American Journal of Epidemiology Advance Access published December 20, 2006

American Journal of Epidemiology
Copyright © 2006 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

DOI: 10.1093/aje/kwk049

Practice of Epidemiology

Method to Assess Identifiability in Electronic Data Files

Holly L. Howe[1], Andrew J. Lake[2], and Tiefu Shen[3]

[1] North American Association of Central Cancer Registries, Inc., Springfield, IL.
[2] Information Management Services, Inc., Silver Spring, MD.
[3] Illinois State Cancer Registry, Springfield, IL.

Received for publication December 19, 2005; accepted for publication July 28, 2006.

The authors developed the Record Uniqueness (RU) software program to assess electronic data files for risk of confidentiality breach based on unique combinations of key variables. The underlying methodology utilized by the RU program generates a frequency distribution for every variable selected for analysis and for all combinations of the variables selected. In addition, the program provides the regression coefficient that designates the relative contribution of each variable to the unique records on the data file. The authors used RU to evaluate a North American Association of Central Cancer Registries research data set with 4.67 million cases from 34 population-based cancer registries for 1995–2001. To illustrate the process and utility of RU, they describe the evaluation process of the confidentiality risk of adding a county-based socioeconomic measure to the research file. The RU method enables one to be assured of record confidentiality, provides flexibility to adjust record uniqueness thresholds for different users or purposes of data release, and facilitates good stewardship of confidential data balanced with maximum use and release of information for research. RU is a useful data tool that can quantify the risk of confidentiality breach of electronic health databases, including reidentifiability of cases through triangulation of information or linkage with other electronic databases.

confidentiality; medical informatics; neoplasms; privacy; regression analysis; social class

Abbreviations: NAACCR, North American Association of Central Cancer Registries; RU, Record Uniqueness; SEER, Surveillance, Epidemiology, and End Results.

Individual health information remains a bastion of privacy, with access protected and protocols ensuring that all releases are ethically and scientifically grounded. Privacy concerns include the capability to identify a patient, the potential to gain new information about a known patient, or the reidentification of a patient through triangulation of information. This balance can be accomplished by omission of personal identifiers or other sensitive variables or by aggregation of variable values to limit indirect disclosure through unique combinations. The release of electronic data files, as well as requests for electronic record linkage, has added to concerns that the risk for potential breaches be known or quantifiable and that suppression based on counts is neither practical (1) nor reasonable.

We developed the Record Uniqueness (RU) software program to assess electronic data files for risk of confidentiality breach based on unique combinations of key variables. We use RU, now publicly available (2), to assess the uniqueness of records included in all Cancer in North America research files (referred to as "CINA Deluxe") produced by the North American Association of Central Cancer Registries (NAACCR).

MATERIALS AND METHODS

Record Uniqueness program

RU was developed to test data files for potential patient identifiability due to small numbers (2), the single most

Reprint requests to Dr. Holly L. Howe, 2121 West White Oaks Drive, Springfield, IL 62704-6495 (e-mail: hhowe@naaccr.org).

important attribute that can identify individuals. Personal identifiers (name, address, or Social Security number) offer absolute record uniqueness. It is not obvious, however, that, when viewed in combination, less personal identifiers (e.g., race, a 5-year age group, a county of residence) can also result in record uniqueness.

RU generates frequencies for every variable and combination of variables. For each frequency distribution, the program counts the number of records with a frequency of one (unique records) and the number of records with a frequency of five or less (unique record sets).

For example, if one were to analyze the uniqueness of race for a file that contained 205 records, of which 150 were coded Chinese, 50 were coded Japanese, four were coded Korean, and one was coded Vietnamese, the RU program would identify one record (0.49 percent) as unique. It would also identify five (four Korean + one Vietnamese) records (2.4 percent) as unique record sets. The complexity of RU grows exponentially as the number of variables (and thus combinations) grows. The number of variable combinations analyzed by RU follows the combinatorial rule and is always $2^N - 1$, where N is the number of variables. Thus, for three variables, seven combinations (or frequencies) are generated; for five variables, 31 combinations; and for nine variables, 511 combinations are generated.

The user defines the variables to assess uniqueness and the variables' precision (regardless of categorical or numerical, e.g., 5-year, age groups or single year of age). The variables that should always be included are age, sex, race, and year of diagnosis, when these variables are in the requested data set and include more than one value. In addition, when a data set contains more than one geographic area (e.g., states within the United States) or more than one disease outcome, then these variables should also be included in the default variable set.

Once the data have been processed, RU outputs the number of records in each analysis and the number and proportion of unique records and unique record sets. The program provides a regression coefficient, that is, the relative contribution of each variable to the unique records. The regression model is *Ln* (proportion of unique records) = $\alpha + \Sigma_i(\beta_\iota \times$ (presence of variable i)), where the presence of variable i is 1 if that variable is present and 0 otherwise. Higher values of the β coefficient reflect a relatively greater contribution (weight) to uniqueness. RU guides the user to decrease the number of unique records, by identifying the variable with the greatest contribution. By collapsing this variable into fewer values, one can achieve the greatest reduction in unique records. RU is a guide, enabling the user to make decisions about the importance and necessary precision of any of the variables for the analysis. Through collapsing values and an iterative process, a user can create a data file that achieves a balance between record confidentiality and information release. If collapsing value categories does not achieve the desired threshold or becomes meaningless for analysis, then, in the RU program, one must omit variables to meet the uniqueness threshold. The suggested threshold (2) for research files is that no more than 20 percent of the variable combinations identify unique record sets based on the default variable set. For public use files, fewer than 5 percent of the variable combinations should identify unique record sets based on the default variable set (including geography).

TABLE 1. Variables with measurement precision used in RU* software analysis

Variable	Values count
Registry	34
Race (White, Black, other, unknown)	4
Age	
5-year age groups	18
20-year age groups (0–19, 20–39, 40–64, ≥65)	4
Modified 5-year groups (0–19, 20–24, 25–29 … ≥85)	15
Sex	2
Year of diagnosis	7
Primary site	
SEER* site recode	78
Major sites	16
Minor sites	45
% county poverty	
Actual	283
Rounded to integers	40
Grouped by 2% intervals	21
Grouped by 3% intervals	15
Grouped by 5% intervals	10
Grouped by 10% intervals (0–10, 11–20, 21–30, >30)	4

* RU, Record Uniqueness; SEER, Surveillance, Epidemiology, and End Results.

Data source

We used RU to evaluate the NAACCR research data file of 4.67 million cancer cases from 34 population-based cancer registries for 1995–2001. Cancer incidence registries participate in national programs: in the United States, the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program or the Centers for Disease Control and Prevention's National Program of Cancer Registries, or both. In Canada, registries participate in the Canadian Cancer Registry. All registries are included in annual updates of the NAACCR file after evaluation that NAACCR high-quality standards for incidence statistics have been met.

Selection of variables and their measurement precision were determined by RU and the research importance of the variable and its precision. The desired threshold was that the file should have fewer than 20 percent unique record sets.

To illustrate RU, we describe the evaluation of adding a county-based socioeconomic measure to the NAACCR research file, thereby potentially decreasing the geography identifier from a state to a county. The purpose of this file is

TABLE 2. RU* software assessment of the NAACCR* research file, 1995–2001, with a county-based socioeconomic measure and various variable recodes, CINA Deluxe,* 1995–2001 (4.67 million cases)

Variable list	RU weight	Unique cases		Unique case sets of ≤5	
		No.	%	No.	%
Registry, race recode, 5-year age recode, site recode, sex, year diagnosis, % poverty		1,049,929	22.4704	2,299,113	49.2053
% poverty	5.15467				
Site recode	4.98515				
5-year age recode	3.44328				
Registry	2.20534				
Race recode	1.82314				
Year diagnosis	1.72377				
Sex	0.74528				
Registry, race recode, 5-year age recode, major site, sex, year diagnosis, % poverty		540,694	11.5719	1,548,756	33.1463
% poverty	5.50175				
5-year age recode	3.58988				
Major site	3.51871				
Registry	2.4479				
Race recode	2.14181				
Year diagnosis	1.83828				
Sex	0.74634				
Registry, race recode, 5-year age recode, minor site, sex, year diagnosis, % poverty		895,096	19.1567	2,090,596	44.7427
% poverty	5.37685				
Minor site	4.61168				
5-year age recode	3.59623				
Registry	2.307				
Race recode	1.90846				
Year diagnosis	1.73937				
Sex	0.73337				
Registry, race recode, 5-year age recode, site recode, sex, year diagnosis, % poverty (integer)		693,494	14.8421	1,754,815	37.5563
Site recode	5.29957				
% poverty (integer)	3.90265				
5-year age recode	3.65113				
Registry	3.01285				
Year diagnosis	2.04606				
Race recode	1.88676				
Sex	0.75015				
Registry, race recode, 5-year age recode, site recode, sex, year diagnosis, % poverty (2% groups)		552,551	11.8256	1,499,591	32.0941
Site recode	5.47403				
Age recode	3.81203				
% poverty (2% groups)	3.33046				
Registry	3.15364				
Year diagnosis	2.05626				
Race recode	2.04164				
Sex	0.78312				
Registry, race recode, 5-year age recode, site recode, sex, year diagnosis, % poverty (3% groups)		470,073	10.0604	1,328,571	28.4339
Site recode	5.52133				
Age recode	3.81503				
Registry	3.1567				
% poverty (3% groups)	3.00068				
Race recode	2.1167				
Year diagnosis	2.06196				
Sex	0.77972				
Registry, race recode, 5-year age recode, site recode, sex, year diagnosis, % poverty (5% groups)		380,107	8.135	1,127,825	24.1376
Site recode	5.47772				
Age recode	3.84986				
Registry	3.25527				
% poverty (5% groups)	2.64693				
Race recode	2.06141				
Year diagnosis	2.0221				
Sex	0.76974				
Registry, race recode, 5-year age recode, site recode, sex, year diagnosis, % poverty (10% groups)		285,701	6.11	903,655	19.34
Site recode	5.9817				
Age recode	4.1135				
Registry	4.0793				
Race recode	2.3318				
Year diagnosis	2.3259				
% poverty (10% groups)	1.7991				
Sex	0.864				

* RU, Record Uniqueness; NAACCR, North American Association of Central Cancer Registries; CINA Deluxe, Cancer in North America research files.

TABLE 3. RU* software assessment of the NAACCR* research file, 1995–2001, with options that meet uniqueness thresholds with a county-based socioeconomic measure (% median poverty integers), CINA Deluxe,* 1995–2001 (4.67 million cases)

Variable omitted	Unique cases		Unique case sets of ≤5		Variable list
	No.	%	No.	%	
None	693,494	14.8421	1,754,815	37.5563	Registry, site recode, sex, race recode, year diagnosis, % poverty, 5-year age recode
Race	522,308	11.1784	1,487,792	31.8415	Registry, site recode, sex, year diagnosis, % poverty, 5-year age recode
Sex	493,796	10.5682	1,390,263	29.7542	Registry, site recode, race recode, year diagnosis, % poverty, 5-year age recode
Site recode	39,237	0.8397	168,028	3.5961	Registry, sex, race recode, year diagnosis, % poverty, 5-year age recode
Age recode	120,560	2.58021	464,425	9.9396	Registry, site recode, sex, race recode, year diagnosis, % poverty
Year of diagnosis	215,902	4.62071	696,925	14.9155	Registry, site recode, sex, race recode, % poverty, 5-year age recode
Registry	169,003	3.617	576,790	12.3444	Site recode, sex, race recode, year diagnosis, % poverty, 5-year age recode

* RU, Record Uniqueness; NAACCR, North American Association of Central Cancer Registries; CINA Deluxe, Cancer in North America research files.

to provide researcher access to an analytical file that would meet the NAACCR Institutional Review Board's definition of a nonidentifiable data set that would meet the file needs of most researchers.

Approach

Table 1 summarizes the variables included in the assessment with the number of values used for each iteration. The table includes all permutations that were tested for age, site, and county-based socioeconomic measure.

After each iteration, the results were examined to determine whether the RU threshold was met. If not, the weight of each variable was used to select variable(s) with the greatest impact on uniqueness, in order to collapse values for the next iteration. Because we wanted the county-based socioeconomic measure to have the greatest precision possible, it was overlooked in early iterations. Only after all iterations involving reaggregation of other variables were exhausted was collapsing county-based socioeconomic measure values attempted.

The analysis began with SEER cancer site groups (3), 5-year age groups, and four race categories. Further iterations of site codes included major site groups only and minor site groups only. Age groups were also collapsed into four 20-year age groups (table 1).

We relied on the work of Krieger et al. (4–7) and Singh et al. (8) to identify a meaningful county-based socioeconomic measure that would be simple, useful, and meaningful across many geographic areas and over time. This measure was the percentage of county residents that lived below the poverty level, data available from the US Bureau of the Census (9). This county-based socioeconomic measure is reported to the tenth decimal place, and this precision was included in the first RU iteration (table 1).

RESULTS

Variable recodes and reaggregation

Adding the county-based socioeconomic measure increased the unique record sets to about half of the records (table 2), with the county-based socioeconomic measure contributing most to uniqueness, followed by SEER site and 5-year age groups. Values of the cancer and 5-year age groups were collapsed over several iterations using several combinations (table 2). None, however, achieved the desired threshold. For example, using major site groups, unique record sets dropped to 33.1 percent. Because the threshold could not be reached, we determined that the county-based socioeconomic measure values had to be collapsed into whole integers and then 2 percent, 3 percent, 5 percent, and finally 10 percent interval groups. Each iteration reduced uniqueness sets from 37.6 percent down to 19.3 percent, using the 10 percent county-based socioeconomic measure intervals.

Variable omission

Collapsed variable combinations met the needs of researchers and, thus, we did not consider omitting variables. However, RU did provide information on the impact of omitting any of the key variables (table 3). If we chose to omit race (31.8 percent unique record sets) or sex (29.8 percent unique case sets), the threshold would still not be met. Eliminating site or age groups had the greatest impact, but without these variables, the file would not have been useful to the purpose. Eliminating either year of diagnosis or registry code would be effective in reducing unique record sets to below the uniqueness threshold (e.g., omitting registry yielded 12.3 percent unique record sets).

DISCUSSION

The RU method enables one to measure record confidentiality in electronic files, provides the flexibility to adjust record uniqueness thresholds for different users or purposes of data release, and facilitates good stewardship of confidential data. RU is useful to assess electronic data files of any size of number of records. Record uniqueness depends on the number of variables, categories within a variable, and the distribution of cases across the categories. RU can assess up to nine variables simultaneously. We believe that, beyond nine variables, one reaches a point where little new information can be gained about a patient, when so much is already known. The choice of the variables to include, collapse, or omit is the decision of the user. RU is sufficiently flexible that these decisions can be made to render the most meaningful data file for the purpose of a particular query or study. The decisions that we described to reduce uniqueness were ours and may not be choices that would be made by others to reduce uniqueness. Nonetheless, RU enables one to make the choices.

Innovative statistical methods and increments in computing speed in the future might enable inclusion of more variables in the program while still disguising unique data elements, specifically files with large numbers of variables sensitive to identifiability.

If records are selected on the basis of geography or specific disease, then this data item is essentially a known, and the probabilities of identifying a record or record set and the contributions of variables to the probabilities change and depend on, again, the number of categories in the remaining key variables and the distribution of cases across those categories. For example, a rare cancer, such as gallbladder, has small numbers of cases, but record uniqueness is more impacted by a rare tumor occurring infrequently in an age group or a small geography than by the rarity of the tumor. Omitting outliers in the frequency distributions can enable release of a data set whose utility is not compromised by extensive aggregation of values or omission of key variables. In a test run of 11,728 gallbladder cancer cases, a rare tumor, and using the same steps outlined above, we could not reach the RU threshold because of age group outliers and one small registry contributing few cases. On the other hand, in a test of 542,008 colorectal cancer cases, a tumor that is more common, with no age outliers and no registries contributing a small number of cases, the RU threshold was achieved with greater precision in the county-based socioeconomic measure (at 1 percent intervals) than was attainable when the cancer type was unknown.

RU is a useful data tool that can quantify risk of confidentiality breach of electronic health databases, including reidentifiability of cases through triangulation of information or linkage with other electronic databases. RU can help users and gatekeepers to produce the most valuable and informative research data files after assessing and protecting patient privacy.

ACKNOWLEDGMENTS

Conflict of interest: none declared.

REFERENCES

1. Office of Management and Budget, Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. Report on statistical disclosure limitation methodology. Statistical policy working paper 22. Springfield, VA: National Technical Information Service, 1994. (NTIS PB94-165305) (www.fcsm.gov/working-papers/spwp22.html).
2. North American Association of Central Cancer Registries, Inc. Record Uniqueness software. Springfield, IL: NAACCR, 2005. (http://www.naaccr.org/).
3. Havener L, ed. Standards for cancer registries. Vol III. Standards for completeness, quality, analysis, and management of data. Springfield, IL: North American Association of Central Cancer Registries, Inc, 2004.
4. The Public Health Disparities Geocoding Project monograph. Geocoding and monitoring US socioeconomic inequalities in health: an introduction to using area-based socioeconomic measures, version 7.04. Boston, MA: President and Fellows of Harvard College, 2004.
5. Krieger N, Chen JT, Waterman PD, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? Am J Epidemiol 2002;156:471–82.
6. Krieger N, Chen JT, Waterman PD, et al. Race/ethnicity, gender, and monitoring area-based socioeconomic measures—the Public Health Disparities Geocoding Project. Am J Public Health 2003;93:1655–71.
7. Krieger N, Waterman P, Lemeieux K, et al. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001;91:1114–16.
8. Singh GK, Miller BA, Hankey BF, et al. Area socioeconomic variations in U.S. cancer incidence, mortality, stage, treatment, and survival, 1975–1999. NCI cancer surveillance monograph series, number 4. Bethesda, MD: National Cancer Institute, 2003. (NIH publication no. 03-5417).
9. US Census Bureau. Small Area Income and Poverty Estimates Program: model-based estimates for states, counties & school districts. Washington, DC: US Census Bureau, 2006. (http://www.census.gov/hhes/www/saipe/saipe.html).