**Instructions for creating data to use for the SEER-Medicare analysis comparative effectiveness analysis.**

*The comparative effectiveness analysis here is drug alone versus drug plus A. For example, gemcitabine alone versus gemcitabine plus erlotinib.*

In order to do this, you will need a few datasets handy:

1. Raw data - contains 2 variables:
   1. ‘id’ – unique patient identifier
   2. ‘timeZero’ – time (date) when follow-up begins and eligibility criteria are met (for your study, when “drug” is initiated)
2. Baseline data – contains:
   1. ‘id’ – unique patient identifier
   2. All baseline covariates. This includes demographics and comorbidities (in the 12 months prior to timeZero). All these variables should have ‘\_b’ as a suffix (e.g. anemia\_b).
3. Time-varying data – contains:
   1. ‘id’ – unique patient identifier
   2. For each time-varying covariate:
      1. ‘symptom’ - binary, 1 if exists, 0 if not
      2. ‘symptomDate’- date when code for symptom existed
      3. For example, if you had anemia as a time-varying covariate, you would have 2 variables in this dataset: ‘anemia’ and ‘anemiaDate’
      4. You want to be sure to keep only time-varying symptom occurrences that happened after ‘timeZero’
4. Treatment data – contains:
   1. ‘id’ – unique patient identifier
   2. ‘A’ – 1 for each instance
   3. ‘ADate’ – date when code for ‘A’ was recorded
      1. For individuals who never are prescribed ‘A’, both ‘A’ and ‘ADate’ are set to missing
5. Other chemo\* data – contains:
   1. ‘id’ – unique patient identifier
   2. ‘otherChemo’ – 1 for each instance
   3. ‘otherChemoDate’ – date when code for ‘otherChemo’ was recorded
      1. For individuals who are never prescribed ‘otherChemo’, both ‘otherChemo’ and ‘otherChemoDate’ are set to missing
   4. Note\*: since your trial is drug alone versus drug plus A, the otherChemo variable should not include instances of administration of drug or A.
6. Outcome/censoring information – contains:
   1. ‘id’ – unique patient identifier
   2. ‘dateDeath’ – date of death (Medicare). This is missing for individuals who do not die during observation.
   3. ‘insDate’ – date of loss of enrollment in Medicare Parts A&B, or enrollment in an HMO (use month and year provided, impute 15 for the day). Missing for individuals who do not die during observation.
   4. ‘admDate’ – date of administrative end of follow-up (for me, December 31, 2013). Same for everyone.

Part I: How to create a long version of the data, 'longdata'

1. Start with a dataset with 2 variables: eligible IDs and their respective time zeros (this is a date).
2. Add the baseline covariates (variables that are constant within each person). This data should be created using demographics, and clinical indications that happen in the 12 months prior to time zero.
3. **Expand data** to full length of follow-up for each individual (say, 60 months per person + baseline = 61 times per individual). Create a variable that runs from 0 to 60 within each individual, called 't'
4. For each month of person-time in your dataset, define a ‘startDate’ and ‘endDate’. Example: for month 0, ‘startDate’ will equal time zero. ‘endDate’ will equal time zero plus 30 days.
5. Fill in 'A' for each of the 61 time points. This amounts to, for each individual, checking if they have ever received treatment before ‘endDate’ for that particular month.
   1. Create a variable 'initA' that contains the date when 'A' was first administered. This is missing for individuals who never initiate 'A'.
   2. Create a variable ‘Alag’ that contains the prior value of ‘A’ (from time t-1). This variable is missing for t=0.
6. Fill in outcome: ‘Tevent’. This is 0 until month of death (death date between 'startDate' and 'endDate'). Once an individual dies, this variable is set to missing.
7. Fill in insurance data 'ins'. This is 1 until month of loss of parts A or B, or enrollment in an HMO.
8. Create an 'otherchemo' variable. This amounts to, for each individual, checking if they have ever received qualifying other chemo\* before 'endtime' for that particular month. (once it switches to 1, it stays at 1 in all future months).
   1. Create a variable 'initOtherChemo' that contains the date when 'otherChemo' was first administered.
   2. Note\*: since your trial is drug alone versus drug plus A, the otherChemo variable should not include instances of administration of drug or A.
9. Create a variable called ‘trt\_b’. This is the treatment group indicator. Look back from the end of the grace period (end of month 3, for example), and determine whether an individual has initiated 'A' ever in that time. This is an individual-fixed variable (constant within individual).
10. In this dataset (longdata), create a variable called 'touseweight' which is 1 for data to use in modeling for the weights, and 0 otherwise. Example code:

data longdata;

set longdata;

touseweight = 1;

if endDate > adm then touseweight = 0; /\* Delete rows from times after administrative end of follow-up \*/

if ins = 0 then touseweight = 0; /\* Delete rows from times when the individual has incorrect insurance \*/

if Tevent = . then touseweight = 0; /\* Delete rows from after individuals die \*/

if initOtherChemo < initA & endDate > initOtherChemo then touseweight = 0; /\* Delete rows where individuals received other chemo before receiving A (but after drug was administered). Leave in time before otherchemo was initiated. \*/

run;

Part 2: How to create the cloned version of the data (‘clonedata’)

1. Copy the long dataset ‘longdata’, to use for the cloning. Example code:

data longdata1;

set longdata;

trt\_clone = 0; /\* Assign all individuals to drug only \*/

cloneDate = .; /\* This is the month individuals would be censored at for violating their treatment strategy. This is set to . here because it exists in the 'initA' variable already for this strategy \*/

maxDate = min(deathDate, insDate, admDate, initA); /\* This is the date (NOT month) when follow-up ends \*/

maxDate\_m = min((maxDate - timeZero) / 30, 60); /\* This is the month when follow-up ends \*/

run;

data longdata2;

set longdata;

trt\_clone = 1; /\* Assign all individuals to drug plus A \*/

cloneDate = 3; /\* This is the month individuals are censored at if they have not initiated A - 'artificial' censoring date \*/

maxDate = min(deathDate, insDate, admDate, initOtherChemo); /\* This is the date (NOT month) when follow-up ends \*/

maxDate\_m = min((maxDate - timeZero) / 30, cloneDate, 60); /\* This is the month when follow-up ends - will censor everyone with 'treated = 0'\*/

run;

data clonedata;

set longdata1 longdata2;

run;

1. Create a variable that indicates whether to use this observation in the outcome regression.

data clonedata;

set clonedata;

touse = 1;

if maxDate\_m < t then touse = 0;

run;

The dataset ‘clonedata’ is now the version that I read into R to run the analyses. You have to further check that the data was read in properly (numerical not string, etc.), but should be straightforward from there.