PharmaSUG 2012 - Paper SP12

A SAS® Macro for Biomarker Analysis using Maximally Selected Chi-Square Statistic With Application in Oncology

Quan Jenny Zhou, Eli Lilly and Company, Indianapolis, IN
Bala Dhungana, Eli Lilly and Company, Indianapolis, IN

ABSTRACT

Biomarker assessment has become an essential tool for evaluating treatment effects in subpopulations of potential drug responders in oncology studies. It is believed that treatment effects can differ between patient subgroups with different genotypes, and therefore biomarkers may help predict treatment effects in these subpopulations. There are many statistical methods for selecting biomarkers as candidate classifiers when identifying subgroups. One approach is to transform each continuous biomarker into a binary covariate by selecting a threshold with certain optimal properties and fit it in the survival model, one biomarker at a time. Selection of a threshold can be done by using maximally selected chi-square statistic, as proposed in Miller and Siegmund, 1982.

In this paper, we demonstrate how to implement the maximum chi-square method with a SAS macro that fits proportional hazards Cox regression models on time-to-event endpoints, determines the biomarker threshold to classify patients into subgroups, and then performs analysis to test for the biomarker effect and treatment effect within and between the biomarker patient subgroups using Kaplan-Meier and Proportional Hazards models. This macro can be applied more broadly to evaluate treatment effects in subgroups formed by a set of continuous covariates in the context of survival analysis. It can also be easily modified to fit logistic models for binary and ordinal outcomes.

INTRODUCTION

Many biomarkers are continuous variables; for example, H-score of an immunohistochemistry assay. Continuous biomarkers are sometimes converted into categorical variables by grouping values into two categories to select genomic patient subgroups. One approach was proposed by Miller and Siegmund, 1982, who derived asymptotic distribution of the maximal chi-square statistic that arises when selecting the cut point to maximize the value of the "standard" chi-square statistic.

This theory can be applied to dichotomize continuous biomarkers into two subgroups when evaluating prognostic and predictive effects of candidate biomarkers. For a given biomarker, patients are initially classified into two groups: "low" as those with biomarker values smaller than a cut point and "high" as those with equal or larger values. The initial cut point is chosen using the ordered biomarker values so that at least a certain proportion, for example, 25%, of patients is in the "low" group. Our strategy is to identify optimal cut points for each biomarker, so as to evaluate their prognostic and predictive effects. To this end, two Cox regression models, one with main effects of treatment and biomarker group (to evaluate prognostic biomarker effect), and the other with additional treatment-by-biomarker group interaction (to evaluate potential predictive biomarker effect), are constructed. For the two models we retain chi-square statistics for the main biomarker effect and treatment by biomarker interaction effect, respectively. Then for each model we optimize the cut point, by repeatedly setting it equal to every observed value and computing the chi-square statistics until the percentage of subjects in the low group reaches to a certain pre-specified upper limit, for example, 75%. In the end, for each model, the maximum value across all the calculated chi-square statistics is identified, and the corresponding cut point is selected as the final dichotomization threshold to select patient subgroups. The final thresholds are used to fit the same two models and the biomarker effect (from the first prognostic model) and treatment effects between the two patient subgroups (from the second predictive model) are tested. Results from both models are generated and stored in a data set ready for reporting.

To implement the above strategy, we developed a macro called *%am_maxchi_coxreg* ( ).

MACRO PARAMETERS AND ASSUMPTIONS

The macro parameters are listed in **Table 1**.

Table 1. Macro Parameters

Name	Default	Description
inset	none	The input data set with biomarker value and time-to-event variable and censoring indicator.
mrkvar	PARAM	The name of the biomarker name/description variable. An example would be the variable of the names of immunohistochemistry biomarkers such as EGFR, TS, TTF etc.
mrknm	&&marker&i	The distinct value of the biomarker name extracted from &mrkvar. It is the description of the biomarker name that appears on the report.
mrkscr	AVAL	The name of the biomarker value variable that is used to determine dichotomization. This is assumed to be a numerical variable that is measured in ordinal scale. An example would be the variable of the H-Score for immunohistochemistry biomarkers.
startpct	0.25	The minimum fraction of patients in the “low” biomarker group when starting to search for the threshold for dichotomization. The initial cut point is determined based on this percentile.
endpct	0.75	The maximum fraction of patients in the “low” group when stopping to search for the threshold for dichotomization. The final cut point in the search for the optimal threshold is determined based on this percentile

The macro assumes that the input data set contains one record per patient per biomarker, with columns for the biomarker variable, the time-to-event and censoring indicator variable. To reduce the number of macro parameters, it also assumes that the time-to-event variable is named as TTERN and the censoring indicator is TTECENSFLG, with 1 being the censored value. The name of the treatment variable is assumed to be TRTSORT with numerical values coded as “1” for the active treatment arm and “0” for the control.

MACRO IMPLEMENTATION DETAILS

The macro reads in a SAS data set that contains one record per patient for a particular biomarker, possibly with multiple biomarkers. It extracts a subset of data for a specific biomarker of interest, sorts the records by the biomarker values in ascending order. It then determines the initial and final cut points to ensure the fraction in the low group is between &startpct and &endpct .

Extract the subset of records for the biomarker of interest and compute the total number of patients in the data set. The number of patients is the number of observations in the extracted data set.

```
%*-- extract individual markers -;
PROC SORT data=&inset out=indmrk_;
  by &mrkscr;
  where &mrkvar = "&mrknm";
RUN;

DATA _null_;
  set indmrk_ nobs = obs;
  %*-- initialize the total number of observation into a macro variable -;
  CALL SYMPUT("total", compress(put(obs,best12.)));
RUN;
```

Determine the start and end points based on &startpct and &endpct, the minimum and maximum fractions of patients allowed in the low group, respectively. If the cut points are calculated incorrectly so that the end point is less than the

start point, or if there are too few subjects in the data set, then the macro will terminate and write a message in the log file. The macro also terminates if the cut-points associated with &startpct and &endpct are identical.

```
%*-- define the start and end points of the observation number -;
%*-- the start point is determined to have a MINIMUM of &startpct fraction of records
in the low group -;
%*-- the end point is determined to have a MAXIMUM of &endpct fraction of records in
the low group -;
%LET start = %SYSEVALF(&total * &startpct + 1, ceil);
%LET end =  %SYSEVALF(&total * &endpct + 1, floor);


%*-- in the case where the start and end points do not make sense, or if there are too
few samples, terminate analysis -;
%*-- Experience tells us that if there are only 12 or less total patients there is no
need to perform such an analysis -;
%IF &start > &end or &total<12 %THEN %DO;
  %PUT for marker=&marker, start and end point not making sense (start=&start
  end=&end), or not enough records (n=&total), terminate analysis;
  %LET stoprun = 1;
  %GOTO exit;
%END;

%*-- get the record number immediately prior to the start point -;
%IF %EVAL(&start-1)>0 %THEN %LET prevrec = %EVAL(&start-1);

DATA _null_;
  set indmrk_ nobs = obs;
  %*-- get the marker value of the previous record -;
  if _n_=&prevrec then CALL SYMPUT("previous",compress(put(&mrkscr,best32.)));
  %*-- get the marker value at the end of the cut point -;
  if _n_=&end then CALL SYMPUT("post", compress(put(&mrkscr,best32.)));
RUN;

%*-- in case where the search starts and ends at the same value, terminate analysis -;
%IF %SYSEVALF(&previous eq &post) %then %do;
  %PUT For marker=&mrknm, start and end results (&previous, &post) are equal,
  terminate analysis;
  %LET stoprun = 1;
  %GOTO exit;
%END;
```

Dichotomize patients into “low” and “high” biomarker groups based on the initial cut point, and fit the two Cox regression models: the “main effects model” and the “interaction model”. The chi-square statistic associated with the interaction term from the interaction model and the chi-square statistic associated with the biomarker effect from the main effects model are retained. The cut point is incremented to the next observed value, and the same steps are repeated until the cut point reaches to the upper limit, retaining the two chi-square values each time. The chi-square statistics associated with all the cut points evaluated from the same Cox model are compared and the cut points corresponding to the maximum chi-square values are passed to macro variables, *&intthold* and *&mainthold*, as the final optimal thresholds for the models to dichotomize the patient population into “low” and “high” biomarker groups.

```
%DO k = &start %TO &end;
  %IF %SYSEVALF(&cut ^= &previous) %THEN %DO;

   DATA cut_;
      set indmrk_;
      by &mrkscr;
      cutpt=_n_;

      %*-- define a temporary variable for high/low group -;
      if cutpt ge &k then tempgrp = 1;
      else tempgrp = 0;
    RUN;
```

```
    %*-- interaction model -;
    ODS OUTPUT parameterestimates=parami_(where=(upcase(variable)='TRT_GRP'));
    PROC PHREG data = cut_;
      MODEL ttern * ttecensflg(1) = trtsort tempgrp trt_grp;
      trt_grp = trtsort * group;
    RUN;

    %*-- main effects model -;
    ODS OUTPUT parameterestimates=paramr_(where=(upcase(variable)='TEMPGRP'));
    PROC PHREG data = cut_;
      MODEL ttern * ttecensflg(1) = trtsort tempgrp;
    RUN;

   /** Insert code to stack parami_ and paramr_ from all the iterations together **/

  %END;

  %LET previous=&cut;

  /** Insert code to compare the chi-square values and pick the maximum one from each
model. The one from the interaction model is saved as &intthold and the one from the
main model is saved as &mainthold. **/

%END;
```

After identifying the optimal threshold biomarker values under the two models, these cut points are applied to dichotomize patient population into “low” and “high” biomarker groups. An output SAS data set is created with two variables, *igroup* and *mgroup*, assuming values 0 and 1 to designate the “low” and “high” groups based on the optimal biomarker cut points from the interaction model and the main effects model, respectively.

```
  %*-- data set to feed into the interaction and the main effects models -;
  DATA anads_;
    set indmrk_;
    length expression $5;

    if &mrkscr >= &intthold then do;
      igroup=1;
      iexpression='High';
    end;
    else do;
      igroup=0;
      iexpression='Low';
    end;

    if &mrkscr >= &mainthold then do;
      mgroup=1;
      mexpression='High';
    end;
    else do;
      mgroup=0;
      mexpression='Low';
    end;

    keep ttern ttecensflg trtsort mgroup mexpression igroup iexpression;;
  RUN;
```

Two sets of Kaplan-Meier analyses on the time-to-event endpoint are performed, stratified by treatment and biomarker group, one for the group from the interaction model and the other for the group from the main effects model. The output data sets are created using appropriate ODS statement to get the summary of the event and its median estimate.

```
%*-- KM analysis - interaction model;
ODS OUTPUT quartiles=mediani_(keep=trtsort group percent estimate lowerlimit
           upperlimit rename = (estimate = median) where = (percent = 50))
          censoredsummary=counti_(keep =  trtsort group total censored failed pctcens
           where = (group in (0, 1)));
PROC LIFETEST data = anads_ ALPHA = .05;
  TIME ttern * ttecensflg(1);
  STRATA group trtsort;
RUN;

%*-- KM analysis - main effect model;
ODS OUTPUT quartiles=medianm_(keep=trtsort group percent estimate lowerlimit
           upperlimit rename = (estimate = median) where = (percent = 50))
          censoredsummary=countm_(keep=trtsort group total censored
           failed pctcens where=(group in (0, 1)));
PROC LIFETEST data = anads_ ALPHA = .05;
  TIME ttern * ttecensflg(1);
  STRATA group trtsort;
RUN;
```

The same two proportional hazards Cox regression models as described above are fit using the biomarker groups defined based on the optimal cut points. Treatment effects within biomarker groups and biomarker effects within treatment groups are examined under the interaction model. Treatment-independent (prognostic) biomarker effect is examined under the main effects model. Output data sets with hazard ratios and associated confidence intervals, as well as Wald Chi-square test statistics and p-values are generated using appropriate ODS statements. The following table (**Table 2**) shows how beta coefficients from the proportional hazards interaction model are used to compute hazard ratios for the contrasts of interest.

Table 2. Computing Hazard Ratios for the Contrasts of Interests from the Estimated Regression Coefficients in the Interaction Model

Interaction model:		treatment	group	treatment*group	HR for	
stratum	effect	β_1	β_2	β_3	contracts	Description
group=1	treatment=1	1	1	1	$\exp(\beta_1+\beta_3)$	Treatment effect within biomarker group = "high"
	treatment=0	0	1	0		
group=0	treatment=1	1	0	0	$\exp(\beta_1)$	Treatment effect within biomarker group = "low"
	treatment=0	0	0	0		
treatment=1	group=1	1	1	1	$\exp(\beta_2+\beta_3)$	Biomarker effect within treatment group = "1"
	group=0	1	0	0		
treatment=0	group=1	0	1	0	$\exp(\beta_2)$	Biomarker effect within treatment group = "0"
	group=0	0	0	0		

```
%*-- interaction model -;
ODS OUTPUT parameterestimates=main_int_(keep=variable chisq hazardratio HRlowerCL
           HRupperCL ProbChiSq)
           testprint1=covmtrx_int_(keep=label col1 col2 rename=(col1=varcov
           col2=betahd))
          teststmts=pval_int_(keep=label waldchisq probchisq);
PROC PHREG data = anads_;
  MODEL ttern*ttecensflg(1) = igroup trtsort trt_igrp / ALPHA=0.05 RISKLIMITS;
  trt_igrp = trtsort * igroup;
  %*-- testing treatment effect when igroup=high -;
  TEST1: test trtsort+trt_igrp /e print;
  %*-- testing group effect when treatment=1 -;
  TEST2: test igroup+trt_igrp /e print;
  /* note that: β(trtsort) is the coefficient of treatment effect when  igroup=low,
     and β(igroup) is the coefficient of group effect when trt=0. */
RUN;
```

```
%*-- main effects model -;
ODS OUTPUT parameterestimates=main_eff_(keep = variable chisq hazardratio HRlowerCL
           HRupperCL ProbChiSq);
PROC PHREG data = anads_;
  MODEL ttern*ttecensflg(1) = trtsort mgroup /ALPHA=0.05 RISKLIMITS;
  /* note that: β(trtsort) is the coefficient for testing treatment effect. */
RUN;
```

The final "reporting-ready" output data set is created by processing and combining outputs from Kaplan-Meier and Cox regression models.

THE OUTPUT

The output from the macro is a SAS data set that is structured for the following reporting table (**Table 3**).

The report contains biomarker effects within each treatment group and treatment effect within each biomarker group as well as the associated p-values and 95% confidence intervals. The p-values for the biomarker effect in the main effect model and the p-values for the treatment by biomarker effect in the interaction model are adjusted for the optimal cut point search using methodology in Miller and Siegmund, 1982 (as documented in the footnote "e" of the report).

Table 3. Example of Output Report Produced by the Macro

	High Biomarker Group [a]		**Low Biomarker Group [b]**	
	Treatment A	**Treatment B**	**Treatment A**	**Treatment B**
Biomarker *xxx*, threshold	x.xxx			
Total number of subjects, N	x	X	x	x
Subjects censored, n(%)	x (x.xx)	x (x.xx)	x (x.xx)	x (x.xx)
Subjects with observed event, n(%)	x (x.xx)	x (x.xx)	x (x.xx)	x (x.xx)
Median *TTE*, months	x	x	x	x
95% CI for median *TTE*	(xx.x, xx.x)	(xx.x, xx.x)	(xx.x, xx.x)	(xx.x, xx.x)
HR [d] (Within Expression Level)	x.xxx		x.xxx	
95% CI for HR[d]	(x.xxx – x.xxx)		(x.xxx – x.xxx)	
Wald chi square	x.xxx		x.xxx	
p-value[e]	.xxx		.xxx	
HR [f] Within Treatment A	x.xxx			
95% CI for HR[d]	(x.xxx – x.xxx)			
HR[g] Within Treatment B	x.xxx			
95% CI for HR[d]	(x.xxx – x.xxx)			
Interaction Wald Chi-square	x.xxx			
Interaction p-value[e]	.xxx			
Main effects model threshold	Xxx			
Total number of subjects, N	x	X	x	x
Patients censored, n(%)	x (x.xx)	x (x.xx)	x (x.xx)	x (x.xx)
Patients with observed event, n(%)	x (x.xx)	x (x.xx)	x (x.xx)	x (x.xx)
Median TTE, months	x	x	x	x
95% CI for median TTE	(xx.x, xx.x)	(xx.x, xx.x)	(xx.x, xx.x)	(xx.x, xx.x)
Treatment independent HR[i] (95% CI)	x.xxx			
95% CI for HR[i]	(x.xxx – x.xxx)			
Wald chi square	x.xxx			
p-value[e]	.xxx			

Footnotes:
(a) Subjects with high relative biomarker expression level.
(b) Subjects with low relative biomarker expression level.
(c) Biomarker value that is the threshold maximizing the effect of interaction between treatment and biomarker groups. A biomarker value at or above this threshold is classified as in the high biomarker group. A biomarker value below this threshold is classified as in the low biomarker group. Range of biomarker values assessed for threshold: *25 – 75th* percentile.
(d) Hazard ratio for *treatment A* vs. *treatment B* within protein expression level.
(e) * Asymptotic probability of the observed maximum chi-square statistic calculated with formula of Miller and Siegmund (1982). **Probability calculated using the ordinary chi square distribution; because the asymptotic p-value calculated from the max chi square distribution was inappropriately smaller. This situation may occur with small chi square statistic values, since the max chi square distribution applies asymptotically as the chi square statistic approaches infinity (i.e. the asymptotic max chi square distribution is most accurate for those values which are most significant).
(f) Hazard ratio for high vs. low biomarker group within *treatment A.*
(g) Hazard ratio for high vs. low biomarker group within *treatment B*.
(h) Threshold H score value identified under a model without an interaction term maximizing the effect of biomarker group (i.e. only main effects for treatment and biomarker group were included). Range of biomarker values assessed for threshold*: 25 – 75th* percentile.
(i) Hazard ratio for high vs. low biomarker group.

EXAMPLE OF MACRO CALL

The following SAS code calls the macro *%am_maxchi_coxreg()*. The code reads in a data set named *admrk* with one record per patient per biomarker variable. Therefore, multiple biomarker variables are stacked within a single SAS data set. The variable in the data set that stores the names of all biomarkers is *param* and the variable *aval* stores the values of biomarkers. For each biomarker, the macro searches the middle 50 percent of the distribution to determine the optimal thresholds for the interaction and the main effects models. The macro dichotomizes the patient population into “high” and “low” groups based on the threshold values and performs Kaplan-Meier and Cox Regression analyses. Finally the output data set named “report” is generated that contains the analysis results from all the biomarkers.

```
%MACRO generator;

  %put NOW EXECUTING MACRO generator;

  PROC SORT data = admrk out = temp0 nodupkey;
    by param;
  RUN;

  DATA _null_;
    set temp0 nobs = n;
    CALL SYMPUT('nummrk',left(put(n,best12.)));
    CALL SYMPUT(compress("marker"||put(_n_,best12.)), left(trim(param)));
  RUN;

  %*-- perform the analysis for all biomarkers -;
  %DO i = 1 %TO &nummrk;

    %am_maxchi_coxreg_(inset=admrk, mrkvar=param, mrknm=&&marker&i, mrkscr=aval,
     startpct=.25, endpct=.75);
  %END;

%MEND generator;

%generator;
```

CONCLUSION/SUMMARY

The SAS macro has been developed that automates a popular biomarker cut point selection strategy, based on maximal Chi-square statistics providing biomarker effects within treatment and treatment effects within biomarker and associated p-values and confidence intervals. The p-values for the biomarker effect and the biomarker-by-treatment interaction are adjusted for multiplicity inherent in selecting optimal cut points for a given biomarker.

We note, however, that the macro does not adjust the individual biomarker p-values against multiplicity in selecting across biomarkers. It also implements one-biomarker-at a time approach which may be not optimal if the true subgroups is formed by a combination of two or more biomarkers. For literature suggesting procedures that simultaneously search for groups formed by a combination of markers, while adjusting for multiplicity in the entire search process and providing the overall type I error control, see Lipkovich et al, 2012 http://onlinelibrary.wiley.com/doi/10.1002/sim.4289/pdf

REFERENCES

[1] Miller R, Siegmund D. (1982), Maximally Selected Chi Square Statistics. *Biometrics*, **38**, 1011–1016

[2] Lipkovich I, Dmitrienko A, Denne J, Enas G. (2011) Subgroup identification based on differential effect search - A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, **30**, 2601 – 2621

ACKNOWLEDGEMENT

The authors would like to thank Dr. Ilya Lipkovich for reviewing the paper and providing insightful comments.

CONTACT INFORMATION

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

Quan Jenny Zhou
Eli Lilly and Company
Lilly Corporate Center
Indianapolis, IN 46285 U.S.A.
Email: zhou_quan_jenny@lilly.com

Bala Dhungana
Eli Lilly and Company
Lilly Corporate Center
Indianapolis, IN 46285 U.S.A.
Email: dhungana_bala@lilly.com