Procedures and Costs for Use of the Research Data Center

National Center for Health Statistics
Centers for Disease Control and Prevention

Last revised: November 7, 2005

Table of Contents

Purpose

This document provides information about the National Center for Health Statistics' (NCHS) Research Data Center (RDC), including how to submit proposals requesting use of data, mechanisms to access the RDC, requirements, use of outside data sets, costs for using the RDC, and other pertinent topics. The Guidelines pertain to use of data produced by NCHS and non-NCHS entities. The guidelines were revised based on comments provided in response to the November 18, 2004, Federal Register Notice, comments from users of the RDC, and recommendations provided in the 2005 Institute of Medicine (IOM) report, Vaccine Safety Research, Data Access, and Public Trust. These revised guidelines, which are in effect now, reflect NCHS' commitment to improving its operations and assisting users in accessing data from the RDC. They will be reviewed on a regular basis so as to be responsive to changes in the environment that affect confidentiality protections. Please check the NCHS Web site or contact the RDC to determine if modifications have been made.

Background

In order to advance knowledge on the health and well-being of the nation and its health care system, NCHS and other organizational entities in the Department of Health and Human Services release statistical micro data containing health and related variables. These files allow outside researchers and analysts to conduct independent analyses and research. The goal of NCHS is to maximize data release through public use files. Public use files provide the greatest access to NCHS data for the largest number of users. However, any release of data, whether micro data files or the results of statistical analyses, must be consistent with the confidentiality provisions under which the data were collected. In the case of data collected or obtained by NCHS, Section 308(d) of the Public Health Service Act (42 U.S.C. 242m(d)) and the NCHS Staff Manual on Confidentiality do not permit the release of data that are either identified or identifiable to persons outside of NCHS. In order to preserve privacy and confidentiality, details that might identify or facilitate the identification of persons and organizations participating in surveys and data systems are suppressed in published data products. Examples of data elements that might be abridged are geographic identifiers, details of sample design, and variables such as age or income that might exist in other databases.

Despite the wide dissemination of data through publications, CD-ROMs, web release and other mechanisms, the inability to release files with, for instance, lower levels of geography, severely limits the utility of some data for research, policy, and programmatic purposes and sets a boundary on one of the goals of the U.S. Department of Health and Human Services, i.e., to increase our capacity to provide state and local area estimates. In pursuit of this goal and in response to the research community's interest in accessing data that cannot be released using public use products because these data could be used to determine the identity of an individual or establishment (these data are referred to as confidential or restricted), NCHS established the Research Data Center (RDC). The RDC provides a mechanism whereby researchers can access detailed data files in a secure environment, without jeopardizing the confidentiality of respondents. Under clearly defined conditions, the RDC provides restricted access to NCHS data. The RDC also houses, under special agreement, data sets resulting from projects that were not conducted by NCHS. The protected environment of the RDC is an efficient way for the project directors to effectively meet the needs of their data users. **Appendix I** contains a listing of some of the NCHS data sets currently available in the RDC. Please note that the list in Appendix I is for illustration purposes only. Virtually all NCHS survey data files, in varying degrees of detail, can be made available in the RDC. The availability of non-NCHS data sets changes over time.

Special requirements for use of non-NCHS data can be found in **Appendix IV**, Project-Specific Requirements.

Research Data Center (RDC) - Operations

The NCHS RDC, located at the NCHS headquarters in Hyattsville, MD, allows researchers meeting certain qualifications, and under strict supervision, to access confidential statistical micro data files. To qualify, researchers must submit a proposal for review and approval. Researchers can use one of three access methods: (1) Direct on-site access; (2) a remote program submission system through which researchers can submit work to be done in the RDC with the output returned to them

by e-mail; or (3) programming services for outside researchers provided by RDC staff. In all three methods, confidential data files remain in the RDC where access to unit records is restricted, and in the case of on-site access or the use of programming services, output is inspected before it leaves the RDC (this review is done automatically as part of the remote access system). Researchers using the RDC conduct their analyses using specially created project-specific data files. In order for the RDC to create the file needed by a researcher, the researcher's proposal must contain sufficient detail regarding data specifications. Creation of the file can be greatly facilitated if the researcher establishes contact with the RDC prior to or during proposal development to ensure that the needed data are available and that RDC requirements are addressed.

In addition, researchers are expected to have competencies in several specific areas. For example, researchers should have the ability to analyze health outcomes data, the ability to analyze large databases, and the ability to use one or more statistical packages.

1. Guest Researcher (on-site)

As currently designed, the NCHS RDC facility in Hyattsville has four user workstations and a secure room for the RDC printer. In addition, there is office space for the RDC staff and long-term outside researchers.

The RDC computers have no electronic link either to the NCHS network, the CDC-NCHS mainframe, or the Internet. The RDC workstations consist of Pentium III 933 MHz computers running Windows 2000. There is sufficient storage on the workstations and the server for any confidential data. SAS, SUDAAN, Watcom Fortran 77, and Stata are installed on the workstations, and additional programming/analytic languages can be added as needed.

The computers have been configured so that removable media such as floppy disks are inaccessible to users. All print output is routed to a central printer which is monitored by RDC staff while the RDC is open to external researchers. Further, the system's workstations are configured such that researchers are given read-only access to requested data files and can write only onto the local workstation's hard disk. These restrictions ensure that users cannot remove information that has not been subjected to a review for confidentiality.

The researcher submits a research proposal to the RDC and, upon approval, conducts his/her research on site at NCHS in the RDC. RDC staff construct the necessary data files before the guest researcher arrives and ensure that no restricted data leave the facility. Data from virtually all of the NCHS data collection systems may be made available through the RDC. Also available are data from other data collection systems.

Researchers may take the results of their analyses off-site only after disclosure review by NCHS RDC staff. Disclosure review consists of looking for tabular cells less than five, tables with geographic variables in any dimension, models with geographic variables (or variables tantamount to geographic variables) as outcome variables, or case listings. In general, disclosure review is consistent with the guidelines published in the NCHS Staff Manual on Confidentiality (see **Appendix II**, Requirements for the Release of NCHS Micro Data Files).

2. Remote Access

Users are able to electronically submit analytical computer programs using SAS as the programming language. After their proposals are approved, researchers are registered with the RDC remote access system and introduced to the procedures and programming limitations to be followed in accessing data. It was recognized at the outset that the availability of a reliable, economical, and flexible tool for remote data access would require certain limitations and constraints in the design of the remote system. Despite these limitations, however, more than 50 percent of all RDC researchers have used the remote system, either singly or in combination with on-site access.

Researchers send programs to the RDC and receive output by e-mail. As is the case for on-site access, RDC staff prepare the requested data files which may consist of confidential data merged with user data. Both submitted programs and output undergo a programmed disclosure limitation review and are also subject to a manual review. Certain procedures and SAS functions are not allowed (see Appendix III, Disallowed SAS Functions, Statements, and Procedures for a complete list). For example, users cannot use PROC TABULATE or PROC IML, nor are functions allowed that are capable of producing listings of individual cases such as LIST and PRINT. Additionally, functions that may select individual cases are not allowed (R--, FIRST., LAST., and others).

The output is scanned for cells containing fewer than five observations. For each one found, not only is that cell suppressed, but several additional cells will also be suppressed (complementary suppression). Alternatively, the researcher may be asked to revise and resubmit his/her analyses. The job log is also scanned with particular attention to certain types of error conditions that may spawn case listings. Some projects are not suitable for the remote access method. Stewards of the file/s in consultation with RDC staff make this determination.

A second generation remote system, one that will offer much more flexibility and utility than the current one, is being developed. For example, the current system supports SAS only. The new system will also support SUDAAN and STATA. In addition, it will provide a Graphic User Interface (GUI) for language free statistical analysis. But, even then, there will continue to be projects which will not be suitable for the remote access method.

3. RDC Staff-Assisted Research

This is mainly useful for those planning to use statistical software not available for the remote system and who are not able to travel to the RDC facility. Under this method, an approved researcher e-mails a statistical software program to the assigned RDC staff person who runs the program and, after disclosure review, provides the output to the researcher by e-mail. More extensive programming services are also available.

Each of the access methods outlined above has an associated cost which includes equipment and space rental, staff overhead, and setup. The staff overhead and setup include the time and resources necessary for monitoring progress, setting up equipment and datafiles, disclosure limitation review, and file management. Since these impose varying demands on resources, accurate cost estimates cannot be

given without complete knowledge of the proposed research. In general, though, the setup fee is $500 per day of effort (see Costs of Using the RDC, below).

Submission of Research Proposals Using NCHS Data

As stated above, researchers must submit proposals that are detailed enough in their data specifications to permit RDC staff to easily determine what data elements are required. Prospective researchers are encouraged to check with RDC staff prior to writing their proposals to ensure that the data of interest can be made available to them. Researchers should develop their proposals in a way that facilitates the ability of the RDC staff to create the analytic files required for the project. Proposals should be explicit regarding the variables needed as well as any case selection required. Only those data items required to conduct the proposed analyses will be included in the analytic data file and the proposals should address why the requested data are needed for the proposed study. Overly large and complex projects or poorly defined projects will require extensive communication between RDC staff and the researchers proposing the project, and this can cause the process to move slowly. Work to prepare data files can be accomplished most expeditiously if large, complex projects are subdivided into manageable parts and requested data are clearly defined.

Researchers wishing to link data in the RDC with external data should provide the external data to RDC staff in advance of their entry to and use of the RDC (a minimum of 7 days prior to the approved date for access to the RDC).

The RDC expects that all researchers will adhere to established standards and principles for carrying out statistical research and analyses. Researchers must conduct only those analyses which have received approval. Failure to comply will result in cancellation of the research activity and potential disbarment from future research activities in the RDC. In the case where Institutional Review Board (IRB) approval is required to conduct research, RDC staff will notify relevant IRBs of infringements of protocol approvals.

Instructions for developing proposals: The format detailed below pertains specifically to use of NCHS data. Some data files have project specific requirements which can be found in **Appendix IV** (Project-Specific Requirements). For example, this appendix contains information on submitting a research proposal requesting use of data from the Vaccine Safety Datalink (VSD) project. If no project specific requirements are provided for non-NCHS data, the format below is to be used.
The research proposal must contain the following information:
A. Cover letter.
B. Project Title.
C. Abstract: approximately 100-300 words summarizing the project.
D. Full personal identification, institutional affiliation, mailing addresses (including overnight express mail address), phone, and e-mail address. Applicants who are students must append a letter from the department chair or advisor stating that the applicant is a student working under the direction of the department.
E. Dates of proposed tenure at the RDC (or use of the remote access system). Proposals requesting remote access should include an appendix describing the computer and e-mail account that will receive output as well as the security provisions established for them.
F. Source of funding for the proposed project.
G. Background of study.

1. Key study questions or hypotheses.
2. Public health benefits.

H. A summary of the data requirements for the proposed research along with an explanation of why the data are needed for the proposed study.

1. Identification of cases to be included in the analytic file.
2. Identification of variables to be included in the analytic file.
3. Data to be supplied by the researcher and merged with NCHS or other data.
4. A description of why publicly available data are insufficient.

I. Methods for the study.

1. Analytic strategy and statistical methods to be used.
2. Software requirements (currently, SAS, Stata, SUDAAN, LIMDEP, HLM, SPSS, and Watcom Fortran 77 are available in the RDC; other languages can be made available with sufficient lead time).

J. A description of the output that the researcher intends to have reviewed for non-disclosure. This should include table shells, model equations, or test statistics of any output that the researcher plans to remove from the RDC. This will help the reviewers to determine the risk of disclosure and plan for the disclosure review.

K. Appendices.

1. A current resume or Curriculum Vitae for each person who will participate in the research activity. Resumes or CVs must specify nationality.
2. A letter from student applicant's department chair or academic advisor stating that student is working under the direction of the department.
3. A data dictionary: a complete listing of the specific data requested--data system, files, years, cases, variables, matching or linking variables, etc.
4. A data dictionary for researcher-supplied data, if any, to be merged with the confidential data. This includes identifying the source of the data, variable names, variable codes or ranges, file layout, number of records, and restrictions on NCHS use of the data (currently the RDC policy prohibits release of merged data to anyone other than the prospective researcher).
5. A description of the computer and e-mail system to be used to receive output from the remote access system as well as the security provisions established for them.

Portions of doctoral proposals or grant applications with appropriate modifications may suffice for the research proposal.

Proposals to use the Research Data Center should be sent to:
Research Data Center, National Center for Health Statistics, 3311 Toledo Road, Suite 4113, Hyattsville, MD 20782, **RDCA@cdc.gov**.

Upon receipt, the Research Proposal will be evaluated by a review committee convened for that purpose. The committee meets on the third Tuesday of each month, but alternate days can be used when necessary. Researchers are contacted within a week after the review is completed. The Proposal Review Committee consists of (at minimum) the director of the NCHS RDC, the RDC staff liaison, the NCHS Confidentiality Officer, and the director (or designee) of the NCHS data division whose data are requested in the proposal. The division director (or designee) serves as the subject matter expert, while others provide expertise in areas such as confidentiality, statistical methodology, data analysis, etc. Proposals for use of non-NCHS data undergo review as determined by the steward/s of those data.

The following criteria apply to proposal review for projects requesting use of NCHS data and other data sets that do not have project specific requirements.

- Scientific and technical feasibility of the project.
- Availability of resources at the RDC.
- Risk of disclosure of restricted information.
- For projects using NCHS data, whether the proposed project is in accordance with the mission of the NCHS, which is to provide statistical information that will guide actions and policies to improve the health of the American people.

Researchers should note that approval of their application does not constitute endorsement by NCHS of the substantive, methodological, theoretical, or policy relevance or merit of the proposed research. NCHS approval only constitutes a judgment that this research, as described in the application, is not an illegal use of the requested data file and that there is high probability that the project can be successfully done in the RDC.

Researcher-Supplied Data

The RDC allows researchers to supply their own data to be linked with RDC data sets to create merged data sets that will be stored in the RDC. The researcher-supplied data may consist of proprietary data collected and "owned'' by the researcher or other publicly available data obtained by the researcher such as census data. Researchers MUST provide RDC staff with complete documentation of any data proposed to be merged with RDC data. Researchers expecting to use merged files are responsible for interacting with RDC staff to ensure that their data can be merged with the data resident at the RDC and the format of the data is consistent with the RDC data. The RDC will accept user data files in SAS, Stata, or ASCII format (flat files) with variables either column-delimited or column-specific. Other formats may also be proposed. RDC staff, prior to the arrival of the researcher, will do the merging of researcher-supplied data with RDC data sets. Identifying information in linking fields will be removed after the merge and will not be made available to the researchers.

Owners or stewards of RDC data sets make the determination of whether and how the resultant merged files will be made available to other researchers. For RDC files that are owned by NCHS, this determination is made by the owners of the researcher-supplied data that will be merged with the NCHS owned RDC files. For files that are NOT owned by NCHS, the determination is made by the stewards or owners of the RDC files. The owners of these files can require that any merged files be made available to all interested researchers or allow this determination to be made by the owners of the researcher supplied data.

The RDC periodically creates and maintains backup copies of all computer files. Backup files are stored in a secure storage area accessible by RDC staff only, although they may be made available to researchers who need to return for additional analyses. These backup files will contain user-supplied data as well as the merged files. These backup files will be destroyed only upon the written request of the user.

General Requirements for Guest Researchers

1. Researchers must work under the supervision of RDC staff and only during normal working hours (Monday-Friday, 8:30 a.m.- 5:00 p.m.). Admittance to the RDC will be limited to the researchers whose names are included in the Research Proposal (Section D). Researchers will be required to show photo identification before admittance. A maximum of 3 collaborating researchers can sit at a computer station in the RDC.
2. Computers will be pre-loaded with the approved datasets by NCHS staff approximately one day prior to the external researcher's use of the RDC. Once the analysis is completed, NCHS staff will remove the datasets from the RDC computer.
3. Guest researchers must be able to conduct their analyses with the software specified in their research proposal.
4. External researchers are not allowed to bring documents, manuals, books, etc., that may enable them to identify and disclose confidential information they access in the RDC. Neither are they allowed to bring into the RDC cell phones, pagers, or other devices which would enable them to communicate with persons outside of the RDC.
5. All logs will be printed or electronically archived and will be kept by the RDC, which will retain only the programs and procedures run by external researchers. The logs will not include results from their research.
6. All computer output generated by statistical programs and all hand-written notes based on such computer output are subject to disclosure review by RDC staff before removal from the RDC. Output is restricted to summary tables of geographic or patient-level data (e.g., line listings of diagnoses by study identifier will be prohibited).
7. Guest researchers may not save output, files, or programs to transportable electronic media. RDC staff can copy output or programs to transportable media, if requested.
8. Researchers proposing multiple analyses that employ multiple data sets will have access to only one dataset at a time. Under no circumstance will researchers be permitted any opportunity to merge datasets on their own.

General Requirements for Remote Access

1. Researchers must register an e-mail address that is credibly secure. Although programs can be sent to the RDC from any address, results will always be returned to the registered e-mail address.
2. Data requests must be in the form of SAS programs (currently version 8.1; version 9.1 is being tested for future installation). However, certain SAS commands/statements are not allowed through remote access. A list of such commands/statements is included in **Appendix III**. This list is periodically reviewed and may be modified as necessary. The SAS program must be in plain ASCII format.
3. During the first week of registration, researchers' data requests are executed in a manual mode, requiring RDC staff to review the program and resulting output before its release. During this period, remote access is available only during normal working hours. After the first week, researchers may submit data requests any time (day or night) and receive prompt response, except when the CDC e-mail system is down or when the remote access system is taken off-line for maintenance.
4. The remote access system does not allow users to write permanent datasets in its disk space. Jobs that attempt to create permanent datasets or files are flagged, terminated, and an error message is sent to the researcher.
5. The remote access system limits researchers' time and storage. No single

program is allowed more than one hour to complete execution or to generate output in access of 1.5 MB.
6. With one exception, macros are not allowed through the remote access system. The exception, GLIMIX, requires special permission.

Also, see RDC Operations in the front of this document for a description of planned upgrades in the remote access system.

Use of the RDC

In order to get access to restricted data files in the RDC, researchers must include in their proposals a signed "Agreement Regarding Conditions of Access to Confidential Data in the Research Data Center for the National Center for Health Statistics'' (**Appendix V**). All researchers participating on an approved project must sign the agreement--which clearly states the penalties for violating the conditions of agreement. In addition, each researcher must sign an "Affidavit of Confidentiality'' (**Appendix VI**). The RDC reserves the right to terminate any project at any time that it deems that an investigator's actions will compromise confidentiality or ethical standards of behavior in a research environment.

Over the next year, NCHS will be phasing in “designated agent” authority for RDC users of NCHS data when appropriate. This authority (included in the Confidential Information Protection and Statistical Efficiency Act of 2002) permits statistical agencies to designate agents who, under supervision and subject to certain limitations and penalties, can perform exclusively statistical activities on data that cannot be released publicly. Implementation of this authority at NCHS will result in changes to some procedures and materials related to use of the RDC.

Statistical micro data files are collections of data from individual units such as persons or providers. Statistical agencies world wide are bound by ethical and legal requirements to preserve the privacy of individual respondents and the confidentiality of data provided to the agency by them or otherwise pertaining to them. As mentioned earlier, confidentiality protection at NCHS is governed by Section 308(d) of the Public Health Service Act (42 U.S.C. 242m).

This section states that:
No information, if an establishment or person supplying the information or described in it is identifiable, obtained in the course of activities undertaken or supported under section 304, 306, or 307 may be used for any purpose other than the purpose for which it was supplied unless such establishment or person has consented (as determined under regulations of the Secretary) to its use for such other purpose and in the case of information 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 has consented (as determined under regulations of the Secretary) to its publication or release in other form.

Having read and familiarized themselves with the Researcher Affidavit of Confidentiality, including Section 308(d) of the Public Health Service Act (42 U.S.C. 242m) (see below), researchers agree:
1. To make no copies of any files or portions of files to which they are granted access

except those authorized by NCHS Research Data Center staff.
2. To return to RDC staff all NCHS restricted materials that they may be provided during the conduct of their research at NCHS and other materials as requested.
3. Not to use ANY technique in an attempt to learn the identity of any person, establishment, or sampling unit not identified on public use data files.
4. To hold in strictest confidence the identification of any establishment or individual that may be inadvertently revealed in any documents or discussion, or analysis. Such inadvertent identification revealed in their analyses will be immediately brought to the attention of RDC staff.
5. Not to remove any printouts, electronic files, documents, or media until they have been scanned for disclosure risk by RDC staff.
6. Not to remove from NCHS any written notes pertaining to the identification of any establishment, individual, or geographic area that may be revealed in the conduct of their research at NCHS.
7. To the inspection of any material they may bring to or remove from the NCHS Research Data Center.
8. To comport themselves in a manner consistent with principles and standards appropriate to a scientific research establishment.

The "Agreement Regarding Conditions of Access to Confidential Data in the Research Data Center of the National Center for Health Statistics," **Appendix V**, signed by all investigators on the project, must be submitted with the initial proposal. Deliberate violation of any of these conditions may result in cancellation of the data access, and the researcher may be escorted from the premises by the duly authorized Federal protection service on duty at NCHS. The researcher may also be barred from any future use of the RDC upon review and determination by the Director of NCHS that this is necessary to protect the integrity and confidentiality of the RDC.

The RDC technical monitor will perform a disclosure review and must provide approval to the researcher before removal of any data from the RDC, whether it is in electronic or paper form. Any violation by the researcher may be punishable by fine or imprisonment for up to 5 years or both under Title 18 U.S.C. 1001.

As noted above, the RDC contains work stations with computers pre-loaded by NCHS staff with the requested dataset(s) to be analyzed with statistical software. External researchers must schedule time for use of the RDC, pay the appropriate user fees, and abide by the standard practices of the RDC. Among the requirements is a restriction on equipment that can be brought into the RDC, signing agreements to maintain confidentiality, and submitting to review of all results for any potential breaches in confidentiality.

Costs for Using the RDC

As the table below indicates, the cost of using the RDC facilities varies according to whether the researcher conducts his or her analyses on site or through the remote access system. Currently, researchers who are geographically close to the RDC facility in Hyattsville, MD, have a greater opportunity to conduct their research on site. NCHS is investigating the feasibility of opening research data centers in other locations to ameliorate this somewhat. The remote access system and the option of using RDC staff also provide alternatives.

Time to conduct research in the RDC can be scheduled in increments ranging from a consecutive 2-day minimum to a consecutive 10-day maximum. Extensions can be negotiated with RDC staff subject to scheduling requirements. Scheduling time at the RDC is on a first-come, first-served basis.

Researchers using the NCHS RDC will be charged for space and equipment rental and staff time necessary for supervision, disclosure limitation review, maintenance of computer facilities (including both hardware and software), and the creation and maintenance of data files required by the researcher. The cost per project (or creation of an analytic file) is given below:

Guest Researcher (on site)..	$200 per day (2-day minimum).
Remote Access.....	$250 per month for mortality files. For all other files: $500 per month for files with less than 130,000 records. $1,000 per month for files with 130,000 records or more. $500 per year for selected standard files.*

* There are selected files that have been developed for repeat and multiple users which require minimal set up procedures and involve minimal content changes to the file when preparing for different users. For that reason, charges for accessing these files are considerably less expensive than the regular fees. Two files fall under this category: the contextual data file for the National Survey of Family Growth (NSFG-CDF) and the Polio file for the National Health Interview Survey (NHIS-Polio).

There is a minimum setup charge of $500 per day ($250 for the Mortality file) for new file creation. An additional $500 per day is charged as needed for file creation and for special handling, such as the merging of additional data or creating custom file formats.

More complex projects may require discussion between the researcher and RDC staff to determine the cost of file creation. Researchers are encouraged to develop their proposals in a way that facilitates the ability of the RDC staff to create the analytic files required for the project. Proposals should be explicit regarding the variables needed as well as any case selection required. Overly large and complex projects will require extensive communication between RDC staff and the researchers proposing the project, and this can cause the process to move slowly. Work to prepare data files can be accomplished most expeditiously if large, complex projects are subdivided into manageable parts.

Payment is expected in advance of the use of the RDC. A cashier's check or money order made payable to NCHS RDC must be received seven business days prior to the start date scheduled for use of the RDC. Payments should be mailed to:
NCHS RDC
Attn: RDC Director
3311 Toledo Road, Suite 4113
Hyattsville, MD 20782

Disclosure Review Process

The disclosure review process in the RDC is centered on a rigorously conducted research base. Briefly, RDC staff, either independently or in collaboration with staff from other areas of the NCHS, other government agencies, and non-governmental researchers, conduct research into the use of technological and statistical advances to develop and refine additional methods to access restricted data such as the use of the internet or encrypted data, assessment of disclosure risk through statistical and automated procedures, and the use of disclosure limitation methodologies (e.g., statistical noise) to enable the release of otherwise restricted data files. The results of these research activities are applied to disclosure review activities in the RDC.

Researchers may take the results of their analyses off-site after disclosure review by RDC staff. Disclosure review consists of looking for tabular cells less than 5, tables with geographic variables in any dimension, models with geographic variables (or variables tantamount to geographic variables) as outcome variables, or line listings. In general, disclosure review is consistent with the guidelines published in the NCHS Staff Manual on Confidentiality (see **Appendix II**, Requirements for the Release of Micro Data Files). RDC staff review data summaries to assure maintenance of respondent confidentiality. In no case may any table contain cells with fewer than 5 observations. If found, these small cells are suppressed, generally by obliterating the cell. To assure that small cells cannot be calculated from the other cells in the same row or column, staff make illegible the totals for the rows and columns corresponding to the small cell. Once disclosure review is completed, researchers receive a photocopy of the final tabulations. RDC staff, when reviewing cross-tabulations for small cells, use the following procedures:

1. Shred all tables having fewer than five total observations (table total).
2. Shred all tables having fewer than five observations in each cell.
3. If the table passes the first two criteria, RDC staff will review the table one row at a time.
4. Make illegible all counts and percents for cells with four or fewer observations.
5. If one row cell is < 5, that cell and at least one other row cell will be suppressed; if two or more row cells are each < 5, each will be suppressed, but the row total need not be suppressed because the suppressed row cells cannot be determined.
6. If one column cell is < 5, that cell and at least one other column cell will be suppressed; if two or more column cells are each < 5, each will be suppressed, but the column total need not be suppressed because the suppressed column cells cannot be determined.
7. Row (or column) total is suppressed ONLY if it (i.e., total) is < 5; since the cells that are < 5 (row or column as appropriate) are suppressed, user cannot determine their values by knowing the row (or column) total.

RDC staff will use best practices in determining whether data are identifiable and will be conservative in their decisions. RDC decisions are final and not subject to negotiation by researchers.

Publication

For NCHS files, any published material derived from the data should acknowledge NCHS as the source and should include a disclaimer that credits any analyses, interpretations, or conclusions reached by the author (recipient of the file) to that author and not to NCHS, which is responsible only for the initial data. Researchers

who want to publish a technical description of the data should make a reasonable effort to ensure that the description is consistent with that published by NCHS.

Appendix I--Examples of NCHS Data Available Through the NCHS RDC

The goal of NCHS is to make as much of its data as possible available through public use files. There are times, however, when data files are more restrictive and can be made available only through the RDC. The following files represent a small sample of the types of data available in the RDC. Virtually all NCHS survey and vital statistics data files, in varying degrees of detail and including, for example, sample design information and lower levels of geography, can be made available as requested and needed in the RDC.

National Health Interview Survey
Data from the core and supplements for survey years 1987-2002 are available for merging user-supplied data at the state and county levels (note that RDC users do not have access to county FIPS codes; these are replaced with randomly assigned dummy codes). Additionally, state data files may be made available for analysis and reporting.

National Survey of Family Growth--Contextual Data File
The 1995 NSFG has available sets of contextual variables at the state, county, census tract, and block-group levels for the residence of the respondents in 1990, 1993, and 1995

Third National Health and Nutrition Examination Survey (1988-1994
Data from NHANES III are available with state and county identifiers (there are restrictions on the use and reporting of geographic units).

Mortality Follow-ups for the NHIS and NHANES

The 1986-2000 NHIS have been linked to mortality data through 2002; NHANES III (BASELINE 1988-1994) has been linked to deaths through 2000.

Appendix II--Requirements for the Release of NCHS Micro Data Files

The following rules apply to all files released by NCHS which contain any information about individual persons or establishments, except where the supplier of information was told, prior to his giving the information, that the information would be made public:

A. Before any new or revised micro data files are published, they, together with their full documentation, must be approved for publication by the Confidentiality Officer who will rely upon assistance from the NCHS Disclosure Review Board in reaching decisions.

B. The file must not contain any detailed information about the subject that could facilitate identification and that is not essential for research purposes (e.g., exact date of the subject's birth, excessive detail for occupation, extreme values of income and age, detailed race or ethnicity for small and highly visible groups--and other characteristics that would make an individual or establishment easier to identify). It is recommended that the following be consulted concerning possible techniques that would permit the maximum amount of information to be released consistent with sound principles of statistical disclosure limitation: The Confidentiality and Data Access Committee's Checklist on Disclosure Potential of Data (**http://www.fcsm.gov/committees/cdac/checklist_799.doc**) and Statistical Policy Working Paper 22, Report on Statistical Disclosure Limitation Methodology. Office of Information and Regulatory Affairs, Office of Management and Budget (**http://www.fcsm.gov/working-papers/wp22.html)**.

C. Geographic places that have fewer than 100,000 people are not to be identified on the file. Depending upon the statistical structure of a file and other circumstances, a higher figure may be employed. It is the responsibility of the program proposing the data release to determine the disclosure risk associated with the proposed minimum size of geographic areas to be identified.

D. Characteristics of an area are not to appear on the file if they would uniquely identify an area of fewer than 100,000 people (e.g., a variable describing the size of a Metropolitan Area in which a respondent was interviewed providing for a category of fewer than 100,000 in a file where Region is also provided).

E. Information on the drawing of the sample which might assist in identifying a respondent must not be released outside the Center. Thus, the identities of primary sampling units are not to be made available outside the Research Data Center except in limited circumstances and as approved by the Confidentiality Officer. When such circumstances require the disclosure of the identity of areas in which data collection activities take place, the survey manager must insure that all information for this survey proposed for release takes into account the greater risk of identification because of this exception. The decision as to whether PSU identities are to be made public should be made before data are collected and plans for data release finalized.

Appendix III--Disallowed SAS Functions, Statements, and Procedures

The list below is used by the RDC remote access system to scan user-submitted programs for functions, statements, and procedures that may result in an unauthorized disclosure. Any user-submitted program that contains one or more of these keywords is automatically rejected, and the user is asked to correct the problem and resubmit the program. Because the remote access system is an automated system, the RDC does not and cannot make any exceptions. This list may change pending development of additional methodologies.

r--word
add
print
obs
firstobs
first.
last.
&
%
nocol
report
pctn
pctsum
tabulate
iml
nofreq
nocum
browse
editor
summary
list
put
file
r--
plot
PROC DATASET:
-Copy
-Delete
-Rename
-Repair
-Append
-List
Compress
Pointobs
multi part data set names

In addition to the above disallowed statements and functions, users of the remote access system cannot use any statements or functions that write permanent data files to the hard disk.

Appendix IV--Project-Specific Requirements

Vaccine Safety Datalink (VSD) Data Sharing Program

Background

The VSD was established to allow the Centers for Disease Control and Prevention (CDC) to carefully monitor vaccine safety in the United States. The VSD, a large, linked database, contains medical and immunization information on more than 5.5 million people annually. Information available in the database includes basic demographic information, managed care organization (MCO) enrollment, dates of vaccination, and medical visits. The VSD is a collaborative project involving CDC and several large MCOs. Information from the VSD is used by CDC to conduct vaccine safety studies.

Recognition of the need for improved monitoring of vaccine safety prompted the CDC to initiate the VSD project in 1990. This project currently involves partnerships with MCOs to continually monitor vaccine safety. All vaccines administered within a MCO are recorded and include vaccine type, date of vaccination, concurrent vaccinations (those given during the same visit), the manufacturer, lot number and injection site. Medical visits are also recorded which can be used to monitor for potential adverse events resulting from immunization. The VSD project allows for planned vaccine safety studies as well as timely investigations of emerging hypotheses. At present, the VSD project is examining potential associations between vaccines and a number of serious conditions. Data from the VSD also are used to test new vaccine safety hypotheses that result from the medical literature, signals from the Vaccine Adverse Events Reporting System (VAERS), changes in the immunization schedule, the introduction of new vaccines, or recommendations from the Institute of Medicine (IOM) and Advisory Committee on Immunization Practices (ACIP) recommendations. This project is a powerful and cost-effective tool for the on-going evaluation of vaccine safety. It should be noted that the MCOs have broad decision-making authority over data release as well as a recognized need and right to protect proprietary data.

In August 2002, CDC's National Immunization Program (NIP) and its managed care partners created a Data Sharing Program to allow limited access to VSD data through the NCHS RDC with confidentiality protection under Sec. 308(d) of the Public Health Service Act (42 U.S.C. 242m). Proposals requesting use of VSD data through the Data Sharing Program undergo a review by the MCOs' Institutional Review Board(s) (MCO IRB) in addition to a review by RDC staff. After approval of their research proposal and payment of fees for the associated costs, researchers are able to independently analyze VSD data through the VSD Data Sharing Program.

Data Sharing Program

There are two types of VSD data which can be accessed at the RDC under the VSD Data Sharing Program:

- a relational database containing data through December 31, 2000 that were obtained from the detailed administrative and medical records held by the MCOs; and

- final datasets that were used for published studies from August 2002 and beyond.

Data from the VSD project collected after December 31, 2000 are not available through the RDC VSD Data Sharing Program. VSD data beyond 2001 can be accessed through a formal collaboration with an MCO and the external researcher must work through MCO procedures. It should be noted that collaboration is at the discretion of the MCO. Such collaboration would be outside the scope of the VSD Data Sharing Program and, therefore, data would not be accessed at the RDC. CDC cannot guarantee external investigators' ability to gain access to the VSD data at the MCOs.

Access to VSD data through the VSD Data Sharing Program

1. The relational database:

The VSD data are comprised of several separate data files derived from computerized data sources from seven participating VSD MCOs. The VSD data files that are available through the VSD Data Sharing Program contain data through December 31, 2000, and include information such as vaccinations, hospital discharge and other diagnoses, and demographic characteristics. With these data, an external researcher may conduct a new vaccine safety study in order to test his/her vaccine safety hypothesis. The external researcher may request only the variables that are found in the VSD data files (as listed in the data dictionary).

To assist researchers, CDC makes available on its website: (1) A list of recommended scientific references relevant to conducting research using large linked databases such as the VSD data files; and (2) a data dictionary that lists all the variables contained in the VSD data files available for new vaccine safety research (**http://www.cdc.gov/nip/vacsafe/vsd/default.htm#data**).

Proposals to conduct new vaccine safety studies using data from the relational VSD database should include only those specific variables that are needed to conduct the proposed analyses, including a brief explanation with justification for use of these variables.

Data collected for the VSD project and that can be accessed through the VSD Data Sharing Program have been created from MCO administrative data which were not solely collected for the purpose of scientific research. It should be noted that the quality of the data from the relational VSD database cannot be guaranteed. Potential data discrepancies and varying degrees of data quality that are specific to such types of data do exist and typically are not resolvable with data that are available in the RDC.

2. Final datasets from published VSD studies:

External researchers who would like to perform a reanalysis of a published VSD study performed by VSD investigators may request the final dataset for the specific study they wish to re-analyze. Data collected for the final datasets of the published studies may include additional variables not listed in the data dictionary that is referenced above; therefore, the RDC will provide the external researcher with the

necessary data dictionary for the requested dataset(s). No additional source or "raw'' data, or earlier versions of the final data set, are available for reanalysis of published VSD studies.

In general, VSD studies published after August 2002 (irrespective of data year) are available for reanalysis. Many studies were published prior to the establishment of the VSD Data Sharing Program; some of the earlier published VSD study datasets may not be available for re-analysis for the following reasons:

- Some IRBs mandate that datasets be destroyed after research is completed.
- The principal investigator may no longer be affiliated with VSD or the collaborating MCOs; therefore, the location of the dataset is unknown.
- Rapidly changing technology can mean that data are on obsolete media.

Documentation for variables and datasets used in VSD studies completed after August 2002 are maintained according to the CDC data sharing policy regarding archival of data that are available on the Web at
http://www.cdc.gov/od/ads/pol-385.htm

Following receipt of a proposal for a reanalysis, the RDC will verify that the data variables requested from the published study are available. If these data are not available (for one or more of the reasons stated above), the RDC will notify the external researcher.

Requirement for Proposals

All proposals requesting use of VSD data through the VSD Data Sharing Program should contain the following information:
A. Project Title.
B. Name of proposed investigator and collaborators (RDC rules limit number of persons at a work station to 3 at a time).
C. Name of point of contact, address, telephone number, and e-mail address.
D. Summary of proposed study (i.e., background, reasons for conducting the study, public health benefits).
E. Specific hypothesis for new vaccine safety studies to be investigated or title of published VSD study to be reanalyzed.
F. Proposed methodology for new vaccine safety studies or the specification of the methods used in published VSD studies:

1. Definition of the study population of interest and type of study to be conducted:
   a. Descriptive studies: specify the variables and values for those variables to be used to select the study population.
   b. Case-control studies: specify criteria for cases and controls.
   c. Cohort studies: specify criteria for the exposed and unexposed population.
   d. For all new vaccine safety studies, please include the following information as part of the definition of the study population of interest:
      i. Adult or Pediatric data (0-17 or 18+).
      ii. Study years of interest (i.e. 199X-2000). Please note the study years available vary by MCO site.
      iii. How the study population will be selected from the VSD data files based on available fields in the VSD data dictionary.
2. Specification of the variables that will be required including:
   a. Exposures: Specific criteria defining exposures based on the VSD data

dictionary should be included. For instance, specific vaccines given within 14 days of the outcome of interest.

b. Outcomes: Specific criteria defining those outcomes based on the VSD data dictionary should be included. For instance, specific ICD-9 codes for outcomes of interest and type of health care encounter (hospitalization, outpatient encounter, emergency room visit).

c. Person Time or Enrollment: Specify criteria to determine calculation of person time, follow-up time, or MCO enrollment restrictions.

d. Confounding or control variables, including:

i. Demographic information.

ii. Pre-existing or co-morbid conditions.

iii. Concurrent vaccinations.

iv. MCO Site.

e. Other required variables to perform the proposed analysis.

G. Proposed analytic strategies and statistical methods to be used including Software requirements (currently, SAS, Stata, SUDAAN, LIMDEP, HLM, SPSS, and Watcom Fortran 77 are available in the RDC; other languages can be made available with sufficient lead time) and a description of the output that the researcher intends to have reviewed for non-disclosure (table shells, model equations, or test statistics of any output that the researcher plans to remove from the RDC.) This will help the reviewers to determine the risk of disclosure and plan for the disclosure review.

Submission of Proposals

All proposals to access VSD data through the VSD Data Sharing Program at the RDC will be evaluated using the same evaluation criteria used by NCHS for other RDC studies. These evaluation criteria include:

- Scientific and technical feasibility of the project.
- Availability of resources at the RDC.
- Risk of disclosure of restricted information.

Technical feasibility is the primary evaluation criterion for the review of proposals submitted to the VSD Data Sharing Program. All proposals are reviewed to determine if the requested data are available within the VSD database.

After completing the review, the RDC staff will notify the external researcher whether his/her proposal meets the evaluation criteria and whether the requested variables are available. If all the requested data variables can be located for the proposed new vaccine safety study or proposed reanalysis, review of the proposal by the appropriate MCO IRBs takes place. In compliance with federal law and regulations, access by external researchers to a portion of the VSD data files or to datasets from VSD published studies requires review and approval by the appropriate IRBs of the relevant MCOs.

IRB applications may require a more detailed description of the proposed vaccine safety study and may vary according to individual IRB requirements. Furthermore, various IRBs may have different time lines for submission of proposals for review. Each IRB may have specific policies or requirements for data sharing that have not been adopted by the other MCO IRBs. These policies may include required collaboration with an MCO investigator, fees associated with the IRB review process, or differing criteria for the IRB review process.

The MCO IRBs have the responsibility to protect the confidentiality and privacy of their members' medical records and to adhere to the rules and regulations applicable to their respective institution(s). Consequently, each of the MCO IRBs must review any request for access to the VSD data files that contain information on its MCO members. Any appeal by the requestor of an IRB decision must follow the national, federal procedures for IRBs. CDC is not involved in the MCO IRB process at any time. General information pertaining to the rules and regulations of IRB submission can be found at **http://www.cdc.gov/od/ads/hsr2.htm/**

MCO IRBs will use their established procedures and time lines to review the proposed research and to consider any appeals. As a rule, IRBs attempt to inform researchers as to the status of their proposals. Approval for access to MCO data contained within the VSD data files does not indicate approval for obtaining additional data contained within the MCO's member medical records or elsewhere, if such data are not contained within the VSD data files that reside in the RDC.

For new vaccine safety studies, it is possible that an external researcher will receive approval for access to VSD data from some, but not all, relevant IRBs. If this occurs, then the dataset(s) needed to conduct the new vaccine safety study will still be created, but only with data from the MCOs whose IRBs approved access. VSD data sets for new vaccine safety studies must contain data from two or more MCOs' data. Access will not be provided to data from only one MCO. For reanalysis of a published VSD study, all relevant IRBs from the MCOs that participated in the published study must approve the proposal for reanalysis; therefore if one or more IRBs do not approve access to VSD data used in the published study, the final dataset cannot be provided.

Once the external researcher has received a response from all of the appropriate IRBs, the RDC will begin the process of creating or formatting the approved dataset(s). The RDC will not create or prepare the dataset(s) until it receives copies of all final IRB dispositions along with other responses directly from the IRBs.

In addition, each proposed investigator must submit a signed copy of the Agreement Regarding Conditions of Access to Confidential Data in the Research Data Center of the National Center for Health Statistics (Appendix V) and the Researcher Affidavit of Confidentiality (Appendix VI).

Following receipt of final IRB dispositions, RDC staff will arrange for access to the RDC as described in the general data sharing document. All rules, procedures, and fees as outlined in the general data sharing document will apply.

Research Findings

Researchers may provide status reports on studies involving VSD data. Such reports should be submitted to the Director of the RDC and can address the type of study being conducted, the results obtained to date, and planned further activities.

When an external researcher has completed his/her work at the RDC and wishes to publish research results and findings using VSD data, there are specific requirements that must be followed:

- A copy of the manuscript must be submitted to NCHS at least 30 days prior to submission to a journal or other print or electronic media.
- External researchers are required to submit a copy of these data sharing guidelines with any manuscript submitted to a journal.
- External researchers are required to submit (to the journal) a copy of the Confidentiality Agreements he/she signed prior to conducting research at the RDC
- Disclaimers must be included in the manuscript which state: The research was conducted using data from the Vaccine Safety Datalink Project, through the VSD Data Sharing Program at NCHS/CDC.
- Any published material using VSD data must acknowledge CDC as the original data source.
- Additionally, disclaimers must be included that state: The analysis, interpretations, and conclusions are the responsibility of the authors and do not represent the views and opinions of the CDC, the Federal Government, or the managed care organizations providing the data.

Appendix V--Agreement Regarding Conditions of Access to Confidential Data in the Research Data Center of the National Center for Health Statistics

I _____ (print name) am aware that the information contained in the _____ (name of data file) has been provided to NCHS in accordance with the provisions of Section 308(d) of the Public Health Service Act (42 U.S.C. 242m), with the assurance that it will be used only for health statistical reporting and analysis and will not be published or released in identifiable form. I am also aware that I can be held legally liable for any harm resulting from my activities at the RDC incurred by individuals or establishments who have provided or are described in the information contained in the above work files to which I will have access.

Having read and familiarized myself with the Researcher Affidavit of Confidentiality, including Section 308(d) of the Public Health Service Act (42 U.S.C. 242m) (attached), I agree:

1. To make no copies of any files or portions of files to which I am granted access except those authorized by NCHS Research Data Center staff.
2. To return to RDC staff all NCHS restricted materials with which I may be provided during the conduct of my research at NCHS and other materials as requested.
3. Not to use ANY technique in an attempt to learn the identity of any person, establishment, or sampling unit not identified on public use data files.
4. To hold in strictest confidence the identification of any establishment or individual that may be inadvertently revealed in any documents or discussion, or analysis. Such inadvertent identification revealed in my analysis will be immediately brought to the attention of RDC staff.
5. Not to remove any printouts, electronic files, documents, or media until they have been scanned for disclosure risk by RDC staff.
6. Not to remove from NCHS any written notes pertaining to the identification of any establishment, individual, or geographic area that may be revealed in the conduct of my research at NCHS.
7. To the inspection of any material I may bring to or remove from the NCHS Research Data Center.
8. To comport myself in a manner consistent with the principles and standards appropriate to a scientific research establishment.

Deliberate violation of any of these conditions may result in cancellation of the data access agreement, and the researcher may be escorted from the premises by the duly authorized Federal protection service on duty at NCHS. The researcher may also be barred from any future use of the RDC upon review and determination by the Director of NCHS that this is necessary to protect the integrity and confidentiality of the RDC.

_____

Researcher's Signature Date

_____

NCHS Witness Date

Appendix VI--Researcher Affidavit of Confidentiality

I certify that no confidential data or information viewed or otherwise obtained while I am a researcher in the National Center for Health Statistics (NCHS) Research Data Center (RDC) will be removed from NCHS. Further, I understand that NCHS will perform a disclosure review and must provide approval to me before I remove any data from the RDC, whether they are in electronic or paper form. I acknowledge NCHS Confidentiality Statute, Sec. 308(d) of the Public Health Service Act (42 U.S.C. 242m) stated below and fully understand my legal obligations to NCHS to protect all confidential data. Further, I understand that any violation may be punishable by fine or imprisonment for up to 5 years or both under Title 18 U.S.C. 1001.

NCHS Confidentiality Statute--No information, if an establishment or person supplying the information or described in it is identifiable, obtained in the course of activities undertaken or supported under section 304, 306, or 307 may be used for any purpose other than the purpose for which it was supplied unless such establishment or person has consented (as determined under regulations of the Secretary) to its use for such other purpose and in the case of information 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 has consented (as determined under regulations of the Secretary) to its publication or release in other form.

Title 18 U.S.C. 1001--Deliberately making a false statement in any matter within the jurisdiction of any Department or Agency of the Federal Government violates Title 18 U.S.C. 1001 and is punishable by a fine or up to 5 years in prison or both.

_____
Researcher's Signature

_____
Date

_____
NCHS Witness

_____
Date