Using an Excel®-based data dictionary to automate production of codebooks for large-scale surveys

Elise Comperchio[1], Caitlin Finan[1], Matthew Kastin[1], Megan Stead[1], Shannon Corcoran[2], Jian Zhang[2], and Michael Slater[2]

[1]NORC at the University of Chicago; [2]Centers for Medicare and Medicaid Services (CMS)

ABSTRACT

The Medicare Current Beneficiary Survey (MCBS) is a continuous, multipurpose survey of a nationally representative sample of the Medicare population, conducted by the Centers for Medicare & Medicaid Services (CMS) through a contract with NORC at the University of Chicago. The MCBS contains approximately 2,000 variables on a variety of topics such as health status and functioning, health care use and expenditures, and health insurance coverage. In this poster, we describe the use of an Excel® -based data dictionary to automate codebook production for a large-scale survey in SAS®. While there is currently no process directly provided by SAS® for producing a survey codebook, other researchers have written survey codebook programs using this software and we build upon these efforts in the development of Excel® -driven codebooks. First, we define the main input columns in the data dictionary: the dataset name, variable name, variable label, question text, code frame, and format name. We then explain the use of SAS® to combine information from these input columns and the datasets to create a data dictionary-driven codebook. This process includes 1) obtaining metadata from the data dictionary for variables of interest; 2) producing summary data for each variable; and 3) formatting the final codebook PDF. We also provide examples of key features of the codebook, such as listings of variable attributes, dataset attributes, and descriptive statistics. Finally, we propose several ideas to further enhance automated codebook production, including providing additional descriptive statistics and generating a codebook quality control (QC) report to expedite manual codebook review.

BACKGROUND

What is the MCBS?

The Medicare Current Beneficiary Survey (MCBS) is a continuous, multipurpose survey of a nationally representative sample of the Medicare population, conducted by the Centers for Medicare and Medicaid Services (CMS) through a contract with NORC at the University of Chicago. The MCBS collects data from Medicare beneficiaries at three points per year for four consecutive years. The data include topics related to health care access, health expenditures, health insurance coverage, health status and functioning, demographics, and satisfaction with health care. The MCBS data structure stores approximately 2,000 variables across 40 analytic data files. MCBS Limited Data Sets are available annually as well as an MCBS Public Use File.

Given the complexity of the MCBS data structure, a logical first step in understanding what is captured by the questionnaire involves becoming familiar with data file characteristics, variable attributes, and response distributions. A codebook is one mechanism to facilitate such documentation and review of large-scale, complex survey data.

Introduction to the codebook

Codebooks generally serve two purposes. First, a codebook documents data file structure, contents, and variable code definitions. Second, a codebook functions as a handbook for survey researchers to use in coding responses (Lavrakas, 2008). Codebooks may be produced using a variety of software platforms, including Excel®, SPSS®, or SAS®.

In this poster, we describe the utilization of both Excel® and SAS® to automate codebook production. While SAS® does not currently have a codebook production capability available, we developed a codebook generation program based on the principles outlined in the University of North Carolina Population Center's SAS® codebook macro article (UNC Carolina Population Center, 2010). The examples presented in this poster are specific to the MCBS; however, this effort can inform codebook production for surveys in general.

METHODS

Codebook inputs

The MCBS data dictionary is an Excel® file that documents key features of analytic data files and corresponding questionnaire information for each analytic variable. Data dictionary fields include variable attribute information, response options and code values, and question text and the questionnaire screen name at which the data are collected. The data dictionary is a road map in understanding how data move from the survey instrument to analytic data files.

The MCBS data dictionary is also used to drive codebook production. The data dictionary contains valuable metadata that provides the key features of a codebook when coupled with the actual data and associated variable formats.

Below is a list of the data dictionary content that serves as inputs into MCBS codebooks:

- Analytic data file name
- Analytic variable name
- Analytic variable label
- Analytic variable type
- Question text
- Response options
- Variable format name

The following information from the analytic data files serves as inputs into MCBS codebooks:

- Number of observations (rows) in the data file
- Frequency of each reported response option

Codebook generation program

We run the codebook generation program as a process designed within SAS Enterprise Guide 7.1®. The codebook generation process is comprised of three steps, outlined below.

An advantage of this codebook generation program is its flexibility. As long as the corresponding data dictionary has the appropriate input columns, the program can be used to create any codebook with minimal modification to the code.

Macro Catalog Compilation

This step compiles general purpose (utility) macros used throughout codebook production. Functions include:

- Parameter assertions (missingness, equality, existence)
- Obtain data file attributes
- Create lists and macro variable arrays

Codebook Generation Macro

This step compiles the macro used to generate the codebook PDF. Steps executed by this macro include:

- Obtain metadata from data dictionary input columns
- Obtain table attributes and variable-level statistics from analytic files
- Combine above input information to create content for codebook report

Run Codebook Program

This portion of the process calls the codebook generation macro and creates the codebook PDF. Steps include:

- Set macro variables
- Format title page and codebook key
- Produce codebook

RESULTS

Table 1. Codebook Fields and Corresponding Sources of Information

Codebook Field	Description	Source
Analytic Table	Name of the analytic file	Data dictionary
Universe Statement	Contents and purpose of the analytic file	Data Dictionary
Number of Observations	Number of observations (rows) in the analytic file	Analytic file
Number of Variables	Number of variables (columns) in the analytic file	Data dictionary
SAS® Variable Name	Variable name in the analytic file	Data dictionary
SAS® Variable Type	Variable type in the analytic file (character or numeric)	Data dictionary
SAS® Label	Variable label in the analytic file	Data dictionary
Question Text	Text displayed on the screen for the questionnaire question(s) tied to this variable	Data dictionary
Value: Description	Possible data values and value definitions	Data dictionary
Frequency	Number of observations that has a given data value	Analytic file

Exhibit 1: Example Table-Level Report*

Medicare Current Beneficiary Survey (MCBS) Community Annual Analytic File Codebook
Detailed Table-Level Report

Analytic Table:	ACCS
Universe Statement:	ACCS contains fall-round information about access to medical care, usual source of medical care and satisfaction with medical care.
Number of Observations:	360
Number of Variables:	302

Exhibit 2: Example Variable-Level Report for a Numeric Variable*

Detailed Variable-Level Report

SAS Variable Name: ADMITHOS
SAS Variable Type: Num
SAS Label: SP ADMITTED TO HOSPITAL OVERNIGHT
Question Text: In the last six months, [were you/was SP] admitted to a hospital overnight or longer?

Value: Description	Frequency	Percent
(01) YES	200	55.5
(02) NO	100	27.8
(.) INAPPLICABLE	50	13.9
(.R) REFUSED	5	1.4
(.D) DON'T KNOW	5	1.4

Exhibit 3: Example Variable-Level Report for a Character Variable*

Detailed Variable-Level Report

SAS Variable Name: GETUSOS
SAS Variable Type: Char
SAS Label: HOW DOES SP USUALLY GET TO DR-OTHER SPECIFY TEXT
Question Text: SOME OTHER WAY (SPECIFY)

Value: Description	Frequency	Percent
OTHER SPECIFY TEXT	300	83.3
( ) INAPPLICABLE	60	16.7

*The data shown in these exhibits are mock data and are not representative of actual MCBS data. Mock data are included for illustrative purposes only.

DISCUSSION

Using metadata from Excel® -based data dictionaries combined with attributes from analytic data files, we create codebooks that provide table-level and variable-level information about data collected in the MCBS. Because the data dictionary serves many functions within the MCBS project, building off of this file for codebook creation is an efficient use of resources. Additionally, implementing an Excel® -driven codebook creates a streamlined, efficient process that facilitates iterative review and updating of codebook input information without making significant updates to SAS® code.

The following are necessary parts of codebook development and creation:

- Plan for multiple iterations of production and review
- Understand the relationship between the data dictionary and the codebook – unintended editing will have ramifications on what appears in the codebook
- Once codebook framework is established, it is easy to refine stylistic features (e.g., text size, wording, etc.)

FUTURE ENHANCEMENTS

- Include additional descriptive statistics for continuous response variables, including mean and median values, minimum and maximum values, and quartile ranges
- Include universe statements that are at the variable-level instead of at the table-level
- Develop cookbook quality control (QC) report to expedite manual review. QC steps may include confirming variable names and labels are not truncated, frequencies and descriptive statistics are not null, and a cross-check of the number of variables in the data dictionary against the number appearing in the codebook
- Increase flexibility of the codebook program to meet the needs of different data delivery products. For example, it may be useful to display other data dictionary columns for different data delivery products.
- Update the formatting of the codebook to ensure 508 compliance

REFERENCES

Lavrakas, P. (2008). Encyclopedia of Survey Research Methods. Retrieved from http://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n69.xml

UNC Carolina Population Center. (2010). SAS codebook macro. Retrieved from http://www.cpc.unc.edu/research/tools/data_analysis/proc_codebook

CONTACT

- Elise Comperchio: Comperchio-Elise@norc.org
- Caitlin Finan: Finan-Caitlin@norc.org
- Matthew Kastin: Kastin-Matthew@norc.org
- Megan Stead: Stead-Megan@norc.org
- Shannon Corcoran: Shannon.Corcoran@cms.hhs.gov
- Jian Zhang: Jian.Zhang@cms.hhs.gov
- Michael Slater: Michael.Slater@cms.hhs.gov

This work is submitted under contract number HHSM-500-2014-00035I, HHSM-500-T0002 with the Centers for Medicare & Medicaid Services, Office of Enterprise Data and Analytics. The opinions and views expressed in this work are those of the authors. No official endorsement by the Department of Health and Human Services or the Centers for Medicare & Medicaid Services is intended or should be inferred.