

## PH-92: SAS Programming And Generic Techniques for Cohort Creation and Consort Diagrams

By Brandy Sinco, BS, MA, MS

### Example: Create Cohort from National Cancer Database

#### Step 1: SAS Data Step to Compute Variables (ends on P. 8).

Note important programming techniques:

- Converting character variable to numeric in SAS (year of diagnosis).
- Create time variable with 0 at origin, YearsAfter2012.
- Naming variable after PI who created the algorithm (Histology\_TW).
- Use of Attribute statements in SAS to definite variable labels and formats.

```
data NCDB_PUF0_CompVar;  
set library.NCDB_PUF;
```

```
Attrib AgeDiag Label="Age at Diagnosis" Format=7.0;  
Attrib Grade_TW2 Label='Tumor Grade, Ton Wang Method' Format=5.0;  
Attrib YearsAfter2012 Label="Years After 2012" Format=7.0;  
Attrib YearsAfter2012_2 Label="Years After 2012 sqr" Format=7.0;  
Attrib YearsAfter2012_3 Label="Years After 2012 cube" Format=7.0;  
Attrib Yr Label='Year of Diagnosis' format=7.0;
```

```
/* Convert Year of Diagnosis and Age from Character to Numeric Variables */
```

```
Yr = YEAR_OF_DIAGNOSIS + 0;  
YearsAfter2012 = yr - 2012;  
YearsAfter2012_2 = YearsAfter2012**2;  
YearsAfter2012_3 = YearsAfter2012**3;
```

```
/* Compute Histology_TW2 */
```

```
if HISTOLOGY in ('8022', '8035', '8230', '8500', '8501', '8502', '8503', '8504', '8507',  
'8508', '8523') then Histology_TW2 =1; /* ductal */  
if HISTOLOGY in ('8520', '8521', '8524', '8525') then Histology_TW2=2; /* lobular. */  
if HISTOLOGY = '8522' then Histology_TW2=3; /* both. */
```

```
if HISTOLOGY not in ('8022', '8035', '8230', '8500', '8501', '8502', '8503', '8504', '8507',  
'8508', '8520', '8521', '8522', '8523', '8524', '8525') then Histology_TW2=9; /* Others */
```

```
/* EXCLUDE phyllodes (9020), EXCLUDE sarcomas (8800, 9120, 9170, 9130, 8900,
8920, 8910, 8901, 8850, 8890, 9180) */
run;
```

**Step 2: Use of Where Statement to Select the Cohort, Based on Computed Variables. Followed by a series of delete statements for the exclusion criteria.**

```
/* Apply inclusion/exclusion criteria. */
Data NCDB_PUF0;
Set NCDB_PUF0_CompVar;
*** Inclusion Criteria, Society for Surgical Oncology Paper ***;
Where ('2012'<=YEAR_OF_DIAGNOSIS<='2020')
  AND SEX='2' /* Female, Criterion 1 */
  AND ('018'<=AGE<='090') /* Age 18+, Criterion 2 */
  AND not (CLASS_OF_CASE='00') /* Care at same facility, Criterion 3 */
  AND ('01'<=REGIONAL_NODES_EXAMINED<='40') /* Criterion 4 */

  /* Clinical N1 or N2, Criterion 5 */
  AND (TNM_CLIN_N in ('1', '2', '2B', 'C1', 'C2', 'c1', 'c1A', 'c2', 'C2A', 'C2B') OR
    AJCC_TNM_CLIN_N IN ('1', 'N1', 'c1a', 'cN1', 'cN1mi', 'cN2', 'cN2a', 'cN2b'))

  /* Include only women who had SLNB (2,4 ), ALND (5), both (6,7) aka Axillary
  Surgery, Criterion 6 */
  AND RX_SUMM_SCOPE_REG_LN_2012 in ('2', '4', '5', '6', '7')

  /* Chemo Before Surgery, Criterion 7 */
  AND ChemoBeforeSurg=1

  /* Pathologic T0-T4 disease, Criterion 8 */
  AND 'T0'<=PathTStage<='T4'

  /* Clinical T1-T4 disease, Criterion 9 */
  AND 'T1'<=ClinTStage<='T4'

  /* HISTOLOGY values Invasive ductal or lobular, Criterion 10 */
  AND HISTOLOGY IN
  ('8000', '8010', '8022', '8032', '8035', '8041', '8050', '8070', '8140', '8200',
  '8201', '8211', '8230', '8246', '8255', '8260', '8290', '8314', '8315', '8401',
  '8410', '8430', '8480', '8481', '8500', '8501', '8502', '8503', '8504', '8507',
```

'8510', '8513', '8520', '8521', '8522', '8523', '8524', '8525', '8530', '8540',  
'8541', '8543', '8550', '8560', '8570', '8571', '8572', '8574', '8575', '8982',  
'8983', '8990');

**\*\*\* Exclusion Criteria \*\*\*;**

/\* Criterion 1: Clinical stage 4 \*/

if (TNM\_CLIN\_STAGE\_GROUP in ('4', '4A', '4A1', '4A2', '4B', '4C') and TNM\_CLIN\_M =  
'p1') then delete;

if AJCC\_TNM\_CLIN\_STG\_GRP = '4' and AJCC\_TNM\_CLIN\_M in ('cM1', 'pM1') then  
delete;

/\* Criterion 2: Inflammatory Breast Cancer \*/

if TNM\_CLIN\_T in ('c4D', 'cT4d') then delete;

if AJCC\_TNM\_CLIN\_T in ('T4d', 'cT4d') then delete;

/\* Criterion 3: Clinical N3 disease, dont need because we included only N1 and N2 \*/

if ClinNStage = 'N3' then delete;

/\* Criterion 4: Clinical T0 disease \*/

if ClinTStage = 'T0' then delete;

Run;

**Step 3: Select Facilities with Average Annual Case Count  $\geq$  10.**

/\* Compute Hospital Volume by Year for New Data \*/

Proc SQL;

Create Table HospitalVolumeYear2012\_2020 AS

SELECT DISTINCT PUF\_FACILITY\_ID, YEAR\_OF\_DIAGNOSIS,  
Count(PUF\_CASE\_ID) as CaseCount label="Case Count" format=7.2  
FROM library.NCDB\_PUF

WHERE (SEX='2' AND '2012'<=YEAR\_OF\_DIAGNOSIS<='2020' AND  
'018'<=AGE<='090') /\* Female, Diagnosis Years \*/

GROUP BY PUF\_FACILITY\_ID, YEAR\_OF\_DIAGNOSIS

ORDER BY PUF\_FACILITY\_ID, YEAR\_OF\_DIAGNOSIS;

Quit;

/\* Compute Average Hospital Volume by Year for New Data \*/

Proc SQL;

Create Table HospitalAveVolume2012\_2020 AS

```
SELECT DISTINCT PUF_FACILITY_ID, Mean(CaseCount) as MeanCaseCount
label="Mean Case Count" format=7.2
FROM HospitalVolumeYear2012_2020
GROUP BY PUF_FACILITY_ID
ORDER BY PUF_FACILITY_ID; Quit;
/* 1329 facilities */
```

```
Data HospitalAveVolume2012_2020;
set HospitalAveVolume2012_2020;
if MeanCaseCount<10 then delete;
Run;
/* 1329 facilities, 12 facilities have ave <10, so 1317 total facilities */
```

```
/* Store hospital case volume dataset for future use */
Proc Copy IN=Work out=Library;
Select HospitalVolumeYear2012_2020 HospitalAveVolume2012_2020;
Run;
```

```
/* Final Cohort, Excluding Facilities with Average Annual Case Volume < 10 */
Data NCDB_PUF;
Merge NCDB_PUF0(IN=N) library.HospitalAveVolume2012_2020(IN=C);
by PUF_FACILITY_ID;
if (C=0 or N=0) then delete;
Run;
```