Drug and Alcohol Services Information System

The DASIS Report

December 10, 2003

Protecting Confidentiality in TEDS

In Brief

- Disclosure analysis involves the careful examination of indirect identifiers such as gender, education, race, and income that could be used in combination to identify a respondent
- Disclosure protection techniques must balance the trade-off between analytic utility and data loss, taking key uses of the data into account
- It is important that disclosure protection procedures be made publicly available so that analysts are aware of the changes to the data

Preserving the confidentiality of survey and administrative data in publicly distributed data files is central to maintaining the trust of study participants and the ongoing ability of Federal agencies to collect important statistical information on State and national trends, service utilization, treatment outcomes, and numerous other topics. How are promises of confidentiality maintained? This report explains the general methods used to protect data and the specific methods applied to the Treatment Episode Data Set (TEDS).

TEDS is an annual compilation of data on the demographic characteristics and substance abuse problems of those admitted for substance abuse treatment. The information comes primarily from facilities that receive some public funding. TEDS records represent admissions rather than individuals, as a person may be admitted to treatment more than one time during a calendar year.

The DASIS Report is published periodically by the Office of Applied Studies, Substance Abuse and Mental Health Services Administration (**SAMHSA**).

Routine Methods

Routine confidentiality protection measures include removing *direct* identifiers such as name, social security and other identification numbers, and address. Routine processes may also include techniques such as:

- Re-coding variables (e.g., using date of birth to calculate age)
- Top- or bottom-coding variables in which the lowest or highest codes could distinguish respondent records (e.g., converting top incomes to "$100,000 or more").

These measures help remove the uniqueness of individual records within a file.

Disclosure Analysis

To determine if there are any remaining threats to confidentiality, a disclosure analysis is conducted. Disclosure analysis involves the careful examination of a data file for *indirect* identifiers that could be used in combination to attempt to re-identify (or "disclose") a respondent.[1] Examples of indirect identifiers are personal characteristics such as gender, education, income, race, and ethnicity or organizational characteristics such as capacity, services offered, and programs for special populations.[2] Disclosure analysis also involves assessment of external databases that may be used to link data.[3]

A critical factor in disclosure analysis is the availability of geographic data. In general, the more specific the geography, the more attention must be paid to disclosure risk. Geographic codes immediately narrow the search for a specific record to the lowest level of geography on the file. Finding a record within a known city is easier than finding a record within a known State, for example.

Also, analysts are often interested in subgroups of survey populations (e.g., pregnant women, racial minorities, and persons with health conditions). Yet characteristics that distinguish these groups are often the very characteristics that create disclosure risk. Disclosure protection techniques must take into account the key uses of the data and balance the trade-off between analytic utility and data loss. The analysis must include the impact on the data: What is lost and gained by the methods proposed? Which analytic capabilities are diminished? Which are preserved? How will the information lost affect data interpretation? Can the lost information be released in some other way?

Application of Disclosure Techniques to TEDS

TEDS data present a disclosure risk because they contain unique records. For analytic purposes, the Primary Metropolitan Statistical Area (PMSA)[4] and detailed race and ethnicity codes are important to retain on the file. Thus, procedures are needed to ensure the protection of records that are unique within these detailed geographic areas.[5]

Disclosure protection methods include removing some of the uniqueness of records by re-coding variables such as age and education into categories. For the unique records that remained, a technique called data swapping is used. As a first step, data swapping involves identifying the set of variables that when taken together could potentially identify an individual. Next, a sample of the unique records is swapped with records taken from elsewhere in the file that match them on essential analytic variables. This process leaves the analyses of these variables unaffected. The following sets of variables are defined for purposes of data swapping in TEDS:

Unique key (the set of potentially identifying variables): Age, gender, methadone planned as part of treatment,[6] race, ethnicity, pregnancy, and veteran status.

Swapping key (variables used to match unique records and deemed important to remain intact for analytic purposes): Race, ethnicity, sex, age, pregnancy, primary substance of abuse, and methadone planned as part of treatment.

Swapping attribute (the variable over which swapping occurs, typically a geographic variable): Matches were first sought within Census division, then Census region, and then across the entire file.

Protected variables: All other variables on the file.

The steps involved are:

1. Apply standard re-codes.
2. Identify unique records.
3. Determine the subset of unique records to be swapped.
4. Randomize all records.
5. Identify pairs of swappable records based on (a) match on swapping key and (b) geographic hierarchy, arranged from smallest to largest.
6. Run swap program; finalize swaps.
7. Review a sample of swapped records to ensure procedures were applied accurately.
8. Re-run critical analyses to ensure the integrity of the results.

Results

Data swapping has several benefits over other disclosure protection options for TEDS:

1. The overall impact on the data is very small; less than 1 percent of the records were impacted by the disclosure procedures;
2. Data for special populations (e.g., minorities, pregnant women) are no more impacted than other data; and
3. The procedures allow detailed codes to remain on the public use file (e.g., PMSA and the original race and ethnicity codes).[7]

Publicly Available Documentation of Procedures

To the extent possible, it is important that disclosure protection procedures be made publicly available so that analysts are aware of the changes to the data. The TEDS codebook details the protection procedures applied to the file, and includes an analysis of the impact to the data. TEDS is distributed through the Substance Abuse and Mental Health Data Archive (SAMHDA). See the SAMHDA Web site at http://www.icpsr.umich.edu:8080/SAMHDA-SERIES/00056.xml for access to TEDS files and online analysis.

End Notes

[1] Rather than applying simplistic techniques that would render the public use data unsatisfactory for key analyses or restrict the availability of the data, the application of disclosure techniques, in most cases, allows dissemination of a public use version of the data that is effective for most analytic purposes.

[2] Disclosure analysis typically focuses on attributes that can be known by an outsider (e.g., age) rather than attitudes or beliefs (e.g., feelings of sadness).

[3] Two files may themselves not contain disclosure risks, but contain a common identifier such as a facility ID that can be used to link the files, and thus create disclosure risk.

[4] TEDS includes codes for Metropolitan Statistical Areas (MSAs), Primary Metropolitan Statistical Areas (PMSAs), and New England County Metropolitan Areas (NECMAs), all under the "PMSA" variable. The Census Bureau provides detailed definitions of these terms on its Web site: http://www.census.gov.

[5] For example, consider how a record in a given PMSA with the following characteristics would stand out in a file: Age: *25-29*; Gender: *Female*; Methadone planned: *Yes*; Race: *Native American*; Ethnicity: *Not Hispanic*; Pregnant: *Yes*; Veteran: *Yes*.

[6] All methadone clinics are federally licensed. Some sparsely populated states have just one or two licensed facilities, and the names of the facilities are publicly available. Therefore, knowing that methadone was planned as part of treatment could indicate the approximate geographic location of the client.

[7] For further discussion of data swapping see (1) Steel, P., and Zayatz, L., "Disclosure Limitation for the 2000 Census of Housing and Population," in Statistical Data Protection: Proceedings of the Conference, Lisbon, 25-27 March 1998, Eurostat, 1999; and (2) Reiss, S., "Practical Data Swapping: The First Steps," ACM Transactions on Database Systems, 9 (March 1984).

The Drug and Alcohol Services Information System (DASIS) is an integrated data system maintained by the Office of Applied Studies, Substance Abuse and Mental Health Services Administration (SAMHSA). One component of DASIS is the Treatment Episode Data Set (TEDS). TEDS is a compilation of data on the demographic characteristics and substance abuse problems of those admitted for substance abuse treatment. The information comes primarily from facilities that receive some public funding. Information on treatment admissions is routinely collected by State administrative systems and then submitted to SAMHSA in a standard format. Approximately 1.6 million records are included in TEDS each year. TEDS records represent admissions rather than individuals, as a person may be admitted to treatment more than once.

The DASIS Report is prepared by the Office of Applied Studies, SAMHSA; Synectics for Management Decisions, Inc., Arlington, Virginia; and RTI, Research Triangle Park, North Carolina.

Access the latest TEDS reports at: http://www.samhsa.gov/oas/dasis.htm

Access the latest TEDS public use files at: http://www.samhsa.gov/oas/SAMHDA.htm

Other substance abuse reports are available at: http://www.DrugAbuseStatistics.samhsa.gov

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Office of Applied Studies
www.samhsa.gov