Paper 163

Same Question, Different Name: How to Merge Responses

Michelle M. Dahnke, DrPH, Florida A&M University; Tyra Dark, PhD, Florida A&M University

ABSTRACT

To correctly use data from the Collaborative Psychiatric Epidemiology Surveys, which joins three individual surveys, users may need to evaluate the cross-survey linking and merge responses. While the codebook identifies the variable name assigned to each question, in some instances the same questions were assigned different names in the three surveys. For example, a question about being diagnosed with high blood pressure was named V04052 and V06677 depending on the survey. This paper demonstrates how to merge response data in circumstances such as this so the user can conduct analysis on the maximum number of valid responses from all three surveys.

INTRODUCTION

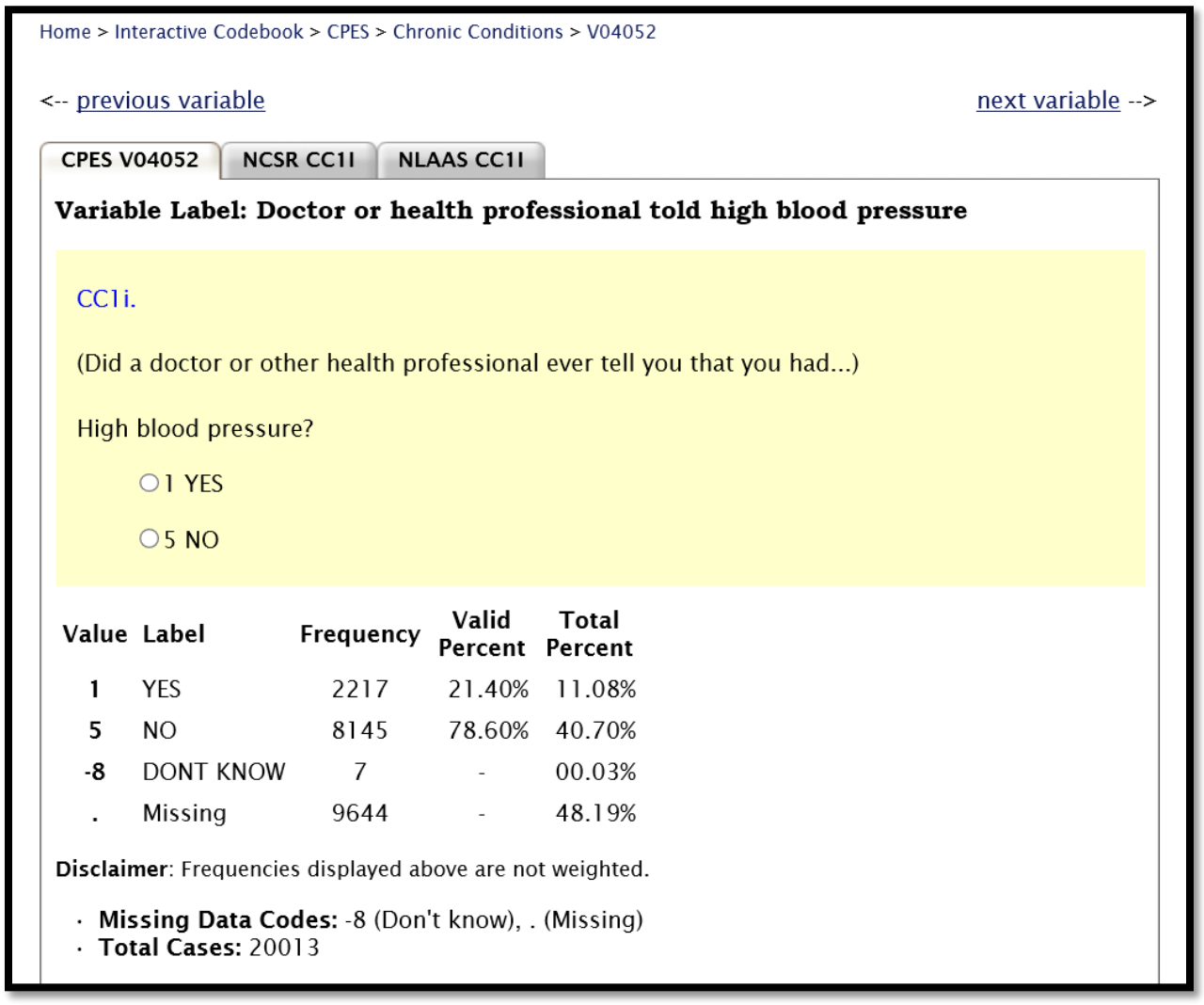
The need to merge data is a common function before conducting statistical analysis. Depending on the source and collection method, questions that ask the same question may have the same variable name or different variable names. The latter is the case when using certain data from the National Institute of Mental Health Collaborative Psychiatric Epidemiology Surveys (CPES).

CPES data are combined from three nationally representative surveys that each focus on the mental health of a specific population. The surveys are the National Comorbidity Survey Replication (NCS-R), the National Survey of American Life (NSAL), and the National Latino and Asian American Study (NLAAS). A notable strength of using the CPES data set is the incredible diversity of its participants, which allows for a unique opportunity to assess variation in outcomes between and among different population subgroups. Respondents were from 252 geographic areas, 50 of which were included in all three surveys. The same data collection methods were used for all three surveys with project managers and support staff from the Survey Research Center (SRC), which is part of the Institute for Social Research at the University of Michigan in Ann Arbor, Michigan. This structure allowed for the data sets to be analyzed individually or as though they were a single, nationally representative study with more than 20,000 participants.

Before combining data, it is important to verify the two variables are capturing the same information and that participants for the two variables to be combined are mutually exclusive. This paper will begin by explaining that process when using CPES data, then will show how to combine the different variables using SAS®.

USING THE CPES INTERACTIVE CODEBOOK TO REVIEW VARIABLES

CPES has an Interactive Codebook online that makes it simple to identify what the variable names are, and if they are different, for questions that capture the same information. Figure 1 below is a screen capture of the Interactive Codebook for the variable named V04052, which shows all the information included in the codebook, specifically detail about the question that was asked to participants, variable label, response value, frequency and valid percent, among other information. The tabs near the top show this variable was only asked of participants in NCS-R and NLAAS.



Home > Interactive Codebook > CPES > Chronic Conditions > V04052

<-- previous variable next variable -->

CPES V04052 | NCSR CC1I | NLAAS CC1I

Variable Label: Doctor or health professional told high blood pressure

CC1i.

(Did a doctor or other health professional ever tell you that you had...)

High blood pressure?

○ 1 YES

○ 5 NO

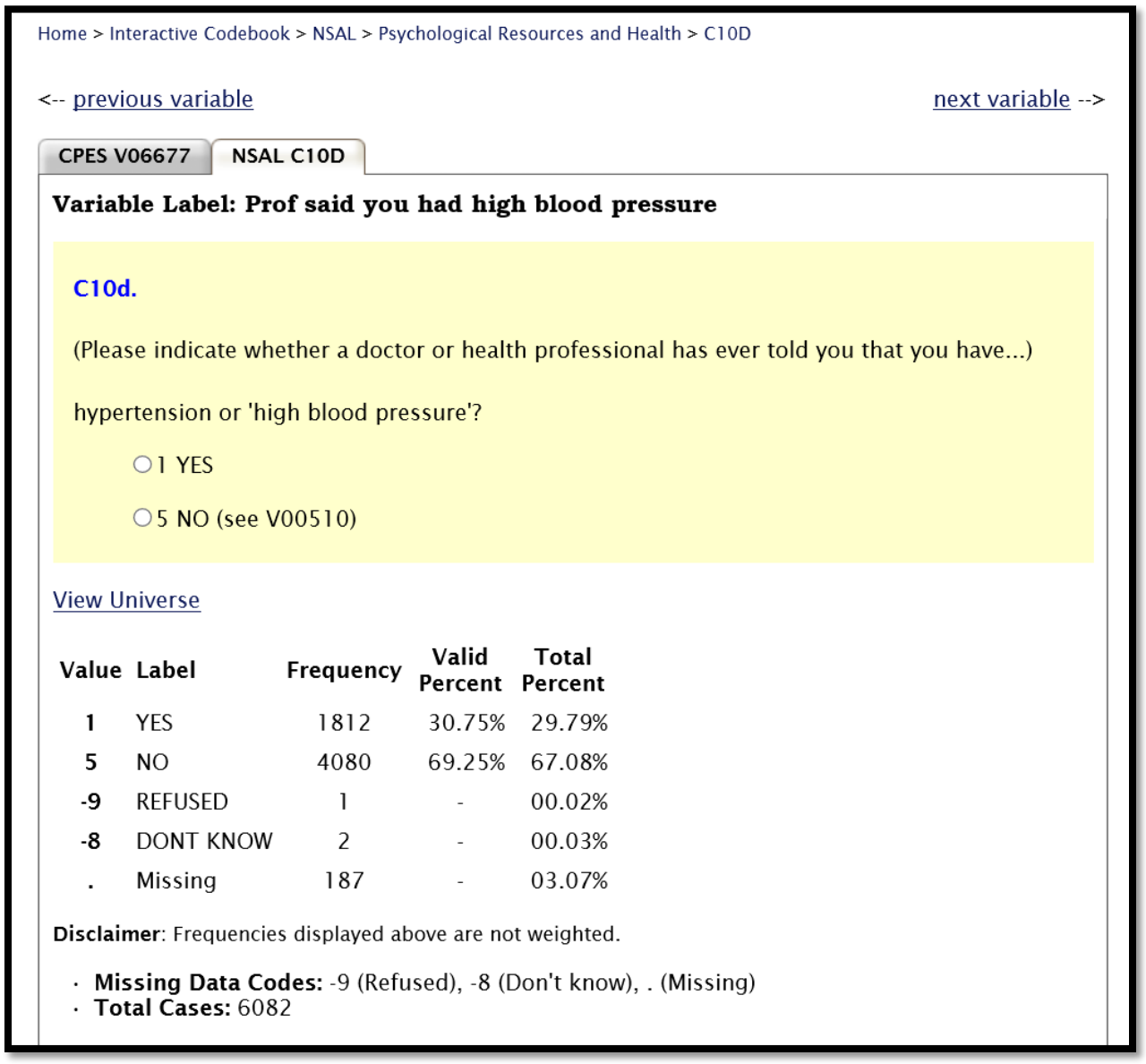
Value	Label	Frequency	Valid Percent	Total Percent
1	YES	2217	21.40%	11.08%
5	NO	8145	78.60%	40.70%
-8	DONT KNOW	7	-	00.03%
.	Missing	9644	-	48.19%

Disclaimer: Frequencies displayed above are not weighted.

- **Missing Data Codes:** -8 (Don't know), . (Missing)
- **Total Cases:** 20013

Figure 1 Screen Capture of Interactive Codebook for Variable V04052

Figure 2 below is a screen capture of the Interactive Codebook for the variable named V06677, which shows similar information for this variable and again includes the tab near the top that identifies this variable was only asked of NSAL participants.



Home > Interactive Codebook > NSAL > Psychological Resources and Health > C10D

<-- previous variable next variable -->

CPES V06677 | NSAL C10D

Variable Label: Prof said you had high blood pressure

C10d.

(Please indicate whether a doctor or health professional has ever told you that you have...)

hypertension or 'high blood pressure'?

○ 1 YES

○ 5 NO (see V00510)

View Universe

Value	Label	Frequency	Valid Percent	Total Percent
1	YES	1812	30.75%	29.79%
5	NO	4080	69.25%	67.08%
-9	REFUSED	1	-	00.02%
-8	DONT KNOW	2	-	00.03%
.	Missing	187	-	03.07%

Disclaimer: Frequencies displayed above are not weighted.

- **Missing Data Codes:** -9 (Refused), -8 (Don't know), . (Missing)
- **Total Cases:** 6082

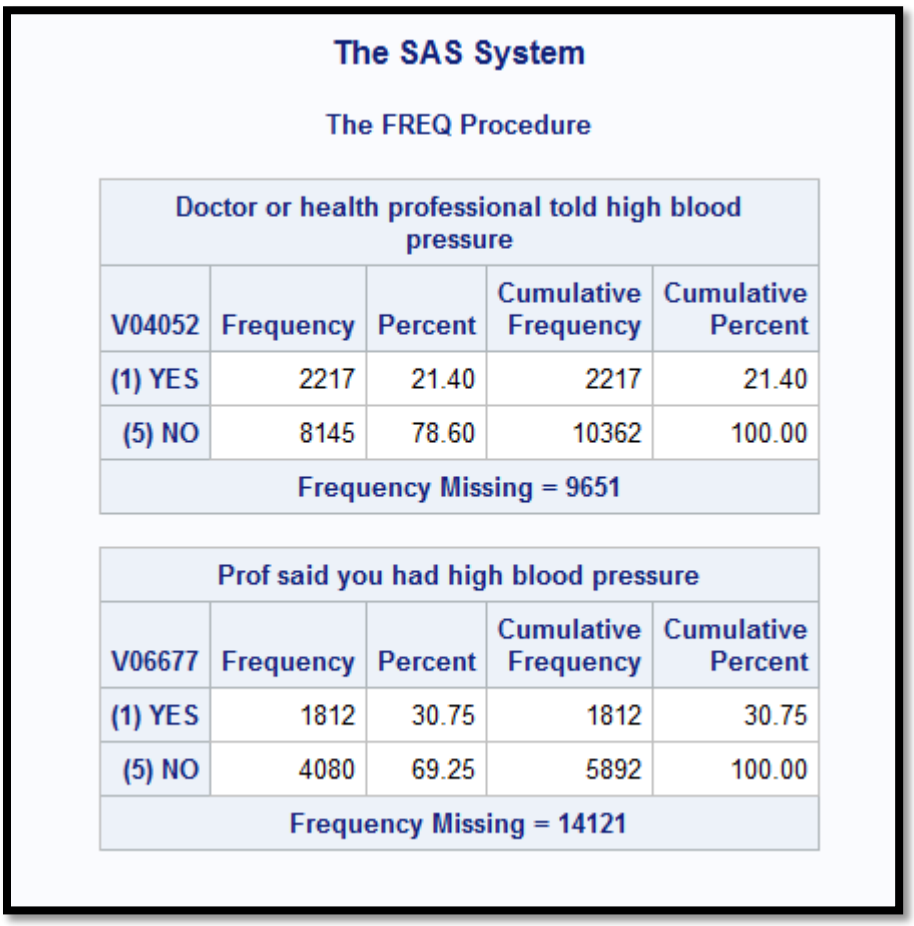
Figure 2 Screen Capture of Interactive Codebook for Variable V06677

Figure 1 and Figure 2 show that valid responses are categorized the same for both variables: “YES” and “NO” are assigned the values “1” and “5”, respectively. This is important to note and will be necessary information for merging the data later using SAS®.

With the exception of the survey specific information, the other data about these variables can be validated by running a PROC FREQ and creating tables. An example of this PROC FREQ code is below.

```
PROC FREQ;
TABLES V04052 V06677;
RUN;
```

Figure 3 shows the output tables, which includes the response, frequency, percent, cumulative frequency and cumulative percent for the variables with their original names.



The SAS System

The FREQ Procedure

Doctor or health professional told high blood pressure				
V04052	Frequency	Percent	Cumulative Frequency	Cumulative Percent
(1) YES	2217	21.40	2217	21.40
(5) NO	8145	78.60	10362	100.00
Frequency Missing = 9651				

Prof said you had high blood pressure				
V06677	Frequency	Percent	Cumulative Frequency	Cumulative Percent
(1) YES	1812	30.75	1812	30.75
(5) NO	4080	69.25	5892	100.00
Frequency Missing = 14121				

Figure 3 PROC FREQ Output for Variables with Original Names

MERGING DATA FROM TWO VARIABLES BY CREATING A NEW VARIABLE

After gaining an understanding of the data as it is available by CPES, the next step in SAS® is to create a new variable that combines data from the two original variables. One way to do this is to create a new variable name using an IF-THEN statement. In this case, the new name will also better describe the variable and the responses as it is renamed “HBP” with responses “YES” and “NO” as opposed to “V04052” and “V06677” with responses “1” and “5”. An example of this code is below:

```
IF (V04052=1) THEN HBP = 'Yes';
IF (V06677=1) THEN HBP = 'Yes';
IF (V04052=5) THEN HBP = 'No';
IF (V06677=5) THEN HBP = 'No';
RUN;
```

This code merges the “1” responses from both of the original variables and the “5” responses from both original variables. It is important to verify the new variable named “HBP” was created and the responses variables were merged correctly.

VERIFYING DATA IN THE NEWLY CREATED VARIABLE “HBP”

A straightforward way to check if the code worked as intended is by running another PROC FREQ. This time, the new variable name, “HBP”, will be used to create the table. An example of this code is below and the output is shown in Figure 4:

```
PROC FREQ;
TABLES HBP;
RUN;
```

The SAS System

The FREQ Procedure

HBP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	12225	75.21	12225	75.21
Yes	4029	24.79	16254	100.00
Frequency Missing = 3759				

Figure 4 PROC FREQ Output Showing Newly Created Variable "HBP"

Figure 4 shows the frequency for “No” responses is 12225, which is the sum of 8145 and 4080, and the frequency for “Yes” responses is 4029, which is the sum of 2217 and 1812. This verifies the data were merged as intended. From here, the user can continue analysis on the new “HBP” variable that contains all of the valid responses from the three CPES surveys.

CONCLUSION

This paper was written to demonstrate the process by which data that has different variable names, but asks the same question, can be merged for use in subsequent analysis.

REFERENCES

Collaborative Psychiatric Epidemiology Surveys. *Welcome to CPES.* July 1, 2015. Available at http://www.icpsr.umich.edu/icpsrweb/CPES

ACKNOWLEDGMENTS

The first author would like to thank the second author, Dr. Tyra Dark, and the other members of her dissertation committee being that this SAS® skill was learned while completing analysis for her dissertation. The remaining committee members are Dr. Gebre Kiros (FAMU), Dr. Saleh Rahman (FAMU), Dr. Hong Xiao (UF) and Dr. Roger Boothroyd (USF).

CONTACT INFORMATION

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

Name: Michelle M. Dahnke, DrPH
Enterprise: Florida A&M University Institute of Public Health
Address: 1515 S. Martin Luther King, Jr. Blvd, SRC207
City, State ZIP: Tallahassee, Florida 32307
Phone: 954.494.7471
E-mail: mdahnke@gmail.com