SENTINEL SAS MACRO TOOLKIT:

De-identification Tool

(%ms_deidentify)

Documentation version: 1.0

Prepared by the Sentinel Operations Center

For use with De-identification Tool version 1.0

March 25, 2016

Sentinel is sponsored by the U.S. Food and Drug Administration (FDA) to monitor the safety of FDA-regulated medical products. Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that complements previously existing methods of safety surveillance. Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

Data De-identification Tool

1. Introduction

This program uses a randomization method to de-identify user-specified fields in a dataset, in order to minimize the risk of sharing or disseminating identifiable, individual-level information. The randomization method works by retaining fields that represents direct identifiers but replacing the original values with randomly-generated values. This program has utility in studies where individual-level data are needed but actual values of person-level identifying variables are required to be masked.

2. Program Objectives

This program creates a de-identified dataset by assigning a randomly generated 'caseid' to one or more user-specified variables in an existing dataset. The program also produces crosswalk files for each variable that is replaced with a random 'caseid'.

Please note that this program is not meant to replace variables for which one or more values are missing. If missing values are encountered in any variables listed in the VARLIST parameter, the program will issue a warning accordingly and stop processing.

It is also important to note that the program can only run on version 9.2 and higher of SAS©, as it makes use of a function (CMISS) that was unavailable in prior software versions.

3. Parameter Specifications

Program Macro Variable Name	Short Description	Long Description
INFILE	Dataset with individual-level data	**Details:** Name of the dataset containing variable(s) to be masked **Input type:** Required, referenced with *LIBREF.SAS-data-set.* **Format**: text **Example**: indata.uncoded

Program Macro Variable Name	Short Description	Long Description
VARLIST	Variable name(s) to be de-identified	**Details:** Name(s) of variable(s) that require replacement with a randomly-generated caseid. If multiple variables are listed, separate variable names with a space (" "). ***Please note that this program is not meant to replace variables for which one or more values are missing. If missing values are encountered in any variables listed in this parameter, the program will issue a warning accordingly and stop processing.*** **Input type:** At least one (1) variable is required. **Format**: text **Example**: var1 var2
XWALKLIB	Directory to store crosswalk file(s)	**Details:** Libname (directory) to which crosswalk files will be saved. **Input type:** Required **Format**: text **Example**: Outdata
OUTFILE	Output file with subset of individual level data	**Details:** Name of the output dataset **Input type:** Required, referenced with *LIBREF.SAS-data-set.* **Format**: text **Example**: outdata.coded

4. Outputs

The output dataset will contain all variables on the original dataset, with the randomly-generated 'caseid' variables replacing the user-specified variables.

Crosswalk output datasets are produced for each variable replaced with a random 'caseid' and contain each original value along with its 'caseid' value.

5. Example

In the following example, the %MS_DEIDENTIFY macro is used to create a de-identified dataset (i.e. &OUTFILE.) from an input dataset called INFILE by assigning a randomly generated 'caseid' to a list of user-specified variables existing in the INFILE. The user-specified variables to be de-identified by the program are PATID and NDC. The parameters used in the macro are:

%ms_deidentify(INFILE= indata.ms_deidentify,
VARLIST= patid ndc
OUTFILE= outdata.ms_deidentify,
XWALKLIB= outdata);

The program first creates intermediate datasets for each variable listed in VARLIST, containing unique values of each variable named on VARLIST along with a unique random number for each record. It then creates a new 'caseid' for each variable in the list of the user specified variables in VARLIST. Using a left outer join, the new 'caseid' variables are then merged to the final de-identified dataset named in the OUTFILE parameter.