Paper 54-2018

Combined Predictor Selection for Multiple Clinical Outcomes Using PHREG

L. Grisell Diaz-Ramirez, University of California, San Francisco
Siqi Gan, Healthcare Department, Philips Research China, Shanghai, China
Sei J. Lee, University of California, San Francisco
Alexander K. Smith, University of California, San Francisco
W. John Boscardin, University of California, San Francisco

ABSTRACT

Like other regression methods in SAS®, the PHREG procedure has built-in options to perform predictor selection using stepwise methods or best subsets. However, these built-in options for predictor selection do not allow the simultaneous selection of predictors for multiple outcomes. The selection of a common set of predictors for multiple outcomes is important in clinical settings where practitioners are frequently interested in predicting multiple outcomes in the same subject, while at the same time obtaining a parsimonious model with appropriate predictive accuracy. In this paper, we describe a SAS Macro for selecting a common set of variables for predicting multiple outcomes. The selection method uses a variant of backward elimination based on the average normalized Bayesian Information Criterion (BIC) across multiple outcomes. The BICs are obtained by fitting multivariable survival models using PHREG in SAS version 9.4, SAS/STAT 14.2. We illustrate the proposed method using the Health and Retirement Study data. We compare the predictive accuracy and parsimony of the final model with the models obtained for each individual outcome. We then test the correct inclusion and correct exclusion of variables in the final model using a simulation study. Our method provides a straightforward approach to obtain a common set of predictors for multiple clinical outcomes without compromising parsimony or predictive accuracy.

INTRODUCTION

When selecting the best set of predictors for individual outcomes, stepwise regression methods like backward elimination, forward selection, and standard stepwise regression are commonly applied since they are easy to use and interpret. Additionally, when there are a limited number of initial predictors, best subset regression can also be easily implemented because it allows the selection of best sets of predictors for a specified number of variables in the model based on statistics like the information criteria.

Some authors have suggested a combination of stepwise regression, information criteria such as the Akaike information criterion (AIC) (Akaike 1974), and the best subset selection for predictor selection in survival and logistic regressions (Shtatland *et al.* 2003, 2005). In their method, Shtatland *et al.* (2003, 2005) first use stepwise selection method to get the sequence of models from the null model with no predictors through the full model with all the predictors. Next, they find the minimum AIC which indicates the AIC-optimal model, and lastly, they use the best subset selection with the score statistic to find the best subsets models around the AIC-optimal model. They call this two-dimensional subset the "stepwise-AIC-best-subset blanket." The Bayesian Information Criterion (BIC) (Schwarz 1978) is related to the AIC but has a larger penalty term for additional parameters than the AIC, resulting in the BIC favoring more parsimonious models than the AIC.

In many clinical settings, competing risk regression is necessary to appropriately account for death when a second outcome (such as nursing-home placement) is of primary interest. Kuk and Varadhan (2013) developed a stepwise regression approach for Fine and Gray Competing-risk regression model based on AIC, BIC, and BICcr (a modified BIC for Competing-risk regression with right censoring). In a simulation study, they found that these selection procedures performed well, the BIC criterion selected the true model more times than the AIC, and, as expected, the BIC chose more parsimonious models than the AIC in the simulations and the application study.

In clinical setting, there is an increase interest in analyzing multiple outcomes as practitioners want to evaluate the effects of risk-factors on multiple health-related outcomes. Although there are many variable-selection algorithms for individual outcomes, there are limited studies on the selection of predictors for multiple outcomes (Sofer *et al*. 2014).

In this paper, we extended the use of the predictor-selection methods mentioned above for individual outcomes to obtain a common set of predictors for multiple outcomes while accounting for competing risks. We used a variant of backward elimination based on the average normalized-BIC across outcomes. We normalized the BIC so that changes in BIC from complex models to simpler models count approximately the same across multiple outcomes.

To illustrate this method, we used the Health and Retirement Study data and selected a common set of health-related and demographic variables to predict 4 clinical outcomes: time to first Activity of Daily Living (ADL) dependence, time to first Instrumental Activity of Daily Living (IADL) difficulty, time to first walking-across-the-room dependence, and time to death.

SAS® MACRO FOR BIC COMBINED BACKWARD ELIMINATION

We performed BIC backward elimination using a SAS macro developed by our group and the 4 time-to-event clinical outcomes and 39 initial predictors obtained from the Health and Retirement Study data. The predictors 'dAGE' and 'SEX' were forced into the model. The predictor ‘dAGE’ was derived using deciles of the variable age.

When performing a survival analysis, it is important to consider what type of event an individual can experience. If there is only one type of event like all-cause mortality, a Cox regression model (Cox 1972) is appropriate. On the other hand, when there are multiple types of events a competing risk event can delay the observation of the event of interest or modify the chance that this event occurs, so we need to appropriately account for it. Fine and Gray (1999) developed a Survival Competing-risk regression method that accounts for one or more competing risks. In our study, we fitted outcomes 1 through 3, namely time to first ADL dependence, time to first IADL difficulty and time to first walking-across-the-room dependence with Fine and Gray Competing-risk regression so that we could appropriately account for the risk of death. We fitted outcome 4 (time to death) with Cox regression.

In this selection method, we used a normalized-BIC to ensure that a change in BIC from a more complex model to a simpler model means roughly the same across multiple outcomes. To compute the normalized-BIC, we divided the BIC of each outcome for a specific model by the difference between the BIC in the full model and the BIC in the best individual model. That is:

$$Normalized\ BIC(k) = \frac{BIC(k)}{(BIC\ full\ model - BIC\ best\ individual\ model)}$$

Where:

$$BIC = -2\log L + k\log n$$

L: the maximized value of the likelihood function of the fitted model

k: number of parameters estimated by the fitted model

n: sample size

For all fitted models, we measured predictive accuracy using the Harrell’s C-statistic (Harrell 1986) obtained from the CONCORDANCE option in the PHREG procedure. The CONCORDANCE option for computing the Harrell’s C-statistic in PROC PHREG does not allow for Competing-risk regression. Therefore, we used Wolbers *et al*. (2009) adaptation of Harrell’s C-statistic to the competing risks setting, where death status is switched to censored and the time to event is equal to the longest possible time to event that any respondent is followed up (e.g. 15 years).

In general, for a given number of initial predictors in the model (p) the SAS macro selects a subset of p-1 predictors that produces the minimum average normalized-BIC across multiple outcomes; the lower the BIC the better the model. In the next step, the method selects a subset of p-2 predictors that again renders the minimum average normalized-BIC across outcomes. The same process is repeated until there are two variables left in the model—namely, 'dAGE' and 'SEX', which are forced into the model.

Lastly, this method selects the final set of predictors that has the minimum average normalized-BIC across subsets of different number of predictors from p-1 to 2.

SAS MACRO IN DETAIL

In the steps below, we describe how the SAS macro works by presenting fragments of the SAS code. The SAS code shown is for the specific case of 4 outcomes and 37 possible predictors, but that can be easily changed. The complete SAS code for the macro is in the Appendix.

The main steps for the BIC combined backward elimination macro are:

1. There are several macro variables that need to be specified:

   a) DATA: name of the input data set
   b) ALLOUTCOME: all status variables
   c) ALLTIME: all time-to-event variables
   d) ALLLABEL: short labels for time-to-event variables
   e) BASE: list of all predictors in the initial full model
   f) DELE: list of predictors that are to be deleted

   In the code shown below there are 39 predictors in the macro variable ‘BASE’ and 37 predictors in the macro variable ‘DELE’ since predictors ‘dAGE’ and ‘SEX’ are forced in.

   Macro ‘DeleteOneVar’ is called within macro ‘best_bic’ for the first time. In this first run, k=1.

```
   %macro best_bic;
    %let DATA=finaldata;
    %let ALLOUTCOME=status_adldepdth status_iadldifdth status_walkdepdth
death;
    %let ALLTIME=time_adldepdth time_iadldifdth time_walkdepdth time2death;
    %let ALLLABEL= adl iadl walk death;
    %let BASE=%sysfunc(compbl(<all 39 initial variables>));
    %let DELE=%sysfunc(compbl(<all 39 variables-2(e.g. ‘dAGE’ and
‘SEX’)=37>));
    %do k=1 %to 37;
     %DeleteOneVar;
```

2. Macro ‘DeleteOneVar’ is executed for the first time for all the outcomes (e.g. i=1 to 4). For each outcome, the macro fits as many survival models as there are number of variables in the macro variable ‘DELE’. In the code shown below, the macro fits 37 survival models (e.g. j=1 to 38-1). The list of variables used in each survival model is generated by removing one variable at a time from the macro variable ‘BASE’ using the total variables available in the macro variable ‘DELE’. This list is defined in the macro variable ‘VARNAME’. In our example, each of the fitted models has 38 variables, including ‘dAGE’ and ‘SEX’.

```
   %macro DeleteOneVar;
    %do i=1 %to 4;
     %let OUTCOME=%scan(&ALLOUTCOME,&i);
     %let TIME=%scan(&ALLTIME,&i);
     %let LABEL=%scan(&ALLLABEL,&i);
     %do j=1 %to %eval(38-&k);
      %let DELEVAR=%scan(&DELE,&j);
      %let VARNAME=%sysfunc(compbl(%sysfunc(tranwrd(&BASE,&DELEVAR,%str(
)))));
```

3. For time-to-event outcomes with competing-risk events (e.g. outcomes: 1 to 3, namely time to first ADL dependence, time to first IADL difficulty, and time to first walking-across-the-room dependence), the macro fits Competing-risk regression models and uses Wolbers *et al*. (2009) adaptation to compute the Harrell’s C-statistic. The log-likelihood, number of observations in model, degree of freedom, C-statistic, and Integrated Area under the Curve (IAUC) are saved in output data sets to be used in subsequent steps.

```
      %if &i ne 4 %then %do;
       proc phreg data = &DATA;
        class &VARNAME;
        model &time*&outcome(0) = &VARNAME / eventcode=1;
        output out=BSOUT xbeta=xb;
        ods output FITSTATISTICS=FITS1 NObs=NOBS GlobalTests=DF;
       run;
       /*Use Wolbers et al. (2009) adaptation of Harrell’s C-statistic*/
       data BSOUT;
        set BSOUT;
        if &outcome=2 then do;
         &outcome=0;
         &time=15.0278689;
        end;
       run;
       proc phreg data = BSOUT CONCORDANCE=HARRELL
rocoptions(method=RECURSIVE iauc);
        class &VARNAME;
        model &time*&outcome(0) = &VARNAME / nofit;
        roc 'CompRiskC' pred=xb;
        ods output CONCORDANCE=concord IAUC=iauc;
       run;
      proc delete data=BSOUT; run; quit;
      %end;
```

4. For time-to-event outcome with no competing-risk events (e.g. outcome 4, namely time to death), the macro fits Cox regression models. The log-likelihood, number of observations in model, degree of freedom, C-statistic, and Integrated Area under the Curve (IAUC) are saved in output data sets to be used in subsequent steps. Using a DATA _NULL_ step, 4 macro variables are created: degree of freedom (DF), number of observations in model (NOB), C-statistic (c), and Integrated Area under the Curve (iauc).

```
      %else %if &i=4 %then %do;
       proc phreg data = &DATA CONCORDANCE=HARRELL
rocoptions(method=RECURSIVE iauc);
        class &VARNAME;
        model &time*&outcome(0) = &VARNAME;
        ods output FITSTATISTICS=FITS1 NObs=NOBS GlobalTests=DF
CONCORDANCE=concord IAUC=iauc;
       run;
      %end;
      data _null_; set DF; call symputx ('DF',DF); run;
      data _null_; set NOBS; call symputx ('NOB', NObsUsed); run;
      data _null_; set concord; call symputx ('c', Estimate); run;
      data _null_; set iauc; call symputx ('iauc', Estimate); run;
```

5. For each outcome, the macro derives as many ‘FITS2’ data sets as survival models are fitted. In this example, there are 37 different data sets ‘FITS2’ derived. Each ‘FITS2’ data set contains results from a different survival model, and each survival model uses the list of variables defined in the macro variable ‘VARNAME’. The ‘FITS2’ data set has 1 row and 7 columns. The 7 columns are:

a) VARINMODEL: list of predictors in the model
b) DELEVAR: predictor deleted
c) DELELIST: list of predictors that are to be deleted in the next run of macro ‘DeleteOneVar’ (e.g. when k=2)
d) AIC
e) normalized-BIC
f) C-statistic
g) IAUC

```
      data FITS2 (keep=VARINMODEL DELEVAR DELELIST AIC_&label BIC_&label
C_&label iauc_&label);
       set FITS1 end=last;
       retain VARINMODEL DELEVAR C_&label iauc_&label AIC_&label LOGL_&label;
       format AIC_&label LOGL_&label 10.4;
       if _N_=1 then do; AIC_&label=.; end;
       VARINMODEL="&VARNAME";
       DELEVAR="&DELEVAR";
       DELELIST=compbl(tranwrd("&DELE","&DELEVAR",' '));
       NOBS=&NOB;
       DF=&DF;
       C_&label=&c;
       iauc_&label=&iauc;
       if CRITERION='-2 LOG L' then LOGL_&label=WITHCOVARIATES;
       AIC_&label=LOGL_&label+(2*DF);
       BIC_&label=LOGL_&label+(DF*log(NOBS));
       if &i=1 then BIC_&label=BIC_&label/333.79;/*adl: BIC_Full-
BIC_BestIndividual=333.79*/
       else if &i=2 then BIC_&label=BIC_&label/349.72;
       else if &i=3 then BIC_&label=BIC_&label/382.71;
       else if &i=4 then BIC_&label=BIC_&label/195.96;
       if last;
      run;
```

6. For each outcome, the macro appends the different data sets ‘FITS2’ (e.g. 37 total) one by one to generate the data set ‘CTABLE_&label’. In this example, the data set ‘CTABLE_&label’ has the results from 37 survival models—one model per row—and each model has 38 predictors (i.e. 39 initial predictors minus 1 predictor).

```
proc append base=CTABLE_&label data=FITS2 force; run;
```

7. The data sets corresponding to each outcome are joined by the variables included in each model. The macro calculates the average normalized-BIC, the average C-statistic, and the average IAUC across outcomes. The macro selects the model with the minimum average normalized-BIC and saves the selected model with the initial number of variables minus 1 variable in the ‘BIC’ data set. In our example, we start with 39 initial predictors so after the first iteration of the BIC combined backward elimination method, the model has 38 predictors.

```
proc sql;
 create table AVGCTABLE as
 select a.*,
 b.C_iadl, b.iauc_iadl, b.AIC_iadl, b.BIC_iadl,
 c.C_walk, c.iauc_walk, c.AIC_walk, c.BIC_walk,
 d.C_death, d.iauc_death, d.AIC_death, d.BIC_death,
 mean (BIC_adl, BIC_iadl, BIC_walk, BIC_death) as BIC_avg,
 mean (C_adl, C_iadl, C_walk, C_death) as C_avg,
 mean (iauc_adl, iauc_iadl, iauc_walk, iauc_death) as iauc_avg
 from CTABLE_adl a inner join CTABLE_iadl b on a.VARINMODEL=b.VARINMODEL
 inner join CTABLE_walk c on a.VARINMODEL=c.VARINMODEL
 inner join CTABLE_death d on a.VARINMODEL=d.VARINMODEL
 order by BIC_avg descending;
quit;
data AVGCTABLE2; set AVGCTABLE end=last; if last=1; run;
proc append base=BIC data=AVGCTABLE2 force; run;
```

8. After the first execution of the macro ‘DeleteOneVar’ is completed, by using a DATA _NULL step, the macro variable ‘BASE’ is redefined with the variables left in the previous step (e.g. 38). The macro variable ‘DELE’ is also redefined with the variables left in the list to delete from (DELELIST)—that is, 36 variables in our example.

```
%do k=1 %to 37;
 %DeleteOneVar;
 data _null_;
  set AVGCTABLE2 (keep=VARINMODEL DELELIST);
  call symputx ('BASE', VARINMODEL);
  call symputx ('DELE', DELELIST);
 run;
```

9. In our example, in the second run of ‘DeleteOneVar’, the macro fits 36 survival models for each outcome, and each survival model has 37 variables. The data sets corresponding to each outcome are joined by the variables included in each model. The macro calculates the average normalized-BIC, the average C-statistic, and the average IAUC across outcomes. The macro selects the model with the minimum average normalized-BIC and saves the selected model with the initial number of variables minus 2 variables in the ‘BIC’ data set—that is, 37 variables in our example.

10. The macro repeats the above steps until there is only 1 variable left to be removed from the model (e.g. when k=37), so that the last iteration of BIC combined backward elimination results in a model that has only the variables that are forced in (e.g. ‘dAGE’ and ‘SEX’).

11. Lastly, this method selects the final set of common predictors for all outcomes by selecting the model with the minimum average normalized-BIC across subsets of different number of predictors (e.g. from 38 through 2 predictors).

```
proc sort data=BIC out=BIC2; by descending BIC_avg; run;
data BIC2; set BIC2 end=last; if last=1; run;
```

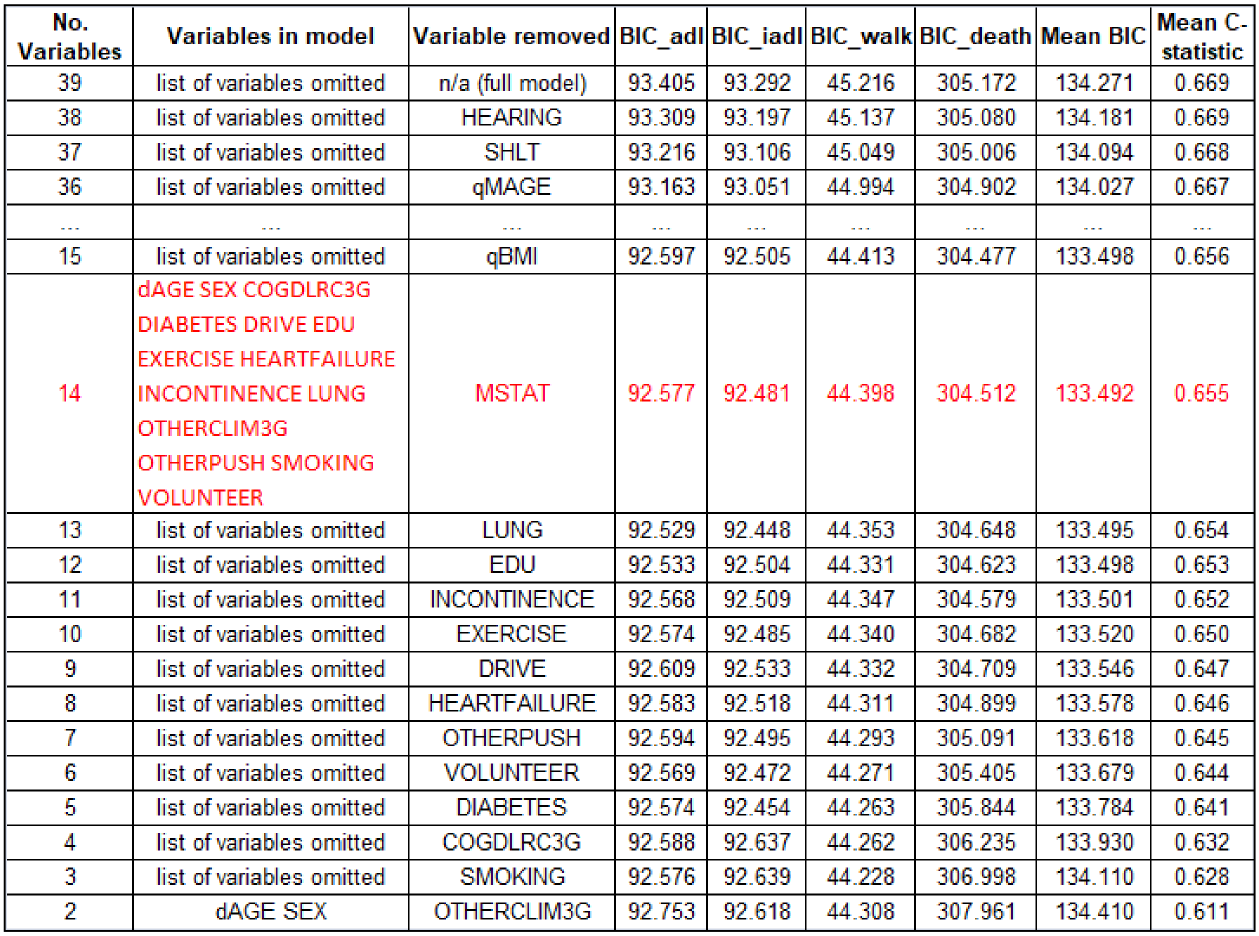
CASE STUDY: HEALTH AND RETIREMENT STUDY DATA

We created a nationally representative cohort of 5,531 community-dwelling seniors enrolled in the Health and Retirement Study (HRS). The HRS is a longitudinal survey of a representative sample of all persons in the United States over age 50. Respondents are interviewed every two years primarily through phone interviews on subjects like health care, housing, assets, pensions, employment, and disability. Our study sample included direct respondents who were 70-years old or older at the time of their interview in 2000 and provided a valid core interview. The follow-up time was up to 2014. In the full model, there were 39 health-related and demographic predictors measured at baseline. The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. We used the public HRS data: Cross-Wave Tracker file and RAND[1] HRS data file.

We used 4 outcomes: time to first ADL dependence (including five ADLs: bathing, dressing, toileting, transferring, and eating), time to first IADL difficulty (including two IADLs: managing money and medication), time to first walking-across-the-room dependence, and time to death. We fitted multivariable Cox proportional hazards models of time to death and multivariable Competing-risk regression models for the rest of the outcomes with death as competing-risk.

Table 1 shows a fragment of the results from BIC combined backward elimination. We highlighted in red the final model with the minimum average normalized-BIC (mean BIC).

[1] The RAND HRS data file is an easy-to-use data set based on the HRS core data. This file was developed at RAND with funding from the National Institute on Aging and the Social Security Administration.



No. Variables	Variables in model	Variable removed	BIC_adl	BIC_iadl	BIC_walk	BIC_death	Mean BIC	Mean C-statistic
39	list of variables omitted	n/a (full model)	93.405	93.292	45.216	305.172	134.271	0.669
38	list of variables omitted	HEARING	93.309	93.197	45.137	305.080	134.181	0.669
37	list of variables omitted	SHLT	93.216	93.106	45.049	305.006	134.094	0.668
36	list of variables omitted	qMAGE	93.163	93.051	44.994	304.902	134.027	0.667
...	...	...	...	...	...	...	...	...
15	list of variables omitted	qBMI	92.597	92.505	44.413	304.477	133.498	0.656
14	dAGE SEX COGDLRC3G DIABETES DRIVE EDU EXERCISE HEARTFAILURE INCONTINENCE LUNG OTHERCLIM3G OTHERPUSH SMOKING VOLUNTEER	MSTAT	92.577	92.481	44.398	304.512	133.492	0.655
13	list of variables omitted	LUNG	92.529	92.448	44.353	304.648	133.495	0.654
12	list of variables omitted	EDU	92.533	92.504	44.331	304.623	133.498	0.653
11	list of variables omitted	INCONTINENCE	92.568	92.509	44.347	304.579	133.501	0.652
10	list of variables omitted	EXERCISE	92.574	92.485	44.340	304.682	133.520	0.650
9	list of variables omitted	DRIVE	92.609	92.533	44.332	304.709	133.546	0.647
8	list of variables omitted	HEARTFAILURE	92.583	92.518	44.311	304.899	133.578	0.646
7	list of variables omitted	OTHERPUSH	92.594	92.495	44.293	305.091	133.618	0.645
6	list of variables omitted	VOLUNTEER	92.569	92.472	44.271	305.405	133.679	0.644
5	list of variables omitted	DIABETES	92.574	92.454	44.263	305.844	133.784	0.641
4	list of variables omitted	COGDLRC3G	92.588	92.637	44.262	306.235	133.930	0.632
3	list of variables omitted	SMOKING	92.576	92.639	44.228	306.998	134.110	0.628
2	dAGE SEX	OTHERCLIM3G	92.753	92.618	44.308	307.961	134.410	0.611

Table 1. BIC Backward Elimination for Combined Outcomes

COMPARISON OF FINAL MODEL FROM BIC COMBINED BACKWARD ELIMINATION WITH MODELS FROM BIC BACKWARD ELIMINATION BY INDIVIDUAL OUTCOMES

We ran the BIC backward elimination macro for each individual outcome and selected a final model for each outcome with the minimum BIC. We created a set of predictors with the union of all predictors in the final model of each outcome, and we compared it with the set of predictors obtained using combined outcomes.

Table 2 shows the result of this comparison. We can see that the combined model is more parsimonious with 14 variables, compared to the model obtained from the union of the final predictors in each individual outcome, which has 22 variables. Also, the predictive accuracy measured using the Harrell's C-statistic in the combined model by outcome is very similar to the predictive accuracy of individual models obtained from BIC backward elimination by outcome and to the predictive accuracy of the union final model by outcome. Variables with gray background in individual models are present in the combined model.

The final combined model has a good balance between parsimony and predictive accuracy. This is important because you want a set of predictors that gives you a model that is simpler to understand and explain (parsimony) while at the same predicts your outcomes well (predictive accuracy).

	time to first ADL dependence (9 variables)	time to first IADL difficulty (8 variables)	time to first walking-across-the-room dependence (7 variables)	time to death (16 variables)	combined (14 variables)	union (22 variables)
Harrell's C-statistic	dAGE SEX DRIVE EDU EXERCISE INCONTINENCE OTHERARM OTHERLIFT OTHERSTOOP	dAGE SEX COGDLRC3G DRIVE EDU HEARAID INCONTINENCE SMOKING	dAGE SEX COGDLRC3G HYPERTENSION INCONTINENCE OTHERLIFT OTHERSTOOP	dAGE SEX COGDLRC3G DIABETES DRIVE EXERCISE HEARTFAILURE HYPERTENSION LUNG MSTAT OTHERCLIM3G OTHERPUSH OTHERWALK qBMI SMOKING VOLUNTEER	dAGE SEX COGDLRC3G DIABETES DRIVE EDU EXERCISE HEARTFAILURE INCONTINENCE LUNG OTHERCLIM3G OTHERPUSH SMOKING VOLUNTEER	dAGE SEX COGDLRC3G DIABETES DRIVE EDU EXERCISE HEARAID HEARTFAILURE HYPERTENSION INCONTINENCE LUNG MSTAT OTHERARM OTHERCLIM3G OTHERLIFT OTHERPUSH OTHERSTOOP OTHERWALK qBMI SMOKING VOLUNTEER
Individual models	0.637	0.635	0.636	0.711		
Combined model by outcome	0.639	0.635	0.642	0.706	0.655 (average 4 models)	
Union model by outcome	0.647	0.638	0.648	0.711	0.661 (average 4 models)	

Table 2. Comparison of Final Predictor Set from BIC Combined Backward Elimination with Final Predictor Sets from BIC Backward Elimination by Outcome

SIMULATION STUDY

We created two sets of 100 simulated data sets with the same sample size (N=5,531) and same number of initial predictors (i.e. 39) with the same distribution as the original data set. In the original data set, the 4 outcomes are highly correlated. Thus, to test whether the correlation among the outcomes impacted the BIC combined selection method, we generated one set of 100 simulated data sets with high correlation among the outcomes and a second set of 100 simulated data sets with low correlation among the outcomes.

We generated the time to event of correlated outcomes from the multivariate normal distribution. First, we used the SIMNORM procedure to generate 4 normal random variables that had means equal to zero, standard deviations equal to 1, and the correlation structure of the original data set. Second, we inverted the random values to probabilities using the function PROBNORM. Finally, we fitted survival models with the predictors in the final combined model and used the simulated probabilities to select the survival probability—Si(t)—and corresponding time to event for each respondent. That is, if at any time point respondents had a survival probability smaller than a specific simulated probability, they were classified as having the event at the last time point where Si(t) is greater than a specific simulated probability. On the other hand, if all their survival probabilities are greater than a specific simulated probability, respondents are censored at their last time to event.

To generate the time to event for uncorrelated outcomes, first, we obtained probabilities from the uniform distribution using the procedure IML, subroutine RANGEN, and distribution UNIFORM. Then, we fitted the survival models with the predictors in the final combined model and used the simulated probabilities to select the survival time and corresponding time to event for each respondent in the same way that we did for correlated outcomes.

We ran the BIC combined backward elimination macro in each of the simulated data sets. For the simulated data sets and outcomes time to first ADL dependence, time to first IADL difficulty, and time to first walking-across-the-room dependence, we simplified the macro by fitting Cox regression models with Wolbers *et al.* (2009) adaptation instead of Competing-risk regression models. The run time to fit Competing-risk regression models is significantly higher than the time for Cox regression (i.e. more than 40 seconds for Competing-risk regression with 39 predictors versus less than 1 second for Cox regression with same number of predictors). Thus, we used Cox regression for all 4 outcomes to avoid the bottleneck in run-time consequence of fitting hundreds of Competing-risk regression models for each outcome and each simulated data set. We do not believe this simplification affected the BIC selection method in the simulated data sets since we obtained the same final set of predictors with and without this simplification in the original data set.

We computed the percentage of times that each of the variables in the final combined model appears in each of the final combined models for the simulated data sets—namely, percentage of correct inclusion—and the percentage of times that each of the variables that are not present in the final combined models do not appear in each of the simulated final combined models—namely, percentage of correct exclusion.

Table 3 shows that in both sets of simulated data sets, with correlated and uncorrelated outcomes, most of the variables in the final combined model have a percentage of correct inclusion of 80% or higher. That is, the majority of these predictors shows up in 80 to 100% of the final combined models using simulated data sets. Except for one variable, most of the variables that are not included in the final combined model do not show up in any of the simulated final combined models. Thus, the percentage of correct exclusion is almost 100%. The variables ‘dAGE’ and ‘SEX’ are not included in this table because they were forced into the model.

Variable	Percentage Correct Inclusion Uncorrelated outcomes	Percentage Correct Inclusion Correlated outcomes
COGDLRC3G	100	100
DIABETES	100	100
DRIVE	95	89
EDU	78	71
EXERCISE	93	91
HEARTFAILURE	94	96
INCONTINENCE	81	78
LUNG	73	74
OTHERCLIM3G	87	82
OTHERPUSH	97	92
SMOKING	100	100
VOLUNTEER	98	98
Mean	91.33	89.25

Table 3. Percentage of Correct Inclusion/Variable Selected in the Final Combined Model using 100 Simulated Data sets with Uncorrelated and Correlated Outcomes

We also computed the mean and 95% Confidence Interval (CI) of the Harrell’s C-statistic and the number of predictors in the final combined models from simulated data sets.

Table 4 shows that the mean of the Harrell’s C-statistic of the final combined models from the simulated data sets is very similar to the C-statistic of the final combined model in the original data. The means of the number of predictors in the final combined models from simulated data sets with correlated and uncorrelated outcomes are slightly smaller compared to the number of final predictors in the original final combined model. This last result may be explained because our attempt to replicate the structure of the original data set did not entirely capture its complexity.

	Harrell's C-Statistic	No. Variables Final Combined Model
Original data set	0.655	14
Simulated data sets with uncorrelated outcomes (mean [95% CI])	0.647 [0.646-0.647]	12.960 [12.793-13.127]
Simulated data sets with correlated outcomes (mean [95% CI])	0.646 [0.645-0.648]	12.720 [12.548-12.892]

Table 4. Mean and 95% CIs of Harrell's C-statistic and Number of Variables in Final Combined Model from 100 Simulated Data sets with Uncorrelated and Correlated outcomes

CONCLUSION

In this paper, we propose a method to select a common set of predictors when analyzing multiple outcomes in the same study. The method produces a final model that is more parsimonious than a model obtained from the union of the best set of predictors in individual outcomes. The predictive accuracy of the models with combined set of predictors is similar to the predictive accuracy of the models using the best set of predictors by outcome. Furthermore, a simulation study of 100 data sets shows a high percentage of correct inclusion per variable selected in the final combined model. Although the method shown here was developed for time-to-event outcomes using BIC as the statistic for selection, it could also be applied to continuous or categorical outcome variables, and other Information Criteria like the AIC can be used.

REFERENCES

Akaike, H. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control*, 19:716-723.

Cox, D. R. 1972. "Regression models and life tables." *Journal of the Royal Statistical Society, Series B*, 34:187-220.

Fine, J. P. and R. J. Gray. 1999. "A proportional hazards model for the subdistribution of a competing risk." *Journal of the American Statistical Association*, 94:496-509.

Harrell, F. E. 1986. "The PHGLM Procedure." In *SUGI Supplemental Library Guide*, *Version 5 Edition*. SAS Institute Inc., Cary, NC.

Health and Retirement Study, (RAND HRS Data, Version P) public use data set. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2016).

Health and Retirement Study, (Cross-Wave Tracker File 2014 Final, Version 1.0) public use data set. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2017).

Kuk, D. and R. Varadhan. 2013. "Model selection in competing risks regression." *Statistics in Medicine*, 32:3077-3088.

RAND HRS Data, Version P. Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration. Santa Monica, CA (August 2016).

Schwarz, G. 1978. "Estimating the dimension of a model." *Annals of Statistics*, 6:461-464.

Shtatland, E. S., Kleinman K., and E. M. Cain. 2003. "Stepwise methods in using SAS® PROC LOGISTIC and SAS® ENTERPRISE MINER for prediction." *Proceedings of the SAS® Users Group International*, Seattle, WA: SAS Institute Inc., Cary, NC. Available at http://www2.sas.com/proceedings/sugi28/258-28.pdf.

Shtatland, E. S., Kleinman K., and E. M. Cain. 2005. "Model building in PROC PHREG with automatic variable selection and information criteria." *Proceedings of the SAS® Users Group International*, Philadelphia, PA: SAS Institute Inc., Cary, NC. Available at http://www2.sas.com/proceedings/sugi30/206-30.pdf.

Sofer, T., Dicker, L., and X. Lin. 2014. "Variable selection for high dimensional multivariate outcomes." *Statistica Sinica*, 24:1633-1654.

Wolbers, M., Koller, M. T., Witteman, J. C., and E. W. Steyerberg. 2009. "Prognostic models with competing risks: methods and application to coronary risk prediction." *Epidemiology*, 20:555-561.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institute on Aging R01AG047897.

CONTACT INFORMATION

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

L. Grisell Diaz-Ramirez
University of California, San Francisco
grisell.diaz-ramirez@ucsf.edu

Siqi Gan
Healthcare Department, Philips Research China
siqi.gan@philips.com

Sei J. Lee
University of California, San Francisco
sei.lee@ucsf.edu

Alexander K. Smith
University of California, San Francisco
alexander.smith@ucsf.edu

W. John Boscardin
University of California, San Francisco
john.boscardin@ucsf.edu

APPENDIX. SAS MACRO TO PERFORM BIC COMBINED BACKWARD ELIMINATION

```
   %macro DeleteOneVar;
    %do i=1 %to 4;
/*Extract the ith status and ith time*/
     %let OUTCOME=%scan(&ALLOUTCOME,&i);
     %let TIME=%scan(&ALLTIME,&i);
     %let LABEL=%scan(&ALLLABEL,&i);
     %do j=1 %to %eval(38-&k);
/*Select the jth word to delete. &DELE is defined in 'best_bic' macro*/
      %let DELEVAR=%scan(&DELE,&j);
/*Select the final set of variables to run model by replacing the deleted variable
with blank*/
      %let VARNAME=%sysfunc(compbl(%sysfunc(tranwrd(&BASE,&DELEVAR,%str(
)))));

/*For model fitted using Competing-risk regression compute: log-likelihood, number of
observations in model, degree of freedom, C-statistic, and Integrated Area under the
Curve (IAUC)*/
      %if &i ne 4 %then %do;
       proc phreg data = &DATA;
        class &VARNAME;
        model &time*&outcome(0) = &VARNAME / eventcode=1;
        output out=BSOUT xbeta=xb;
        ods output FITSTATISTICS=FITS1 NObs=NOBS GlobalTests=DF;
       run;
/*The status and time are changed to apply Wolbers et al. (2009) adaptation of
Harrell’s C-statistic*/
       data BSOUT;
        set BSOUT;
        if &outcome=2 then do;
         &outcome=0;
         &time=15.0278689; /*maximum follow-up time*/
        end;
       run;
       proc phreg data = BSOUT CONCORDANCE=HARRELL
rocoptions(method=RECURSIVE iauc);
        class &VARNAME;
        model &time*&outcome(0) = &VARNAME / nofit;
        roc 'CompRiskC' pred=xb;
        ods output CONCORDANCE=concord IAUC=iauc;
       run;
      proc delete data=BSOUT; run; quit;
      %end;

/*For model fitted using Cox regression compute: log-likelihood, number of
observations in model, degree of freedom, C-statistic, and IAUC*/
      %else %if &i=4 %then %do;
       proc phreg data = &DATA CONCORDANCE=HARRELL
rocoptions(method=RECURSIVE iauc);
        class &VARNAME;
        model &time*&outcome(0) = &VARNAME;
        ods output FITSTATISTICS=FITS1 NObs=NOBS GlobalTests=DF
CONCORDANCE=concord IAUC=iauc;
       run;
      %end;
      data _null_; set DF; call symputx ('DF',DF); run;
      data _null_; set NOBS; call symputx ('NOB', NObsUsed); run;
      data _null_; set concord; call symputx ('c', Estimate); run;
      data _null_; set iauc; call symputx ('iauc', Estimate); run;

/*Create data set with information of variables in the model, deleted variable, list
of variables to delete from, AIC, BIC, C-statistic, and IAUC*/
      data FITS2 (keep=VARINMODEL DELEVAR DELELIST AIC_&label BIC_&label
C_&label iauc_&label);
       set FITS1 end=last;
       retain VARINMODEL DELEVAR C_&label iauc_&label AIC_&label LOGL_&label;
       format AIC_&label LOGL_&label 10.4;
       if _N_=1 then do; AIC_&label=.; end;
       VARINMODEL="&VARNAME"; /*variables in the model*/
       DELEVAR="&DELEVAR"; /*deleted variable*/
       DELELIST=compbl(tranwrd("&DELE","&DELEVAR",' '));
/*'DELELIST' contains the list of variables to delete from in subsequent runs. In our
example, the second time macro 'DeleteOneVar' runs there are 36 variables in macro
variable 'DELE' instead of 37. At this step, we call it 'DELELIST' and it contains the
'DELE' list minus the variable deleted in this run. Later, 'DELELIST' is redefined as
'DELE'*/
       NOBS=&NOB;
       DF=&DF;
       C_&label=&c;
       iauc_&label=&iauc;
       if CRITERION='-2 LOG L' then LOGL_&label=WITHCOVARIATES;
       AIC_&label=LOGL_&label+(2*DF); /*AIC=-2ln*(L)+2k, where -2ln*(L)=LOGL*/
       BIC_&label=LOGL_&label+(DF*log(NOBS)); /*BIC=-2*ln(L) + k*ln(n)*/
       if &i=1 then BIC_&label=BIC_&label/333.79; /*Normalized Outcome 1 (ADL):
BIC_Full-BIC_BestIndividual=333.79*/
       else if &i=2 then BIC_&label=BIC_&label/349.72; /*Normalized Outcome 2
(IADL): BIC_Full-BIC_BestIndividual=349.72*/
       else if &i=3 then BIC_&label=BIC_&label/382.71; /*Normalized Outcome 3
(Walk): BIC_Full-BIC_BestIndividual=382.71*/
       else if &i=4 then BIC_&label=BIC_&label/195.96; /*Normalized Outcome 4
(Death): BIC_Full-BIC_BestIndividual=195.96*/
       if last;
      run;

/*SASFILE statement with the LOAD option opens ‘CTABLE_&label’ data set, allocates the
buffers, and reads the data into memory. This improves performance*/
      %if &j=1 %then %do;
       proc append base=CTABLE_&label data=FITS2 force; run;
       sasfile WORK.CTABLE_&label load;
      %end;
      %else %do;
       proc append base=CTABLE_&label data=FITS2 force; run;
      %end;
      proc delete data=concord iauc FITS1 FITS2 NOBS DF; run; quit;
     %end; /*end j do loop*/
     sasfile WORK.CTABLE_&label close;
    %end; /*end i do loop*/
```

```
/*Join 4 data sets corresponding to each outcome by the variables included in the
model. Calculate the average normalized-BIC, the average C-statistic, and the average
IAUC across outcomes. Order data set ‘AVGCTABLE’ in descending order so that the
predictor subset with minimum average normalized-BIC is the last one.*/
    proc sql;
     create table AVGCTABLE as
     select a.*,
     b.C_iadl, b.iauc_iadl, b.AIC_iadl, b.BIC_iadl,
     c.C_walk, c.iauc_walk, c.AIC_walk, c.BIC_walk,
     d.C_death, d.iauc_death, d.AIC_death, d.BIC_death,
     mean (BIC_adl, BIC_iadl, BIC_walk, BIC_death) as BIC_avg,
     mean (C_adl, C_iadl, C_walk, C_death) as C_avg,
     mean (iauc_adl, iauc_iadl, iauc_walk, iauc_death) as iauc_avg
     from CTABLE_adl a inner join CTABLE_iadl b on a.VARINMODEL=b.VARINMODEL
     inner join CTABLE_walk c on a.VARINMODEL=c.VARINMODEL
     inner join CTABLE_death d on a.VARINMODEL=d.VARINMODEL
     order by BIC_avg descending;
    quit;
/*Select the subset of predictors with the minimum average normalized-BIC*/
    data AVGCTABLE2; set AVGCTABLE end=last; if last=1; run;
    proc delete data=CTABLE_adl CTABLE_iadl CTABLE_walk CTABLE_death
AVGCTABLE; run; quit;
   %mend DeleteOneVar;

   %macro best_bic;
    %let DATA=finaldata;
    %let ALLOUTCOME=status_adldepdth status_iadldifdth status_walkdepdth
death;
    %let ALLTIME=time_adldepdth time_iadldifdth time_walkdepdth time2death;
    %let ALLLABEL= adl iadl walk death;
    %let BASE=%sysfunc(compbl(<all 39 initial variables>));
    %let DELE=%sysfunc(compbl(<all 39 variables-2(dAGE SEX)=37>));
/*In our example, 'DeleteOneVar' macro runs for predictor subsets that ranges from 38
through 2 variables*/
    %do k=1 %to 37;
     %DeleteOneVar;
     data _null_;
      set AVGCTABLE2 (keep=VARINMODEL DELELIST);
/*Redefine the ‘BASE’ macro variable with the variables left in the previous run of
the macro*/
      call symputx ('BASE', VARINMODEL);
/*Redefine the ‘DELE’ macro variable with the variables left in the 'DELELIST'*/
      call symputx ('DELE', DELELIST);
     run;
     %if &k=1 %then %do;
/*BIC is the final data set with results of all the models obtained from performing
BIC combined backward elimination. In our example, we have 39 initial predictors and 2
predictors forced in, so the BIC data set has results from models with 38 predictors
through 2 predictors*/
      proc append base=BIC data=AVGCTABLE2 force; run;
      sasfile WORK.BIC load;
     %end;
     %else %do;
      proc append base=BIC data=AVGCTABLE2 force; run;
     %end;
     proc delete data=AVGCTABLE2; run; quit;
    %end;
   %mend best_bic;
```