THE MCHP SAS PC TUTORIAL: APPENDICES

I. SIMULATED MANITOBA HEALTH DATA

- Prepare the data
    - Version 1 (unformatted)
- Record layout (variable definitions)
- Related files (formats and labels)
- Practice questions

Sample Data Sets: Simulated MB Health - Version 1

Two versions were created of the program that creates a temporary SAS data set called "test" from the simulated Manitoba Health data available at MCHP, this data can be used to answer the accompanying questions.

Version 1, shown below, does not include labels or formats, so the output shows a list of unlabelled variables (in the CONTENTS output) and unlabelled values (in the PRINT output)

Information regarding record layouts (for raw files) and formats and labels (for all types of files) is normally available to the user as well. Version 2 adds labels (for labelling variables) and formats (for labelling values) from several files . The formats and labels specific to the simulated Manitoba Health raw data set are based on a record layout that describes all 33 variables.

The following steps are required to create the output for Version 1:

1. Ensure that the file called "fivet93m.raw" containing values for 5,000 observations is accessible (e.g., on A:\ or C:\ drive).
2. Enter the following program directly into the SAS Program Editor window (the comments do not need to be entered). Two changes should be made:
   - The statement *options linesize=min;* does not need to be entered; this was used to reduce the width of the output so that it would fit on the website.
   - The FILENAME statement should be changed to reflect the actual location of the data on your system.

```
options linesize=min;

**================================== **
** f=readmb.sas                      **
**                                   **
** This program creates a temporary  **
** SAS data set from the simulated   **
** Manitoba Health data              **
**==================================**;

    /* Assign a name ("rawdata") to the path and file */

filename rawdata
'c:\My Documents\My SAS Files\sasmanual\data\fivet93m.raw';

                    /* Begin the DATA step */

  DATA TEST;              /* Create "test" data set */
   INFILE RAWDATA;        /* Read the external data */
```

```
        /* Assign variable names, types & locations */
    INPUT
        NCASE     $  1-5
        GENDER    $    6
        AGE          7-9
        LOS        10-13
        NDAYICU    14-17
        TRANADM   $   18
        TRANDIS   $   19
        OP01      $ 20-23  /* This is numeric 0 */
        DIAG01    $ 24-28  /* This is numeric 0 */
        DIAG02    $ 29-33  /* This is numeric 0 */
        TOTCHAR    34-35
        CHARYES    $ 37
        SCHEDULE   $ 38
        RISKDRG    $ 39
        @40 DRGWT 6.4  /* 6 columns, incl. 4 decimal*/
        DRG       $ 46-48
        WBURG      $ 49
        DISCRE     $ 50
        REGIONRE   $ 51
        REGIONH    $ 52
        INCDR      $ 53
        TYPEHSP    $ 54
        DEATHSEP   55-58
        DRGRGN    $ 59-63
        SEVERDRG   $ 64
                @65 RDRGWT 6.4
        CMG       $ 71-73
                @74 RIW 7.4
        TREATY     $ 81
        ABSTYPE    $ 82
        DAYSFRPR   83-85
        DAYSTOR    86-88
        @89 ICD17BRK $char2. /* Keep leading spaces */
;                    /* End of INPUT statement*/
run;                   /* End of DATA step      */

     /* Obtain general information about the data set */
proc contents data=test;      /* Begin PROC step       */
run;                    /* End PROC step         */

/*Obtain values of all variables for the 1st 10 records*/
proc print data=test (obs=10);  /* Begin PROC step     */
run;                     /* End PROC step       */
```

3. Save the program (e.g., *mb_cr.sas*; do not save it as fivet93m.raw because this would overwrite the raw data set).
4. Submit the program for processing.
5. Check both log and output windows to ensure the program ran accurately. Debug the program, if necessary, saving and submitting it again (clearing the log and output windows first so that only the most recently-submitted versions will appear).

The following computer files are available from MCHPE. Note that results may occasionally display truncated labels, depending on the procedure. It is important to ensure, within the SAS program, that the file name is correctly specified to permit SAS to read the file.

File Name	Size	Format Name	Description
fivet93m.raw	460,000	not applicable	data set containing values for 5,000 observations and 33 variables.
dx93xfmt.sas	34,039	$DIAGXL	value labels for 3-digit ICD-9-CM diagnoses (1980).
op93fmt.sas	29,371	$OPL	value labels for 4-digit ICD-9-CM procedures (1980).
cmg93fmt.sas	23,285	$CMGL	value labels for CMGs (Case Mix Groups) (1991).
drg93fmt.sas	21,058	$DRGL	value labels for DRGs (Diagnosis Related Groups) (5th revision).
lbls93.sas	2,270	not applicable	variable labels
fmts95.sas	3,132		assorted value labels for other variables i the data set.

Exercises on simulated MB Health data

Prior to completing the exercises, several steps are needed to prepare the data:

- Open the lbls93 file in the Program Editor window to comment out the FORMAT statement (i.e., add * to the beginning of the statement). Save the revised file and clear the Program Editor window. The original values, rather than the formatted values of variables can thus be referenced in the programs (it is simpler, for example, to refer to *regionre='1'* rather than *regionre='central Manitoba'*).
- Open the program that creates the temporary SAS data set "test" from the simulated Manitoba Health data set into the Program Editor window. No changes need to be made to this program.
- Submit the program and check the log for messages; the log should indicate that a temporary SAS data set called "test" (in the WORK library) was created for use for this SAS session.

It is assumed that the questions are completed during the course of one SAS session. If not, the data set must be re-created for the next SAS session, as well as the formats (the record layout shows which formats correspond with each of the variables in the data set).

Programs can be developed and tested in a number of different ways. If the programming for all the questions below is saved into one file, the user might, rather than submitting the entire file to test only portions of code, instead highlight the portion to be tested before pressing the submit key. The resulting log and output can thus be checked to ensure the code is accurate, before keeping it as part of the larger program.

For each of the questions, add: 1) a title descriptive of the data set being used, and 2) either a second title or a footnote indicating the question number. The same title can be used for each question, so there is no need to repeat the TITLE1 statement for the other questions (SAS will automatically keep the same title for the duration of the SAS session unless instructed otherwise).

1. Produce the following listings of data:

   - For the first 20 observations, specify the following variables to be shown on the output (original values): *gender*, *age*, *los*, *op01*, *diag01*, and *diag02*.
   - Sort the data by *gender* and *regionre* and produce a listing of the first 40 observations. Display only *ncase*, *gender*, *regionre*, and *icd17brk* in the output. This time display the formatted, or labelled, values rather than the original values for all except *ncase*.

2. For a later exercise, utilization for Winnipeg vs non-Winnipeg residents will be compared. Create two formats, one that will be used to group *regionre* into new values and one that will be used to label the new values:

- Name the grouping format *$wpgf*; this format should be able to group the Winnipeg value into '1' and non-Winnipeg values into '0',
- Name the labelling format *$wpgl*; this format should be able to label each of the two new values.
  Although this question could be done using only one format (i.e., specifying the label 'Winnipeg' in the first format instead of '1'), the two-step process is typically used, for example, to simplify specification of values of the new variable within a SAS program - e.g., to be able to use '1' within a line of code rather than 'Winnipeg' to reference Winnipeg records.

3. Obtain information on the number of observations and the mean, minimum, and maximum values, setting maximum decimal places to 2 for the following:

   - The variables for age, length of stay, and days to death.
   - Note the skewed results for *deathsep*. The value of 9999 actually refers to those still alive. Run a program for this variable only, including a WHERE statement to keep only the values which are less than 9999.
   - The variables for age and length of stay, this time showing the results by region of residence. Use the region format to attach labels to region of residence.

4. How does the distribution of hospital discharges for selected categories of ICD-9-CM diagnoses *icd17brk* differ by gender *gender*? Display the information using original values and again using formatted values.

5. Examine the relationship among variables for the following:
   - Is the presence of high-risk diagnoses on admission *charyes* associated with neighbourhood income level *incdr*? Display the formatted values for both variables.
   - How does the relationship between these two variables differ by gender (use the formatted value for this variable as well)?

6. Develop a program that will create the following new variables (always within a data step):

   - *loswks* - a numeric variable that has values of length of stay calculated in weeks.
   - *losgroup* - a character variable that groups length of stay into 3 categories (0 to 30 days, 31 to 365 days, and 366+ days). A grouping format can be created, and a labelling format; this PROC FORMAT step will need to go before the data step creating these variables.
   - *wpgres* - a character variable created from region of residence that uses the previously created $wpggrp format.
   - *diag3x* - a character variable created from *diag01* that will include only the first 3 digits.
   - *op2x* - a character variable created from *op01* that will include only the first 2 digits.

Also create labels for each of the 5 new variables within the same data step. Before submitting the DATA step, add the PROCs in the next question to the program.

7. Check the new variables:
   - For *losgroup* and *wpgres*, use a side-by-side listing (PROC FREQ) to compare original variables against the new variables, ensuring that labelled values are used for the 3 character variables. Both comparisons can be run within the same PROC FREQ.
   - For *loswks* (the only new numeric variable) and *los*, run a PROC MEANS.
   - For the remaining two character variables, run a PROC PRINT for the first 30 observations, showing both original and new variables (i.e., output for a total of 4 variables).
   - Do a PROC CONTENTS on the data set to ensure the new variables were properly labelled.

The program, log, and output are all available for the above questions. For additional practice, another set of more research-focused questions has been developed.

Record Layout for the Simulated Manitoba Health Data Set

This record layout provides the following information, in alphabetical order, for the 33 variables in the simulated Manitoba Health data set.

- Names assigned to variables.
- Column numbers, reflecting how much space is required for the values of each variable.
- Type of variable, denoting whether the values should be read in as numeric (N) or alpha/numeric (A/N).
- General description of variable.
- Data values for each variable, including format names and files, where applicable.

Unless otherwise specified, the formats are available in the file titled *fmts95*.

Variable Name	Columns	Type	Variable Description	Data Values
ABSTYPE	82	A/N	Abstract type	1 - Adult/child 2 - Pediatric 3 - Obstetric 4 - Newborn/stillborn 5 - Rehab/respiratory Format: $ABSTYPEL
AGE	7-9	N	Patient age at discharge	Range: 0 to 103
CHARYES	36-37	N	Had a high-risk diagnosis (Charlson Index - Charlson et al., 1987)	0 - No 1 - Yes Format: $CHARL
CMG	71-73	A/N	CIHI's Case Mix Group - a Canadian version of DRG (somewhat parallel)	Range: 1 to 999 Format labels are in CMG93C.FMT file
DAYSFRPR	83-85	N	Interval in days from discharge date of previous admission to current admission	0 to 365 - number of days from previous admission 999 - no previous admission within the last year
DAYSTOR	86-88	N	Interval in days from current admission to readmission	0 to 364 - number of days to readmission 999 - no readmission within the subsequent year

DEATHSEP	55-58	N	Days from hospital separation to death	0 - died in hospital 1 to 449 - days to death 9999 - alive as of 500 days after discharge
DIAG01	24-28	A/N	ICD-9-CM diagnosis code: most responsible diagnosis	Range: 001 to 999.9 Format labels are in DX93.FMT file
DIAG02	29-33	A/N	ICD-9-CM diagnosis code: second diagnosis	Same as for DIAG01
DISCRE	50	N	Discretionary diagnoses (Anderson et al., 1986)	1 - discretionary 0 - other Format: $DISCREL
DRG	46-48	A/N	Diagnosis-Related Group	Range: 1 to 477 Format labels are in DRG93C.FMT file
DRGRGN	59-63	A/N	DRG Refinement Group Number (RDRG) - a 3-digit DRG followed by a "class" digit	Range: 0010 to 9230
DRGWT	40-45	N	DRG weight: relative weights based on a mainly 65+ population, scaled so that 1.00 is standard (in the U.S., this determines how much hospitals are paid for patients of this type).	Range: 0 to 7.5631
GENDER	6	A/N	Patient gender	1 - male 2 - female Format: $GENDERL
ICD17BRK	89-90	A/N	ICD-9-CM diagnoses grouped into 17 categories (for the actual diagnostic codes, see DIAG01 and DIAG02).	Range: 1 to 17 (includes ICD-9-CM diagnoses 1 to 999.9, excluding V- and E-codes) Format: $ICD17L

INCDR	53	A/N	Neighbourhood income quintile, based on mean household income assigned to postal code from aggregate enumeration area data	1-less than or equal to $22,300 2-$22,400 to $27,500 3-$27,600 to $33,300 4-$33,400 to $41,200 5-greater than $41,200 Format: $INCDRL
LOS	10-13	N	Length of hospital stay	Range: 1 to 902
NCASE	1-5	A/N	Case number (unique identifier)	Range: 1 to 5000
NDAYICU	14-17	N	Number of days in Intensive Care	Range: 0 to 73
OP01	20-23	A/N	ICD-9-CM procedure code: primary operative procedure	Range: 01 to 99.99 Format labels are in OP93.FMT file
RDRGWT	65-70	N	Refined DRG weight: based on LOS data for each DRG, but this has not been scaled (like DRGWT), and is based on a large 1986 US sample mainly age 65+.	Range: 0.1433 to 5.515 (the higher the value, the longer the expected stay).
REGIONH	52	A/N	Hospital region	1 - central Manitoba 2 - eastern Manitoba 3 - Interlake 4 - northern Manitoba 5 - Parkland 6 - Thompson 7 - Westman 8 - Winnipeg Format: $REGIONL
REGIONRE	52	A/N	Patient's region of residence	Same as REGIONH
RISKDRG	39	A/N	Score identifying risk of adverse event associated with admission (see info sheet)	0 - not classified 1 - low risk of adverse event 2 3 4 - high risk of adverse event

RIW	74-80	N	Resource Intensity Weight: based on 1985 New York cost data, to be used in conjunction with CMGs	Range: 0.1366 to 10.729
SCHEDULE	38	A/N	Scheduled admission (usually elective surgery)	1 - not scheduled 2 - scheduled Format: $SCHEDL
SEVRDRG	64	A/N	DRG severity code (the last digit of the DRGRGN variable)	0 - no or minor complications/ comorbidities 1 - moderate c/c 2 - major c/c 3 - catastrophic c/c Format: $SEVRL
TOTCHAR	34-35	A/N	Charlson Index score: weighted score on high-risk diagnosis index (Charlson et al., 1987)	Range: 0 to 8
TRANADM	18	A/N	Patient transferred from other institution	0 - no transfer/other 1 - hospital 2 - personal care home (PCH) Format: $TRNADML
TRANDIS	19	A/N	Patient transferred to other location	0 - no transfer/other 1 - another hospital 2 - personal care home (PCH) 3 - died in hospital Format: $TRNDISL
TREATY	81	A/N	Treaty First Nations: defined using municipal code, does not necessarily denote residence	0 - non-Treaty 1 - Treaty Format: $TREATYL
TYPEHSP	54	A/N	Type of hospital	1 - Winnipeg teaching 2 - Winnipeg non-teaching 3 - other Manitoba Format: $TYPEHL

WBURG	49	A/N	Wennberg's variation code (Fisher et al., 1992)	1 - low-variation medical condition 2 - surgery 3 - high-variation medical condition 4 - obstetric (DRG 370-379) Format: $WBURGL