MWSUG 2012 - Paper PH02-2012

Random assignment of proxy event dates to unexposed individuals in observational studies: An automated technique using SAS®.

Raymond Harvey, Dana Drzayich Jankus, David Mosley, UnitedHealthcare®

ABSTRACT

Limited technical resources are available to automate the random assignment of proxy exposure dates to unexposed populations (i.e., controls) in matched retrospective cohort studies of health services. Using disease management programs as an example, exposed and unexposed individuals have distinct eligibility dates that identify the onset of risk for exposure to the intervention. In addition to this eligibility date, exposed individuals also have unique treatment start dates. The time difference between the eligibility date and the exposure date has a frequency distribution. Based on this distribution, each unexposed individual is randomly assigned to a time difference in order to create a proxy exposure date (eligibility date + time difference=proxy exposure date). The resulting distribution of the time difference among unexposed individuals will replicate the observed distribution of the time difference of the exposed. This paper presents code to automate the random assignment of a proxy exposure date to unexposed individuals using Base SAS and SAS/STAT®.

INTRODUCTION

In a retrospective database study of a disease management program, both exposed and unexposed individuals will have a date associated with meeting the eligibility criteria for exposure to the intervention. This date could be a date of diagnosis, an admission date, or even a birth date. It is the date that they met the criteria for the particular intervention. For the purpose of this paper we will refer to this as their eligibility date (*elig_date*). A distinction between exposed and unexposed in a retrospective matched cohort study is exposed individuals have an actual date that they were exposed to the intervention (*event_date*). In this example the exposure date corresponds to a treatment or intervention start date. In the process of selecting comparable unexposed individuals, a proxy or simulated exposure date must be established. Establishing this proxy or simulated exposure data in the unexposed population creates a basis of comparison for the intervention period between exposed and unexposed individuals. By automating the creation of this random exposure date, based on the distribution of the difference between eligibility date and exposure date of the exposed, we have increased the efficiency at which we can assign proxy dates, reduced the potential for human error, and also reduced a potential source of bias when executing retrospective analyses.

METHODS

Exposed and unexposed individuals will need to be initially maintained in separate data sets. For ease, these will be referred to as exposed and unexposed. The paper will focus on the demonstration data sets that include 500 exposed and 2,500 unexposed individuals which can be reproduced with the sample code provided in the appendix.

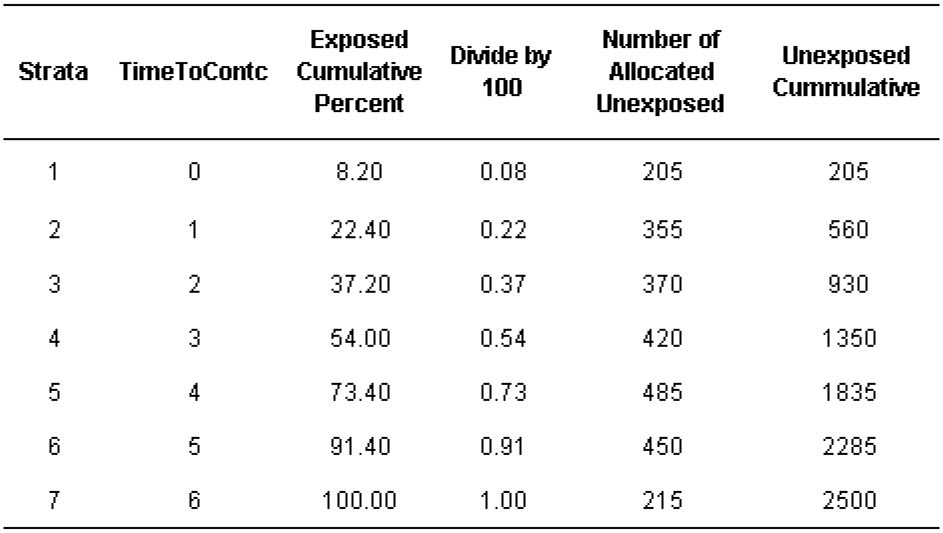
Using the INTCK function, the number of months between the eligibility date and the exposure date is calculated (*timetocontc*) for the exposed group. The frequency distribution for this time-to-contact variable is produced using the FREQUENCY procedure and a data set is created using ODS OUTPUT (table 1.). This frequency distribution will be used to determine how many unexposed individuals are required in each frequency bin (strata).

Based on the sample data, there are seven distinct strata (0, 1, 2, 3, 4, 5, 6), representing the seven intervals of time, or months, between the exposed eligibility date and exposure date. The unexposed data set, which contains 2,500 individuals, must have the same number of strata and the same percentage of observations within the strata (cut points). For example, the time between eligibility for the intervention and intervention initiation was one month for 14.2% or 71 of the 500 exposed (table 1). We want to randomly assign the same one month interval between eligibility and proxy exposure for the unexposed group or 355 of the 2,500 unexposed.

CUT POINT LOGIC

To achieve this, we will create cut points that represent the number of unexposed that are needed within each strata by dividing the cumulative frequency by 100 and multiplying by the number of unique observations in the unexposed data set (n=2,500). The output from the PROC FREQ contains the total number of strata and the respective cumulative percentages. Using the SQL procedure we create a GLOBAL MACRO (*&num*) variable representing the total number of unexposed individuals (n=2,500) and complete the calculations to create these cut points in a DATA step. This logic is demonstrated in table one.

Table 1. Demonstrating Cut Point Logic



Strata	TimeToContc	Exposed Cumulative Percent	Divide by 100	Number of Allocated Unexposed	Unexposed Cummulative
1	0	8.20	0.08	205	205
2	1	22.40	0.22	355	560
3	2	37.20	0.37	370	930
4	3	54.00	0.54	420	1350
5	4	73.40	0.73	485	1835
6	5	91.40	0.91	450	2285
7	6	100.00	1.00	215	2500

To assign 8% of the unexposed to the first strata we will randomly allocate 205 unexposed members, 355 individuals to strata two, 370 individuals to strata three, 420 to strata four, 485, 450, and 215 individuals to strata five, six and seven respectively.

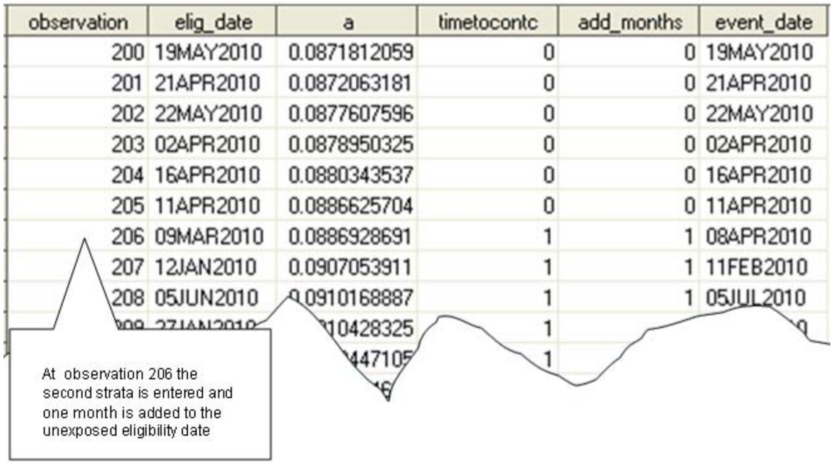
PREPARING THE DATA AND EXECUTING THE DYNAMIC ARRAY TO DISTRIBUTE PROXY ASSIGNMENT DATES ACROSS CUT POINTS

The output data set that contains the frequency for the time between eligibility and exposure for the exposed as well as the data set containing the cut points will be restructured from long to wide using the TRANSPOSE procedure. Once transposed, the two data sets will be merged together using a dummy variable to form one row containing the cut points and seven strata.

Using the DATA step and the UNIFORM function, a pseudo-random number and dummy variable will be created for each observation in the unexposed data set. The data set will be sorted by the random number using the SORT procedure and merged by the dummy variable to the cut point and strata data set to create the matrix for the dynamic array to process.

PROC SQL will be used to define the dimension of the array by creating a GLOBAL MACRO variable that corresponds to the number of strata (*&strata*) from the original frequency distribution data set.

The array will assign the appropriate number of unexposed individuals to the appropriate strata; executing until all unexposed have been assigned to a strata. Each stratum is associated with a number of months, representing time between eligibility and the exposure to the intervention, which will be added to the unexposed eligible date to create a proxy exposure date. The first 205 observations will have zero months added to their eligibility date, observations 355 through 369 will have one month added to their eligibility date. Figure one is an excerpt of the results of the dynamic array; note that the array is adding zero months to the unexposed eligibility date until observation 260 then it adds one month to the eligibility date as expected.



observation	elig_date	a	timetocontc	add_months	event_date
200	19MAY2010	0.0871812059	0	0	19MAY2010
201	21APR2010	0.0872063181	0	0	21APR2010
202	22MAY2010	0.0877607596	0	0	22MAY2010
203	02APR2010	0.0878950325	0	0	02APR2010
204	16APR2010	0.0880343537	0	0	16APR2010
205	11APR2010	0.0886625704	0	0	11APR2010
206	09MAR2010	0.0886928691	1	1	08APR2010
207	12JAN2010	0.0907053911	1	1	11FEB2010
208	05JUN2010	0.0910168887	1	1	05JUL2010
209	27JAN2010	10428325	1		
		447105	1		

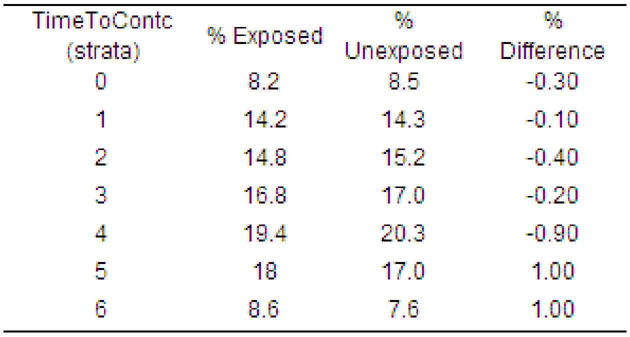
At observation 206 the second strata is entered and one month is added to the unexposed eligibility date

Figure 1. Excerpt of Array Results

COMPARE FREQUENCY DISTRIBUTION OF TIME FROM ELIGIBILITY DATE TO EVEN DATE BETWEEN GROUPS

The final step is to compare the frequency distribution (percent of individuals) of time, in months, from eligibility to real or proxy exposure date between the exposed and unexposed groups. This is done using PROC FREQ, PROC SORT and the DATA step. We would expect that there would be little variance between the percentages of individuals within the strata (table 2.). We may be interested in exploring the difference between the exposure dates between the exposed and unexposed; however, the variance of this comparison may not align as nicely as the comparison of the time between the two dates. Our interest and purpose was to ensure that the unexposed were allocated based on the time between the eligibility date and the exposure date.

Table 2. Comparison of Frequency Distribution Between Exposed and Unexposed



TimeToContc (strata)	% Exposed	% Unexposed	% Difference
0	8.2	8.5	-0.30
1	14.2	14.3	-0.10
2	14.8	15.2	-0.40
3	16.8	17.0	-0.20
4	19.4	20.3	-0.90
5	18	17.0	1.00
6	8.6	7.6	1.00

CONCLUSION

The information contained between the time of risk (eligibility) and the time of exposure to a specific treatment (event/intervention) contains important characteristics and may be used as baseline information for matching or model adjustment. Unexposed individuals do not have a specific date when exposure was initiated. Using their eligibility date alone would be introducing systematic bias to the analysis. The authors have presented one technique to mitigate this potential source of bias.

Using basic procedures such as DATA step programming and array logic, an end user can easily execute the code necessary to randomly allocate proxy exposure dates to an unexposed sample. The code may be modified to work at the level of measurement contained in a data set (hour, day, quarter). Additionally, the code provided below may also be used to create a MACRO to sequentially execute across multiple samples.

ACKNOWLEDGEMENTS

The authors would like to thank Muthia Kachiryan for her guidance in creating the dynamic array.

CONTACT INFORMATION

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

Name: Raymond Harvey
Enterprise: UnitedHealthcare
E-mail: raymond_a_harvey@uhc.com
Web: www.linkedin.com/pub/ray-harvey/1b/9a/793

Name: Dana Drzayich Jankus
Enterprise: UnitedHealthcare
E-mail: dana_drzayich_jankus@uhc.com
Web: www.linkedin.com/pub/dana-jankus/11/1a7/26

APPENDIX: SAMPLE DATA AND PROGRAM

```
/****************************************************************************/

/**   Create random dates for demonstration
**/
/**   Exposed individuals will have an eligibility date (elig_date) and a treatment
      exposure date (exposure_date) **/
/**   Unexposed individuals will only have an eligibility date (elig_date)
**/

/****************************************************************************/

DATA exposed(keep=grp elig_date exposure_date)  unexposed(keep=grp
elig_date);
      study_st='01Jan2010'd; /*observation (study) start date*/
      study_end='31Dec2011'd;/*observation (study) end date*/

      offset = (study_end-study_st+1)/4;/*offset only used for date
generating */

      do i = 1 to 500;
            grp=1;
                  Elig_date=study_st + int(ranuni(21)*offset);
                        Exposure_date=Elig_date + int(ranuni(321)*offset);
      output exposed;
            end;

      do i = 1 to 2500;
            grp=0;
                  Elig_date=study_st + int(ranuni(21)*offset);
      output Unexposed;
            end;

      format study_st study_end Elig_date Exposure_date date9.;
      RUN;


/**   Step 1:  Calculate time between eligiblity and treatment exposure 
(TimeToContc) for exposed                 **/

DATA Exposed_;
      set exposed;
            TimeToContc=intck('month',elig_date,exposure_date);

RUN;

/**   Step 2:  Output oneway Frequency table for event date and TimeToContc 
for exposed                                                       **/

PROC FREQ DATA=exposed_;
      table timetocontc;
            ods output onewayFREQs=exp_time;
RUN;
PROC FREQ DATA=exposed_;
      format exposure_date MONYY.;
            table exposure_date;
                  ods output onewayFREQs=exposed_edate;
RUN;


/*    Step 3:  Create global macro variable representing number of 
observations in unexposed DATA set                **/

PROC SQL noprint;
      select count(*)
            into: num
                  from unexposed;
quit;

%put &num; /* check number of unexposed */

/* Step 4:  Determine number of unexposed individuals per bins (cutpoints) 
based on the distribution of the exposed
                  using the number of unexposed as the denominator 
                                                                        **/

DATA exp_time_;
      set exp_time;
            cutpt=round((CumPercent/100)*&num);/* this creates the number of 
unexposed within each FREQuency bin based
                                                                  on the 
number of exposed in the ds */
                  keep cutpt;
RUN;

/**   Step 5:  SORT, TRANPOSE, and merge nunmber of unexposed needed per bin 
(cutpoints) to bins created from the
                   exposed Frequency distribution                                                    
**/
```

```
PROC SORT DATA=exp_time_;
      by cutpt;
RUN;
PROC TRANSPOSE DATA=exp_time_ out=trans_exptime(drop=_name_) prefix=cutpt;
      var cutpt ;
RUN;
PROC TRANSPOSE DATA=Exp_time out=trans_Expcontact(drop= _name_)
prefix=c_timetcntc;
      var timetocontc;
RUN;
DATA Exp_pct;
      set trans_exptime;
            dummy=1;
RUN;
DATA Exp_contact;
      set trans_Expcontact;
            dummy=1;
RUN;

DATA trans_merge_exp;
      merge exp_pct exp_contact;
            by dummy;
RUN;
PROC SORT;
      by dummy;
RUN;

/**   Step 6:  Randomly merge bins and number of unexposed needed per bin to
the unexposed Dataset                        **/

DATA unexposed_premerge;
      set unexposed;
            dummy=1;
                  a = UNIFORM(-32);
RUN;

PROC SORT DATA=unexposed_premerge;
            by a dummy ; /* SORTing the obs by the random number */
RUN;

/**    Step 7:  Create global macro variable for the number of Frequncy bins
(strata)                                           **/

            proc sql noprint;
                  select count(*)
                        into: strata
                        from exp_time;

                        quit;
                        %let strata=&strata;

%put &strata; /* check macro variable */

/**   Step 7:  Using a dynamic array randomly assign months between time of
eligibility to proxy exposure to create
                   a proxy exposure date for the unexposed group
                                                                    **/
```

```
DATA final_unexposed;
      merge trans_merge_exp unexposed_premerge;
            by dummy;

            b=_n_; /* creating a var that corresponds to observation number
for ease of qa*/
array c[*] c_timetcntc1 - c_timetcntc&strata;
array k[*] cutpt1 - cutpt&strata;
if b <= k[1] then add_months = c[1];
else if b GT k[dim(k) - 1] then add_months = c[dim(k)];

else
do _n_ = 2 to dim(k) - 1;

   if k[_n_ - 1] LT b LE k[_n_] then do;
      add_months = c[_n_];
      leave;
   end;
end;
                        exposure_date=elig_date+( add_months*30);/* create
proxy control start date from
                                                case distribution of time to
contact */
                        exposure_month=month(exposure_date);
                        TimeToContc=intck('month',elig_date,exposure_date);
                        format elig_date exposure_date date9.;
*drop cutpt1--cutpt&strata c_timetcntc1--c_timetcntc&strata;

RUN;



/** Step 9:  Compare Frequency distribution of time between eligibility and
exposure (timetoc) and
                   exposure dates between exposed and unexposed
                                                              **/

PROC FREQ DATA=final_unexposed;
      table timetocontc;
            ods output onewayFREQs=unexposed_check;
RUN;
PROC SORT DATA=unexposed_check out=unexpcheck(keep=TimeToContc Percent
      rename=(Percent=co_pct));
            by timetocontc;
RUN;


PROC SORT DATA=exp_time out=expcheck(keep=TimeToContc Percent
      rename=(Percent=Ca_pct));
            by timetocontc;
RUN;

DATA QA_TIMETOC;
      merge expcheck unexpcheck;
            by timetocontc;
                        pct_diff=ca_pct-co_pct;
                        format pct_diff 8.1;
RUN;

PROC SORT DATA=exposed_edate out=exp_edate (keep=f_exposure_date percent 
rename=(percent=ca_pct));
      by f_exposure_date;
      RUN;
PROC FREQ DATA=final_unexposed;
      format exposure_date MONYY.;
      table exposure_date;
      ods output onewayFREQs=unexposed_edate;
RUN;

PROC SORT DATA=unexposed_edate out=unexp_edate(keep=f_exposure_date percent 
rename=(percent=co_pct));
      by f_exposure_date;
RUN;


DATA QA_event_date;
      merge exp_edate unexp_edate;
            by f_exposure_date;
                  pct_diff=ca_pct-co_pct;
                        format pct_diff 8.1;
RUN;
```