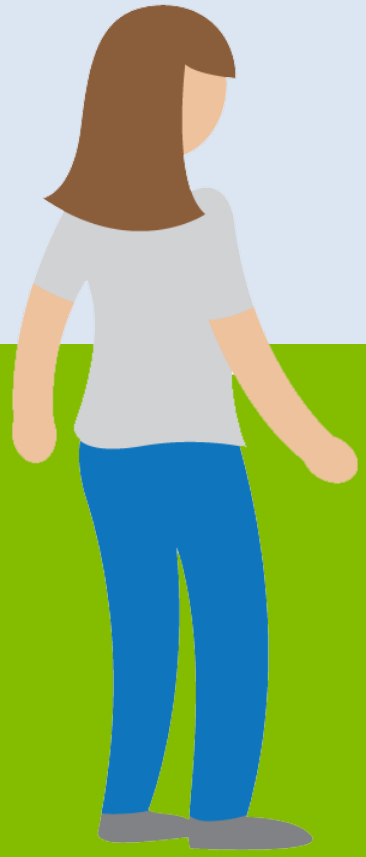
SAGE

Vetting Requirement Checklist for Research Data

Version 1.1



PolicyWise
for Children & Families

Revision History

Version	Date	Produced by	Change
V 1.0	June 21 2016	Lucie Richard	New document
V 1.1	Aug 17 2016	Lucie Richard	Revised after discussion with John Marcotte (ICPSR)
V 1.2	June 22 2017	Robert Jagodzinski	Revised to reflect more discretion around small cell sizes

Contents

Introduction

As part of the process to safeguard the privacy of research participants during secondary use of data assets, SAGE staff reviews (or "vets") all research outputs to be publically shared (ie. Shared with individuals without permission to access the data). Reviews of research outputs are only required for data accessed from SAGE`s physical enclave (Level 1 Security Designation) or Virtual Research Environment (Level 2 Security Designation). The following checklist describes the requirements each type of output must meet in order to be brought outside of SAGE`s Physical Enclave or Virtual Research Environment.

Please note that these are standard requirements only. SAGE will inform you in advance if any data asset to be used in your project, or the linkage of certain data assets to be used in your project, requires additional, fewer or different vetting standards. SAGE may also impose additional requirements depending on your specific output request.

Also, please note that SAGE does *not* vet the accuracy of your program or outputs.

SECTION I: General requirements

Regardless of the type of output to be vetted, as a general rule SAGE requires that they be clean and organized. Some standard things to check include the following:

i. Be sure your outputs are of a permitted type. Permitted outputs include
   a. program syntax or logs of programs that were run;
   b. descriptive results
   c. analytic results
   d. figures or graphs

ii. Use a consistent naming convention for files to be vetted, and where possible make file names descriptive. SAGE staff should know at a glance what they are about to review;

iii. Although you may refine the formatting of your outputs after export, outputs must be clean and organized enough for SAGE staff to understand and interpret (this includes syntax!);

iv. Consider submitting a "read me" file along with your outputs to be vetted to help describe the outputs and how they relate to one another (if applicable);

v. All outputs for a project should be vetted at once (in one review);

vi. There is no maximum number of pages or files that can be reviewed in a vetting request; however, be mindful to only export what is needed for your report or manuscript as this will affect the amount of time required before your files are available to you;

vii. Use discretion where outputs run the risk of reidentifying research subjects. When multiple variables are combined they can reveal certain characteristics about research participants. Reidentification is particularly a risk when reporting rare or highly sensitive conditions (such as ALS or HIV) along with small geographic areas.

viii. Each set of outputs will be evaluated on a case by case basis at the discretion of SAGE staff. Some of the following vetting rules are meant to be guidelines where suppression and aggregation of results may be required. Be careful to ensure your analyses minimize the risk of reidentification of research participants. Outputs should be aggregated at a level no smaller than the minimum level to answer your research questions.

ix. Never release anecdotal information about particular individuals in research studies. Many of SAGE's data assets contain comment fields where respondents have revealed sensitive information. Never disclose the individual responses. If analyzing these responses is relevant to your research question use qualitative analysis methods (eg. content analysis) to process these data.

SECTION II: Syntax documents

A syntax document refers to any document that records the exact procedure used to extract, transform and/or analyze a cut of research data. Syntax documents are usually created and used within a statistical package (eg. SAS, STATA, SPSS, R), but can also be the logged results (including warnings, errors, etc) of running a program in a statistical package. It is strongly recommended that you generate clear and organized syntax document(s) for your research project. You are further encouraged to export syntax documents once a research project is complete to serve as a record of how the data was manipulated and analyzed, in the event questions pertaining to this are raised during report writing or peer-review of a manuscript.

Syntax documents must meet the following vetting requirements to be exported by SAGE staff:

i. Commented-out areas should pertain directly to the purpose and functioning of the syntax. No information about specific data records, outliers or results of programming that identify a group of 10 or fewer records are permitted;

ii. At no point should it be possible to identify a small cell (10 or fewer records) using the syntax, particularly if the syntax document is also a log. For example, it is not permissible to export a log that reveals that a filter by an attribute/set of attributes reduces the sample size by 10 or fewer records;

iii. Syntax documents must be provided to SAGE staff in a txt or PDF format;

SECTION III: Descriptive or Results tables

Descriptive or Results tables are aggregated outputs of data manipulation which either describe the sample analyzed (such as with baseline characteristics tables) or the results of analysis performed on the sample (such as regression outputs, test results), usually in a table format. They provide answers to the research question, and are usually exported for use in peer-reviewed publications, presentations or reports.

Descriptive or Results tables must meet the following vetting requirements to be exported by SAGE staff:

i. Output tables must only contain aggregate results (No individual level data);

ii. In some instances cell sizes less than or equal to 10 may need to be suppressed (usually replaced with "<=10"). This depends on what variables are being combined. If the variables crossed run the risk of reidentifying research participants they must be suppressed. This holds when crossing data on rare or highly sensitive conditions (such as ALS or HIV) with geographic data. When multiple variables are combined identifying very small subgroups cell sizes under 10 must be suppressed. For example, felons under the age of 18 with cancer.

iii. When suppression is required it also applies to zeroes. Rates representing suppressed cells of less than or equal to 10 must also be suppressed, as must any adjacent or related cells which could be used to re-calculate a small cell. For example, if one cell in a row of data is suppressed, but the total of the row and other cells are present, the small cell value can easily be recalculated. In such a case other cells in the row must be suppressed so as to make it impossible to calculate the small cell value.

iv. Those wishing to export tables with rates or non-count data must, in addition to the table, also provide a version of the table containing the count-equivalent of the cells.

SECTION IV: Figures

Figures are graphs or other visual outputs of data manipulation. They are similar to Descriptive or Results tables in purpose but are more visual in function. As such they have slightly different vetting requirements to be exported by SAGE staff:

i. Figures may not show individual-level data (eg. a scatterplot showing sex and age of individuals); This includes visuals of outliers (eg. in a boxplot);

ii. Labels, titles, legends and other marking up of figures must not reveal any identifying information;

iii. When analyses run the risk of reidentifying individuals categorical groups should be aggregated. In these instances counts within categories must be greater than or equal to 10, or missing. For example, the number of individuals represented by a slice in a pie chart cannot be 10 or fewer. For analyses where there is a high risk of reidentification groups less than or equal to 10 in number (including 0) may need to be suppressed. Groups may also be aggregated to create a new total that is greater than or equal to 10. When suppression or aggregation is applied all corresponding rates must be handled similarly.

iv. Those wishing to export figures with rates or non-count data must, in addition to the figure, also provide a version containing the count-equivalent of the groups represented.

v. Figures must be provided to SAGE staff in JPEG, PNG or PDF format.