Unique Record Identification on Public Use Files as Tested on the 1994-1998 CINA Analytic File

Holly Howe Andy Lake Mel Lehnherr Dave Roney and Members of the CINA Deluxe Advisory Committee* *Brenda Edwards, John Fulton, Erich Kliewer, Carol Kosary, Chuck Lynch, Phyllis Wingo, William Wright

Background

An important use of cancer registry data is research and public information. However, patients' right to privacy is an ongoing concern, such that an appropriate balance is always sought between the usefulness of information being released and the potential for a breach in confidentiality. Not only of concern is patient identifiably from the file; but also patient re-identification through linkage of the registry file with other electronic files. Confidentiality is compromised when records can be uniquely identified so that, either through data elements on the file or linking records with external files, the identifying information sought by an invader becomes known. The key to preventing a confidentiality breach is to reduce the uniqueness of individual records. This is accomplished through omission of personal identifiers to eliminate direct identification, omitting variables, and aggregating variable values to diminish indirect disclosure through unique combinations of data values.

The Illinois State Cancer Registry developed a SAS program to test all data files for the potential for confidentiality breaches prior to their release as public use or analytic files. The SAS program evaluates the data file for re-identifiability, since confidential data elements (i.e., name, address, social security number) are omitted from the files. The program identifies when any combination of designated variables results in a unique record or in five or fewer unique records. Using combinations of age, sex, race, and cancer site, the Illinois registry evaluates each data file for the proportion of unique records and aggregates data within each record until the desired threshold of unique records is achieved. They have found that the proportion of unique records increases as the number of variables involved in the combinations increases. The threshold for public use files is that fewer than 5% of the records are unique.

Purpose

The purpose of this analysis was to apply the Illinois algorithm to the 1994-1998 CINA data analytic file to determine both the frequency of unique records on the entire file and the frequency for each registry. The NAACCR CINA analytic file is the data file used by NAACCR committees for research studies. The data are embedded in SEER*Stat. All variables submitted in the NAACCR call for data are included on this file when a registry meets the gold standard of data quality and when it provides consent for inclusion. At this time (July 2002), this analytic file is only released for approved NAACCR Committee research projects. However, it is the prototype for developing CINA Deluxe, the research file that will be made available under discretionary release to any NAACCR researcher. These files and the process for access are described elsewhere (packet IV of the NAACCR annual call for data). In addition, the analysis will also provide information on the usefulness of the file, by indicating the frequency of cell suppression.

Method

Source of Data

Twenty-seven registries met the high data quality inclusion criteria for the 1994-1998 CINA analytic file and gave consent to use the file for analysis and evaluation by NAACCR committees. The file consisted of 3,346,878 records. The registries included the states of Arizona, California, Colorado, Connecticut, Delaware, Florida, Hawaii, Idaho, Illinois, Iowa, Kentucky, Louisiana, Montana, Nebraska, New Hampshire, New Jersey, New Mexico, North Carolina, Pennsylvania, Rhode Island, Utah, Washington, West Virginia, Wisconsin, and Wyoming as well as the metropolitan areas of Atlanta, Georgia and Detroit, Michigan.

Variable Precision

Most variables in the CINA analytic file are in the same format as originally submitted (codes are not aggregated or grouped). However, age was aggregated into the standard 18 age groups, 0 to 4, 5 to 9.... 75+, and cancer site was recoded into the standard SEER groups. The patient identification number is suppressed and is not a unique identifier for the file (only useful to identify a particular case and link the record back with the original registry report submitted).

Approach

The Illinois algorithm uses a SAS program to identify the percent of unique records. Three iterations were run for the whole data set and then again for each registry in the combined data set:

- 1. The variables for age, sex, race, diagnosis year, and registry were included.
- 2. All the variables in number 1 above, plus cancer site group, were included.
- 3. All the variables in number 2 above, plus a 4-digit morphology code, were included.

The output identifies the percent of unique records on the file, based on the variables included for each of the three iterations. It also provides the percent of combinations that would produce fewer than 6 cases. This latter situation would be an indicator of the level of cell suppression that would occur during analysis and thus address the utility of the data file and the potential for a confidentiality breach.

The Illinois algorithm also examines, through multiple regression, the independent contribution of each involved variable to the proportion of unique records,. This estimate can be used to guide decisions about further data aggregation or variable omission, should a decrease in record uniqueness be desired.

Potential Standard

No standard has been established of an acceptable level of unique records in pubic use files, or even analytic data files released to researchers (discretionary release under more controlled circumstances). The rule of thumb in presenting data in tabular formats is that an average individual should have less than a 20% chance of being unique in any cell (thus the corresponding rule that cell sizes are suppressed when five or fewer cases occur).

Based on the Illinois standard for their discretionary file releases to researchers and the general rule of thumb for tabular data, a 20% unique record frequency was used to evaluate whether adequate protection against re-identifiability was present in the data file. This decision was based on the considerations that CINA Deluxe is a discretionary released file and it will be released only under signed agreements and only to NAACCR researchers.

Results

The table below summarizes the results of the three iterations to identify unique records for all 3.3 million records on the CINA data analytic file. In the first iteration, virtually no unique records were identified, while about 3% were found in iteration 2 and about 12% in iteration 3. Cell suppression would occur in less than 0.5% of analyses or inquiries involving the variables in iteration 1; fewer than 11% in iteration 2; and about 28% in iteration 3/

Frequency of Record Uniqueness for CINA Analytic File, 1994-1998							
Iteration	One Unique Record		Five or fewer Unique Records				
	Count	Percent	Count	Percent			
1. Age, sex, race, dx year,	1774	0.05	11695	0.35			
registry							
2. 1 above + site	99630	2.98	357585	10.68			
3. 2 above + morphology	396839	11.85	930922	27.8			

In the first iteration, the variables that contributed most to identifying a unique record were age, race, and year of diagnosis. Age contributed to a 5.5% increase in the proportion of uniqueness; race, 4.8%; and year of diagnosis, 2.3%. In the second iteration, the variables that contributed most to an increase in the proportion of unique records were site (6.1%), followed by registry (4.4%), and then age (4.3%).

The table below summarizes the registry-specific ranges in frequency of unique records within a specific registry.

Registry-specific Ranges for Unique Records, 1994-1998						
Iteration	One Unique Record		Five or fewer Unique Records			
	Low Range	High Range	Low Range	High Range		
1. Age, sex, race, dx year	0.001	0.78	0.01	2.68		
2. 1 above + site	1.03	16.66	4.62	47.1		
3. 2 above + morphology	6.83	37.5	17.9	67.9		

Registries with a low frequency of one unique record tended to be the larger registries. (Registryspecific results are available upon request). The frequencies of five or fewer unique records did not follow registry size, in fact some of the high ranges occurred among the largest registries and some of the low ranges occurred in the smallest registries. This might suggest that with more cases, there is a greater probability and greater frequency of very rare combinations.

In the first iteration, year of diagnosis contributed most to the increase in unique records for 16 registries, with race or age being first for 4 registries each, and sex contributing most to the increase in unique record frequency for two registries. In the second iteration, the site recode contributed most to the increase in unique records for all registries. Similarly, histology contributed most to the increase in unique records for all registries.

Conclusion

The purpose of this analysis was to apply the Illinois algorithm to the 1994-1998 CINA data analytic file to determine both the frequency of unique records on the entire file and the frequency for each registry. In addition, the analysis will also provide information on the usefulness of the file, by indicating the frequency of cell suppression.

The variables available in the NAACCR public use file, CINA+ Online 1994-1998, resemble the variables used for iteration 2. With the cell suppression function built into the query system, the design provides good protection for minimizing the potential for a confidentiality breach. Further, with cell suppression of all cells with fewer than 5 cases, fewer than 11% overall of all potential queries are suppressed. This file was judged to be a good balance between confidentiality protection and file usefulness.

The consensus of the CINA Deluxe Advisory Group was that all three iterations to identify unique records in the 1994-1998 file resulted in a very good balance between confidentiality protection and file utility, in conjunction with the NAACCR discretionary process in giving NAACCR researchers access to the data file. Further researchers have the option within SEER*Stat to combine categories to further minimize suppression of counts. Researchers also sign agreements attesting that there will be no linkage of the files and no attempt will be made to identify individuals.

The Group recommended that this assessment be conducted annually on each new data set. The assessment will include evaluation of the CINA combined set and a registry-specific evaluation. Each data set will be evaluated under each of the three iterations to identify unique records as described by the methods above. The group concluded that if the numbers produced were similar to those for the 1994-98 data file, then NAACCR should feel comfortable that they have made a very reasonable attempt to protect confidentiality and make a useful data file available to NAACCR researchers that sign its research agreements (and follow the protocol for using CINA data). In short, the data file has achieved a reasonable balance between file utility and confidentiality protections. In the future, establishing standards of acceptable levels of unique records in public use or analytic files may be warranted.

Finally, the Group believed that the Illinois algorithm to identify unique records and to identify the variable contributing most to the increase in the frequency of unique records was a tool that had the potential to be useful to each registry. However, several suggestions for algorithm program modification were made: (1) rewrite it in a user-friendly, flexible program; (2) summarize the output in a user-friendly format; and (3) provide program documentation to accompany the algorithm. Further development of Illinois algorithm that addresses these suggestions has been taken under advisement.

References

Confidentiality of Public Use Cancer Files. Illinois Health and Hazardous Substances Registry Newsletter . Springfield (IL): Illinois Department of Public Health, Spring 1999. pp. 1-2.

Shen T, Lehnherr M, Howe HL. Evaluation Levels of Confidentiality breaches in public use cancer files. In Howe HL. Report of the CINA Deluxe Beta Test Results. Sacramento, CA: North American Association of Central Cancer Registries, March 2000, p.4.