Minimizing Disclosure Risks when Using the NHATS-CMS Standard Linked Files in MedRIC's HaAD Enclave

Access to NHATS-CMS files is limited to users that meet requirements of NHATS and the National Institute on Aging (NIA). The linked data are accessed through the Health and Aging Data (HaAD) enclave managed by MedRIC. Details are available on the NHATS website. Questions about the process should be directed to NHATS Staff at Johns Hopkins University: nhats-cms-data@jh.edu.

This document reviews steps that NHATS takes to minimize the risks of identification (or "disclosure") of respondents or the areas in which they live for researchers are using NHATS-CMS files. Specifically, NHATS limits geographic data in the HaAD enclave and places limits on the files that are allowed in and out of the enclave. We also have a series of requirements to review materials that researchers want to take out of the enclave. We refer to this review process as "vetting." Vetting rules for the NHATS-CMS Standard Files are also provided in the appendix table. (Because there may be additional risks of disclosure on the NHATS-CMS Provider Files, a separate document provides vetting rules for those files.)

NOTE: Requests to bring files in or out of the HaAD enclave should go to the NHATS compliance officer (CO) at JHU using the HaAD communications portal.

A. Limits on using NHATS-CMS files with geographic data. Hospital Referral Region (HRR), Census division, and metro/non-metro area are available in the HaAD enclave for use with NHATS linked CMS files. At this time, NHATS allows only very limited use of additional geographic information (available through the NHATS Restricted Data Repository) with CMS files (e.g. limited contextual variables at one geographic level e.g. at the state, county, or tract level). If you would like to use contextual variables with NHATS-CMS data should review "Instructions to Link Geographic-based Contextual Variables to NHATS-CMS Linked Files," available on the NHATS website. Please email nhats-restricted-data@umich.edu and nhats-cms-data@jh.edu early in the application process for approval. Additional justification will be required.

B. Limits on Files Allowed Into the Enclave. All materials that are requested for import into the enclave will be reviewed for disclosure risks by NHATS. The following files are allowed, pending review:

a. Non-identifiable data sets listed on your NHATS restricted data application File Request Form and described in your research plan; and
b. Statistical code (e.g. .do or .sas files) and other supporting documents. You must include a description of the file and justification of need.

Procedure:

- Access the Communications Portal
- In the Files tab for the "To Enclave" topic, click Library.
- At the top of the Files table, click the Upload Files button.
- Click the Choose File button. Select the file(s) you want to transfer into the Enclave.
- Enter a description of the file's contents in the Description field.
- Click Upload.

The NHATS Compliance Officer will review your list of approved files to make sure the file is listed* and then vet the import file(s) for content. Once approved, the requested file(s) will be moved to the requester's folder. Once the process is complete, the researcher will receive notification via email.

*If the file is not listed, you will be asked to update your list of approved files and research plan.

C. Limits on Files Allowed Out of the Enclave. Prior to removal, all files will be reviewed for disclosure risk by NHATS (see appendix for specific rules).

Files ALLOWED out of the enclave, pending review:

a. Well-labelled tables and graphs that meet vetting requirements described below. Files will be prioritized and typically reviewed within 5 business days, if possible.
b. Statistical code (e.g. .do or .sas files) without notes about sample sizes.
c. Only in exceptional circumstances will log/output files be reviewed or allowed to be removed from the enclave. Such files will take lower priority and may take longer than average to review. Strong justification is needed for such requests.

Files NOT allowed out of the enclave:

a. Microdata files
b. Geographic visualizations (maps)
c. Visualizations that show individual observations (sequence analysis, scatter plots)

Procedure:

- Create a well-labelled table or graph from your output. Include a description of the sample and the sample size for the table (unweighted).
    - For descriptive tables, provide the unweighted sample size for each cell.
    - For models, provide the frequency of the underlying variables in the models.
    - For statistical code (.do or .sas files), review to make sure none of the notes in the file refer to sample size or include any restricted output.
- Before requesting review by NHATS Repository staff, researchers should self-vet their own output according to guidelines in the appendix.
- Once self-vetting is complete:
    - In your project's Desktop Session or Stat Interactive Applications Session, double-click the Transfer Directory icon. Double-click the File Auditing and Security Tool application.
    - In the Home page, click HaAD {YourProjectID} File Transfer to CP (Communications Portal). Click the New Transfer button.
    - Provide a reason for requesting to transfer the file and upload one or more files to be transferred.
    - Attest to the transfer. Click the submit button.
- A reviewer will vet the files to be exported and may also review your research plan to ensure the exported file(s) is in compliance.
- Once approved, the requested file(s) can be downloaded from the Communications Portal.

NHATS Staff, in consultation with the NHATS PIs, have the final decision over whether a given set of results may be exported. More details on Rules that NHATS uses for vetting is found in the appendix tables.

D. Limits on Sharing/Publishing The Fact that a Cell Size is <11.

As a general rule, researchers should not be reporting in publications or presentations or otherwise sharing that a cell size for a particular group is <11. Researchers will be asked to collapse cells if a cell size is <11.

E. Vetting Requirements

NHATS takes steps to protect both the identity of participants (including NSOC participants) and of places where NHATS draws its sample. Users may not remove any tables that could potentially identify either directly or indirectly an NHATS/NSOC respondent, NHATS sampling information, or NHATS/NSOC geographic areas below the level of Census Division.

We therefore have adopted the following rules:

- Tabulations with cells/strata < 11, including minimum and maximum values (ranges) for variables may not be removed.
- Additional cells must be suppressed if they may lead to uncovering cells/strata size <11 through subtraction.
- Tabulations and/or visual representations or coefficients based on geographic areas below the Census Division may not be removed from the enclave.
- Categorical contextual variables (e.g. HRR characteristics in categories) have additional limitations. Researchers must make sure there are at least 3 places (in this example HRRs) in each category. This rule may require additional tabulations of approved contextual data.
- Files with hidden information (Excel files with hidden rows and columns; SPSS "spv" files) may not be removed

Users should self-vet their own output before requesting review by NHATS Repository staff.

NHATS Repository staff, in consultation with the NHATS PIs, have the final decision over whether a given set of results may be exported. More details on Rules that NHATS uses for vetting is found in the appendix tables.

APPENDIX: NHATS VETTING RULES FOR NHATS-CMS FILES

General Rules		
Rule	**if this…**	**then…**
G1. Only analyses at the individual level are allowed	The unit of analysis must be individuals. Other samples, such geographic units (e.g. tracts, counties, states, HRRs) or facilities (hospitals, nursing homes, assisted living), are not allowed to be the unit of analysis.	Inform researcher that an analysis will not be released
G2. Researcher must provide understandable output in tabular or graphic form	Output is not clear (you cannot tell what researcher has done) or is a log file	Return to researcher for clarification
G3. Program files must not contain notes about sample sizes	Program files include a note about number of cases (e.g. dropped n=8 cases; kept n=99 cases)	Return to research for editing
G4. No hidden information allowed in files	Excel files with hidden rows or columns; SPSS "spv" files	Return and request PDF of file
Descriptive Tables (Frequencies)		
Rule	**if this…**	**then…**
T1. Table title must make clear the sample and type of results presented	Sample / type of results unclear	Return to researcher to make title clearer
T2. All table rows and columns must have understandable labels	Label unclear or uses variable names	Return to researcher to make row/column labels clearer.
T3. Minimums and maximums (e.g. top and bottom of ranges for variables) must have at least 11 cases if shown	Minimums or maximums do not include the number of cases at the value	Return to the researcher to add the number of cases
T3. All table cell counts (unweighted) must be shown	Cell counts not shown	Return to researcher to add cell counts
T4. All categories of a variable must be included (cannot drop a category because it is too small; should be collapsed)	Table does not include all categories	Return to researcher for complete table with collapsed cells
T5. Table cells should sum to total sample	Table cells do not sum to totals	Return to researcher for complete table
T6. All table cell counts must be at least 11	Table cell has less than 11 cases	Return to researcher to combine cells
T7. For means of binary variables, numerators and denominators must be shown	Mean for binary variables shown without numerators and denominators	Ask for table with numerators and denominators
T8. For means of binary variables, numerators must be at least 11 (and sample n minus numerator must be at least 11)	Numerator for binary variable is less than 11 cases (or sample n minus numerator is less than 11 cases)	Return to researcher to combine cells
T9. Redaction of a single cell, column or row is not acceptable; omitted group is known to be small by omission	Single cell, column or row redacted	Return to researcher to combine rows and columns until all cells have at least 11 cases
T10. Empty cells as the result of sampling are not allowed	Cell with zero frequency that is possible	Return to researcher to combine row and columns
T11. Text may not report which specific cells are masked because of small sample size	Text says which specific cells were masked	Inform researcher that text is not allowed. Return to researcher to

		combine rows and columns until all cells have at least 11 cases. Do not name specific cell that is redacted.
T12. Specific geographic areas (e.g. states, county, tract, HRR) and types of places (e.g. hospital, assisted living facility, nursing home) may not be listed in tables	Tables have disallowed variables	Inform researcher that table will not be released
T13. Output may not list or print cases	Output contains list/print cases	Inform researcher that output will not be released
T14. Categorical contextual data must have at least 3 places per category	Categorical contextual data is presented in a table	Request a tabulation of the contextual data showing at least 3 places per category
Charts and Plots		
Rule	**if this…**	**then…**
C1. Scatterplots are not allowed (show individual observations)	Scatterplots	Inform research that scatterplots are not allowed
C2. Histograms must meet frequency requirements.	Histogram does not include the number of cases for each bar or has <11 cases for a bar	Request that researcher provide the frequencies in a table or label the bars. Ask researcher to collapse bars if < 11.
C3. Line charts must meet frequency requirements at each point	Number of cases with minimum and maximum values not shown or < 11	Ask researcher to provide the number of cases at the extremes. If < 11, collapse distribution tail until at least 11 cases in each extreme
C4. Line charts must meet frequency requirements at each point	Number of cases at each point not shown or < 11	Ask researcher to provide the number of cases at each point in a table
C5. Categorical contextual data must have at least 3 places per category	Categorical contextual data is presented in a chart or plot	Request a tabulation of the contextual data showing at least 3 places per category
Regressions (Linear, Logistic non-Linear, Multi-level)		
Rule	**if this…**	**then…**
R1. Table title must make clear the sample and type of results presented	Sample / type of results unclear	Return to researcher to make title clearer
R2. All rows and columns must have understandable labels	Label unclear or variable names used	Return to researcher to make row/column labels clearer
R3. Minimums and maximums (e.g. top and bottom of ranges for variables) must have at least 11 cases if shown	Minimums or maximums do not include the number of cases at the value	Return to the researcher for the number of cases
R4. Counts (unweighted) for all variables in models must be shown	Cell counts not shown	Return to researcher to add cell counts or create separate table with cell counts
R5. Logistic regression: coefficients should not exceed 6 (-6) and OR should not exceed 16 (e^6) (this may indicate a small cell)	Logistic regression coefficient >6 or <-6 or OR>16 or <.06	Request that researcher provide crosstabs of outcome by each predictor
R6. Regression equations must not be able to replicate the data	A linear regression has an R-squared greater than .8 (without lagged variable)	Inform researcher that output will not be released

R7. Predicted values of a variable must have cell sizes of at least 11	Predicated values of variable have less than 11 cases	Return to researcher to collapse predicted categories
R8. Multilevel models with individuals nested in places may not include predictors of places	Multilevel model with stage 1 of the model has places (e.g. states, counties, tracts, HRR) or type of place (e.g. hospital, assisted living, nursing home)	Inform researcher that output will not be released
R9. Categorical contextual data must have at least 3 places per category	Categorical contextual data is presented in a regression	Request a tabulation of the contextual data showing at least 3 places per category