Paper AD08_05

# Using MS-ACCESS® Metadata to Drive Automated SAS® Data Processing

Gary N. Weeks, Centers for Disease Control, Atlanta, Georgia


## ABSTRACT

This paper describes the automation of data processing for complex survey data using metadata in a MS Access database. The Pregnancy Risk Assessment Monitoring System (PRAMS) run by CDC Division of Reproductive Health in conjunction with state health agencies uses mail and phone questionnaires to collect pregnancy and birth outcome data to identify pregnancy risk. PRAMS has grown from three participating states in 1988 to 32 participating agencies in 2004 and continues to grow. When the number of participating states was small data were processed with 10 manually run SAS programs for each state then manually cleaned. As the PRAMS expanded an automated macro driven SAS data processing system was implemented that effectively reduced the processing time but did not address data cleaning which was still labor intensive. For the 2004 survey year PRAMS began a new phase that required rewriting the questionnaires, concurrently a data processing and validation system was developed to use metadata in a Microsoft Access database to drive processing. The MS-Access metadata database allowed data managers without SAS programming skills to dynamically add, remove, and alter variables and their attributes and ranges as well as confirm skip pattern validity by using dynamically created integrity constraints while reducing the need for specialized SAS programming.


## INTRODUCTION

The Pregnancy Risk Assessment Monitoring System (PRAMS) is an ongoing state- and population-based surveillance system designed to monitor selected self-reported behaviors and experiences that occur before, during, and after pregnancy among women who deliver a live-born infant in 31 states and New York City. PRAMS, which has grown from 3 participating agencies in 1988 to 32 in 2004, continues growing and anticipates additional participation in the near future. PRAMS began a new phase in 2004, and the projected increase in complexity and volume of information necessitated a new data processing approach. The goal of this project was to automate data processing using a metadata database that could be modified to handle future changes without significant changes in the existing code.

## DATA

PRAMS employs a mixed-mode data collection methodology. Mail surveys are used as the primary mode of data collection, followed by a telephone interview for non-responders. The information is entered using a data entry program that generates a SAS data set of questionnaire responses. Additional data are generated to track sampling size and efficiency. The states send an average of 12 batches of data a year to CDC for additional processing, cleaning, and weighting prior to release for analysis. Each batch contains flat files, ancillary to the questionnaire responses, containing birth certificate, sampling frame, comment, operations, and survey tracking data. The SAS data consist of categorical numeric variables, date values (not in SAS date/time format), and text. The previous processing system used a SAS framework of macro programs and variables to provide the flexibility for one program to handle data whose contents varied by source. Although the existing system was functional, significant programming and testing would be required to modify the existing code for use in the new phase of PRAMS. To reduce future maintenance and increase flexibility for future changes, a new metadata-based processing system was implemented.

## PROCESSING

The states access a secure Web site to submit their batches as a compressed archive that contains data files and related reports. The data are deposited into a common folder accessible on the Local Area Network. A Windows scripting program searches the folder for data submissions, and if a new batch is found, it is placed in an archive and a copy is made for processing. The script unzips the archive, checks its contents, and reports missing files via email. If all data files are present, the data are processed; otherwise, no processing is done. To initiate processing the scripting program starts a SAS batch session that runs the control program macro. The necessary parameters are passed by the script to SAS as a SYSPARM option. A libname statement is used to read the metadata from the MS Access database which is subset into a SAS dataset using parameters passed from the controlling program to filter for data specific only to the submitting state.

Flat files are read into SAS datasets using column input positions, variable attributes, and formats read from the metadata. Standard procedure is to put all of the submitted data into SAS data sets that are held in a directory along

with data generated during processing. Having access to these data as SAS data sets facilitates debugging and data correction. Some of the data in the flat files are also used to augment the final questionnaire data set.

Data quality and integrity are validated using SAS integrity constraints to test for valid ranges and skip pattern legitimacy. In both cases the metadata are used to dynamically write PROC DATASET code that creates integrity constraints. Because the goal is to report data errors and not to correct them, two copies of the questionnaire data are generated, one without any observations and one with all observations. The integrity constraints are applied to the empty data set, and the complete data are appended to the empty copy. Any violations of the integrity constraints result in an error message that is captured in a SAS data set and subsequently examined by the data managers to determine the proper method of resolution.

The data entry application used by the states to enter the survey responses is occasionally updated, which may cause the metadata associated with a variable to change. To track changes in the data entry program, a revision variable is included in the data set and incremented with each change in the data entry application. To accommodate processing batches that contain data from more than one revision, the data are divided into different temporary data sets, based on the revision of the data entry program, and processed using metadata specific to their revision.

Range checks are straightforward. A null data step, with the metadata table as the input data, is used to read the variable names and valid ranges, which are then written to a text file in the form of a series of PROC DATASETS procedures that build the integrity constraints. The program is then inserted into the processing run using a %INCLUDE statement. Writing a program with put statements and including it in the run creates a record of the range checks for later reference.

Checking skip patterns is a little more complex. Responses that are skipped have a special missing value of "._" in the raw data and are later recoded to the special missing value of ".S". The data entry application is programmed to automatically skip questions if any preceding questions contain a value that would make it unnecessary to answer the question. Some skips may be triggered by responses to multiple questions or a cascading sequence of question responses. To validate correctly skipped responses, the value of all relevant preceding questions must be checked. Consider the following questions:

24. Did you have any of these problems during your most recent pregnancy? For each item, circle Y (Yes) if you had the problem or circle N (No) if you did not.

| | | No | Yes |
|---|---|---|---|
| a. | High blood sugar (diabetes) that started before this pregnancy | N | Y |
| b. | High blood sugar (diabetes) that started during this pregnancy | N | Y |
| c. | Vaginal bleeding | N | Y |
| d. | Kidney or bladder (urinary tract) infection | N | Y |
| e. | Severe nausea, vomiting, or dehydration | N | Y |
| f. | Cervix had to be sewn shut (incompetent cervix) | N | Y |
| g. | High blood pressure, hypertension (including pregnancy-induced hypertension [PIH], preeclampsia, or toxemia) | N | Y |
| h. | Problems with the placenta (such as abruptio placentae or placenta previa) | N | Y |
| i. | Labor pains more than 3 weeks before my baby was due (preterm or early labor) | N | Y |
| j. | Water broke more than 3 weeks before my baby was due (premature rupture of membranes [PROM]) | N | Y |
| k. | I had to have a blood transfusion | N | Y |
| l. | I was hurt in a car accident | N | Y |

If you did not have any of these problems, go to Question 26.

25. Did you do any of the following things because of these problems? For each item, circle Y (Yes) if you did that thing or circle N (No) if you did not.

| | | No | Yes |
|---|---|---|---|
| a. | I went to the hospital or emergency room and stayed less than 1 day | N | Y |
| b. | I went to the hospital and stayed 1 to 7 days | N | Y |
| c. | I went to the hospital and stayed more than 7 days | N | Y |
| d. | I stayed in bed at home more than 2 days because of my doctor's or nurse's advice | N | Y |

Each pregnancy problem in question 24.a.–24.l. is a separate variable in the questionnaire. If the response to each problem is "No," then questions 25.a.–25.d. are skipped. Otherwise the interviewee answers questions 25.a.–25.d.. It is important that the skip pattern for question 25 be followed appropriately. Where any variable in question 25 has a skip value, it is required that each variable in question 24 be checked to ensure that all responses were negative ("No"). In a manner similar to the range checks, a null data step with the metadata as an input data set is used to generate a series of PROC DATASETS procedures with put statements that are subsequently included in the processing.

PRAMS collects survey data both by mail and telephone. Missing data, refusal to answer, and "don't know" responses are recoded to different special missing values depending on the mode (mail or telephone) of data collection. Missing values in mail surveys are coded as ".B" to indicate a blank because it is unknown if the interviewee intentionally left a question blank or unanswered. Unlike in the mail survey, in a telephone survey the interviewee has the option to answer "refused" or "don't know" to every question. In telephone surveys a value of ".R" is returned in the data set for refusal to answer during the interview. The data set contains a variable that indicates the mode of data collection, and the metadata have a column of values that are recoded to ".B" if the survey was returned by mail. Because of existing (and possible future) incongruities between the output of the data entry program and the desired data values, other variables require recoding. These variables are recoded using the put function and formats found in a column in the metadata. There is a table in the metadata database, maintained by the data managers, that is used as a CNTLIN= option in a PROC FORMAT procedure. These formats are generated for each batch to ensure that the most recent changes are available at processing time.

## CONCLUSION

Using a Microsoft Access database as a repository for the metadata provides a system for data mangers with knowledge of the data to add, remove, or augment data processing parameters without having to be skilled SAS programmers. This approach provides flexibility for future changes in the PRAMS project without having to make significant changes to the processing code.


## ACKNOWLEDGMENTS

The entire PRAMS team in the Division of Reproductive Health at the CDC was an invaluable and willing resource without whose help this project could not have been completed. A special thanks to Ms. Karen Colberg for extending her career to help bridge the divide between the past and the future.


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Gary N. Weeks
CDC-Division of Reproductive Health
4770 Buford Highway, MS K-21
Atlanta, Georgia 30341-3717
Work Phone: (770) 488-6321
Fax: (770) 488-6354
Email: giw9@cdc.gov