Paper 252-2009

Test for linearity between continuous confounder and binary outcome first, run a multivariate regression analysis second

Jiming Fang, Peter C. Austin, Jack V. Tu
Institute for Clinical Evaluative Sciences, Toronto, ON, Canada

ABSTRACT

Previous statistical studies have indicated that dichotomizing a continuous confounding variable in multivariate regression analyses can lead to biased estimation of the effect of exposures, treatments, and risk factors on outcomes. We suggest that, prior to entry in the multivariate analysis, one should test whether or not the continuous confounding variable is linearly related to log-odds of the binary outcome or hazard ratios of the time-to-event binary outcome. If there is a linear relationship, we encourage that the variable not be dichotomized. We illustrate this issue using clinical data that we recently published in the New England Journal of Medicine (NEJM). The linearity assumption is tested by restricted cubic splines using SAS/Stat® procedures.

INTRODUCTION

In clinical studies, investigators are frequently interested in determining the association between risk factors and binary outcomes (e.g., dead or alive at discharge) or time-to-event binary outcomes (e.g., time to death, time to readmission) after adjusting for confounding variables. Continuous confounders (such as patient age, blood pressure, glucose, etc.) are often dichotomized prior to entry in the multivariate regression model (e.g., old vs. young, high vs. low blood pressure). Dichotomization is widespread in clinical studies (Del Priore 1997), as such simplicity simplifies the statistical analysis and leads to easy interpretation and presentation of the results.

Dichotomization assumes that the relationship between the predictor and the outcome is flat within intervals, i.e., a discontinuity in outcome as interval boundaries are crossed, however, this assumption is far less reasonable than a linearity assumption in most cases. From a statistical point of view, dichotomization reduces statistical power, primarily due to the reduction in the inherent variability of the predictor variable. It is known that dichotomizing continuous predictor variables may result in biased estimation, either in ordinary linear regression (Cumsille et al. 2000), or in logistic regression (Becher 1992; Schulgen et al. 1994), or in Cox proportional hazards regression (Altman et al. 1994; Buettner et al. 1997; Royston et al. 2006). Furthermore, it may result in inflation of the type-I error of the risk factor (Altman et al. 1994; Austin and Brunner 2004). In this paper, we emphasize this issue based on a previous clinical study of time-to-event outcome of heart failure that we conducted (Bhatia et al. 2006).

CASE STUDY: OUTCOME OF HEART FAILURE WITH PRESERVED EJECTION FRACTION

We conducted a study to compare the features and outcomes of patients with heart failure with preserved ejection fraction (n=880) with those of patients with heart failure with reduced ejection fraction (n=1570).

We used Cox proportional-hazards regression analysis to identify factors associated with an increased risk of death after hospitalization for heart failure. Backward variable elimination was used to create a parsimonious model for predicting mortality. Systolic blood pressure was one of the variables selected for inclusion in the final Cox model. By forcing the variable denoting ejection-fraction group into the final

Paper 252-2009

model, we determined the adjusted hazard ratio for death among patients with reduced ejection fraction as compared with those with preserved ejection fraction.

Systolic blood pressure is a continuous variable. However, it is often dichotomized to determine its association with mortality outcomes in cardiovascular studies. An interesting finding is that if we dichotomized systolic blood pressure (>120 versus <=120 mmHg) in the Cox model, then we find that the adjusted hazard of mortality for patients with reduced ejection fraction was significantly higher than that of patients with preserved ejection fraction (hazard ratio: 1.202, 95%CI, 1.004-1.440, p-value=0.0445, Table 1). However, if systolic blood pressure was treated as a continuous variable in the Cox model, then the adjusted hazard rate was not significantly different (hazard ratio: 1.131; 95%CI, 0.943-1.355, p-value=0.1847). Thus, the final conclusion about the differences in the risk of death between the two heart failure groups may depend on how systolic blood pressure was treated in the regression model. Which hazard ratio should we report?

Table 1. Adjusted hazard ratio of 1-Year mortality in heart failure patients with reduced ejection fraction versus preserved ejection fraction

Adjusted Confounder	Hazard Ratio (95% Confidence Interval)	P-value
Adjusted for dichotomized systolic blood pressure (>120 versus <=120 mmHg)	1.202 (1.004-1.440)	0.0445
Adjusted for dichotomized systolic blood pressure (>140 versus <=140 mmHg)	1.173 (0.979-1.405)	0.0830
Adjusted for systolic blood pressure, age as continuous scale	1.131 (0.943-1.355)	0.1847
Adjusted for dichotomized age (>70 versus <=70 years old)	1.086 (0.907-1.301)	0.3702

LINEARITY ASSUMPTION TEST FOR CONTINUOUS PREDICTORS: RESTRICTED CUBIC SPLINES

To address this problem, we tested the linearity assumption of the relationship between systolic blood pressure and the log-hazard of mortality. The linear relationship was assessed by the use of restricted cubic splines (Harrell 2001). A Wald chi-square test showed there was a linear relationship between systolic blood pressure and the log-hazard of mortality (Wald Chi-square=5.6547, p=0.1297, Figure 1). Therefore, we reported the hazard ratio 1.131 in our NEJM paper (Bhatia et al. 2006, Table 1), stating that the survival of patients with heart failure with preserved ejection fraction was similar to that of patients with reduced ejection fraction.

Paper 252-2009

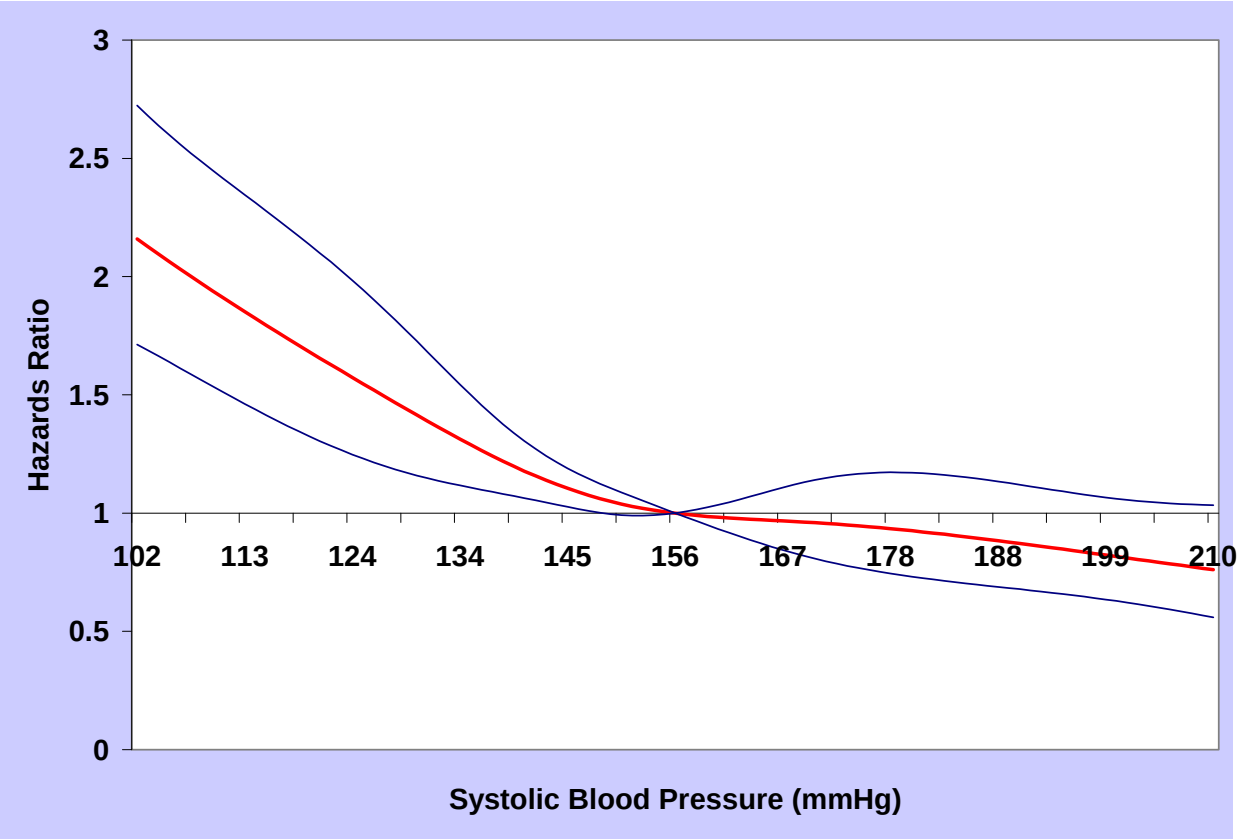

Figure 1. Functional relationship (and 95% pointwise confidence band) between systolic blood pressure and the hazard ratios of 1-year mortality in heart failure patients, estimated by a restricted cubic spline function with 5 knots at 120, 130, 150, 170 and 210 mmHg

SAS MACRO: %RCS

Restricted Cubic Splines can be done easily using Heizel and Kaider's SAS macro, %RCS (Heinzl and Kaider 1997). Visit www.meduniwien.ac.at/msi/biometrie/programme/Rcs.htm (accessed on January 23, 2009) to download the macro and its manual.

Before running the %RCS, we selected 5 knots with fixed percentiles (Harrell 2001). The following SAS program shows how to create a macro variable (&SBP) for the five knots of systolic blood pressure.

```
proc univariate data=chf;
  var SysBP;
  output out=BP pctlpts=5 27.5 50 72.5 95 pctlpre=bp;
run;

proc transpose data=BP out=BP;
run;

proc sql;
  select col1 into: SBP separated by " "
  from BP;
run;
```

Paper 252-2009

We then called %RCS like this.

```
%include "c:/SASmacros/rcs.sas";

%rcs(
title=%str(restricted cubic splines: entire cohort),
data=CHF,
dirdata=%str(C:/SASData/CHF),
program=%str(C:/SugiPaper/sugi_paper_rcs.sas),
time=survival_1yr,
status=mortality_1yr,
cov1=SysBP, what1=0, knots1=&SBP,
graph=both,
timeunit=mmHg
);
```

Running this macro will generate another SAS program, called sugi_paper_rcs.sas. Its first part is for the linearity assumption test.

```
TITLE ' restricted cubic splines: entire cohort ';

PROC PHREG DATA=CHF COVOUT OUTEST=__RCS;
 MODEL survival_1yr*mortality_1yr(0) =  SysBP  __1_1  __1_2  __1_3  /RL;

 ********** spline modelling of fixed covariate SysBP;
 ********** with 5 knots located at;
 ********** 102 130 150 170 210;
 __1_1=((SysBP-102)**3)*(SysBP>102)
     -((SysBP-170)**3)*(SysBP>170)
     *(210-102)/(210-170)
     +((SysBP-210)**3)*(SysBP>210)
     *(170-102)/(210-170);
 __1_2=((SysBP-130)**3)*(SysBP>130)
     -((SysBP-170)**3)*(SysBP>170)
     *(210-130)/(210-170)
     +((SysBP-210)**3)*(SysBP>210)
     *(170-130)/(210-170);
 __1_3=((SysBP-150)**3)*(SysBP>150)
     -((SysBP-170)**3)*(SysBP>170)
     *(210-150)/(210-170)
     +((SysBP-210)**3)*(SysBP>210)
     *(170-150)/(210-170);

 *--------- Testing variable: vs_sysbp ---------;
 EFFECT1: TEST  SysBP, __1_1, __1_2, __1_3;
 NONLIN1: TEST  __1_1, __1_2, __1_3;
 RUN;
*==================== End of PROC PHREG ====================;
```

Paper 252-2009

The following SAS output shows linearity between systolic blood pressure and 1-year mortality.

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
VS_SYSBP	1	-0.01409	0.00464	9.2041	0.0024
__1_1	1	-1.2346E-6	3.16071E-6	0.1526	0.6961
__1_2	1	0.0000108	0.0000144	0.5567	0.4556
__1_3	1	-0.0000193	0.0000222	0.7583	0.3839

Analysis of Maximum Likelihood Estimates

Variable	Hazard Ratio	95% Hazard Ratio Confidence Limits		Variable Label
VS_SYSBP	0.986	0.977	0.995	Systolic blood pressure on admission
__1_1	1.000	1.000	1.000	
__1_2	1.000	1.000	1.000	
__1_3	1.000	1.000	1.000	

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
EFFECT1	72.1239	4	<.0001
NONLIN1	5.6547	3	0.1297

Paper 252-2009

Run the second part of sugi_paper_rcs.sas is to draw the graph of relative hazard ratio function (see Figure 1).

```
*-------------------- Graph for SysBP --------------------;
PROC IML;
 NPOINTS=101;   * Number of points to build the graphic;
 LOWEREND=102;     *Smallest value for X-axis;
 UPPEREND=210;     *Largest value for X-axis;
 REF=(102+210)/2;     *Reference value for X-axis;
 X=T(DO(LOWEREND,UPPEREND,(UPPEREND-LOWEREND)/(NPOINTS-1)));
 S1=((X-102)##3)#(X>102)
     -((X-170)##3)#(X>170)
     #(210-102)/(210-170)
     +((X-210)##3)#(X>210)
     #(170-102)/(210-170)
     -((REF-102)##3)#(REF>102)
     +((REF-170)##3)#(REF>170)#(210-102)/(210-170)
     -((REF-210)##3)#(REF>210)#(170-102)/(210-170);
 S2=((X-130)##3)#(X>130)
     -((X-170)##3)#(X>170)
     #(210-130)/(210-170)
     +((X-210)##3)#(X>210)
     #(170-130)/(210-170)
     -((REF-130)##3)#(REF>130)
     +((REF-170)##3)#(REF>170)#(210-130)/(210-170)
     -((REF-210)##3)#(REF>210)#(170-130)/(210-170);
 S3=((X-150)##3)#(X>150)
     -((X-170)##3)#(X>170)
     #(210-150)/(210-170)
     +((X-210)##3)#(X>210)
     #(170-150)/(210-170)
     -((REF-150)##3)#(REF>150)
     +((REF-170)##3)#(REF>170)#(210-150)/(210-170)
     -((REF-210)##3)#(REF>210)#(170-150)/(210-170);
 XMAT=(X-REF)||S1||S2||S3;
 HV={ SysBP __1_1 __1_2 __1_3 };
 USE __RCS;  READ ALL VAR HV INTO C;
 READ ALL VAR { _NAME_ } INTO HC;  CLOSE __RCS;
 B=C[1,]` ;  HC=REPEAT(HC,1,NCOL(HV));
 HV=REPEAT(HV,NROW(HC),1);
 HV=(upcase(HC)=upcase(HV))[,+];
 HV=LOC(HV#(1:NROW(C))`);  C=C[HV,];
 F=XMAT*B; FU=XMAT*C*XMAT`; FREE XMAT;
 FU=SQRT(VECDIAG(FU));  FO=F+1.96*FU;  FU=F-1.96*FU;
 Z=J(NROW(F),1,1)//J(NROW(F),1,2)//J(NROW(F),1,3);
 F=F//FO//FU; FE=EXP(F);  X=REPEAT(X,3,1);
 CREATE __RCS1 VAR { F FE Z X };  APPEND;  CLOSE __RCS1;
 QUIT;

 SYMBOL1 C=RED  L=1 I=JOIN WIDTH=5;
 SYMBOL2 C=BLUE L=2 I=JOIN WIDTH=5;
 SYMBOL3 C=BLUE L=2 I=JOIN WIDTH=5;

PROC GPLOT DATA=__RCS1;
 PLOT FE*X=Z / VREF=1 LV=3 NOLEGEND;
LABEL FE=HAZARD RATIO;
RUN;
QUIT;
```

Paper 252-2009

EXTENSIONS

The decision of whether or not to dichotomize a continuous predictor variable does not always lead to differences in the significance of a given risk factor. Similar to systolic blood pressure, the confounding variable age was also linearly associated with the log-hazard of mortality in the present case study (Wald Chi-square=1.2807, p=0.7337). However, the 95% confidence interval of the hazard ratio after adjusting for the dichotomized age variable (cut-off at 70 years old) still crossed 1 (Table 1).

In addition, if the systolic blood pressure was dichotomized at 140 mmHg instead of 120 mmHg, the hazard ratio turned out to be not significantly different from 1 (Table 1), suggesting that the results of an analysis may vary depending on the cutoff points that are used, and the use of different cutpoints may make comparisons across studies extremely difficult (Buettner et al. 1997; Royston et al. 2006). The simplicity achieved by dichotomization is gained at a cost; dichotomization may create rather than avoid problems (Royston et al. 2006). See Altman and Royston (2006) for a brief review on the cost of dichotomizing a continuous variable.

CONCLUSIONS

The present case-based study suggested that when a continuous independent confounding variable is dichotomized prior to entry in a multivariate regression analysis, the conclusion about the effect of a risk factor on the outcome might be biased, especially when the confounder had a linear relationship to the log-odds of binary outcome or the hazard ratios of time-to-event binary outcome. Investigators are encouraged to first test the linearity of the relationship between the confounder and the binary outcome prior to dichotomization. If there is a linear relationship, treat the confounder as a continuous variable. Failure to do so could result in an inflated assessment of the statistical significance of the association between one of the variables and the outcome of interest.

REFERENCES

Altman DG, Lausen B, Sauerbrei W, Schumacher M (1994) Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 86:829-835.

Altman DG, Royston P (2006) The cost of dichotomising continuous variables. *British Medical Journal* 332:1080.

Austin PC, Brunner LJ (2004) Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine* 23:1159-1178.

Buettner P, Garbe C, Guggenmoos-Holzmann I (1997) Problems in defining cutoff points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma. *Journal of Clinical Epidemiology* 50:1201-1210.

Becher H (1992) The concept of residual confounding in regression models and some applications. *Statistics in Medicine* 11:1747-1758.

Bhatia RS, Tu JV, Lee DS, Austin PC, Fang J, Haouzi A, Gong Y, Liu PP (2006) Outcome of heart failure with preserved ejection fraction in a population-based study. *New England Journal Medicine* 355:260-269.

Cumsille F, Bangdiwala SI, Sen PK, Kupper LL (2000) Effect of dichotomizing a continuous variable on the model structure in multiple linear regression models. *Communications in Statistics – Theory and Methods* 29:643-654.

Del Priore G, Zandieh P, Lee MJ (1997) Treatment of continuous data as categoric variables in obstetrics and gynecology. *Obstetrics and Gynecology*. 89:351-354.

Harrell, FE (2001) Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. Springer-Verlag New York, Inc. New York, USA.

Heinzl H, Kaider A (1997) Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine* 54:201-208.

Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 25:127-141.

Schulgen G, Lausen B, Olsen JH, Schumacher M (1994) Outcome-oriented cutpoints in analysis of quantitative exposures. *American Journal of Epidemiology* 140:172-184.

ACKNOWLEDGEMENTS

This study used data from the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) study. The EFFECT was supported by a grant to the Canadian Cardiovascular Outcomes Research Team from the Canadian Institutes of Health Research and the Heart and Stroke Foundation.

CONTACT INFORMATION

Jiming Fang, PhD
Institute for Clinical Evaluative Sciences
G106, 2075 Bayview Avenue
Toronto, ON M4N 3M5,
Canada
Tel: 416-480-6100 Ext. 3613
Fax: 416-480-6048
Email: jiming.fang@ices.on.ca
ICES Web Site: www.ices.on.ca

Peter C Austin, PhD
Institute for Clinical Evaluative Sciences
G106, 2075 Bayview Avenue
Toronto, ON M4N 3M5,
Canada
Tel: 416-480-6131
Fax: 416-480-6048
Email: peter.austin@ices.on.ca
ICES Web Site: www.ices.on.ca

Jack V. Tu, MD, PhD, FRCPC,
Institute for Clinical Evaluative Sciences
G106, 2075 Bayview Avenue
Toronto, ON M4N 3M5,
Canada
Tel: 416-480-6038
Fax: 416-480-6048
Email: tu@ices.on.ca
ICES Web Site: www.ices.on.ca