National Center for Health Statistics
RESEARCH DATA CENTER
Guidelines for Proposal Submission

Last revised: June 6, 2008

Background

NCHS releases a range of statistical data products on the health and well-being of the nation and its health care system. Statistical tabulations (tables) present data in predetermined categories such as by age, race, sex or geographic region that are important to describe health status and trends. In addition, statistical microdata containing health and related variables are published so that outside researchers and analysts may also conduct original research and special studies to address issues of public health science and policy. However, Section 308 (d) of the Public Health Service Act 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 NCHS surveys and data systems are suppressed in published data products. Examples of data elements that might be abridged or suppressed are geographic identifiers, genetic data, details of sample design, and variables like age or income that might exist in other databases.

Despite the wide dissemination of NCHS data through publications, web releases, etc., the inability to release files with these sensitive variables limits the utility of NCHS data for research, policy, and programmatic purposes and set a boundary on one of the Department's goals, i.e., to increase our capacity to provide state and local area estimates. In pursuit of this goal and in response to the public research community's interest in restricted data, NCHS established the NCHS Research Data Center (RDC), a place where non-NCHS researchers can access detailed data files in a secure environment, without jeopardizing the confidentiality of respondents. The RDC provides restricted access to NCHS data and some non-NCHS data. Special requirements for use of non-NCHS data can be found in Appendix III, Project-Specific Requirements.

Research Data Center (RDC)

The NCHS RDC is a secure research facility located at NCHS headquarters in Hyattsville, MD, where researchers meeting certain criteria are allowed access, under strict supervision, to restricted statistical microdata files. To gain access to the RDC, researchers must submit a proposal for review and approval.

Any NCHS data can be made accessible through the RDC. The National Health Interview Survey, Linked Mortality Files, the National Survey of Family Growth, and the Pre-2000 Vaccine Safety Datalink are stored on RDC computers while all other NCHS data systems provide in-house data on an as needed basis. Whether a data set is housed in the RDC or it is provided to the RDC upon the project's approval is the prerogative of the Director of that data system.

As currently designed, the NCHS RDC has six user workstations. In addition, there is office space for the RDC staff and NCHS Guest Researchers. The RDC user workstations are "stand alone" and have no link to the NCHS network, the CDC-NCHS mainframe, or the internet. There is sufficient storage on the workstations and the server for any confidential data. PC-SAS, SUDAAN, and STATA are installed on the workstations and additional programming/analytic languages can be added as needed.

Drives on the workstations for removable media such as USB ports are configured so as to be inaccessible to users. The workstations are configured such that users 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.

Three Methods to Access Data

1. **NCHS Guest Researcher** – Researchers conduct their research on-site at NCHS in the RDC. NCHS RDC staff constructs the necessary data files before the guest researcher arrives and ensures 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 other data sets, such as the pre-2001 Vaccine Safety Data. Prospective users should check with NCHS RDC staff prior to writing their proposals to ensure that the data of interest can be made available to them. Researchers are able to take the results of their analyses off-site only after disclosure review by an NCHS RDC Staff Analyst.

2. **Remote Access** – Through remote access researchers are able to electronically submit analytical computer programs using SAS and SUDAAN as the programming languages. 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. Then RDC staff prepares the requested data files which may consist of confidential data merged with user data. Researchers send programs to the NCHS RDC and receive output back by e-mail. Their programs execute on a computer in the RDC. Both submitted programs and output are subjected to a programmed disclosure limitation review and could also be subjected to a manual review. Certain procedures and SAS functions are not allowed (see Appendix II, 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 which may select individual cases are not allowed (R_, FIRST., LAST., and others). The output is scanned for cells containing less than five observations. If any are found, not only is that cell suppressed, but several additional cells will also be suppressed (complementary suppression). 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. This determination is made by stewards of the file(s) in consultation with RDC staff.

3. **Census Guest Researcher** – Researchers can have the same access that is available to them at NCHS at one of the nine Census RDCs. Analytic data sets are constructed at the NCHS RDC according to specifications included in the research proposal and are then securely transferred to the Census data processing facility in Bowie, Maryland. Users can then view the data using "front end dumb terminals" at a Census RDC. The data do not leave the Bowie facility. The researchers output is sent to the NCHS Hyattsville RDC for disclosure review and then sent to the researcher via email after it has been approved. A listing of available locations can be found here: http://webserver02.ces.census.gov/index.php/ces/researchlocations

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 III. 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 (summary)**.**
    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).
    3. A brief explanation of how you will be using sample weights and accounting for the complex sample design if applicable.
- **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.

Question or completed proposals to use the Research Data Center should be sent electronically to: Pmeyer1@cdc.gov

Proposal Review

Upon receipt, the Research Proposal will be evaluated by a review committee convened for that purpose. The Proposal Review Committee consists of (at minimum) the Director of the NCHS RDC, the RDC Staff Analyst, the NCHS Confidentiality Officer, and a representative of the data producing program, often the director (or designee) of the NCHS data division whose data are requested in the proposal. Proposals for use of non-NCHS data undergo review as determined by the owner(s) of those data. The review takes 6-8 weeks.

The following criteria apply to proposal review:

1. Appropriate use of the data and concurrence with the intended use for which it was collected.
2. Scientific and technical feasibility of the project.
3. Availability of resources at the RDC.
4. Risk of disclosure of restricted information.
5. Assurance that the use of the data is in accordance with the informed consent procedures associated with the collection of the data.
6. For projects using NCHS data, whether the proposed project is in accordance with the mission of the NCHS "...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. Rather, NCHS approval constitutes a judgment that this research, as described in the application, is not an illegal or unethical use (as determined by the informed consent and original reason for collecting the data)of the requested data file and does not jeopardize confidentiality of the data.

Researcher-Supplied Data

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 their RDC Staff Analyst with complete documentation of any data proposed to be merged with NCHS data. The NCHS RDC requires that researchers provide all proprietary or public use files that will be used in their analytic data set seven days prior to the approved start date of the analysis. Researchers expecting to use merged files are responsible for interacting with RDC staff to ensure that their data can be merged with the NCHS data and the format of the data is consistent with the NCHS data. The NCHS

RDC will accept researcher data files in SAS, STATA, or ASCII format (flat files) with variables either column delimited or column specific. Other formats may also be proposed. The merging of researcher-supplied data with NCHS in-house data will be done by NCHS RDC staff prior to the arrival of the researcher. Identifying information in linking fields will be removed after the merge and will not be made available to researchers. Current RDC policy is to treat researchers as the owners of the researcher supplied data and the dataset(s) created for them. The RDC may retain copies of datasets at the request of the user but they will not be made available to anyone other than the owner without the owner's written permission. These retained copies of analytic data sets will be stored on a single use computer until the file has been inactive for a year.

General Rules and Regulations for Guest Researchers

1. Researchers may work at the RDC only under supervision of RDC staff and only during normal working hours (Monday-Friday, 9:00am-5:00pm). Admittance to the RDC will be limited to the researchers included in the Research Proposal (Sections D and E). 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 guest 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 analysis with the software specified in their Research Proposal.
4. Guest 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 cell phones, pagers, or other devices into the RDC which would enable them to communicate with persons outside of the RDC.
5. All logs will be electronically archived and will be kept by NCHS. NCHS will retain only the programs and procedures run by guest researchers. The logs will not include researchers' results.
6. All computer output generated by statistical programs and all hand-written notes based on such computer output are subject to disclosure review by NCHS staff before removal from the RDC.
7. Guest researchers may not save output, files, or programs to transportable electronic media. RDC staff may copy output or programs to transportable media, if requested.
8. Researchers' analytic data set will be specified thoroughly in the research proposal. The analytic data set for a project may include multiple cycles of a survey or variables from multiple sources. However, researchers may not access multiple data sets during a project that have not been merged together by an RDC staff member. Under no circumstance will researchers be permitted any opportunity to merge datasets on their own.

General Rules and Regulations for Remote Access

1. Researchers must register an email 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 (Version 9.1). However, certain SAS commands/statements are not allowed through remote access. A list of such commands/statements is included as Appendix II. This list is subject to periodic review and may be modified. The SAS program must be in plain ASCII format.
3. During the first week of registration, user 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 get prompt response, except when the CDC email 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 excess of 1.5 MB.
6. With one exception, macros are not allowed through the remote access system. The exception, GLIMIX, requires special permission.
7. Researchers should contact their RDC Staff Analyst immediately if they have inadvertently produced output that could be used to identify subjects/respondents or if they cannot complete their analysis due to automated disclosure protocols. The Staff Analyst will provide reasonable assistance in completing the analysis while still protecting confidentiality.

Legal and Ethical Issues

In order to get access to restricted data files in the RDC, researchers must sign an NCHS Designated Agent Agreement (Appendix IV), 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). This includes any member of the research team that works directly with the data. NCHS reserves the right to terminate any project at any time that it deems that an investigator's actions will compromise confidentiality, the ethical standards of behavior in a research environment, and/or protocols developed by NCHS to protect the data itself. The researcher may also be barred from future use of the RDC.

As mentioned earlier, confidentiality protection at NCHS is governed by Section 308(d) of the Public Health Service Act, PHSA, (42 USC 242m). Specifically,

> "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 Sections 304, 305, 306, 307, or 309 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 (1) 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 Designated Agent Agreement and understanding the legal framework under which NCHS operates, including Section 308(d) of the Public Health Service Act, 308d (for all data) and the Confidential Information Protection and Statistical Efficiency Act (for all data collected since January 1, 2003), the 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. Not to use any technique to circumvent suppression algorithms or other disclosure minimization protocols developed by the RDC even if the intent is not to re-identify study subjects or respondents.
3. To return to RDC staff all NCHS restricted materials with which they may be provided during the conduct of their research at NCHS and other materials as requested.
4. Not to use <u>any</u> technique in an attempt to learn the identity of any person, establishment, or sampling unit not identified on public use data files.

5. 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.
6. Not to remove any printouts, electronic files, documents, or media until they have been scanned for disclosure risk by RDC staff.
7. 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.
8. To the inspection of any material they may bring to or remove from the NCHS RDC.
9. To submit to NCHS RDC staff for disclosure limitation review any papers or reports submitted for publication.
10. To comport themselves in a manner consistent with principles and standards appropriate to a scientific research establishment.

Any violation the researcher may perform that is punishable under Confidential Information Protection and Statistical Efficiency Act of 2002 carries a fine of up to $250,000 and up to 5 years in prison.

The NCHS 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 received approval. Failure to comply with RDC rules and regulations will result in cancellation of the research activity and potential disbarment from future research activities in the RDC. In the case where Ethics Review Board (ERB) approval is required to conduct research, NCHS will notify relevant ERBs of infringements of protocol approvals.

Disclosure Review Process

Researchers are able to take any results of their analyses off-site that have passed disclosure review by NCHS RDC staff. Disclosure review can include 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 case listings. In general, disclosure review is consistent with the guidelines published in the NCHS Staff Manual on Confidentiality.

RDC staff review data summaries to assure maintenance of respondent confidentiality. Tables containing cells with fewer than 5 observations may not be released to the data user. These cells will be suppressed. If a researcher requires output of an intermediary nature that contains counts of less than five and believes that the release would not compromise confidentiality, they should contact their assigned RDC Staff Analyst or the Director. To assure that small cells cannot be calculated from the other cells in the same row or column, the totals for the rows and columns containing the small cell are also suppressed. Once disclosure review is completed, researchers receive electronic copies of the final tabulations.

Disclosure Risk Analysis

Output generated through RDC access mechanisms will be subject to a review that will include, but not be limited, to the following procedures:

1. In no table should all cases of any line or column be found in a single cell.
2. In no case should the total figure for a line or column of a cross-tabulation be less than 5. One acceptable way to solve the problem is to use a statistical disclosure limitation technique such as rounding.
3. In no case should a quantity figure be based upon fewer than five cases.
4. In no case should a quantity figure be released to the researcher if one case contributes more than 60 percent of the amount.
5. In no case should data on an identifiable case, nor any of the kinds of data listed in preceding items

A-D, be derivable through subtraction or other calculation from the combination of output on a given study.
6. Data will not be disclosed when used in combination with other known data.
7. Low level geography will not be included in output provided to the researchers.

The reviews will all be performed at the NCHS Hyattsville RDC by a Staff Analyst who is trained statistics and statistical disclosure limitation. Currently, the personnel authorized to perform disclosure risk analyses at the RDC are:

- Negassi Beyene, M.A., M.Sc., Statistician
- Karen Davis, MA, Mathematical Statistician
- Robert Krasowski, M.A., M.S., Statistician
- Peter Meyer, M.A., M.P.H., Director
- Christopher Rogers, Ph.D., Statistician
- Deborah Rose, Ph.D., M.P.H., Epidemiologist

Costs for Using the RDC

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:

File creation fees

- **New file creation**
  There is a minimum setup charge of $500 per day ($250 for the Linked Mortality files). 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 (e.g., those employing the VSD files) may require discussion between the researcher and RDC staff to determine the cost of file creation.

On site fees

- **Daily programming costs**
  $200 per day (consecutive 2-day minimum and 10-day maximum, with extensions negotiated subject to scheduling requirements). Time on-site in the RDC can be scheduled in daily increments but the minimum reservation is 2 consecutive days. Scheduling time at the RDC is on a first-come, first-served basis.

Remote access fees

- **Linked Mortality files**
  $250 per month for a single survey cycle
  $125 per month for each additional survey cycle
- **National Survey of Family Growth Contextual Data File (NSFG-CDF)**
  $500 per year
- **Polio file for the National Health Interview Survey (NHIS-Polio)**
  $500 per year

- **For all other files:**
  $500 per month for a single survey wave file
  $250 per month for each additional survey cycle

Payment is expected in advance of the use of the RDC. A check, money order, or Interagency Agreement payable to DHHS Statistical Services must be received 7 business days prior to the scheduled start date of use of the RDC. Payments should be mailed to:

Research Data Center
Attn: Peter Meyer
National Center for Health Statistics
3311 Toledo Road, Suite 4113
Hyattsville, MD 20782

Appendix I – Examples of Data Available Through the NCHS RDC

National Health Interview Survey – Data from the core and supplements for survey years 1987-2002 are available for merging researcher-supplied data at the state or 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 1995, 1993, and 1990.

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

NCHS Restricted-use Linked Data Resources- NCHS has developed a record linkage program designed to maximize the scientific value of the Center's population-based surveys. NCHS has linked various NCHS surveys with death certificate records from the National Death Index (NDI), Medicare enrollment and claims data from the Centers for Medicare and Medicaid Services (CMS), and Retirement, Survivor, and Disability Insurance (RSDI) and Supplemental Security Income (SSI) benefit data from the Social Security Administration (SSA). These linked data files can be merged with their respective NCHS public-use survey file in the RDC for analysts to examine factors that influence disability, chronic disease, health care utilization, morbidity, and mortality.

- **NCHS Linked Mortality Data:** Linkage of the NCHS survey participants with the NDI provides the opportunity to investigate the association of a wide variety of health factors with mortality. Additional information can be found at: http://www.cdc.gov/nchs/r&d/nchs_datalinkage/data_linkage_mortality.htm .
- **NCHS Linked Medicare Enrollment Data:** CMS provided NCHS with Medicare benefit claims data for 1991 through 2000 for all successfully matched NCHS survey participants. These linked survey files provide the opportunity for analysts to answer vital research questions by profiling Medicare service use and assessing health care costs. Additional information can be found at: http://www.cdc.gov/nchs/r&d/nchs_datalinkage/data_linkage_cms.htm .
- **NCHS Linked Social Security Benefit History Data:** SSA provided NCHS with benefit history data from 1962 through December 2003 for RSDI recipients and data from 1974-2003 for SSI recipients. The linked survey files create an analytic tool for the evaluation of the needs of the elderly and other persons with disabilities. Additional information can be found at: http://www.cdc.gov/nchs/r&d/nchs_datalinkage/data_linkage_ssa.htm .

Restricted-use Linked Data is available for the following NCHS surveys:

- **1994-1998 National Health Interview Survey (NHIS)**
- **NHANES I Epidemiologic Follow-up Study (NHEFS)**
- **Second National Health and Nutrition Examination Survey (NHANES II)**
- **Third National Health and Nutrition Examination Survey (NHANES III)**
- **The Second Longitudinal Study of Aging (LSOA II)**
- **1985 National Nursing Home Survey (NNHS)**

The Linked Mortality data for NHANES III, NHIS (1986-2000), and LSO2 are also available in a public-use format that has been perturbed to protect confidentiality.

Centers for Disease Control and Prevention/Immunization Safety Office Vaccine Datalink Project- The Vaccine Safety Datalink (VSD) Project is a collaborative project involving the Centers for Disease Control and Prevention (CDC) and several large managed care organizations (MCOs). The VSD was established primarily to assess vaccine safety in the United States through analyses of large-linked databases collected at the MCOs as part of their routine administration of health services. Additional background information regarding the VSD can be found at http://www.cdc.gov/od/science/iso/research_activties/vsdp.htm

In 2002, the VSD established a data sharing program to allow external researchers to conduct new studies of vaccine safety or to reanalyze study-specific datasets from published VSD studies. Access to approved VSD data will be via the Research Data Center at NCHS. As noted above, the process for reviewing proposals for access to VSD data differs slightly from that for access to NCHS data in that the MCO Institutional Review Boards (IRBs) conduct reviews in accordance with their standard procedures. See Attachment IV for more details pertaining to VSD proposals. Inquiries about VSD proposals may be sent to the VSD email box at: pmeyer1@cdc.gov.

In short, any and all NCHS data can be accessed through the RDC if the proposal is approved. NCHS does not approve or disapprove RDC proposals based on the scientific merit of the project. Whether or not a proposal is approved depends on the impact of the project in terms of confidentiality, technical feasibility, whether the proposed use of the data is consistent with the purpose for which the data were collected, and consistent with the original consent. This determination is made by the Proposal Review committee and is final.

Appendix II – 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, exceptions cannot be made. This list may change as additional methodologies are developed.

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
options
sleep

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 III – Project-Specific Requirements

Vaccine Safety Datalink (VSD) Project

The Vaccine Safety Datalink (VSD) Project is a collaborative project involving the Centers for Disease Control and Prevention (CDC) and several large managed care organizations (MCOs). The VSD was established primarily to assess vaccine safety issues in the United States through analyses of large-linked databases collected at the MCOs as part of their routine administration of health services. Additional background information regarding the VSD can be found at http://www.cdc.gov/od/science/iso/vsd/vsdatashraing.htm

In 2002, the VSD established a data sharing program to allow external researchers to conduct new studies of vaccine safety from the VSD data files available at CDC or to reanalyze study-specific datasets from published VSD studies. The VSD data sharing program is a three-step process:

1. Submission of proposals to CDC's RDC at NCHS
2. Submission of proposals to MCO Institutional Review Boards (IRBs)
3. Use of CDC's RDC at NCHS

Access to VSD data through the VSD Data Sharing Program:

1. New vaccine safety studies:

   To conduct his/her own vaccine safety study, an external researcher may use VSD data files that reside at the RDC. Requests for data for examining new hypotheses are limited to only the variables found in the VSD data files (as listed in the data dictionary). **Please note**: data for new vaccine safety studies are available only through December 31, 2000. Therefore, the RDC will not accept proposals requesting VSD data after December 31, 2000 for new vaccine safety studies. Access to MCO data after December 31, 2000 for new vaccine safety studies may be accessed through establishing a formal collaboration with the MCO(s); however, such collaboration is at the discretion of the MCO and outside the scope of the RDC data sharing program and CDC authority.

   To assist researchers CDC makes available at its website a list of recommended scientific references relevant to conducting research using large linked databases such as the VSD data files and a data dictionary that lists all the variables in the VSD data files available for new vaccine safety research. http://www.cdc.gov/od/science/iso/research_activties/vsdatasharing.htm. Consistent with the Health Insurance Portability and Accountability Act (HIPAA) regulations, proposals for new vaccine safety studies 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 accessible through the data sharing program have been created from MCO administrative data and were not solely collected for the purpose of scientific research. It should be noted that the quality of the data from the VSD for new vaccine safety study cannot be guaranteed. Because there are potential data discrepancies and the quality of the data vary, the data that have been collected by the MCOs and still under their control are typically not resolvable with the data that are available at the RDC.

   2. Re-analyses of published VSD studies:
   External researchers who would like to perform a reanalysis of a 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 dataset are available for reanalysis of published VSD studies.

In general, VSD studies published after August 2002 are available for re-analysis. However, since many studies were published prior to the establishment of the CDC data sharing policy, some of the earlier published VSD study datasets may not be available for re-analysis for the following reasons:

- The principal investigator may no longer be affiliated with VSD or the collaborating MCOs; therefore; the location of the dataset is unknown
- Some IRBs mandate that datasets be destroyed after research is completed
- 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. Refer to the CDC data sharing policy on the web at: http://www.cdc.gov/od/foia/policies/sharing.htm

Following receipt of a proposal for a re-analysis, CDC verifies that the data variables requested from the published study are available. If these data are not available (for the reasons stated above), CDC will notify the external researcher.

CDC will notify the external researcher whether his/her proposal is complete and if the requested variables are available. If all the requested data variables can be located for the new vaccine safety studies or proposed re-analyses, review of the proposal by the appropriate MCO IRB(s) takes place. In compliance with federal 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.

Requirements of Proposals:
All proposals requesting use of VSD data 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)
C. Name of point of contact, address, telephone number, and email address
D. Summary of proposed study (i.e., background, reasons for conducting the study, public health benefits)
E. Specific hypothesis of vaccine safety study to be investigated or title of published VSD study to be reanalyzed
F. Proposed methodology for new vaccine safety study to be investigated or the specification of methods used in published study.
   Abstracts of published studies will not be accepted as proposed methodology from the external researcher.
   1. Definition of the study population of interest and type of study to be conducted
      a. For all new vaccine safety studies, please include the following information as part of the definition of the study population of interest:
         - Adult or Child data (0-17 or 18+)
         - Study years of interest (i.e. 199X-2000). Please note the study years

available vary by MCO site.
- How will the study population be selected from the data files based on available fields in the VSD data dictionary?

b. Descriptive studies: specify the variables and values for those variables to be used to select the study population
c. Case-control studies: specify criteria for cases and controls
d. Cohort studies: specify criteria for exposed and unexposed population

2. Specification of the variables that will be required including:
   a. Exposures: specify the exposures to be studied. 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 of interest: specify the outcomes to be studied. Specific criteria defining those outcomes based on the VSD data dictionary should be included. For instance, 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 the calculation of person-timer, follow-up time, or MCO enrollment restrictions
   d. Confounding or control variables, including
      - Demographic information
      - Pre-existing or co-morbid conditions
      - Concurrent vaccinations
      - MCO site

3. Detailed proposed analytic strategies and statistical methods. VSD data are in SAS format**.** 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 the 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 criteria include:

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

The RDC will review proposals to ensure that the proposal is complete and then determine whether the proposed project is both scientifically and technically feasible. Determination of scientific and technical feasibility include whether data are available at the RDC for new vaccine safety studies or if requested final dataset is available for reanalysis. If these data are not available (for the reasons stated above), CDC will notify the external researcher and the proposal will be need to be resubmitted.

After completing the review, RDC staff will notify the external researcher whether his/her proposal meeting the evaluation criteria. If all the requested data variables can be located for the new vaccine safety studies or proposed re-analyses, review of the proposal by the appropriate MCO IRBs takes place. At least two of the MCO IRBs must approve the proposal for the file to be accessed. In compliance with federal 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.

Review of a proposal submitted by an external researcher by a MCO IRB does not imply that CDC approves or endorses the external researcher's proposed research. 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.

MCO IRBs have the responsibility to protect the confidentiality and privacy of their members' medical records and to adhere to the rules and regulations of 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 included in the MCO IRB process at any times. 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. 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 NCHS RDC.

For new vaccine safety studies, it is possible that an external researcher may receive approval for access to VSD data from some but not all relevant IRBs. If this occurs, then the datasets(s) needed to conduct the new vaccine safety study will still be created, but only with data from the MCOs whose IRBs approved access. For re-analysis of a published VSD study, all relevant IRBs from the MCOs that participated in the published study must approve the proposal for re-analysis; 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, NCHS RDC will begin the process of creating or formatting the approved dataset(s) NCHS 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 Research Data Center of the National Center for Health Statistics and the Researcher Affidavit of Confidentiality.

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.

Publication of Research Using VSD Data

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

- 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 Agreement

he/she signed prior to conducting research at the RDC.

- Disclaimers must be included in the manuscript which state that "the research was conducted using data from the Vaccine Safety Datalink Project, through the data sharing program at the Centers for Disease Control and Prevention." 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 MCOs."

Appendix IV:

Designated Agent – NCHS Research Data Center (RDC)

To be completed by the NCHS staff member and the researcher

As required by the Privacy Act of 1974, the personal information being requested will be kept confidential and will be used only for the purpose of identifying a researcher who may be granted designated agent status. Providing the information is strictly voluntary; however, not providing it will prevent you from being considered for agent status.

PART A

Name (last, first, middle): _____

Date of Birth (month, day, year): _____

Social Security Number: _____

Citizen of the United States: Yes: ___ No: ___

If not, citizen of which country: _____

Local home address
(street, city, state, zip code): _____

_____

Legal address, if different
(street, city, state, zip code): _____

_____

Telephone numbers. Home: _____

Cell: _____

Designated Agent's employer: _____

Name of supervisor: _____

Work address: _____

_____

Work Telephone numbers: Agent: _____

Supervisor: _____

PART B
Reason for presence at NCHS – Title of Project, as listed in the approved project proposal, and up to 5 keywords that describe it.

Title: _____

Keywords: _____

Time period agent expects to work in the NCHS RDC:

From (month/day/year): ____/ ____/ 20____ To (month/day/year): ____ / ____ / 20____

PART C

Affidavit of Non-Disclosure

I, (name) _____, do solemnly swear (or affirm) I will observe all policies and procedures to protect the confidentiality of data to which I will have access in the RDC as set forth in the attached NCHS Research Data Center Procedures and Costs for Use of the Research Data Center and that I will not disclose confidential information, either while an agent or after, contained in data files, lists, or reports created using National Center for Health Statistics data, as specified under section 308 (d) of the Public Health Service Act and under penalties* set forth in §513 of the Confidential Information Protection and Statistical Efficiency Act of 2002 (PL 107-347, title V), 44 USC 3501 note.

Signature of Designated Agent: _____

Subscribed and sworn (or affirmed) before me this ____ day of _____, 20___.

At (city) _____ (state) _____

[SEAL] (Notary Public Signature) _____

My commission expires: _____ Title (Officer/Notary Public): _____

RDC employee supervising the Designated Agent:

Printed Name: _____ Signature: _____

NCHS Confidentiality Officer:

Alvan O. Zarate Signature: _____

Note: The oath of non-disclosure must be administered by a person specified in 5 U.S.C. §2903. The word "swear," wherever it appears above, should be stricken out when the appointee elects to affirm rather than swear to the affidavit; only these words may be stricken, and only when the appointee elects to affirm the affidavit.

*Whoever, being an officer, employee, or agent of an agency acquiring information for exclusively statistical purposes, having taken and subscribed the oath of office, or having sworn to observe the limitations imposed by section 512, comes into possession of such information by reason of his or her being an officer, employee, or agent and, knowing that the disclosure of the specific information is prohibited under the provisions of this title, willfully discloses the information in any manner to a person or agency not entitled to receive it, shall be guilty of a **class E felony** and **imprisoned for not more than 5 years**, or fined not more than **$250,000**, or both.

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 (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 Signature: _____ Date: _____

NCHS Witness Signature: _____ 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 Signature: _____ Date: _____

NCHS Witness Signature: _____ Date: _____