td>
</tr>
<tr>
<td>MelissaData but not by visual inspection</td>
<td>3</td>
<td>1.4</td>
</tr>
<tr>
<td>Accurint compared with visual inspection</td>
<td>1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

A summary of the characteristics of these three sources of information is described in table 3. We recommend combining MelissaData (17) and Accurint (16) to obtain a sample of households as a first step for identifying a control group, as described in table 4.

**DISCUSSION**

We conclude that combining data from MelissaData (17) and Accurint provides the most accurate list of addresses and current occupants. Using both methods eliminated inaccurate addresses but took advantage of the occupant information available from Accurint (16). We found that using these combined resources was as accurate as visual inspection, but much less costly. Both Accurint (16) and MelissaData (17) are inexpensive, and minimal staff time is required to obtain the information.

Because the six streets we randomly selected included few apartments, the completeness of information for multiple-occupant dwellings from MelissaData (17) and Accurint (16) is unknown. The six streets that we investigated are in a suburban area of a larger metropolitan area, and the accuracy and completeness of information for large urban areas or rural areas are also not known. However, the US Postal Service Delivery Sequence File, a source from which Accurint and MelissaData draw their data, has been successfully used to identify occupancies in metropolitan areas (21).

Since these Internet resources contain data from across the United States, we assume that they could be used to find control populations or samples of any size in any geographic area, but this assumption needs further investigation. We also assumed that the sources described here could be used to attach names and addresses to RDD numbers for those with listed telephone numbers.

In this small study, about 80 percent of the subjects had telephone numbers. This percentage would likely be lower in other areas of the country and may be lower in highly urbanized areas, so a combination of repeated contact by letter and visits to the address may be necessary to achieve adequate response rates.

Additional research is needed to determine the response rates from participants identified by this method. One of the strengths of using this method is that it allows for advanced mailings prior to telephone contact (14). It has been reported that response rates are improved by as much as 18 percent when advanced mailings are used (13). Although response rates using the method reported here have not been studied in the selection of population-based controls, we have used this method in our cohort study to reinitiate contact with cohort members 18 years after previous contact with a high level of response. Accurint and other sources of information were used to locate cohort members, and advanced mailings were used to initiate contact. We located 96.5 percent of the original cohort and, among those contacted, 94.5 percent agreed to examination. Although the response rates may be higher in this cohort because of prior participation, we believe that, since the cohort had not been contacted in 18 years, our experience may well mirror control selection from a general population. This suggests that the method presented here has the potential to yield higher response rates than are obtained by RDD. This method may offer

---

**TABLE 2. Name information on occupants in the Accurint database compared with that by visual inspection, Magna, Utah, 2005**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>No.</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last name or first and last name</td>
<td>31</td>
<td>79.5</td>
</tr>
<tr>
<td>Last name but not first initial</td>
<td>1</td>
<td>2.6</td>
</tr>
<tr>
<td>Mismatch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previously listed resident</td>
<td>6</td>
<td>15.4</td>
</tr>
<tr>
<td>No known resident</td>
<td>1</td>
<td>2.6</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>100.1</td>
</tr>
</tbody>
</table>


**TABLE 3. Characteristics of visual inspection, MelissaData, and Accurint**

<table>
<thead>
<tr>
<th>Source of information</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual inspection</td>
<td>Complete address ascertainment</td>
<td>Expensive, limited information on occupant</td>
</tr>
<tr>
<td>MelissaData</td>
<td>Complete address ascertainment, inexpensive</td>
<td>No information on occupant</td>
</tr>
<tr>
<td>Accurint</td>
<td>Inexpensive, complete information on occupants (names of all household members, address, phone number, birth month and year, dates of occupancy)</td>
<td>Excess addresses, have to identify information on current occupant from that on historical occupants</td>
</tr>
</tbody>
</table>


**TABLE 4. Steps for using Internet databases to identify control subjects**

1. Define geographic area (i.e., carrier route, ZIP Code, county, etc.).
2. Obtain addresses from www.melissadata.com/ for the defined geographic area.
3. Randomly select the required number of addresses.
4. Look up each address in www.accurint.com/ to obtain information on occupants.
5. Use other Internet database resources to evaluate historical information on occupants to identify the current occupant.
6. Contact the current occupant by telephone or mail. (We prefer mail first, as it legitimizes the study and subsequent telephone call.)
epidemiologic researchers a potentially accurate and inexpensive method of locating population-based controls and follow-up of participants in longitudinal studies.

ACKNOWLEDGMENTS

The authors wish to thank the following individuals for their contributions to the research described here: Judy Allred, Keith Carney, Toni Fightmaster, Nikki Konesavanh, Noemi Monge, and David Stone. Conflict of interest: none declared.

REFERENCES

17. Melissa Data Corp, a commercial corporation whose online products are based on information from the United States Postal Service. Rancho Santa Margarita, CA: Melissa Data Corp, 2005. (http://www.melissadata.com/).