# Confidentiality of Public Use Cancer Files

Public use cancer files are generated and disseminated by health agencies to inform the public and to support cancer research. In the current electronic age, a concern often arises as to the confidentiality of individuals in the file. The concern is understandable because technological advances have made matching records increasingly easy and identifying individuals more possible. Confidentiality of a public use file is compromised whenever a tangible proportion of records can be uniquely identified so that, either through data elements in the file or through record linking with external files, the information sought by an invader of these records becomes known. The key to preventing the confidentiality of public files to be breached is to reduce the uniqueness of individual records. This includes deleting specific personal identifiers (e.g., name, social security number, and address) to eliminate direct identification and aggregating variables to decrease indirect disclosure through unique combinations of data values.

The public use cancer files released by the Division of Epidemiologic Studies were constructed with these considerations in mind. Apart from cancer site/stage information, there are three variables in the 1991-1995 ZIP code level file (age, sex, and ZIP code), six in the county level file (age, sex, race, alcohol, tobacco, and county), and eight in the state level file (age, sex, race, Hispanic ethnicity, alcohol, tobacco, year of diagnosis, and birth place), which can form, respectively, seven, 63, and 255 variable combinations. By plotting the proportion of unique records under each combination against the number of combinations, we produced a graph of cumulative percentage of unique records. Clearly, as the number of variables involved in the combination increased, the proportion of unique records became greater. When all variables were used, there were 0.6, 2.4, and 9.36 percent of records unique for ZIP code, county, and state files, respectively (Figure 1). The independent contribution of each involved variable to the proportion of unique records was examined using multiple regression. In the state file, for example, birth place turned out to be the most revealing variable (Table 1). As compared with excluding the variable, including birth place in the file resulted in a 150 percent increase in the percentages of unique records across combinations. This estimate provided a clear guidance for choosing birth place to aggregate if such an action is warranted.

There has been no “standard” for an allowable proportion of unique records in public use cancer files. In the 1994 SEER file, six demographic, geographic, and date of diagnosis variables identified more than 30 percent of records as unique. The rule of thumb for tabular files is that an average individual should have less than a 20 percent chance of being unique in any cell (corresponding to the rule that cells with five or fewer cases are suppressed). The Illinois public use cancer files are judged to be safe because none of the disclosable proportions exceeded 10 percent (comparable to 10 or more cases in each cell). An automated procedure has been developed by division staff to calculate proportions of unique records for all variable combinations and to pinpoint variables that have the highest influence over the unique proportion.

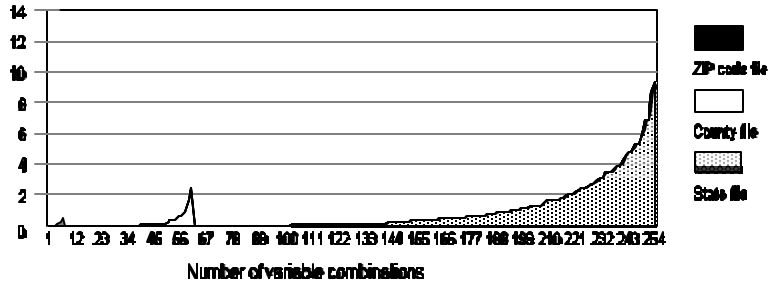


Figure 1. Cumulative percentage of identification

**Table 1. Average Percentage Changes (APC) in Proportions of Unique Records Resulting from Inclusion of Variables in State Level File**

| Variable | Coding | APC^ |
|---|---|---|
| Sex | Male/Female | 0.80 |
| Hispanic | Yes/No | 1.05 |
| Race | White/Black/Other | 1.77 |
| Alcohol | Current/Past/Never/Does not/Unknown | 2.39 |
| Year of diagnosis | Single calendar year (1991-1995) | 2.60 |
| Tobacco | Current/Past/Never/Does not/Unknown | 2.82 |
| Age | 18 five-year age groups | 8.39 |
| Birth place | 233 county/state/region codes | 150.02 |

^ Estimated by regressing natural logarithms of the proportion of unique records on the presence of variables in combinations. All estimates were statistically significant at á=0.01. $R^2$=0.92.