Confidentiality

1. General Definition and Description

Confidentiality is the maintenance of privacy concerning personally-identifiable information or, conversely, that which is not publically disseminated. Vital records contain information that can identify a person either directly, such as a name or social security number, or indirectly, such as parents' names or mailing addresses. Such information should be suppressed or not disseminated if confidentiality of an individual is to be maintained as part of a public use data set. Most vital records registration areas have statutes or regulations to protect confidentiality, especially to prevent illegal use through fraud or identify theft. Indirectly, another way confidentiality can be compromised is when vital statistics are cross-tabulated in such a way that rare situations are revealed via very small numbers of cases, which in turn can be linked to an individual ("re-identify") by matching to other, non-vital records information.

2. Application to Vital Statistics

On vital records, common identifiers are any names, social security number, date of birth, GIS coordinates or addresses, telephone number and other unique variables. Each of these, sometimes after linking with other commonly available information, may reveal the identity of an individual. Increased sophistication of matching software and ready availability of electronic records of many types and sources make linking vital records information to these data increasingly easier and successful. Sterilized vital records information, formats and aggregated files, from twenty or even ten years ago often are found now to be re-identifiable and, therefore confidentiality cannot always be assured.

Sometimes an individual data item on a vital record, such as birth date, is not re-identifying in and of itself, but it becomes so when combined with other information from the same record. For example, a birth date combined with hospital and race category in some cases could be enough information to identify an individual's birth record.

Individual records with obvious identifiers removed that become part of a large data file does not in and of itself ensure confidentiality of any particular record. The interplay of individual records with the power of computer software may result in re-identification, as highlighted by examples in the next section.

3. Examples

The following examples are common ways confidentiality can be protected or disclosure risk reduced within records or cross-tabulated data. One way is simple suppression, where any potential identifiers have restricted access or are not given out in any data sets or abstracted information meant for the public. Another way is to reduce specificity or detail among certain data elements. For example, only month and year could be included for birth dates, suppressing the day of the month. Another example is to categorize discreet variables, such as birth weight, age of mother or causes of death. Birth weight might be combined into categories of 250 gram

intervals or age of mother into five-year groups. This also can include top and/or bottom coding. For example, all very low birth weights might be combined into an under 1,500 gram category. Causes of death can be combined into major categories to avoid the tabulation of rare occurrences. Another way to reduce specificity is to aggregate time period and/or geography (e.g., quarters into a year, years into five-year periods and geography into counties or multi-county regions).

Other examples include the perturbation of data or adding noise to a data set, which is completely possible but used less often in vital records applications. This is where original records or data are changed to protect confidentiality without widespread impacts on changing what the data characterize in a public use data set. Sometimes this is done by "data switching," where several records of data for selected variables are swapped among them. In this way, totals or characteristics of variables remain unchanged within a certain period and/or geography.

Among cross-tabulated vital statistics, resulting cells below a certain size (such as less than 5 or 10 cases) can be unilaterally suppressed. A more sophisticated approach is to also consider the time period, geography and specificity of the variables. For example, if a cross-tabulation reveals that there is only one death due to a rare form of cancer in a given year in a large state, there is little reason to suppress this cell. There is nothing confidential being revealed because there is not enough information to re-identify this person. Even if it is well known, say via the news media, that a person named X died of this rare form or cancer, it is very likely that the year and the state name do not reveal new information about the person named X that is not already known. However, if this same form of cancer is cross-tabulated by age at death, county, and race category, it may be possible to determine who the individual might be.

Here is a specific example to illustrate some of these points.

Crosstab #1 – Mortality Counts for Selected Variables by County: 2008

Geography	All Cancers	Cervical Cancer	Age 25-29	API* race
State Q	24,000	184	700	55
County X	150	10	30	3
County Y	50	5	5	1
County Z	100	12	15	0

Crosstab #2 – Mortality Counts for Cervical Cancer by Age by API* Race by County: 2008

Geography/Age	All Races	API* Race
State Q		
Age 25-29 years	5	1
Age 30-49 years	80	5
County X		
Age 25-29	2	1
Age 30-49	8	2
County Y		
Age 25-29	0	0
Age 30-49	4	0

County Z		
Age 25-29	1	0
Age 30-49	8	0

*API = Asian and Pacific Islander race category based on the 1977 OMB definition

Crosstab #1 shows separate variables singly cross-tabulated with state and county. This results in information that compromises very little any individual confidentiality due to large cell numbers (e.g., those under All Cancers) and/or the underlying population or universe from which it is drawn is large (e.g., all cervical cancer deaths or all API race category deaths). In counties X and Y where there are 3 and 1 deaths to people who were classified as API race, these might be suppressed for reasons of confidentiality if there was one API family or household in each of these counties or an otherwise small underlying population.

Crosstab #2 shows a multiple cross-tabulation of five of the variables in crosstab #1: cervical cancer death by age by API race by geography by calendar year (with a sixth implied variable, sex). Each additional variable in the cross-tabulation increases specificity. The 1 API death for age 25-29 years in County X, especially if this is a small population county, presents the most risk of re-identification. The same 1 death for all of State Q, especially if this is a large population state (AND without any known county-level data, like shown here), presents much less risk of re-identification. The other non-zero values in crosstab #2 represent varying risk levels of re-identification.

Both of these examples illustrate that there are degrees of confidentiality compromise or risk of re-identification. A number of the aggregation and suppression techniques mentioned above could mitigate confidentiality issues in these examples. All data releases involving individual records or aggregations based on individual records are a compromise between complete confidentiality (i.e., data never divulged) and complete transparency (i.e., open records).

4. Technical Notes

- Breach of confidentiality can sometimes occur in unusual ways. For example, a detailed urban block map in a research paper showing a plot of case occurrences to highlight distribution or clusters can be tantamount to publishing street addresses!
- There often are jurisdictional laws that protect confidentiality. Some states have “open records” policies and others are very restrictive.
- Generally, electronic systems containing vital event data have confidentiality safeguards. For example, allowing only multiple-year aggregates of data years can be a safeguard against re-identification by reducing temporal specificity. Electronic systems also recode or do not include variables that could be used for linkages to other datasets. The linkages can result in unique profiles of variables that can lead to re-identification of individuals.
- Sometimes having the same records in more than one public data file increases the likelihood of re-identification. If a high-visibility/unique record (or table cells) in one data file (or report) has more detailed or overlapping categories for a shared variable in another data file, the two files or tables can be compared to determine the value for that variable more precisely, possibly defeating categorization, top/bottom coding or other

aggregation techniques.

5. References and Resources

U.S. Census Bureau paper on using noise:
http://www.census.gov/srd/papers/pdf/bte9601.pdf

U.S. Census Bureau paper on data swapping:
http://www.census.gov/srd/papers/pdf/rr96-4.pdf

Illinois Dept of Public Health guidelines on generating and maintaining public use files:
http://www.idph.state.il.us/about/epi/pdf/PUDEpiReport.pdf

Article on identifying variables and characteristics in public use datasets that contribute to re-identification via record uniqueness:
Holly L. Howe, Andrew J. Lake and Tiefu Shen, "Method to Assess Identifiability in Electronic Data Files," American Journal of Epidemiology Vol. 165, No. 5: 597-601 (2001).

NCHS Staff Manual on Confidentiality