SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL

Moving Along in Health Research:

Applying PROC EXPAND to Medical Encounter Data

USERS PROGRAM

Moving Along in Health Research: Applying PROC EXPAND to Medical Encounter Data

Rayna K. Matsuno, PhD, MPH

Henry M. Jackson Foundation, Bethesda, MD

ABSTRACT

PROC EXPAND is very useful when handling time series data – a series of successive, equally spaced data points – and is commonly used in fields such as finance or economics, but it can also be applied to medical encounter data within a health research setting. Medical encounter data includes detailed information about health care services provided to a patient by a managed care entity and is a rich resource for epidemiologic research. Specific data items include, but are not limited to, dates of service, procedures performed, diagnoses, and costs associated with services provided. Because epidemiologic studies generally focus on a particular health condition, a researcher might wish to distinguish individuals with the health condition of interest by identifying encounters with a defining diagnosis and/or procedure. In this presentation, I will provide two examples of how cases can be identified from a medical encounter database. The first utilizes a relatively simple case definition, and then I will EXPAND the example to a more complex case definition.

ABOUT MEDICAL ENCOUNTER DATA

What is medical encounter data?

Medical encounter data includes information about an insured person's eligibility and their health care claims. It provides details about that person's engagement with the health care system, including such "encounters" as visits with their doctor or drug prescriptions. Although medical encounter data is typically used by payors to track utilization of services and health care costs, it is increasingly being used for research purposes, particularly in the field of epidemiology.

How is medical encounter data used in epidemiological studies?

Encounter data has been used to conduct studies on patterns of care, health outcomes, and cost-effectiveness. Specific studies might include describing the diagnosis and management of chronic back pain; the evaluation of surgical complications among endometrial cancer patients undergoing minimally invasive versus open surgery; or the cost-effectiveness of total mastectomy versus lumpectomy + radiotherapy among early stage breast cancer patients. In these examples – as is often the case in epidemiological studies – the study population includes individuals with a specific health condition of interest. Encounter data can be used to identify patients who meet a specified case definition.

How is medical encounter data typically structured?

Generally, encounter data sets are long files, with each row corresponding to a single encounter.

WHY USE PROC EXPAND?

PROC EXPAND can be used to interpolate missing data, collapse higher frequency intervals to lower frequency intervals, transform data, and much more. For example, in a finance setting, one might use PROC EXPAND to collapse monthly sales data to quarterly sales data. When defining individuals with a health condition using medical encounter data, equal spacing between encounters is not important. In many cases, what matters is that a certain number of encounters with a defining diagnosis occurs within a specified time frame. For instance, as we will see in Example 1, a case might be defined as having two encounters with a defining diagnosis within 180 days of each other. While the LAG or LEAD functions could be used within a DATA STEP, additional programming would be necessary to handle the crossing from one BY-group (i.e., patient) into the next. PROC EXPAND is a very efficient alternative and can easily be used to add earlier or later observations from a long dataset on the same line within a BY-group to allow for computations.

EXAMPLES

What are real-world examples of case definitions?

The Defense Health Agency, Armed Forces Health Surveillance Branch (AFHSB), conducts medical surveillance research on U.S. military and military-associated populations. Many studies use data from the Military Health System Data Repository (MDR), an administrative database of medical encounters for eligible service members and their families, and rely on well-defined case definitions in order to describe trends in frequencies and rates of selected health conditions.

Example 1: Post-traumatic Stress Disorder (PTSD)

Two outpatient medical encounters, within 180 days of each other, with any of the defining diagnoses of PTSD in the *first or second* diagnostic position.

(ICD-9-CM code 309.81)

Example 2: Lung Cancer

One hospitalization with any of the defining diagnoses of lung cancer in the *primary* diagnostic position; or

One hospitalization with a V-code indicating a radiotherapy, chemotherapy, or immunotherapy treatment procedure in the *primary* diagnostic position; AND any of the defining diagnoses of lung cancer in the *secondary* diagnostic position; or

Three or more outpatient medical encounters, occurring *within a 90-day period*, with any of the defining diagnoses of lung cancer in the *primary or secondary* diagnostic position.

(ICD-9-CM codes for lung cancer: 162.2 to 162.9 // ICD-9-CM codes for treatments: V58.0, V58.1)

Moving Along in Health Research: Applying PROC EXPAND to Medical Encounter Data

Rayna K. Matsuno, PhD, MPH

Henry M. Jackson Foundation, Bethesda, MD

EXAMPLE 1: Identify Patients with PTSD

Each row corresponds to a single outpatient encounter

ID	date	dx1	dx2
1	28-Mar-14	29630	30981
1	1-Dec-14	29630	30981
1	3-May-15	29630	
2	10-Aug-11	71945	
2	11-Oct-11	71945	30981
2	18-Jan-12	71945	30981
4	12-Nov-07	30023	
5	24-Oct-11	30981	
6	8-Jan-13	30981	29633
6	15-Jan-13	30981	29633
6	22-Jan-13	30981	29633
6	29-Jan-08	30981	29633
9	10-Mar-06	30482	
9	16-Mar-06	30482	

```
** Subset relevant outpatient encounters **;
data outpat2; set outpat1;
  if substr(dx1,1,5) in('30981') |
     substr(dx2,1,5) in('30981') then ptsd=1;
  if ptsd=1 then output;
  keep id date dx1 dx2 ptsd;
run;
proc sort data=outpat; by id date; run;
```

Now we have a subset of patients with the defining ICD-9-CM code

Two outpatient medical encounters, within 180 days of each other, with any of the defining diagnoses of PTSD in the *first or second* diagnostic position.

ICD-9-CM code 309.81

ID	date	dx1	dx2	ptsd
1	28-Mar-14	29630	30981	1
1	1-Dec-14	29630	30981	1
2	11-Oct-11	71945	30981	1
2	18-Jan-12	71945	30981	1
5	24-Oct-11	30981		1
6	8-Jan-13	30981	29633	1
6	15-Jan-13	30981	29633	1
6	22-Jan-13	30981	29633	1
6	29-Jan-08	30981	29633	1

```
** Apply diagnostic criteria and flag **;
proc expand data=outpat2 out=expanded1 method=none;
  by id;
  convert date = date_lead1 / transformout=(lead 1);
run;
data expanded2; set expanded1;
  flag=0;
  if date_lead1 ne . & (date_lead1-date<= 180) then flag=1;
  drop dx1 dx2 ptsd;
run;
```

ID	TIME	date	date_lead1	flag
1	0	28-Mar-14	1-Dec-14	0
1	1	1-Dec-14	.	0
2	0	11-Oct-11	18-Jan-12	1
2	1	18-Jan-12	.	0
5	0	24-Oct-11	.	0
6	0	8-Jan-13	15-Jan-13	1
6	1	15-Jan-13	22-Jan-13	1
6	2	22-Jan-13	29-Jan-08	1
6	3	29-Jan-08	.	0

→ Patients 2 and 6 are cases

Moving Along in Health Research: Applying PROC EXPAND to Medical Encounter Data

Rayna K. Matsuno, PhD, MPH

Henry M. Jackson Foundation, Bethesda, MD

EXAMPLE 2: Identify Patients with Lung Cancer

ID	date	dx1	dx2	lung	source
1	28-May-08	1625	51881	1	I
2	5-Dec-07	1623	5990	1	I
2	26-Jun-10	1623		1	I
2	2-Jul-10	1623		1	O
3	3-Oct-07	1625	V1582	1	O
3	4-Oct-07	1629	7862	1	O
3	10-Oct-07	1629		1	O
3	11-Oct-07	1629		1	O
3	15-Oct-07	1629	27542	1	O
3	16-Oct-07	1629		1	O
4	3-Jan-11	1629		1	O
4	16-Mar-11	1629		1	O
4	7-May-11	1629		1	O

One hospitalization with any of the defining diagnoses of lung cancer in the *primary* diagnostic position

-OR-

One hospitalization with a V-code indicating a radiotherapy, chemotherapy, or immunotherapy treatment procedure in the *primary* diagnostic position; AND any of the defining diagnoses of lung cancer in the *secondary* diagnostic position

-OR-

Three or more outpatient medical encounters, occurring *within 90-day period* with any of the defining diagnoses of lung cancer in the *primary or secondary* diagnostic position.

ID	TIME	ID	date	TIME	date_lag1	date_lead1	date_lead2	source	source	flag	flag
1	0	1	28-May-08	0	28-May-08	.	.	I	I	1	1
2	0	2	5-Dec-07	0	5-Dec-07	26-Jun-10	2-Jul-10	I	I	1	1
2	1	2	26-Jun-10	1	26-Jun-10	2-Jul-10	.	I	I	1	1
2	2	2	2-Jul-10	2	2-Jul-10	.	.	O	O	0	0
3	0	3	3-Oct-07	0	3-Oct-07	4-Oct-07	10-Oct-07	O	O	1	0
3	1	3	4-Oct-07	1	4-Oct-07	10-Oct-07	11-Oct-07	O	O	1	1
3	2	3	10-Oct-07	2	10-Oct-07	11-Oct-07	15-Oct-07	O	O	1	1
3	3	3	11-Oct-07	3	11-Oct-07	15-Oct-07	16-Oct-07	O	O	1	1
3	4	3	15-Oct-07	4	15-Oct-07	16-Oct-07	.	O	O	0	1
3	5	3	16-Oct-07	5	16-Oct-07	.	.	O	O	0	0
4	0	4	3-Jan-11	0	3-Jan-11	16-Mar-11	7-May-11	O	O	0	0
4	1	4	16-Mar-11	1	16-Mar-11	7-May-11	.	O	O	0	0
4	2	4	7-May-11	2	7-May-11	.	.	O	O	0	0

```
** Apply diagnostic criteria with lag1 (ahead) **;
proc expand data=alldata out=expanded1 method=none;
  by id;
  convert date = date_lag1 / transformout=(lag 1);
  convert date = date_lead1 / transformout=(lead 1);
run;
data expanded2; set expanded1;
  flag=0;
  if source='I' then flag=1;
  if source='I' then flag=1;
  if (date_lead1 - date_lag1 <= 90) then flag=1;
  end;
run;
```

ICD-9-CM codes for lung cancer: 162.2 to 162.9

ICD-9-CM codes for treatments: V58.0, V58.1

→ Patients 1, 2, and 3 are cases

Moving Along in Health Research: Applying PROC EXPAND to Medical Encounter Data

Rayna K. Matsuno, PhD, MPH

Henry M. Jackson Foundation, Bethesda, MD

CONCLUSION

PROC EXPAND is extremely useful and has applications in epidemiologic research that utilizes medical encounter data, which can be treated as time series with unequal spacing between data points (i.e., encounters). Using the CONVERT statement within PROC EXPAND is an efficient way to identify encounters that occur within a certain time frame. As more and more epidemiologic research takes advantage of medical encounter data, PROC EXPAND may prove to be an excellent addition to a SAS programmer's tool box.

REFERENCES

SAS References

SAS/ETS® 9.3 User's Guide Documentation:

http://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_expand_sect001.htm

AFHSB case definitions (http://www.health.mil/Military-Health-Topics/Health-Readiness/Armed-Forces-Health-Surveillance-Branch/Epidemiology-and-Analysis/Surveillance-Case-Definitions)

RESOURCES

Chronic Conditions Data Warehouse (CCW)

A research database that includes Medicare and Medicaid beneficiary, claims, and assessment data for specific, predefined cohorts with a chronic health condition of interest. There are 27 chronic condition categories, including asthma, certain cancers, diabetes, glaucoma, and heart failure.

URL: https://www.ccwdata.org

https://www.ccwdata.org/web/guest/condition-categories

SEER-Medicare linked data

Includes health data for Medicare beneficiaries with cancer through a linkage between the Surveillance, Epidemiology and End Results (SEER) and Medicare claims databases.

URL: https://healthcaredelivery.cancer.gov/seermedicare/

Military Health System Data Repository (MDR)

Centralized data repository for U.S. Defense Health Agency health care data. Includes medical encounter data for military service members, their families, and other TRICARE beneficiaries.

URL: http://www.health.mil/Military-Health-Topics/Technology/Clinical-Support/Military-Health-System-Data-Repository

Armed Forces Health Surveillance Branch (AFHSB)

AFHSB is a branch of the Public Health Division within the U.S. Defense Health Agency. Researchers conduct health surveillance of the U.S. Armed Forces population as well as global emerging infections surveillance. AFHSB also maintains the *Medical Surveillance Monthly Report (MSMR)*, a monthly peer-reviewed journal.

URL: http://www.health.mil/afhsb

Moving Along in Health Research: Applying PROC EXPAND to Medical Encounter Data

Rayna K. Matsuno, PhD, MPH

Henry M. Jackson Foundation, Bethesda, MD

FULL CODE FOR EXAMPLE 1 (PTSD)

```
*****************************************************
 1: Subset relevant outpatient encounters
*****************************************************;
data outpat2;
  set outpat1;
  if substr(dx1,1,5) in('30981') |
     substr(dx2,1,5) in('30981') then ptsd=1;
  if ptsd=1 then output;
  keep id date dx1 dx2 ptsd;
run;
proc sort data=outpat; by id date; run;

*****************************************************
 2: Apply diagnostic criteria and flag
*****************************************************;
proc expand data=outpat2 out=expanded1 method=none;
  by id;
  convert date = date_lead1 / transformout=(lead 1);
run;
data expanded2; set expanded1;
  flag=0;
  if date_lead1 ne . & (date_lead1-date<= 180) then flag=1;
  drop dx1 dx2 ptsd;
run;
```

FULL CODE FOR EXAMPLE 2 (LUNG CANCER)

```
*****************************************************
 1A: Subset relevant inpatient encounters
*****************************************************;
data inpat;
  set med.inpatient;
  lung=0;
  diag1=substr(dx1,1,4);    *dx in primary position;
  diag2=substr(dx2,1,4);    *dx in secondary position;
  if '1622' <=: diag1 <=: '1629' then lung=1;
  else if (diag1 in('V580','V581')) and
          ('1622' <=: diag2 <=: '1629') then lung=1;
  if lung=1 then output;
  keep id date_inpat dx1 dx2 lung;
run;
proc sort data=inpat; by id date_inpat; run;

*****************************************************
 1B: Subset relevant outpatient encounters
*****************************************************;
data outpat;
  set med.outpatient;
  lung=0;
  diag1=substr(dx1,1,4);    *dx in primary position;
  diag2=substr(dx2,1,4);    *dx in secondary position;
  if ('1622' <=: diag1 <=: '1629') |
     ('1622' <=: diag2 <=: '1629') then lung=1;
  if lung=1 then output;
  keep id date_outpat dx1 dx2 lung;
run;
proc sort data=outpat; by id date_outpat; run;
```

```
*****************************************************
 1C: Stack inpatient and outpatient encounters
*****************************************************;
data alldata;
  merge inpat2 (in=i drop=dx1 dx2 rename=(date_inpat=date))
  outpat2 (in=o drop=dx1 dx2 rename=(date_outpat=date));
  by id date;
  if i then source='I'; *some pts have I and O on same day;
  else if o then source='O';
  format date DATE9.;
run;
proc sort data=alldata; by id source date; run;

*****************************************************
 2: Apply diagnostic criteria
*****************************************************;
proc expand data=alldata out=expanded1 method=none;
  by id;
  convert date = date_lag1 / transformout=(lag 1);
  convert date = date_lead1 / transformout=(lead 1);
/*convert date = date_lead2 / transformout=(lead 2);*/
run;
data expanded2; set expanded1;
  flag=0;
  if source='I' then flag=1;
  else if source='O' then do;
    if (date_lag1 ne . and date_lead1 ne .) and
       (date_lead1 - date_lag1 <= 90) then flag=1;
/*  if (date_lag2 ne .) and
       (date_lead2 - date <= 90) then flag=1;*/
  end;
run;
```

SAS® GLOBAL FORUM 2017

April 2 – 5 | Orlando, FL