

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint ECE/Eurostat Work Session on
Statistical Data Confidentiality**
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 24
English only

Topic (i): new applications of disclosure control methods

**NATIONAL CENTER FOR HEALTH STATISTICS APPROACHES TO PROTECTION
AND RELEASE OF MICRODATA**

Submitted by the US National Center for Health Statistics¹

Contributed paper

I. INTRODUCTION

1. The National Center for Health Statistics (NCHS) is the agency with the primary responsibility for the collection, analysis and dissemination of health and health-related data in the United States. Examples of data collected are: the National Health Interview Survey (NHIS) is a large sample of about 40,000 households (about 110,000 persons) and collects a wide range of information on both individuals and families including health status measures, acute and chronic conditions, health insurance, and family resources; the National Health and Nutrition Examination Survey (NHANES), a sample of over 30,000 persons which conducts both an in-person interview and a medical examination consisting of numerous tests and measurements on its subjects; death certificate information on all deaths occurring in the United States each year; information on all births in the United States; the National Survey of Family Growth (NSFG) which asks questions related to reproductive issues on about 10,000 women of childbearing age. Data from these and other NCHS data collection systems are used for research, policy, and programmatic purposes at the national level. There is both a desire and a need to access, analyze, interpret and publish estimates from these data sources for sub-national geographical entities such as states, counties, or even sub-county areas because many of the parameters that affect health in the United States such as laws, reimbursement policies, and health care providers operate at the local level. Additionally, social science research in recent years has demonstrated the value of contextual analyses of data such as that collected by the NCHS for assessing risk, program development, and evaluation. Contextual analyses require the use of variables at the state, county, census tract, or block-group level to be effective.

2. Many of the surveys and data collection systems of the NCHS are capable of producing estimates at lower than national levels of geography such as state, county, or city and could provide an invaluable source of data for local areas. However, the NCHS collects its data under Section 308(d) of the Public Health Services Act which provides the legal requirements for protecting NCHS records as well as allowing external researchers access to unidentifiable microdata. Specifically, 308(d) states: "No information, if an establishment or person supplying the information or described in it is identified, obtained in the course of activities undertaken or supported under section 304, 305, 306, 307, or 309 may be used for any purpose other than the purposes for which it was supplied unless such establishment or person has consented to its use for such other purpose and in the case of information

¹ Prepared by John Horm.

obtained in the course of health statistical or epidemiological activities under section 304 or 306, such information may not be published or released in other form if the particular establishment or person supplying the information or described in it is identifiable unless such establishment or person had consented to its publication or release in other form.” This law prevents the public release of NCHS data that could, either directly or inferentially, identify the respondents to its surveys whether the respondents be individuals or establishments. The NCHS has the policy of not releasing the identity for areas with a total population size of less than 100,000 persons and in actual practice rarely releases the identity of places with less than 500,000 population. Furthermore, the NCHS Staff Manual on Confidentiality prohibits releasing the identity of the PSU's used in its surveys. The Public Health Service Act and the NCHS Staff Manual on Confidentiality do not allow the NCHS to release data files with local identifiers and hence limit the utility of the data.

II. DIRECT AND INFERENTIAL IDENTIFICATION

3. Prevention of direct identification is readily accomplished by removing individual identifiers such as names, social security numbers, addresses, telephone numbers, and other direct identifiers. The prevention of inferential identification is much more difficult. Inferential identification takes several forms. If local area identifiers such as states or counties are included on micro-data files, profile vectors including such variables as age, race, sex, education, income, occupation, family structure and perhaps others could be formed from the data for individual respondents that could then be matched to exogenous local data files such as disease registries or program recipients. Another type of inferential identification that need consideration is the inferential identification of local areas based on forming profile vectors from clustering variables included on the micro-data files. A prime example of a clustering variable would be a randomized code indicating which respondents belong together (cluster) in the same primary sampling unit (PSU). If one then has the knowledge that PSU's consist of counties, one could then form profile vectors for each PSU from the sample data and attempt to match those profile vectors to similar profile vectors from exogenous data files, perhaps census data, and then place direct area identifiers on the sample data.

III. GENERAL METHODS USED FOR DISCLOSURE LIMITATION OF MICRO-DATA FILES

4. **Perturbation of the data** - There are a number of general statistical approaches used by data collection and dissemination organizations for the protection of micro-data files. The NCHS uses a number of these methods for special situations. The reader should note that it is likely that no single method will provide adequate protection by itself and to achieve adequate protection, more than one of these methodologies will have to be employed. The reader should also note that the goal of adding some sort of perturbation to the data file is to protect the file against either intentional or unintentional disclosure of a subject's identity. This goal must be considered in conjunction with the analytic validity of the resultant data file. It is likely that any disturbance of the data will result in a data file that is less useful than the original file however, estimates from the perturbed data file must be valid. A judgement must be made regarding the validity, utility, and disclosure risk of the perturbed file. In all cases the choice and evaluation of the method is multivariate, a variable that poses little risk by itself, may result in an exact disclosure when taken in conjunction with other variables.

5. **Statistical (random) noise** - Random noise may be added to continuous or pseudo-continuous variables (ordinal or interval variable with sufficient categories that they may be considered continuous). Often the noise may not be from a uniform distribution but may belong to a long-tailed distribution or a platykurtic normal distribution. The form of the random noise depends on the distributional characteristics of the variable, at least within the range being considered for additive

noise. Statistical noise should also honor natural cut points or groupings in the data. For example, in the United States, the public school age is from 6 to 18, college from 19 to 24, and the traditional retirement age is 65, thus if noise were to be added to respondents age, it should be added within and not cross over these groupings. One must be careful in the addition of random noise because it can affect both the utility and validity of the resultant file. If too much noise is added, the validity and hence the utility of the data can be destroyed.

6. **Top and bottom coding** - Top- and bottom-coding is useful in those situations where sensitive variables that could be used for identification suffer from outliers at either the upper or lower sides of the distribution. In this case one can choose a cut-point such that all cases below (above) the cut-point are not unique and pose little risk of disclosure. The choice of the upper and lower cut-points is entirely dependent on the distributional characteristics of the data.

7. **Grouping and rounding (limiting detail)** - Grouping and rounding have traditionally been used as first stage approaches to disclosure limitation of micro-data files. Grouping and rounding suffer from at least two problems that limit their utility as disclosure limitation techniques. Firstly, data users can readily discern from the data which variables have been affected and secondly, the data user knows the range of values which contains the true value. The first problem may direct the user away from attempting to use that variable as an identifier and the second problem may direct the user into using that variable as a blocking variable and may enhance the identification of the resultant data file.

8. **Variable suppression** - Some variables are either so sensitive or not amenable to other types of disturbances that the only solution is to remove them from any file considered for public release. This form of disclosure limitation clearly affects the utility of the resultant data but has little if any effect on the validity of the file. The exception is where a critical covariate has been suppressed.

9. **Adding/removing records** - In some cases there are entire records that are not amenable to any of the above methods. These records may be removed from the public use data files. In other cases there may be a set of records for which removal would be undesirable. These records may be duplicated either in whole or in part with some values changed so the data user would not know which records have been duplicated. The adding and removal of records has the advantage that multivariate correlations are not affected but estimates may be affected if entire classes of cases are removed (duplicated).

10. Adding and removing records is a form of sub-sampling and super-sampling which, when taken to the extreme, results in an entirely new data file. The NCHS has not yet produced any data files for public release based on sub- and super-sampling.

11. **Evaluation** - The results of adding disturbances to micro-data files must be evaluated both in terms of the protection afforded and in terms of the validity and utility of the resultant data file. Unless population data are available to allow testing the identifiability of the proposed file, the best that can be done is make guesses, albeit educated guesses, based on the determination of outliers and unique values in the sample and estimate the probability that these cases would also be unique in the population.

12. It is equally difficult to assess the validity and utility of the disturbed data file since the agency developing the file usually does not know what uses will be made of the file. Based on substantive knowledge of the data and assumed analyses, the agency can produce sets of univariate and multivariate estimates from both the original file and the disturbed file and make comparisons between the estimates. In addition to point estimates, multivariate correlations should be generated from both forms of the data file. These comparisons provide some elementary information on which to base a decision about the validity of the disturbed file.

13. Assessment of the analytic utility of the disturbed file generally requires compromises to be made between the desired utility and that required by confidentiality constraints. At one extreme, maximizing utility, the original file would be released to the public. However, this may afford little if any protection from disclosure risk. The other extreme, maximizing the protection from disclosure would result in not releasing the file. This would reduce the utility to zero. Within these two extremes there may lie an acceptable compromise, that of releasing a data file with some disturbances, perhaps by using one or more of the approaches outlined above which provides adequate protection from disclosure and provides an analytically useful data file.

IV. THE NCHS RESEARCH DATA CENTER

14. While the methods above are in use and serve to satisfy a segment of the research community, there are limitations to the use of disturbed data files. For example, the NCHS can not release to the public the identity of the primary sampling units in its surveys; it can not release the identity of areas with a population of less than 100,000 persons, this includes many counties along with census tracts, block-groups, and blocks. The identities of small areas is needed because social science research has shown the value of using contextual variables in attempts to model risks, behaviors, and outcomes of individuals. Indeed, the National Survey of Family Growth (NSFG) received funding to create the "contextual data file" as an adjunct to the survey data. This contextual data file consists of about 300 contextual variables at each of the state, county, census tract, and block-group levels for a total of approximately 1,300 contextual variables. Research indicated that if the NSFG survey data were linked with the contextual data and a file was to be released without direct identifiers for the local geographic areas, a subset of the contextual variables themselves could be used to form unique identifiers for each of the geographic areas. Contextual variables need not be continuous to cause identification problems, even binary variables are capable of causing problems. Consider the case where a researcher has amassed a national dataset of local ordinances governing smoking behavior at the county level. These ordinances may be expressed as a set of binary variables indicating the permissibility of smoking in restaurants, lounges, theaters, in the workplace, in public, etc. Twelve binary variables can mathematically have 4,096 (2^{12}) unique patterns. There are 3,141 counties in the United States. Clearly another approach to accessing data files with local area identifiers was needed if the NCHS was desirous of maximizing the use of its data. The solution was to develop, staff, and equip a Research Data Center (RDC) which opened in 1998.

15. The NCHS RDC is a secure monitored facility where external researchers may be allowed access to internal restricted data files for approved projects. Restricted data files are those which contain information, such as lower levels of geography (e.g., state, county, or lower), but do not contain direct identifiers (e.g., name or social security number). Restricted data files may be used in the RDC by researchers wishing to control for geographic area in their models or they may be used to merge additional data onto the NCHS collected data files for enhanced analyses (e.g., the NSFG contextual data file).

16. Two primary routes of access have been developed by RDC staff: remote access and onsite. In both routes, strict procedures govern the use of the RDC:

- researchers must submit a research proposal;
- no materials may be brought into the RDC;
- no materials, printed or electronic may leave the RDC without a disclosure review;
- researchers must sign a Researcher Affidavit of Confidentiality;
- the RDC is open only when staff are available for supervision;
- use of the RDC is subject to space availability, consistency with the NCHS mission, and the feasibility of the proposed project.

17. Except for very unusual circumstances, researchers are not allowed access to files with direct geographic identifiers. Should a researcher request an NCHS data file merged with external data, RDC staff will merge the files then remove the geographic identifiers leaving the researcher access to a files that consists of the NCHS data merged with the additional data. Should the researcher need clustering variables to stratify on geography, RDC staff will construct a set of dummy geographic indicators.

18. **Remote access** - The remote access system allows researchers to submit analytic computer programs written in the SAS language by e-mail without having direct access to internal NCHS data files. Generally the RDC staff will construct a dummy data file configured exactly like the real data (univariate distributions are the same, variable locations and lengths are the same, and paths are the same) that the researcher can use for developing and debugging programs prior to sending them to the remote access system. The use of the dummy data file results in fewer iterations on the remote access system thus increasing overall efficiency. The remote access system operates entirely automatically, the system scans the e-mail for arriving computer programs, validates the user, scans the program for non-allowable commands (such as those which could result in a case listing), verifies that it is not trying to access unauthorized data files and, if no problems are found, executes the program against the real data. After execution, the system scans the analytic output generated by the users program for disclosure problems. Questionable output is routed to an RDC staff person for manual resolution. Users can submit requests to the remote access system 24 hours a day although output is only returned during normal working hours because staff randomly spot check the system to ensure that the system is working properly in all respects. Generally users receive their output within a few hours of submitting the e-mail.

19. SAS was chosen as the analytic language because it is in wide use and is sufficiently well structured that an automated scanning system could be used. A number of functions available in SAS have been disabled because they are capable of producing unstructured output that can not be readily scanned in an automated system or present an unreasonable risk of disclosure. Disabled functions include PROC TABULATE, PROC IML, PROC PRINT, LIST, and others.

20. The current remote access system operates by e-mail but an internet-based system is under development and testing. The internet-based system offers a more user friendly interface and is capable of improved turn-around time.

21. **Onsite access** - The NCHS RDC has been able to accommodate a large range of users but there remain a set of users that need hands on access to the data. A laboratory has been developed that permits researchers to conduct their work onsite. This laboratory has its own computer system with no connections to any other computer system and has been designed with a number of firewalls and failsafe mechanisms that only allows the onsite researchers access to authorized data. In fact, the system does not contain any data that is not being actively used, archived data files and inactive data files are kept on magnetic tape in a secure room that is accessible only by RDC staff. The user workstations have had the floppy disk drives and parallel ports disabled and all print outs are routed to a single printer in another room that can only be accessed by staff.

22. Through the onsite access laboratory, many different types of research projects can be addressed whether large or small, user-supplied data can be merged with NCHS data (although merging is done by NCHS staff), short-term as well as long-term projects are acceptable and virtually all NCHS data (without direct identifiers) can be made available.

V. CONCLUSIONS

23. The National Center for Health Statistics in the United States has taken a number of new and creative measures to enhance access to type of data that have not until now been available to researchers outside the NCHS. It is our belief that as additional researchers take advantage of this new opportunity, new and exciting discoveries in the health and health status of populations will emerge.
24. The NCHS continues to develop, modify and enhance its research data center to take advantage of new technologies and statistical developments in the fields of data access and release.

References

Producing a Public Use File: A Case Study. Rasinski, Kenneth, Timberlake, Jeffrey, Lee, Lisa, Porras, Javier, National Opinion Research Center and Mulrow, Jeri, Ernst & Young. In press, 1997.

National Health Interview Survey State Data Files: Compromises, Tradeoffs, and Concessions. Horm, John, Zarate, Al and Botman, Steve. Presented at the Public Health Conference on Records and Statistics, Washington, D.C., 1997.

State Data Files from the National Health Interview Survey: Protecting Respondent Confidentiality and Maintaining Analytic Utility. Horm, John, Zarate, Al, Botman, Steven, and Bolesta, Monica. Presented at the Joint Statistical Meetings, Anaheim, Ca. August, 1997.

NCHS Staff Manual on Confidentiality. Department of Health and Human Services, Public Health Service, National Center for Health Statistics, Hyattsville, Maryland. September, 1997.

Producing Public Use Microdata that are Analytically Valid and Confidential. Winkler, William E., Bureau of the Census. In press, 1998.

Statistical Policy Working Paper 22, Report on Statistical Disclosure Limitation Methodology. Prepared by Subcommittee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. May, 1994.

The Use of Statistical Noise to Minimize Disclosure Risk and Preserve Analytic Validity in Public Use Files. Tishkoff, Amanda and Horm, John. In press, 1999.