Identifiability in Microdata Files

Federal Committee on Statistical Methodology
Confidentiality and Data Access Committee

Data holders frequently wish to release information in electronic form without unacceptable risk of identification of the persons or establishments who provided it or to whom it pertains. Usually, but not always, an assurance of confidentiality has been provided that the data would not be shared with anyone, *in identifiable form*, except as consented to by the data provider.

It is not always clear, however, what in a microdata file[i] makes respondents identifiable. This document provides an understanding of what variables and types of data might make individual respondents identifiable in a microdata file. It is intended not only for those who are considering the release of microdata to the public but for data stewards and others holding files who need to be sure as to whether such information should be protected; for analysts who may be unsure of how to handle files they have been provided or created; for members of institutional review boards or other review panels in assessing risk of a protocol; and for administrators who need to establish policy concerning the release and handling of data collected by their organizations.

Identifying respondents on a microdata file: Identifiable information refers to data which can be used to identify individuals or establishments, whether *directly* - using items such as name, address or unique identifying number - or *indirectly* - by linking data about a respondent with other information that uniquely identifies that respondent.

Direct identification: Information such as names, addresses, social security numbers, or establishment identification numbers (see Appendix A for more examples) can be used to identify an individual or establishment with a high degree of certainty. This is information of a type uniquely or closely associated with only one person and little, if any, additional information is required to determine the exact identity of the individual to whom they apply.

When direct identification can be made, in addition to identifying the individual or establishment, the information that is associated with that individual or establishment is also disclosed. Information concerning other family members (family income, genetic characteristics, housing) or study participation by neighbors (in clustered samples) may be revealed as well.

In practice, an attempt to identify a respondent based on direct identifiers in a national or even a State file, might require several steps. Given the amount of information available in some microdata files, however, unless measures are taken to reduce disclosure risk (see "Further Information") this is a relatively simple task.

Indirect identification: An internal file stripped of names or other direct identifiers may nonetheless remain identifiable, if sufficient data are left on the file with which to match with information from an external source that *also contains names or other direct identifiers*[ii]. In such a case, the identity, as well as all information in the file associated with that person has been

disclosed. Appendix A provides examples of data items that may be used in the matching process.

Matching, The General Case: Matching (in some settings referred to as "linking") involves establishing a correspondence between variables shared in common by two or more files. Suppose items such as county or town of residence, exact date of birth, and detailed education and occupation, were included in a file being considered for release, together with information concerning race, gender and place of birth. These are items that are collected routinely in many surveys and *which also appear on files available externally.* If this file were to be released to the public, it would not be difficult to match its contents with those found in data sources containing names or other direct identifiers.

There are a number of situations in which matching presents a risk, and we discuss them below. Risks are associated with matching to files that are:

(a) available to the general public or privately held;
(b) from the same investigation, previously released;
(c) from a different, but related investigation; and
(d) held by other agencies or institutions.

The basic process by which comparisons or matching of records held in two files can lead to identification are illustrated in Figure 1.

Figure 1. Schematic representation of identification of respondents after direct identifier has been deleted

Removal of direct identifier	File content
1. File with direct identifier	**name** abcdefghijkl
2. File with direct identifier deleted	abcdefghijkl
Matching of files with items held in common	
2. Public use file	**abcdef**ghijkl
3. Material available in external file	**abcdef**mnopq **name**

Where:
a = Day, month, year of birth
b = Gender
c = State of residence
d = County of residence
e = Occupation
f = Race

Matching with publicly available or privately held files:

Among publicly available information sources (often found in data base form) that could be used for matching - and which contain names - are: voter registration lists; occupational licensing registries; drivers license records; city directories; property records; crime/court records; corporate proxy statements; stock holding reports; and birth, death and marriage records. In addition, private data bases held by credit or marketing agencies should not be overlooked. Publicly available data bases are often available for small geographic units (counties, voting districts, zip codes), but can be aggregated to larger areas.

In most cases several people may meet the criteria for a match. However, individuals can be easily "singled out" by the matching process if a person or establishment has rare or highly visible characteristics that make them unique in the internal file and which also make them unique in an external database. If the later includes the entire population in the area to which the data pertain, then a disclosure will have been effected.

High visibility characteristics may contribute to the identification of an individual – with or without sophisticated matching methodologies. A member of, for example, one of the Asian subgroups would be easily observable in an external data base if located outside of the areas of heavy concentration for that group. Indeed, members of the community in which the ethnic group is located – (e.g. an Aleut in an Ohio County, a Filipino family in a small township in Maine) could identify them even without the use of another data base. Identification would be even more likely if an individual had an unusual combination of characteristics - e.g. a 20 year-old Ph.D. The same would be true of those with very extreme values of certain variables, e.g. income, rent, height or rare health conditions. All of these situations are more likely to result in identification if they are observed for small geographic areas.

Similarly items such as exact date of birth or death could be used to match records to data bases such as newspaper obituary and birth announcement data bases, and voter registration lists.

Matching with previously released files:

a) From the same investigation: Often, consideration is given to the release of a file from an investigation after the release of another file from the same investigation (common examples are files from longitudinal or "panel" studies, multipurpose surveys, and special purpose files). While each of these files, considered separately, may be considered safe to release, there may be enough information in the two files *combined* to bring about a disclosure. For example, suppose that identities of certain metropolitan areas have been provided in a previously-released file and that the inclusion of information concerning race, occupation, and other demographic variables was considered safe as long they were coded at a high level of aggregation. Consideration is then given to the inclusion of those same variables coded with much greater detail in a new file but with no identities for metropolitan areas.

Even though the two files contain no individual identifiers or no internal identification number with which to match individual files, it must be remembered that if they are based on the same survey, they will have much *data* in common (probably coded the same way) that can be used to match many of the records of the two files. Unless deliberate steps are taken to change the certainty with which each data string characterizes the same respondent in both files, the data themselves may be combined to provide details for socio-economic variables at lower than desired levels of geography. (In our example, the protection afforded by a high level of aggregation of socio-demographic variables in the first file would have been removed after a match with the second filed was carried out.)

b) From a different but related investigation: A somewhat different situation arises where respondents to one survey are re-contacted in a subsequent investigation. For example, both the National Center for Health Statistics (NCHS) Longitudinal Study on Aging and the Agency for Health Care Quality's Medical Evaluation Panel re-interview respondents from the NCHS National Health Interview Surveys. Key items of information (e.g. geography) released from the second survey cannot be more detailed than that released for the first survey if data from the two files can to be linked.

Matching with files held by other agencies or institutions: An enormous amount of information is gathered by federal agencies, universities and research institutions for a variety of purposes. Some research is only feasible because respondents are assured that information about them would never be released to a federal agency that might use it for other than research (administrative) purposes. In other cases, a university may have conducted similar research among the same respondents as those selected in another survey. In either case, it is necessary to be alert to the possibility that matching information may be contained in other data bases.

Other sources of risk: In addition to considering the risk associated with matching to existing files, there are additional sources of risk that should be considered when planning the release of public use microdata files. Some issues to consider are discussed below and a more complete discussion can be found in the *Checklist on Disclosure Potential of Proposed Data Releases*.

Special problems associated with sampling information and sampling design: Unless consent has been obtained from respondents, it is not permissible to share identifiable information with the organization supplying the sampling frame. For this reason, organizations such as Dun & Bradstreet, the American Hospital Association, the American Medical Association, and SMG should not be able to link information they hold (the original frame) with survey data made publicly available. Details permitting such links would make the survey information in a file identifiable to those organizations *and to anyone with whom they share such information*.

Special risks are associated with "complex" surveys: Complex survey designs must be evaluated for risk potential, especially those for which the sample design involves multiple groupings of the populations studied, at least one of which is a geographic entity. If the information provided to researchers is too detailed, knowledge of which geographic entities were utilized in the drawing of the survey sample could be derived and ultimately used to link an individual's record with information for that individual contained in external data bases. To eliminate this possibility, survey sample designs should receive special attention as a potential source of

geographic information that would compromise disclosure principles and guidelines. In some cases, sample units or strata may be combined or regrouped to prevent their identification. Whatever changes are made, consideration should be given to achieving minimal impact on statistical precision consistent with confidentiality protection.

Disclosure risks associated with "contextual" data: Finally, it is increasingly feasible to enhance population-based survey microdata files with information concerning the geographic areas within which survey respondents reside - i.e., contextual, or ecological information. A wide variety of data are available from such data bases as the Area Resource File, Census County-City Data files, Centers for Disease Control and Prevention (CDC) Sexually Transmitted Diseases (STD) files, and FBI Uniform Crime Reports. When such information is added to the great amount of data already contained in a microdata file, the probability that a survey respondent is easily distinguishable from other respondents and from other persons in the general population increases sharply[iii]. Furthermore, unless the added contextual data are altered in some way, the chances of matching the contextual information for a respondent with that in the source file are nearly certain. If that source file contains either area names or additional information useful for identifying areas, disclosure risk will increase, particularly for small areas.

Further Information:

There are a number of methods for reducing the risk of disclosure in an electronic file. An excellent primer on statistical disclosure limitation techniques can be found in Section H of:

1) Federal Committee on Statistical Methodology, Statistical Policy Working Paper 22

The remainder of this report contains more advanced statistical techniques as well as a description of federal agencies practices as of the early 1990s.

A more recent publication describing recent advances and emerging issues can be found in:

2) Doyle, Pat, Julia I. Lane, Jules J.S. Theeuwes, and Laura V. Zayatz (eds.). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam: North-Holland, 2001*

A very useful (and generally user friendly) "checklist" for preparing tabular as well as microdata for public release is available in:

3) Confidentiality and Data Access Committee, Checklist on Disclosure Potential of Proposed Data Releases

Additional information and related material is available at the web sites of two committees:

4) Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee, and

5) The American Statistical Association, Privacy and Confidentiality Committee

Appendix A

Direct Identifiers, consist of information that is uniquely or closely associated with individuals or establishments.

Examples of direct identifiers include:

- Complete name
- Street addresses including zip codes
- Social Security number
- Establishment number
- Medicare or Medicaid number
- Telephone numbers including area codes
- Email address
- Drivers License numbers
- Other linkable information (e.g., ID numbers, certification numbers, etc.)

Indirect Identifiers, are variables, which, when combined, may uniquely characterize individual respondents. These combinations may be used to secure a match with external data sources in which respondents are also uniquely characterized - and which contain names or other direct identifiers.

Examples of potential indirect identifiers include:

- Geographic areas (city, county, zip code, institution) with fewer than100,000 residents (this may be higher depending upon the amount of detail in the file and other restrictions imposed on the data)
- Sample information which contains geographic detail (names of primary sampling units)
- Organizations to which respondents belong or have belonged to (e.g. Veteran status)
- Detailed race
- Occupation coded above at 3 digits or higher
- Health condition coded at 3 digits ICD or higher
- Cause of death coded at 3 digits ICD or higher
- Exact date of entry into health care, or other institution/treatment
- Exact (day, month, year) date of vital event (death, birth, marriage)
- Exact date of beginning and ending of employment
- Ungrouped (continuous) values for age (i.e. single years from 0 - highest age)
- Detailed age (months, days) for those under age 1
- Ungrouped (continuous) values for: income, value or purchase price of property, amount of rent or mortgage, property taxes
- Detailed educational level
- With data from multiple persons in a household, unique multi-racial composition, atypical number and ages of children, extreme reproductive age, significant differences in spouses' ages

- Any combination of variables whose cross classification results in cells with a very small n, especially when the members of the cell are rare and highly visible in the population in which they are found (e.g., the Eskimo in Nebraska, or very young minority female with high income or advanced degree).

Warning: Even though a data file contains less detail than that provided by the above variables, a broader code structure with larger categories might still represent a disclosure risk *if it characterizes only a very small number* of study participants. Two considerations are paramount:

(1) will the classification or information permit the isolation of respondents - does it produce "outliers"? and
(2) is this information likely to be found in an external data source that also contains direct identifiers?

Endnotes

[i] Also referred to as "line listed" or "person" records, these are electronic files consisting of individual records each containing values of variables for a single person, business establishment, or other unit.

[ii] This match need not be exact. In the absence of a unique identifier, linkage is of a "probabilistic" rather than "deterministic" nature owing to the many variations and defects in correspondence even with commonly accepted personal items (names, addresses, etc.) In addition, matches related to identification do not always involve records for individuals. For example, an intruder may want to match the characteristics of all respondents within the same sample unit with similar information from an external source of data, such as the Census. It would not be necessary to have exact data to distinguish among certain counties or cities because even within fairly broad limits many small geographic units are easily distinguishable on the basis of socioeconomic characteristics (e.g. those with high concentrations of certain racial and/or ethnic groups). Once an individual's records are associated with a small geographic area, the possibility of identification is greatly increased.

[iii] The ability to "single out" a respondent within a survey is the first step in identifying the person. Unpublished research has shown that the values for as few as six variables have been shown to be enough to uniquely describe each respondent in a survey. Because most surveys contain hundreds of variables, the addition of many more variables from an external source means that the odds that one individual is described uniquely are astronomically high and the chance of identifying that person is virtually certain.

IdentCDAC3.doc at http://www.fcsm.gov/cdac/index.html
(as of July 5, 2002)