
DEPARTMENT OF HEALTH AND HUMAN SERVICES

Centers for Disease Control and Prevention

Procedures and Costs for Service at the Research Data Center

Last revised March 5, 2013

AGENCY: National Center for Health Statistics, Centers for Disease Control and Prevention (CDC), Department of Health and Human Services (HHS).

ACTION: Notice and request for comments.

AUTHORITY: Section 306 of the Public Health Service Act, as amended (42 U.S.C. 242k) and Pub. Law 103-333.

SUMMARY: This notice provides information about the Research Data Center (RDC) operated by the National Center for Health Statistics (NCHS) within the Centers for Disease Control and Prevention (CDC). The Research Data Center was established in 1998 to provide a mechanism whereby researchers can access detailed data files in a secure environment, without jeopardizing the confidentiality of respondents.

Historically, the data files accessed in the RDC have consisted of NCHS survey data and vital statistics. RDC has recently begun accepting data files that were not produced from NCHS survey data. In order to assure that all data files are processed in a consistent manner, the original guidelines for accessing files in the RDC are being reviewed and revised as necessary.

Last updated March 7, 2013

As part of the revision process, potential users are being given the opportunity to provide input on how the procedures of the RDC can best serve their research needs. This notice describes how to submit proposals requesting use of the data, mechanisms to access the RDC, requirements, use of outside data sets, costs for using the RDC, and other pertinent topics. We are seeking comments on these procedures and will post the final procedures on the NCHS Web site.

ADDRESSES: Send comments concerning this notice to Peter Meyer, National Center for Health Statistics, 3311 Toledo Road, Room 4113, Hyattsville, MD 20782, or e-mail to pmeyer1@cdc.gov

FOR FURTHER INFORMATION CONTACT: Peter Meyer 301-458-4375.

SUPPLEMENTARY INFORMATION:

Operational Procedures for Use of the Research Data Center; National Center for Health Statistics; Centers for Disease Control and Prevention

Table of Contents

Background

Methods of Access to Data

Submission of Research Proposals Using NCHS Data

Proposal Review

Researcher Supplied Data

General Procedures for Onsite Access

General Procedures for Remote Access

Confidentiality and Human Subjects Protection

Disclosure Review Process

Costs for Using the RDC

National Center for Health Statistics Research Data Center Procedures

Background

The National Center for Health Statistics (NCHS) releases and hosts 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 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 analysts may conduct original research and special studies to address issues of public health science and policy. 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 entities participating in NCHS surveys and data systems either owned or hosted by NCHS are not released in published data products. Examples of data elements that might be abridged or suppressed to prevent re-identification are geographic identifiers, genetic data, details of sample design, and variables such as 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 sets a boundary on one of the Department of Health and Human Service's goals: 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 Centers (RDCs), a place where researchers can access detailed data files in a secure environment, without jeopardizing the confidentiality of respondents. Access is regulated by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) and other Federal statutes. The RDCs provide restricted access to NCHS data and external data. Researchers function under the supervision of NCHS employees and are subject to the same provisions of law with regard to confidentiality as NCHS employees. Instructions for developing a research proposal can be found in Appendix II. Special requirements for use of external data can be found in Appendix III, Project-Specific Requirements.

Methods to Access Data

Restricted NCHS data or data hosted by NCHS can be made accessible through the RDC. To gain access to these data, researchers must submit a proposal for review and approval. Once the proposal is approved,

Researchers meeting certain criteria are allowed access, under strict supervision, to restricted statistical microdata file(s). There are four modes of access: 1) NCHS RDC, 2) Remote Access System, 3) Census RDC, 4) Staff-Assisted.

1. NCHS RDC - Researchers conduct their research on-site at one of the NCHS RDCs. The NCHS RDCs are secure research facilities located at NCHS headquarters in Hyattsville, MD and at the Centers for Disease Control and Prevention in Atlanta, GA, where researchers meeting certain criteria are allowed access, under strict supervision, to restricted statistical microdata file(s). The NCHS RDC 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. PC-SAS, SUDAAN, and STATA are installed on the workstations as well as additional programming/analytic languages. Drives on the workstations for removable media such as USB ports are configured so as to be inaccessible to users. Therefore, researchers must provide programs or external data to the RDC before each visit. 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 drive. These restrictions ensure that users cannot remove information that has not been subjected to a review for confidentiality. Researchers are emailed the results of their analyses only after it has undergone disclosure review by an NCHS RDC Analyst.

2. Remote Access - Through remote access researchers are able to electronically submit analytical computer programs using SAS and SUDAAN. After the proposal is approved, researchers are registered with the RDC remote access system and are required to accept the procedures and programming limitations to be followed in accessing data. 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). Researchers send programs to, and receive output from, the remote system through a secure communication network. Their programs execute on a computer in the RDC. Both submitted programs and output are subjected to a programmed disclosure review and may also be subjected to a manual review. For example, 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 .log file 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. Researchers should consider the programming limitations of the remote access system when choosing this method of access. However, the data stewards and RDC staff may also deem the project inappropriate for remote access during the review process.

3. Census RDC - Researchers have the same access that is available to them at NCHS at one of the 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. Users then view the data using "front end dumb terminals" at a Census RDC. The data do not leave the processing facility. The researcher's output is sent via a secure communication network to RDC staff for disclosure review. Once the output has been approved for release, it is sent via email to the researcher. A listing of available Census RDC locations can be found here:

<http://webserver02.ces.census.gov/index.php/ces/researchlocations>

4. RDC Staff-Assisted Research - This option is for researchers planning to use statistical software programming languages other than SAS or who are not able to travel to the RDC facility. Under this method, an approved researcher e-mails statistical programs to the assigned RDC Analyst who runs the program and, after disclosure review, provides the output to the researcher via a secure communication network. More extensive programming services are also available. This option is subject to the availability of RDC Analyst.

Each access modes has an associated cost which offset equipment, space rental, and staff overhead. The staff overhead includes the time and resources necessary for creating the analytical file, monitoring progress, setting up equipment and data files, disclosure limitation review, and file management. Since these reflect varying demands on resources, accurate cost estimates cannot be given without complete knowledge of the proposed research.

Submission of Research Proposals Using NCHS Data

To access restricted data through the RDC, researchers must first submit a proposal. The proposal serves four main purposes:

1. To ensure that researchers have a defined research question.
2. To determine restricted variables needed to complete the project.
3. To assess disclosure risk based on the types of output and requested restricted variables.
4. To determine the mode of access and the required software.

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 by 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. The RDC allows researchers to supply external sources of data to be merged with RDC data. These external sources of data supplied by the researcher may consist of proprietary data collected and "owned" by the researcher or other publicly available data obtained by the researcher such as Census data.

Proposal Review

Upon receipt, the RDC Director will assign the proposal to an RDC Analyst who will review the proposal for completeness and feasibility. Then the RDC Analyst will distribute the proposal to the other members of the Review Committee which consists of (at minimum) the Director of the NCHS RDC (or his designee), the RDC Analyst, the NCHS Confidentiality Officer, and a representative of the data producing program.

The following criteria apply to proposal review:

1. Risk of disclosure of restricted information.
2. Appropriate use of the data and concurrence with the intended use for which it was collected. Including assurance that the use of the data is in accordance with the informed consent procedures associated with the collection of the data.
3. Scientific and technical feasibility of the project.
4. Availability of resources at the RDCs.

The review usually takes 6-8 weeks. The exact amount of time is dependent on a number of factors including the complexity of the proposal and availability of the review committee. The Review Committee can make one of three decisions: approve, resubmit, or disapprove. 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 the confidentiality of the data. Approval of a proposal does not explicitly or implicitly guarantee that all output generated by the analysis will be released. Output that poses a disclosure risk will be suppressed.

Public Data

The researcher may supply two types of data: 1) publically available NCHS data and 2) external sources of data. Researchers must supply these data in advance. The RDC Analyst 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 an NCHS RDC Analyst prior to the arrival of the researcher. Merging variables may or may not be removed from the final analytical data set. For instance, if state and county are used to add Census variables to an NCHS data set, state and county will be removed after the merge unless otherwise specified.

Many projects will require researchers to download public files from the internet and create an extract that includes only the variables required for the project. There are a few exceptions that the RDC Analyst will discuss as needed with the researcher.

Key points:

- The public-use file can only include those variables required for analysis. Do not send the entire public-use files.
- Original NCHS Variables must have the name they are given in the public-use data set. Researcher will have the opportunity to rename once they have been granted access to the dataset.
- Public-Use Mortality Variables: Do not include any public-use mortality variables or variables derived from the public-use mortality data if the project involves restricted mortality variables.

- Any attempt to include variables that may lead to re-identification of subjects or establishments is considered a disclosure violation and will result in the cessation of your project and possible legal actions.

The external data may consist of proprietary data collected and owned by the researcher or other publicly available data obtained such as Census data. Researchers are responsible for working with RDC Analyst to ensure that the data can be merged with the NCHS data and the format of the data is consistent with it. Researchers are also responsible for ensuring that the data they provided has been consented for merging.

General Procedures for Onsite Access

1. Researchers may work at the NCHS RDCs only under supervision of RDC staff and only during normal working hours (Monday-Friday, 9:00am-5:00pm). Admittance to the RDC is limited to the researchers included in the Research Proposal. Researchers are required to show photo identification before admittance. A maximum of 3 collaborating. Researchers can sit at a computer station in the RDC.
2. 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. Under no circumstance will researchers be permitted any opportunity to merge datasets on their own.
3. Computers will be pre-loaded with the approved datasets by NCHS staff approximately one day prior to the researcher's use of the RDC. Once the analysis is completed, the data and programs will be available for one year after the project has been completed or expired.
4. Researchers must be able to conduct their analysis with the software specified in their research proposal.
5. 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.
6. All computer output generated by statistical programs and all handwritten notes based on such computer output are subject to disclosure review by NCHS staff before removal from the RDC.
7. Researchers may not save output, files, or programs to transportable electronic media.

General Procedures for Remote Access

1. Researchers must register an email address that is credibly secure.
2. Data requests must be in the form of SAS programs. However, certain SAS commands/statements are not allowed through remote access.
3. The remote access system does not allow users to write permanent datasets. Jobs that attempt to create permanent datasets or files are flagged, terminated, and an error message is sent to the researcher.
4. 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 5 MB.
5. Researchers should contact their RDC 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 RDC Analyst will provide reasonable assistance in completing the analysis while still protecting confidentiality.

Confidentiality and Human Subjects Protection

In order to access restricted data files in the RDC, researchers must sign an NCHS Designated Agent Agreement (Appendix IV and the Agreement Regarding Conditions of Access to Confidential Data in the Research Data Center of the National Center for Health Statistics (Appendix V.. All members of the research team that work directly with the data must sign these forms. NCHS reserves the right to terminate any project at any time if it deems that an investigator's actions may 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 RDCs.

As mentioned earlier, confidentiality protection at NCHS is governed by Section 308(d) of the Public Health Service Act, PHSA, and (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, edited, linked, merged, transformed, or manipulated at NCHS in any way 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 any 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 Analyst 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 willful disclosure of confidential statistical information by the researcher is punishable under CIPSEA and carries a fine of up to \$250,000 and up to 5 years in prison.

The NCHS RDCs expect 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 RDCs. 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

All output will undergo disclosure review by an RDC Analyst and/or the remote access system. 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 researchers require 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 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.

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. Low level geography will not be included in output provided to the researchers.

The reviews will all be performed by an NCHS RDC Analyst who is trained in statistics and statistical disclosure limitation. For more information consult the Report on Statistical Disclosure Limitation Methodology:
<http://www.fcs.m.gov/working-papers/wp22.html>

Service Costs for Using the RDC

Researchers using the NCHS RDCs 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:

Set-up

- New file creation there is a minimum setup charge of \$750 per day. An additional \$750 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.

On site

- Daily programming costs \$300 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.

Staff- Assisted

- \$750 per day

Remote

- \$750 per month

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