Assessment of a Commercial Searchable Population Directory as a Means of Selecting Controls for Case-Control Studies

Sabeena Chintapalli, MPH[a,b]
Michael Goodman, MD, MPH[a,c]
Mark Allen, PhD[d]
Kevin Ward, PhD[a,d]
Jonathan Liff, PhD[a,d]
John Young, DrPH[a,d]
Paul Terry, PhD[a]

SYNOPSIS

We explored the feasibility of using SalesGenie®, a commercially available database, as a potential alternative to traditional methods of selecting controls for population-based case-control studies. An attractive feature of this particular database is that it permits a search within specific age ranges, geographic locations, and household income.

Information on 1,068 cases reported to the California Cancer Registry between 2001 and 2005 was entered manually into the SalesGenie Web-based search engine. The frequency of Registry-to-SalesGenie matches was then compared with the frequency of matching the registry data to the California Department of Motor Vehicles (DMV) records. Our findings indicate that the SalesGenie database is currently less comprehensive than DMV records. Nevertheless, Web-based population data sources may provide a potential alternative for population-based studies when used in conjunction with other methods, particularly in states where DMV records are not accessible to researchers.

[a]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA

[b]Texas Department of State Health Services HIV/STD Information and Projects Group, Austin, TX

[c]Georgia Center for Cancer Statistics, Atlanta, GA

[d]California Cancer Registry, Sacramento, CA

Address correspondence to: Michael Goodman, MD, MPH, Department of Epidemiology, Emory University Rollins School of Public Health, 1518 Clifton Rd. NE, Atlanta, GA 30322; tel. 404-727-2734; fax 404-909-5155; e-mail <mgoodm2@sph.emory.edu>.

Population-based controls in case-control studies are sampled directly from the source population.[1] However, to ensure control selection that is representative of the general population, it is important to have confidence that the population data source used is reasonably complete. Among the most common sources of data for population controls are Department of Motor Vehicles (DMV) records, voter registration lists, and telephone directories for random-digit dialing.[2]

Each data source has its limitations. The use of DMV records is not allowed in some states due to the Federal Driver's Privacy Protection Act;[3] random-digit dialing is increasingly subject to selection bias due to the advent of new technologies;[4,5] and voter registration lists typically do not provide a representative sample of the population.[6] Although potentially more complete, other possible sources of population controls (e.g., neighborhood surveys or area sampling) may be prohibitively expensive, as they require substantial fieldwork and specialized skills.[7]

In view of methodological and logistical difficulties associated with traditional methods of control selection, there is a need for new sources of population-based controls that may be met through the use of Internet-based commercial databases that can generate names and addresses of participants for population-based studies.

The purpose of this study was to evaluate the completeness of one such database, SalesGenie®, a Web-based data service from InfoUSA (InfoUSA, Inc., Omaha, Nebraska). The SalesGenie population database, which is updated every month, covers the entire United States and uses information merged from multiple sources, such as utility directories, public (e.g., tax) records, telephone directories, survey data, U.S. Census Bureau files, and postal service change-of-address lists. Unlike many other population data sources, the SalesGenie service offers search engines that allow identification of individuals based on a number of criteria, such as gender, five-year age group, and county (or zip code, city, or street) of residence.[8] This feature makes this database potentially attractive for control selection in case-control studies, which often use individual or frequency-based matching of controls to cases.

METHODS

For the purposes of this study, we received from the California Cancer Registry a random sample of 3,000 records for cases diagnosed between January 1988 and October 2005. Analysis was restricted to 1,068 individuals who were reported to the registry during the most recent five-year period, 2001–2005.

The cancer patients in this study were not meant to represent cases in a case-control study, but merely served as a sample of the population. We decided to use a random sample of the cancer registry records for our study because all cancer patients, as opposed to those with specific cancers (e.g., lung or breast cancer), are expected to be reasonably representative of the general population[9] and because cancer registries always collect demographic and address information, which is required for our analyses. California was selected for this study because unlike some states (e.g., Georgia), its DMV records are available for research purposes, and because the California Cancer Registry routinely links all of its records to the DMV database.

All records used in this study were electronically linked to the California DMV records using a process previously described elsewhere.[10,11] To summarize, records are linked by first name, middle initial, last name, and date of birth. Records that remain unmatched are checked for errors and spelling mistakes. In addition, the initial linkage often results in multiple duplicates due to errors or insufficient information; these duplicates are evaluated and reconciled. In some cases, several possible versions of the name (e.g., William, Billy, Bill, Will) are generated to maximize data linkage outcomes.[11] A record is considered a match if the first name, middle initial, last name, and date of birth are the same in both the DMV and the registry databases. The DMV linkage process does not allow for probabilistic (i.e., partial) matches.

Records for use in this study remained in the metropolitan Atlanta and rural Georgia Surveillance Epidemiology and End Results (SEER) Registry in accordance with data security and confidentiality protocols, which include restricted access at all entry points, data storage on a dedicated server in a secured server room, and password-protected access to the data. The study protocol was approved by the Emory University Institutional Review Board.

All contact information for the cases (first names, last names, addresses, and phone numbers) was entered manually into the SalesGenie search engine, one person at a time, in an attempt to locate each individual in the database. The resulting matching status was categorized either as a complete match, a partial match, or a non-match. A complete match was defined as full agreement between the registry information and the SalesGenie data. The partial matches included three types of results: (*1*) there was agreement between the registry and the SalesGenie database with respect to the

patient's last name, address, and/or phone number, but not the first name; (*2*) there was agreement between the registry and the SalesGenie database with respect to the patient's first initial, last name, address, and/or phone number; and (*3*) a patient could not be identified by the address or phone number, but shared the same five-year age category, zip code, and full name with someone in the SalesGenie database. All other results were categorized as non-matches.

Frequency analyses and cross tabulations were conducted to examine the distributions of full and partial SalesGenie matches across various patient characteristics. These characteristics included age, gender, year of diagnosis, level of urbanization, area-based measure of socioeconomic status (SES), and race/ethnicity.

For these analyses, age was divided into three broad categories: 40–64 years, 65–74 years, and ≥75 years. The race/ethnicity variable included the following categories: non-Hispanic white, non-Hispanic black, Hispanic, Asian/Pacific Islander, and other/unknown. The year of diagnosis was used to determine whether SalesGenie database completeness changed over time. The levels of urbanization and the area-based measures of SES were examined to determine whether SalesGenie completeness is different in major metropolitan centers (e.g., Los Angeles) compared with small urban or rural areas of California, and whether poor areas are less represented than more affluent neighborhoods. The level of urbanization was categorized as either metropolitan or non-metropolitan using Beale code information. Beale codes classify all the counties in the U.S. as metropolitan, urban (reserved for small towns and cities), or rural.[12]

The SES was expressed as the percentage of the census tract population living below the federal poverty level (FPL). In the absence of individual SES information, census tract poverty has been shown to serve as a meaningful alternative.[13] Federal standards define census tracts with ≥20% of the population living below the FPL as "poverty areas."[14] The designation of poverty level varies with family size, income, and year (e.g., $17,463 for a family of four in calendar year 2000). The percentage of the population living below the FPL was classified into three groups: 0%–9.9% (low poverty), 10%–19.9% (moderate poverty), and 20%–100% (high poverty). The frequencies of SalesGenie matches across various patient characteristics were compared with those from DMV matches using Chi-square tests.

Multivariate logistic regression analyses were conducted using two dichotomous versions of the dependent variable: (*1*) full match vs. no match or partial match and (*2*) any match (full or partial) vs. no match. The independent variables in these models included age, race/ethnicity, gender, level of urbanization (metropolitan vs. non-metropolitan), percentage of census tract population living below FPL, and year of diagnosis. All models were assessed for interactions among variables. All analyses were performed using SAS®[15] and OpenEpi.[16]

RESULTS

The search of DMV records found 85% of the patients in this study, while the query of the SalesGenie database was able to fully match 56% and partially match an additional 11% of the patients. As shown in Table 1, about 50% of the patients were found in both DMV and SalesGenie databases; nearly 6% of patients were found in SalesGenie but not in DMV records, and 35% of patients were found in DMV records alone. Only about 9% of patients were not located through either method; conversely, about 91% were found in one or the other, or both. The DMV matches and DMV non-matches were similar with respect to age, gender, level of urbanization, area-based measures of SES, and year of diagnosis. However, the likelihood of a DMV match was substantially lower for non-Hispanic black individuals (81%, 95% confidence interval [CI] 68, 90) and in particular for Hispanic individuals (72%, 95% CI 65, 78) compared with white individuals (89%, 95% CI 86, 91).

A corresponding evaluation of the SalesGenie data demonstrated that a complete match was significantly more common among men (61%, 95% CI 56, 65) compared with women (52%, 95% CI 47, 56) and among people residing in census tracts with at least 20% of the population living below FPL (47%, 95% CI 39, 56) compared with those who live in census tracts with a low (<10%) proportion of poor residents (61%, 95% CI 57, 65).

Table 2 presents the results of the logistic regression analyses comparing complete SalesGenie matches to partial or non-matches and any (complete or partial) matches to non-matches. People aged 65 to 74 years were the least likely to have a complete match with the SalesGenie database (odds ratio [OR] = 0.62; 95% CI 0.44, 0.85) when compared with the youngest (<65 years) age group. The oldest age group was also less likely to be found in the SalesGenie database than the youngest age group, although this association was somewhat less pronounced (OR=0.74, 95% CI 0.55, 0.99). Females were 36% less likely than males to be completely matched to the SalesGenie database (OR=0.64, 95% CI 0.50, 0.83). Non-Hispanic white individuals were most likely to be completely matched to the SalesGenie database, but there was

Table 1. Selected characteristics of the study population[a] by match status using the SalesGenie® database

Characteristics	Total	SalesGenie match status: Complete or partial match — Number	Complete or partial match — Percent (95% CI)	Complete match only — Number	Complete match only — Percent (95% CI)	Non-match — Number	Non-match — Percent (95% CI)	Match against DMV records — Number	Match against DMV records — Percent (95% CI)
Age group (in years)									
40–64	413	284	69 (64, 73)	245	59 (55, 64)	129	31 (27, 36)	353	85 (82, 89)
65–74	263	170	65 (59, 70)	136	52 (46, 58)	93	35 (30, 41)	228	87 (82, 90)
≥75	392	265	68 (63, 72)	217	55 (50, 60)	127	32 (28, 37)	327	83 (80, 87)
Race/ethnicity									
Non-Hispanic white	734	541	74 (70, 77)	454	62 (58, 65)	193	26 (23, 30)	652	89 (86, 91)
Non-Hispanic black	52	24	46 (33, 60)	22	42 (30, 56)	28	54 (40, 67)	42	81 (68, 90)
Hispanic	169	93	55 (47, 62)	72	43 (35, 50)	76	45 (38, 53)	121	72 (65, 78)
Non-Hispanic Asian/Pacific Islander	88	47	53 (43, 64)	36	41 (31, 51)	41	47 (36, 57)	73	83 (74, 90)
Unknown/other	25	14	56 (36, 74)	14	56 (36, 74)	11	44 (26, 64)	20	80 (61, 92)
Gender				598					
Male	521	358	69 (65, 73)	316	61 (56, 65)	163	31 (27, 35)	447	86 (83, 89)
Female	547	361	66 (62, 70)	282	52 (47, 56)	186	34 (30, 38)	461	84 (81, 87)
Level of urbanization									
Metropolitan	1,023	684	67 (64, 70)	572	56 (53, 59)	339	33 (30, 36)	868	85 (83, 87)
Non-metropolitan	45	35	78 (64, 88)	26	58 (43, 73)	10	22 (12, 36)	40	89 (77, 96)
Percent of census tract population living below poverty level									
<10	504	362	72 (68, 76)	307	61 (57, 65)	142	28 (24, 32)	438	87 (84, 90)
10–19.9	389	255	66 (61, 70)	209	54 (49, 59)	134	34 (30, 39)	325	83 (79, 87)
≥20	152	88	58 (50, 66)	71	47 (39, 56)	64	42 (34, 50)	124	82 (75, 87)
Not known	23	14	61 (39, 80)	11	47 (27, 69)	9	39 (20, 61)	21	91 (72, 99)
Year of diagnosis									
2001	238	157	66 (60, 70)	136	57 (51, 63)	81	34 (28, 40)	199	84 (79, 88)
2002	220	154	70 (64, 76)	127	58 (51, 64)	66	30 (24, 36)	181	82 (77, 87)
2003	248	166	67 (61, 73)	138	56 (49, 62)	82	33 (27, 39)	210	85 (80, 89)
2004	270	185	69 (63, 74)	150	56 (50, 61)	85	31 (26, 37)	241	89 (85, 93)
2005	92	57	62 (52, 71)	47	51 (41, 61)	35	38 (28, 48)	77	84 (75, 90)
Overall	1,068	719	67 (65, 70)	598	56 (53, 59)	349	33 (30, 36)	908	85 (83, 87)

[a]The study population consisted of 1,068 individuals randomly selected from the California Cancer Registry.

DMV = Department of Motor Vehicles

CI = confidence interval

little difference found among Asian/Pacific Islander, Hispanic, and non-Hispanic black individuals, with all ORs in the 0.42–0.48 range. The corresponding results for the level of urbanization, the area-based level of SES, and the year of diagnosis were not statistically significantly different from the null.

The multivariate analyses comparing all matches (complete and partial combined) with non-matches were generally consistent with the previous results, but with two notable differences. First, the previously observed association between female gender and full SalesGenie match was no longer significant when partial matches were included as matches (OR=0.84, 95% CI 0.64, 1.10). Second, compared with the previous analyses, the difference between non-Hispanic white and non-Hispanic black individuals was more pronounced (OR=0.33, 95% CI 0.19, 0.60).

DISCUSSION

In this study, we confirmed that the drivers' license data were considerably more complete than the data in the SalesGenie database. However, because DMV records are not available in many states, there remains a need to determine whether commercially available databases represent a useful alternative. It is disappointing that <60% of individuals in our sample were in the directory, which reportedly included so many sources of information. Our finding is likely explained by the fact that the data sources comprising the SalesGenie directory are both incomplete and overlapping, thereby limiting the number of unique records. By contrast, those data sources that are likely to include a greater proportion of the population (e.g., DMV records or credit histories) appear to be difficult to access and are less likely to be included in publicly available secondary databases.

Our results also indicate that not all sociodemographic groups and geographic areas are equally represented in the SalesGenie database. Groups that had a low likelihood of being found in the SalesGenie directory included people >65 years of age, women, black or Hispanic people, as well as individuals residing in census tracts designated as poverty areas. Not all

Table 2. Multivariate logistic regression analyses of the association between SalesGenie® database match and selected characteristics of the study population[a]

Variable	*OR[b] (95% CI) for complete match[c]*	*OR (95% CI) for any match[c]*
Year of diagnosis (continuous)	0.95 (0.86, 1.04)	0.98 (0.88, 1.09)
Age group (in years)		
<65	Ref.	Ref.
65–74	0.62 (0.44, 0.85)	0.74 (0.53, 1.05)
≥75	0.74 (0.55, 0.99)	0.86 (0.63, 1.17)
Gender		
Male	Ref.	Ref.
Female	0.64 (0.50, 0.83)	0.84 (0.64, 1.10)
Race/ethnicity		
Non-Hispanic white	Ref.	Ref.
Non-Hispanic black	0.47 (0.26, 0.84)	0.33 (0.19, 0.60)
Hispanic	0.48 (0.34, 0.69)	0.47 (0.33, 0.68)
Asian/Pacific Islander	0.42 (0.27, 0.67)	0.41 (0.26, 0.65)
Other/unknown	0.71 (0.31, 1.62)	0.45 (0.20, 1.01)
Level of urbanization		
Metropolitan	Ref.	Ref.
Non-metropolitan	0.95 (0.51, 1.62)	1.46 (0.71, 3.02)
Percent of census tract population living below poverty level		
<10	Ref.	Ref.
10–19.9	0.83 (0.63, 1.09)	0.81 (0.60, 1.09)
≥20	0.70 (0.47, 1.03)	0.71 (0.47, 1.05)
Not known	0.58 (0.24, 1.36)	0.64 (2.63, 1.55)

[a]The study population consisted of 1,068 individuals randomly selected from the California Cancer Registry.

[b]Partial matches are considered non-matches.

[c]OR<1 indicates lower likelihood of being in the SalesGenie database.

OR = odds ratio

CI = confidence interval

Ref. = reference group

of these factors were independently associated with SalesGenie match status: in the multivariate analyses, the effect of low SES after controlling for age, race/ethnicity, and gender was no longer present, indicating that the association between poverty and underrepresentation in the SalesGenie database is likely explained by other factors.

Contrary to our expectations, geographic locations comprising rural or small urban (non-metropolitan) areas did not appear to be underrepresented compared with metropolitan population centers. There was also no evidence that completeness of the SalesGenie database increased over time, with the probability of the matches in 2004 and 2005 showing no improvement compared with earlier study years.

An attractive feature of the SalesGenie application is that it permits a search within specific age ranges, geographic locations, and household incomes. However, SalesGenie does not allow searching by race/ethnicity, so it cannot be used to obtain a sample stratified on race (e.g., 1,000 white and 1,000 African American people).

Limitations

One of the weaknesses of this study was that the analysis was restricted to California, whose population may differ in several ways from those of other U.S. states. Moreover, the majority of the patients came from metropolitan Los Angeles, and only one patient had a true rural address. For these reasons, the completeness of the SalesGenie database as it relates to other states or geographic areas may still be unknown.

It also should be noted that the completeness of the SalesGenie database may have been underestimated in this study, as we did not distinguish among different subtypes of partial matches. For instance, partial matches that occurred when there was agreement with respect to the patient's first initial, last name, address, and/or phone number may carry more weight because they are closer to a full match than the other subtypes. However, even if all partial matches were treated as being complete, the DMV records would still be more comprehensive.

CONCLUSION

Commercial databases may provide a less expensive alternative to conventional methods such as random-digit dialing, but they are not as complete as DMV records. In the absence of comprehensive population databases, commercial directories may be helpful when applied in conjunction with other independently assembled databases.

REFERENCES

1. Rothman KJ, Greenland S. Modern epidemiology. Philadelphia: Lippincott Williams and Wilkins; 1998.
2. Brogan DJ, Denniston MM, Liff JM, Flagg EW, Coates RJ, Brinton LA. Comparison of telephone sampling and area sampling: response rates and within-household coverage. Am J Epidemiol 2001;153:1119-27.
3. U.S. Code. Drivers Privacy Protection Act. 18 U.S.C. § 2721 et. seq. [cited 2009 Jan 8]. Available from: URL: http://www.accessreports.com/statutes/DPPA1.htm
4. Curtin R, Presser S, Singer E. The effects of response rate changes on the index of consumer sentiment. Public Opin Q 2000;64:413-28.
5. Tuckel P, O'Neill H. The vanishing respondent in telephone surveys. J Adv Res 2002;42:26-48.
6. Census Bureau (US). Reported voting and registration of the total voting-age population, by sex, race and Hispanic origin, for states. November 2004.
7. Mandel JS, Korelitz JJ, O'Fallon WM. Identification of neighborhood controls and the location and recruitment of all study subjects. J Clin Epidemiol 1993;56:304-9.
8. SalesGenie. SalesGenie consumer database. San Carlos (CA): SalesGenie; 2007.
9. Checkoway H, Pearce N, Kriebel D. Research methods in occupational epidemiology. New York: Oxford University Press; 2004.
10. Cockburn MG, Hamilton AS, Zadnick J, Cozen W, Mack TM. Development and representativeness of a large population-based cohort of native Californian twins. Twin Res 2001;4:242-50.
11. Hser YI, Evans E. Cross-system data linkage for treatment outcome evaluation: lessons learned from the California Treatment Outcome Project. Eval Program Plann 2008;31:125-35.
12. Department of Agriculture (US). Rural-urban continuum codes. Washington: Department of Agriculture; 2003.
13. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area-based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: the Public Health Disparities Geocoding Project (US). J Epidemiol Community Health 2003;57:186-99.
14. Census Bureau (US). Statistical brief: poverty areas. Washington: Department of Commerce (US); 2002.
15. SAS Institute, Inc. SAS: Version 9.1 for Windows. Cary (NC): SAS Institute, Inc.; 2005.
16. Dean A, Sullivan K, Soe M. OpenEpi: Version 2.2.1. 2008.