

Extract Information from Large Database Using SAS® Array, PROC FREQ, and SAS Macro

Lifang Zhang, Department of Biostatistics and Epidemiology, Augusta University

ABSTRACT

SAS® software offers the statistician and programmer a lot of ways to extracting useful information from large SAS data sets for statistical analysis and report. The techniques discussed include the use of base SAS software i.e. Common used PROC IMPORT, SAS ARRAY, PROC FREQ and OUPUT, SAS MACRO, and PROC EXPORT to extract useful information for future analysis and report.

This presentation provides a detailed blueprint for 1) Using PROC IMPORT to import Excel data; 2) Using ARRAY to detect specific diseases (Diagnostic code); 3) PROC FREQ and OUTPUT to generate indicate variables and data sets; 4) Merging them together to produce one summary report or dataset. 5) Using SAS MACRO to repeat the step 2 to 4; 6) Export the final data sets or tables to define destination.

The paper use mimic diabetes clinic data from 2012 to 2014. ICD-9 Diabetes diagnostic codes included in range 250.00 - 250.80. The paper will show 1) how to extract each diagnostic code by each year for all visits. 2) Count all diabetes diagnostic codes for each unique patient for each year. In other words, each patient can only count one time in each year.

The Excel tables will be created and exported in a defined destination.at the end of program.

INTRODUCTION

Working with large amounts of data is a challenging and especially time-consuming job. How can the desired information be extracted effectively while reducing the amount of time required to retrieve the data? This paper explains some of the easy and useful data extraction tips using commonly used SAS procedures and techniques.

1, USING PROC IMPORT TO READ THE EXCEL DATA INTO SAS

We used mimic diabetes clinic data from 2004 to 2012. Following is the PROC contents output of the data as shown in Table 1. There were a total of 6222 observations, containing 509 subjects. Variables diagnosis_1 to diagnosis_15 were used to record some patients who had more than one diagnosis at each visit.

Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
1	Patient_Number	Num	8			Patient Number
2	Encounter_Number	Num	8			Encounter Number
3	BIRTH_DATE	Num	8	DATE9.	DATE9.	BIRTH DATE
4	DIAGNOSIS_1	Char	6	\$6.00	\$6.00	DIAGNOSIS_1
5	DIAGNOSIS_2	Char	6	\$6.00	\$6.00	DIAGNOSIS_2
6	DIAGNOSIS_3	Char	6	\$6.00	\$6.00	DIAGNOSIS_3
7	DIAGNOSIS_4	Char	6	\$6.00	\$6.00	DIAGNOSIS_4
8	DIAGNOSIS_5	Char	6	\$6.00	\$6.00	DIAGNOSIS_5
9	DIAGNOSIS_6	Char	6	\$6.00	\$6.00	DIAGNOSIS_6
10	DIAGNOSIS_7	Char	6	\$6.00	\$6.00	DIAGNOSIS_7
11	DIAGNOSIS_8	Char	6	\$6.00	\$6.00	DIAGNOSIS_8
12	DIAGNOSIS_9	Char	6	\$6.00	\$6.00	DIAGNOSIS_9
13	DIAGNOSIS_10	Char	6	\$6.00	\$6.00	DIAGNOSIS_10
14	DIAGNOSIS_11	Char	6	\$6.00	\$6.00	DIAGNOSIS_11
15	DIAGNOSIS_12	Char	6	\$6.00	\$6.00	DIAGNOSIS_12

16	DIAGNOSIS_13	Char	6	\$6.00	\$6.00	DIAGNOSIS_13
17	DIAGNOSIS_14	Char	6	\$6.00	\$6.00	DIAGNOSIS_14
18	DIAGNOSIS_15	Char	6	\$6.00	\$6.00	DIAGNOSIS_15
19	ADMIT_DATE	Num	8	DATE9.	DATE9.	ADMIT DATE

Table 1. Proc Contents for mimic diabetes clinic data from 2004 to 2012

There are several ways of reading Excel data into SAS system. The one of the most efficient method is to use PROC IMPORT as below.

```
PROC IMPORT OUT= WORK.one
  DATAFILE= "C:\SESUG\DataReport\Diabwork.xls"
  DBMS=EXCEL REPLACE;
  RANGE="Deidentified$";
  GETNAMES=YES;
  MIXED=YES;
  SCANTEXT=YES;
  USEDATE=YES;
  SCANTIME=YES;
RUN;
```

Before using PROC IMPORT, check the Excel data and make sure that the first observation in the data set is not missing. If there are some missing variables in the first observation row, add an artificial observation into your data with the right format for that specific variable. After importing to data into the SAS system, delete that specific observation. This will ensure that all variables are entered in the format you want.

2, USE ARRAY AND MACRO TO CREATE INDEX VARIABLES FOR EACH DIABETES DIAGNOSTIC CODE

We first generate the index variable "CODE250_1" for code 250.1 as follows:

```
CODE250_1=0;
ARRAY C1{15} DIAGNOSIS_1 DIAGNOSIS_2 DIAGNOSIS_3 DIAGNOSIS_4 DIAGNOSIS_5
DIAGNOSIS_6 DIAGNOSIS_7 DIAGNOSIS_8 DIAGNOSIS_9 DIAGNOSIS_10 DIAGNOSIS_11
DIAGNOSIS_12 DIAGNOSIS_13 DIAGNOSIS_14 DIAGNOSIS_15;
DO I=1 TO 15;
  IF C1{I} = "250.1" THEN CODE250_1=1;
END;
RUN;
```

After creating the index variable for ICD-9 code 250.1 for "DIABETES WITH KETOACIDOSIS", we use PROC FREQ to count the number of ICD-9 code 250.1 frequencies by each year. The year variable is created by using the YEAR function (Adm_Year= Year(ADMIT_DATE)). We then output the frequencies by year into a data set named "t1".

```
PROC FREQ DATA=TWO;
  TABLE ADM_YEAR* CODE250_1/NOCOL NOROW NOPERCENT OUT=T1;
RUN;
PROC SORT DATA=T1;
  BY ADM_YEAR;
RUN;
```

At this stage, we print out the data t1 to make sure the coding and frequency count are correct for code 250.1. Once the first ICD-9 code index and frequency count is correct, we can start to use MACRO for all other diabetes ICD-9 codes.

```
%MACRO APPLE (CODE=, ANAME=, ID=, ICD=);
&CODE=0;
ARRAY &ANAME{15} DIAGNOSIS_1 DIAGNOSIS_2 DIAGNOSIS_3 DIAGNOSIS_4 DIAGNOSIS_5
DIAGNOSIS_6 DIAGNOSIS_7
DIAGNOSIS_8 DIAGNOSIS_9 DIAGNOSIS_10 DIAGNOSIS_11 DIAGNOSIS_12 DIAGNOSIS_13
DIAGNOSIS_14 DIAGNOSIS_15;
DO &ID=1 TO 15;
  IF &ANAME{&ID} ="&ICD" THEN &CODE=1;
END;
%MEND;
```

```

%APPLE (CODE=CODE250_0, ANAME=C0, ID=J, ICD=250)
%APPLE (CODE=CODE250_1, ANAME=C1, ID=K, ICD=250.1)
%APPLE (CODE=CODE250_2, ANAME=C2, ID=L, ICD=250.2)
%APPLE (CODE=CODE250_3, ANAME=C3, ID=M, ICD=250.3)
%APPLE (CODE=CODE250_4, ANAME=C4, ID=N, ICD=250.4)
%APPLE (CODE=CODE250_5, ANAME=C5, ID=O, ICD=250.5)
%APPLE (CODE=CODE250_6, ANAME=C6, ID=P, ICD=250.6)
%APPLE (CODE=CODE250_7, ANAME=C7, ID=Q, ICD=250.7)
%APPLE (CODE=CODE250_8, ANAME=C8, ID=I, ICD=250.8)
;

%MACRO PEAR ( CODE, NM=, );
PROC FREQ DATA=TWO;
TABLE ADM_YEAR* &CODE /NOCOL NOROW NOPERCENT OUT=&NM;
RUN;
PROC SORT DATA=&NM;
BY ADM_YEAR;
RUN;
%MEND;

%PEAR (CODE=CODE250_0, NM=T0)
%PEAR (CODE=CODE250_1, NM=T1)
%PEAR (CODE=CODE250_2, NM=T2)
%PEAR (CODE=CODE250_3, NM=T3)
%PEAR (CODE=CODE250_4, NM=T4)
%PEAR (CODE=CODE250_5, NM=T5)
%PEAR (CODE=CODE250_6, NM=T6)
%PEAR (CODE=CODE250_7, NM=T7)
%PEAR (CODE=CODE250_8, NM=T8)
;

```

In the first macro (named “apple”), we generated 9 variables named code250_0, code250_1, code250_2, code250_3, code250_4, code250_5, code250_6, code250_7, code250_8. They represent 9 different diabetic complications 250.0 to 250.8 as below.

```

ICD 250.0 = "DIABETES WITHOUT MENTION OF COMPLICATION"
ICD 250.1 = "DIABETES WITH KETOACIDOSIS"
ICD 250.2 = "DIABETES WITH HYPEROSMOLARITY"
ICD 250.3 = "DIABETES WITH OTHER COMA"
ICD 250.4 = "DIABETES WITH RENAL MANIFESTATIONS"
ICD 250.5 = "DIABETES WITH OPHTHALMIC MANIFESTATIONS"
ICD 250.6 = "DIABETES WITH NEUROLOGICAL MANIFESTATIONS"
ICD 250.7 = "DIABETES WITH PERIPHERAL CIRCULATORY DISORDERS"
ICD 250.8 = "DIABETES WITH OTHER SPECIFIED MANIFESTATIONS"

```

For example, if a subject had one or more diagnostics 1 to 15 valued 250.1 (diabetes with ketoacidosis), the new variable code250_1 will have value 1 for that subject at that visit. If not, the variable code250_1 will have value 0.

After reading all 15 diagnoses for each subject for all three years, the second macro “pear” will count the frequencies of each diagnosis by admission year. The macro variable &nm is used to represent the output data sets t0 to t8 for each diagnosis code. These data sets contain the frequency counts of each diabetic complication code.

3, MERGE THE DATA SETS FOR REPORT

Merge all 9 outputted data sets t0 to t8 by admission year, and drop the variables code250_0 to code250_8.

```

DATA THREE;
MERGE T0 (WHERE=(CODE250_0=1) RENAME=(COUNT=ICD250_0) DROP=PERCENT )
      T1 (WHERE=(CODE250_1=1) RENAME=(COUNT=ICD250_1) DROP=PERCENT )
      T2 (WHERE=(CODE250_2=1) RENAME=(COUNT=ICD250_2) DROP=PERCENT )
      T3 (WHERE=(CODE250_3=1) RENAME=(COUNT=ICD250_3) DROP=PERCENT )
      T4 (WHERE=(CODE250_4=1) RENAME=(COUNT=ICD250_4) DROP=PERCENT )
      T5 (WHERE=(CODE250_5=1) RENAME=(COUNT=ICD250_5) DROP=PERCENT )
      T6 (WHERE=(CODE250_6=1) RENAME=(COUNT=ICD250_6) DROP=PERCENT )
      T7 (WHERE=(CODE250_7=1) RENAME=(COUNT=ICD250_7) DROP=PERCENT )

```

```

T8 (WHERE=(CODE250_8=1) RENAME=(COUNT=ICD250_8) DROP=PERCENT )
;
BY ADM_YEAR;
DROP CODE250 CODE250_1 CODE250_2 CODE250_3 CODE250_4 CODE250_5 CODE250_6
CODE250_7 CODE250_8;

LABEL
  ICD250_0="DIABETES WITHOUT MENTION OF COMPLICATION"
  ICD250_1="DIABETES WITH KETOACIDOSIS"
  ICD250_2="DIABETES WITH HYPEROSMOLARITY"
  ICD250_3="DIABETES WITH OTHER COMA"
  ICD250_4="DIABETES WITH RENAL MANIFESTATIONS"
  ICD250_5="DIABETES WITH OPHTHALMIC MANIFESTATIONS"
  ICD250_6="DIABETES WITH NEUROLOGICAL MANIFESTATIONS"
  ICD250_7="DIABETES WITH PERIPHERAL CIRCULATORY DISORDERS"
  ICD250_8="DIABETES WITH OTHER SPECIFIED MANIFESTATIONS"
;
PROC PRINT DATA=THREE LABEL;
RUN;

```

During the data merge, only keep the codes 250_0 to code 250_8 which equal 1 and rename the count variable to the corresponding ICD code and drop the variable percent. Now the merged data contains only admission year and frequency count of each diagnosis code. The print-out will look like the following figure 1.

4, USE PROC EXPORT TO GENERATE THE EXCEL DATA, OR USE TAGSETS.EXCELXP AND PROC REPORT TO GENERATE FINAL REPORT.

Export to Excel for further cost analysis. The Exported Excel spreadsheet looks like the figure 1:

```

PROC EXPORT DATA= WORK.THREE
  OUTFILE= "C:\SESUG\DATAREPORT\ALLENCounter.XLS"
  DBMS=EXCEL REPLACE;
  SHEET="ALL";
RUN;

```

Adm_Year	ICD250_0	ICD250_1	ICD250_2	ICD250_3	ICD250_4	ICD250_5	ICD250_6	ICD250_7	ICD250_8
2004	8	220	37	268	15	6	18	70	2
2005	14	248	34	338	18	4	14	61	4
2006	15	307	30	383	35	1	2	3	5
2007	18	307	23	338	26	2	1	7	5
2008	27	318	12	374	13	4		5	2
2009	21	206	28	443	10	6	17	5	5
2010	27	195	29	488	3	2	3	1	3
2011	41	243	28	582	3	1	2	3	5
2012	19	133	32	322	3	1	1	2	

Figure 1. THE Excel table from PROC EXPORT

Export as final report:

```

ODS TAGSETS.EXCELXP FILE="C:\SESUG\DATAREPORT\ALL.XML" STYLE=PRINTER;
TITLE '2004 - 2012 DIABETIC PATIENT HOSPITAL VISIT BY EACH YEAR';
ODS TAGSETS.EXCELXP OPTIONS(EMBEDDED_TITLES='YES'
  SHEET_NAME='ALLENCounter'
  ABSOLUTE_COLUMN_WIDTH='15');

PROC REPORT NOWINDOWS DATA=WORK.THREE SPLIT='*' STYLE(summary)=header ;
COLUMNS Adm_Year ICD250_0 ICD250_1 ICD250_2 ICD250_3 ICD250_4 ICD250_5
ICD250_6 ICD250_7 ICD250_8 ;

DEFINE Adm_Year / DISPLAY CENTER 'Admission*Year' ;
DEFINE ICD250_0 / DISPLAY CENTER 'DIABETES WITHOUT*MENTION OF*
COMPLICATION';

```

```

DEFINE ICD250_1 / DISPLAY CENTER 'DIABETES WITH*KETOACIDOSIS';
DEFINE ICD250_2 / DISPLAY CENTER 'DIABETES WITH*HYPEROSMOLARITY';
DEFINE ICD250_3 / DISPLAY CENTER 'DIABETES WITH*OTHER COMA';
DEFINE ICD250_4 / DISPLAY CENTER 'DIABETES WITH*RENAL*MANIFESTATIONS';
DEFINE ICD250_5 / DISPLAY CENTER 'DIABETES WITH*OPHTHALMIC*
                                MANIFESTATIONS';
DEFINE ICD250_6 / DISPLAY CENTER 'DIABETES WITH*NEUROLOGICAL*
                                MANIFESTATIONS';
DEFINE ICD250_7 / DISPLAY CENTER 'DIABETES WITH*PERIPHERAL*
                                CIRCULATORY*DISORDERS';
DEFINE ICD250_8 / DISPLAY CENTER 'DIABETES WITH*OTHER SPECIFIED*
                                MANIFESTATIONS';

RUN; QUIT;

ODS TAGSETS.EXCELXP CLOSE;

```

2004 - 2012 Diabetic Patient Hospital Visit By Each Year

Admission Year	DIABETES WITHOUT MENTION OF COMPLICATION	DIABETES WITH KETOACIDOSIS	DIABETES WITH HYPEROSMOLARITY	DIABETES WITH OTHER COMA	DIABETES WITH RENAL MANIFESTATIONS	DIABETES WITH OPHTHALMIC MANIFESTATIONS	DIABETES WITH NEUROLOGICAL MANIFESTATIONS	DIABETES WITH PERIPHERAL CIRCULATORY DISORDERS	DIABETES WITH OTHER SPECIFIED MANIFESTATIONS
2004	8	220	37	268	15	6	18	70	2
2005	14	248	34	338	18	4	14	61	4
2006	15	307	30	383	35	1	2	3	5
2007	18	307	23	338	26	2	1	7	5
2008	27	318	12	374	13	4		5	2
2009	21	206	28	443	10	6	17	5	5
2010	27	195	29	488	3	2	3	1	3
2011	41	243	28	582	3	1	2	3	5
2012	19	133	32	322	3	1	1	2	

Figure 2. PROC REPORT Generated Formal Report for All Patient Visit Count Yearly in Excel

5, COUNT DIABETES DIAGNOSTIC CODES FOR EACH UNIQUE PATIENT FOR EACH YEAR.

To count each diagnosis code only once at each year for each patient no matter how many times this patient has been diagnosed that year, we first do the exact same steps as we have done.

5.1, Create a macro “APPLE” to count all 9 diabetic codes as we did in section 2.

5.2, Before running macro “PEAR”, we create a macro “PEACH”, which will count each patient diagnostic code by admission year. The output data will have a new variable “&ICD” to indicate if the count is larger or equal than 1. If the count is larger or equal than 1, the indicator variable “&ICD” will have value 1. This would make sure that no matter how many visits one subject had in that year, it is only counted as 1.

```

%MACRO PEACH (CODE=, PTCODE=, ICD=);
PROC FREQ DATA=THREE NOPRINT;
BY ADM_YEAR;
TABLE PATIENT_NUMBER* &CODE/OUT=&PTCODE NOCOL NOROW NOPERCENT SPARSE;
RUN;

```

```

DATA &PTCODE;
SET &PTCODE ;
IF &CODE=1;
IF COUNT>=1 THEN &ICD=1;
ELSE &ICD=0;
%MEND;

```

```

%PEACH (CODE=CODE250, PTCODE=PT250, ICD=ICD250)
%PEACH (CODE=CODE250_1, PTCODE=PT250_1, ICD=ICD250_1)
%PEACH (CODE=CODE250_2, PTCODE=PT250_2, ICD=ICD250_2)
%PEACH (CODE=CODE250_3, PTCODE=PT250_3, ICD=ICD250_3)
%PEACH (CODE=CODE250_4, PTCODE=PT250_4, ICD=ICD250_4)
%PEACH (CODE=CODE250_5, PTCODE=PT250_5, ICD=ICD250_5)
%PEACH (CODE=CODE250_6, PTCODE=PT250_6, ICD=ICD250_6)
%PEACH (CODE=CODE250_7, PTCODE=PT250_7, ICD=ICD250_7)
%PEACH (CODE=CODE250_8, PTCODE=PT250_8, ICD=ICD250_8)

```

```
;
```

5.3, Run the “PEAR” macro to count the frequencies. This time, macro “PEAR” will run for 9 different data sets that were created in the “PEACH” macro. This will generate 9 data sets containing frequency counts of each diabetic complication code for each year.

```

%MACRO PEAR ( PTCODE, ICD=, NM=, );
PROC FREQ DATA=&PTCODE NOPRINT;
TABLE ADM_YEAR* &ICD /NOCOL NOROW NOPERCENT OUT=&NM;
RUN;
PROC SORT DATA=&NM;
BY ADM_YEAR;
RUN;
%MEND;

%PEAR (PTCODE=PT250, ICD=ICD250, NM=P0)
%PEAR (PTCODE=PT250_1, ICD=ICD250_1, NM=P1)
%PEAR (PTCODE=PT250_2, ICD=ICD250_2, NM=P2)
%PEAR (PTCODE=PT250_3, ICD=ICD250_3, NM=P3)
%PEAR (PTCODE=PT250_4, ICD=ICD250_4, NM=P4)
%PEAR (PTCODE=PT250_5, ICD=ICD250_5, NM=P5)
%PEAR (PTCODE=PT250_6, ICD=ICD250_6, NM=P6)
%PEAR (PTCODE=PT250_7, ICD=ICD250_7, NM=P7)
%PEAR (PTCODE=PT250_8, ICD=ICD250_8, NM=P8)
;

```

5.4 Using TAGSETS.EXCELXP and PROC REPORT to generate final report. The result show as figure 3 in below.

*2004 - 2012 Diabetic Patient Hospital Visit By Each Year
Each Patient Count Only Once for ICD-9 Code and Each Year*

Admission Year	DIABETES WITHOUT MENTION OF COMPLICATION	DIABETES WITH KETOACIDOSIS	DIABETES WITH HYPEROSMOLARITY	DIABETES WITH OTHER COMA	DIABETES WITH RENAL MANIFESTATIONS	DIABETES WITH OPHTHALMIC MANIFESTATIONS	DIABETES WITH NEUROLOGICAL MANIFESTATIONS	DIABETES WITH PERIPHERAL CIRCULATORY DISORDERS	DIABETES WITH OTHER SPECIFIED MANIFESTATIONS
2004	3	79	13	63	8	5	15	30	2
2005	9	99	12	71	5	3	13	38	4
2006	9	112	13	86	7	1	2	3	4
2007	11	112	7	81	8	2	1	6	5
2008	13	109	5	86	6	4		4	2
2009	9	96	11	86	6	6	8	4	3
2010	15	94	13	99	3	2	3	1	3
2011	20	121	14	102	3	1	2	3	4
2012	10	74	18	87	2	1	1	2	

Figure 3. PROC REPORT Generated Formal Report for Unique Patient Count Yearly in Excel

CONCLUSION

The above examples have shown that using the combination of SAS PROC IMPORT, SAS ARRAY, PROC FREQ and OUPUT, SAS MACRO, and PROC EXPORT is very powerful and efficient in extracting information from large data sets for statistical analysis and reports.

CONTACT INFORMATION

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

Name: Lifang Zhang
 Enterprise: Department of Biostatistics and Epidemiology
 Address: 1120 15th Street, AE-1013
 City, State ZIP: Augusta, GA 30912
 Work Phone: 706-7214453
 Fax: 706-7216294
 E-mail: lizhang@augusta.edu
 Web: <http://www.augusta.edu/mcg/biostatapi/>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.