

COX REGRESSION USING DIFFERENT TIME-SCALES

Alison J. Canchola¹, Susan L. Stewart¹, Leslie Bernstein², Dee W. West¹, Ronald K. Ross², Dennis Deapen², Richard Pinder², Peggy Reynolds³, William Wright⁴, Hoda Anton-Culver⁵, David Peel⁵, Al Ziogas⁵, and Pamela L. Horn-Ross¹
¹Northern California Cancer Center, Union City, CA; ²Keck School of Medicine, University of Southern California, Los Angeles, CA; ³Environmental Health Investigations Branch, California Department of Health Services, Oakland, CA; ⁴Cancer Surveillance Section, California Department of Health Services, Sacramento, CA; ⁵School of Medicine, University of California, Irvine, CA

ABSTRACT

Typically in cohort studies, the time-scale used in Cox regression models is time-on-study, adjusting for age as a covariate. However, age can also be used as the time-scale, where subjects enter the analysis at their baseline age (left-truncation) and exit at their event/censoring age. Using SAS[®] PROC PHREG, we compared five methods using time-on-study and age as the time-scales. We used a subset of data (n=85,882) from the large California Teachers Study cohort, with five years of follow-up, invasive breast cancer as the outcome (n=1,428) and alcohol consumption (≥ 20 grams/day) as the risk factor of interest. Using a time-on-study time-scale adjusting for age continuous in years produced a slight overestimate of the alcohol effect compared with the more accurate but computationally more intensive age time-scale methods. However, allowing for different age effects for younger and older ages gave virtually identical results as using an age time-scale.

INTRODUCTION

Typically in cohort studies, the time-scale used in Cox regression models is time-on-study (i.e. follow-up time or time since baseline), adjusting for baseline age as a covariate in the model. However, age can also be used as the time-scale, where subjects enter the analysis at their baseline age (left-truncation) and exit at their event/censoring age. Models with time-on-study as the time-scale can be adjusted for age in different ways, and models with age as the time-scale can be adjusted for calendar effects. Using SAS[®] PROC PHREG, we compared five Cox regression methods (using time-on-study and age as the time-scales) and examined their effect on risk factor estimates for a particular model.

We used data from the California Teachers Study (CTS), which is a large statewide cohort of 133,479 female California public school teachers and administrators. Most participants filled out the baseline questionnaire in late 1995/early 1996. For this analysis, we had follow-up through 12/31/2000; the outcome of interest was incident invasive breast cancer; and the risk factor of interest was alcohol consumption of ≥ 20 grams/day, which is about 2 or more drinks per day.

Studies have shown a positive association between alcohol consumption and breast cancer risk (Horn-Ross 2002; Chen 2002), with the suggestion of a threshold-type effect at about ≥ 20 grams/day. Age is associated with both breast cancer risk and alcohol consumption, so it is important to adjust for it properly in the analysis.

Cox regression is a semi-parametric time-to-event analysis method. The hazard function for an individual i at time t with a set of covariates, x_p , is written as

$$h_i(t) = h_0(t) \exp\{\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}\}$$

where $h_0(t)$ is a baseline hazard function which is left unspecified and $\exp\{\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}\}$ is a linear function of the covariates which is exponentiated. The hazard rate ratio of any two individuals, i and j , is then

$$\frac{h_i(t)}{h_j(t)} = \exp\{\beta_1(x_{1i} - x_{1j}) + \dots + \beta_p(x_{pi} - x_{pj})\}.$$

The baseline hazard function does not need to be specified in order to compute the hazard rate ratio, because it drops out of the equation (Allison 1995; Collett 1994).

Cox regression models use a partial likelihood method. At each event time, a likelihood function is estimated. This likelihood function is the hazard rate for the subject who experienced the event at time t , divided by the sum of the hazard rates for all subjects under follow-up and at risk at that event time (i.e., with follow-up $\geq t$). A product is then taken over all of these individual likelihood functions, and this function is maximized with respect to the parameter estimates (Allison 1995). The time-scale is important because which individuals are considered at risk and thus contribute to the sum of the hazard rates at a particular event time depends on the time-scale used.

METHODS

To be eligible for this analysis, subjects had to (hierarchically): be living in California at baseline (n=124,613 of the total cohort of 133,479); have no history of breast cancer at baseline (n=118,346); be less than age 85 at baseline (n=116,352); and have

valid/non-missing alcohol data for the year before baseline (n=102,224). For Cox regression models, subjects were required to have complete data for the confounders included in the analysis (race, daily caloric intake, family history of breast cancer in a first degree relative, age at menarche, nulliparity/age at first full term pregnancy, physical activity, estrogen therapy use/duration, body mass index (weight in kg/(height in m)²) and menopausal status) (n=85,882).

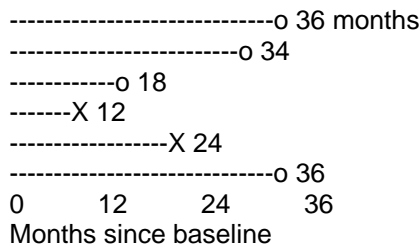
The outcome was invasive breast cancer, which was identified through linkage with the California Cancer Registry (CCR) and diagnosed in 1,428 women after baseline and through 12/31/2000. Women who were not diagnosed with invasive breast cancer during follow-up were censored at the earliest of: death, estimated date of moving out of California, diagnosis of in-situ breast cancer or 12/31/2000. Hazard rate ratios (HRs) and 95% confidence intervals were computed for alcohol consumption in the year before baseline, which was categorized into non-drinkers and 5 levels of consumption. A test for trend was calculated across these categories treating the categories as an ordinal variable.

We compared five Cox regression methods:

1. Time-on-study as the time-scale, i.e. time since baseline, with age at baseline (continuous in years) included as a covariate in the model.

Using time-on-study as the time-scale, all subjects enter the risk set for the model together, at time 0 (baseline). In our cohort, most subjects filled out the baseline questionnaire in late 1995/early 1996, so this time-scale is a proxy for calendar time.

Example of time-on-study as the time-scale:
o=censored X=event



At the first event time in the example above, at 12 months, all 6 subjects are in the risk set. At the second event time at 24 months, 4 subjects are in the risk set, because 2 subjects are no longer under follow-up (one eliminated due to an event, and another due to censoring).

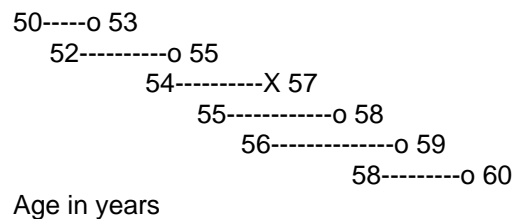
With this time-scale, at an event occurring at time *t*, the model computes risk estimates based on all those subjects whose follow-up durations were $\geq t$. In the

example above, at 24 months, we obtain risk estimates based on those subjects who were followed ≥ 24 months. The risk set includes women of all different ages, so we adjust for age as a covariate (continuous in years) in the model. With this method, the effect of age is modeled linearly (i.e., the log of the hazard function is a linear function of age). Thus the increase in risk of the outcome with each year of age is required to be the same for each year, although the underlying risk may actually change with age (which is true for cancer outcomes).

2. Age as the time-scale, i.e. subjects enter and exit at their age.

Using age as the time-scale, subjects enter the risk set at the age they filled out the baseline questionnaire and exit at their event/censoring age. When age is the time-scale, time 0 would actually be birth. But since people are not followed since birth, we have them enter the risk set at the age they filled out the baseline questionnaire, given that they have lived to this age with no diagnosis of breast cancer. This is called left truncation or late entry into the risk set. Left truncation allows us to put subjects in risk sets only when they are actually under follow-up by the study, which is necessary when age is the time-scale.

Example of age as the time-scale:
o=censored X=event



In this example, at the event at age 57, only the 3 subjects who are under follow-up at age 57 are in the risk set when the model calculates the risk estimates.

There are several advantages to using age as the time-scale:

a) First, notice an important difference between the approaches in #1 and #2. Using time-on-study as the time-scale (method #1), at the event at 24 months, the model is computing risk estimates for those with ≥ 24 months of follow-up, whether they were 30 or 60 years old. Using age as the time-scale (method #2), at the event at age 57, the model is computing risk estimates for those who were 57 years old (whether they were 57 at baseline or in 2000). Korn et al. (1997) argue that for outcomes like breast cancer, 'we would expect the hazard to change more as a function of age than as a function of time-on-study'. For

example, we would expect more of a difference in breast cancer risk between a 30 and 60 year old with the same amount of follow-up, than between two 57 year olds with a different amount of follow-up (i.e., who were 57 in different calendar years). So, Korn et al. recommend using age as the time-scale, which puts subjects with similar risks (i.e., the 57 year olds) in the risk set together; or even better, a variation on this, method #4 (which is discussed below).

b) It allows for a completely nonparametric age effect. For these two reasons, age as the time-scale is the more accurate way to analyze cohort data of this type.

However, using age as the time-scale does have some disadvantages:

a) It is computationally intensive. Using PC-SAS v9.0, to run the two models of interest (6 levels of alcohol consumption and a trend test), it took 8 minutes to run method #2, compared with 13 seconds for method #1.

(Also, on Unix SAS v6.12, it took 45 minutes to run #2, compared with about 1 minute for #1.)

b) Calendar time is important, especially as we accrue more years of follow-up. For example, due to cultural, lifestyle, and medical changes, a woman who was 60 years old in 1996 may be different from a woman who is 60 years old in 2006. However, this can be adjusted for using method #4 (discussed below).

Moreover, according to Korn et al., there are some situations when using time-on-study versus age as the time-scale will give similar results:

- a) When the baseline hazard function is approximately exponentially distributed with age.
- b) When the risk factors of interest are independent of age.

3. Time-on-study as the time-scale, with current age (continuous in years) as a time-dependent covariate in the model.

This sounds like an improvement over method #1, but as detailed by Allison (1995, pg 142), by definition this approach gives the exact same results as method #1.

When the effect of age is modeled linearly, then age as a time-dependent covariate and age as a fixed covariate are equivalent, since if the effect per year of age is constant, then the effect of a given age difference remains the same over time.

4. Age as the time-scale, stratifying the model by birth cohort (5-year intervals), which adjusts for calendar effects.

This approach is the same as method #2, using age as the time-scale, but we stratify the model by birth year in 5-year intervals, which adjusts for calendar effects. Specifically, we used the intervals: 1910-

1914, 1915-1919, 1920-1924, 1925-1929, ..., 1970-1974.

This is the method that Korn et al. recommend using for follow-up of cohorts of healthy people, when calendar effects may be influencing the outcome and risk factors.

5. Time-on-study as the time-scale, with two slopes for age, to allow different effects for younger and older ages.

Here, like method #1 and #3, age is modeled linearly, but the slope is allowed to differ for younger and older ages. To do this, we define two age terms: $AGELT50 = AGE - 50$, where $AGE - 50 < 0$ and 0 otherwise; $AGEGE50 = AGE - 50$, where $AGE - 50 \geq 0$ and 0 otherwise.

For example,

AGE	AGELT50	AGEGE50
37	-13	0
45	-5	0
52	0	2
65	0	15

RESULTS AND DISCUSSION

The hazard rate ratios (HRs) for alcohol consumption using all of these methods are fairly similar (Table 1). Results from methods #1 and #3 are exactly the same by definition. Results from methods #2 and #4 are the same, indicating that calendar effects are not an issue at this point with relatively short follow-up. The HRs for method #1 (time-on-study time-scale) are slightly higher than for method #2 (age time-scale), but the conclusions remain the same. There are several possible reasons why the results from methods #1 and #2 are similar:

a) The follow-up period is still relatively short. Perhaps as the follow-up period increases, the time-scale will be more of an issue.

b) Using method #2, a plot of the cumulative hazard function by age looked approximately exponentially distributed. According to Korn et al., this is one situation where a time-on-study versus age time-scale will not make a big difference to the HR estimates. In fact, Ingram et al. (1997) point out that most outcomes of typical epidemiologic studies have an underlying hazard function that is approximately exponential. They reanalyzed 4 studies from the NHANES I Epidemiologic Follow-up Study that had originally used time-on-study as the time-scale in Cox regression models; using age as the time-scale, they found similar results in all cases. Thus, while they feel that using age as the time-scale may have some

theoretical advantages over time-on-study, in practice, it makes little difference.

Then why might there be a slight overestimate of the alcohol effect using time-on-study versus age as the time-scale (#1 versus #2)? First, alcohol consumption is not independent of age; older women consumed higher levels of alcohol (Table 2). Second, the incidence rate of breast cancer increases faster in younger than older women. When only one age slope is estimated (as in method #1), it is shallower than the slope among women under 50 but steeper than the slope among women over 50. Thus, we tried a time-on-study model allowing for different age effects for younger and older ages (method #5). In method #1, the HR for a one-year increase in age was 1.03; in method #5, it was 1.12 for those under 50 and 1.02 for those over 50. Using one age slope, the age effect is underestimated among the younger women. Thus when older ages are compared with younger ages, the difference in risk is underestimated. Because alcohol use is greater at older ages, many of the events will be among women whose age and alcohol use are both high. With one age slope, due to the limitations imposed on the age effect (i.e. the difference in risk is underestimated), there are more events at older ages than would be expected due to age, so the partial likelihood is maximized by increasing the alcohol effect. This may explain why the HRs for alcohol for method #1 are slightly higher than method #2 (where we allow a completely non-parametric age effect). However, by simply letting the linear age effect differ for younger and older ages with a time-on-study time-scale (method #5), we obtain similar results to #2.

CONCLUSIONS

In conclusion, with cohort data of this type, using age as the time-scale gives more accurate results because it puts similar subjects in the risk set together and allows a completely non-parametric age effect. However, it is computationally intensive and in practice, the results are often similar to those obtained using a time-on-study time-scale.

When alcohol consumption is the risk factor of interest, compared with using age as the time-scale, using time-on-study with one age slope gives a slight overestimate of the alcohol effect, whereas allowing for different age effects for younger and older ages gives virtually identical results and is computationally less intensive.

In general, perhaps for most Cox regression models with incident breast cancer as the outcome and a relatively short follow-up period, the time-scale does not have a major impact. However, it may in some

situations. Thus, the choice of the proper time-scale and the best way to adjust for age should be considered.

REFERENCES

Allison PD. *Survival Analysis Using the SAS System : A Practical Guide*, Cary, NC: SAS Institute, Inc., 1995. 292 pp.

Chen WY, Colditz GA, Rosner B, Hankinson SE, Hunter DJ, Manson JE, Stampfer MJ, Willett WC, Speizer FE. Use of postmenopausal hormones, alcohol, and risk for invasive breast cancer. *Ann Intern Med* 2002; 137(10):798-804.

Collett D. *Modelling Survival Data in Medical Research*, Chapman & Hall, London, 1994. 347 pp.

Horn-Ross PL, Hoggatt KJ, West DW, Krone MR, Stewart SL, Anton-Culver H, Bernstein L, Deapen D, Peel D, Pinder R, Reynolds P, Ross RK, Wright W, Ziogas A. Recent diet and breast cancer risk: the California Teachers Study (USA). *Cancer Causes and Control* 2002; 13:407-415.

Ingram DD, Makuc DM, Feldman JJ. Re: "Time-to-event analysis of longitudinal follow-up of a survey: choice of time-scale." *Am J Epidemiol* 1997; 146:528-529.

Korn EL, Graubard BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of time-scale. *Am J Epidemiol* 1997; 145:72-80.

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

ACKNOWLEDGEMENTS

We would like to thank Jesse Canchola at the University of California, San Francisco, for his help reviewing this paper.

This work was supported by grant RO1 CA77398 from the National Cancer Institute and contract 97-10500 from the California Breast Cancer Research Fund.

AUTHOR CONTACT INFORMATION

Alison Canchola, MS
Northern California Cancer Center
32960 Alvarado-Niles Road, Suite 600
Union City, CA 94587
(510) 429-2533
acanchol@nccc.org

SAS PROGRAM

AGEATQ = age at baseline, continuous in years
PMTHS = person-months of follow-up
EVCA3 = diagnosis of invasive breast cancer during follow-up (1/0)
STDT15 = date filled out baseline questionnaire, mm/15/yy
ENDDT15 = date of end of follow-up, mm/15/yy
ALCYR1-5 = dummy variables for non-drinkers (reference) and 5 levels of alcohol consumption

```
data stend;  
  set alc;
```

```
* Calculate age entered and exited follow-up;  
* SAS handles left-truncation in Cox regression models as (enter,exit], open on the left, so subjects are not in the risk set when they enter, although they should be. So, subtract a tiny amount from each enter age, so subjects are entering the risk set at the age they filled out the baseline questionnaire;  
  birth15=mdy(month(birthx),15,year(birthx));  
*Start age;  
stage2=((stdt15 - birth15)/365.25) - 0.0001;  
*End age;  
endage=(enddt15 - birth15)/365.25;
```

```
* Calculate 2 age slopes, for younger and older ages;  
if ageatq - 50 < 0 then do;  
  age150=ageatq-50; agege50=0; end;  
else if ageatq - 50 >= 0 then do;  
  age150=0; agege50=ageatq - 50; end;  
run;
```

```
* 1. Time-on-study, adjusting for age at baseline.;  
proc phreg data=stend;  
  model pmths*evca3(0) =  
    alcyr1 alcyr2 alcyr3 alcyr4 alcyr5  
    ageatq nonwhite dt_kcal famhx bmic2 bmic3  
    menoc2 menoc3 b2m2 b2m3 b3m2 b3m3  
    menar12 fftp0 fftp2529 fftp30 spmp3yr  
    ertle5 ertgt5 /rl;  
run;
```

```
* 2. Age as the time-scale (i.e. enter and exit at age);  
proc phreg data=stend;  
  model (stage2,endage)*evca3(0) =  
    alcyr1 alcyr2 alcyr3 alcyr4 alcyr5  
    ageatq nonwhite dt_kcal famhx bmic2 bmic3  
    menoc2 menoc3 b2m2 b2m3 b3m2 b3m3  
    menar12 fftp0 fftp2529 fftp30 spmp3yr  
    ertle5 ertgt5 /rl;  
  title Using age as the time scale (i.e. enter and exit at age);  
run;
```

```
* 3. Time-on-study with current age as a time-dependent covariate.;  
proc phreg data=stend;  
  model pmths*evca3(0) =  
    alcyr1 alcyr2 alcyr3 alcyr4 alcyr5  
    currage nonwhite dt_kcal famhx bmic2 bmic3  
    menoc2 menoc3 b2m2 b2m3 b3m2 b3m3  
    menar12 fftp0 fftp2529 fftp30 spmp3yr  
    ertle5 ertgt5 /rl;  
  currage=  
    int((((stdt15 - birth15)/30.4375) + pmths)/12);  
  title Current age as a time-dependent covariate;  
run;
```

```
* 4. Age as the time-scale, stratifying by birth cohort in 5-year intervals (which adjusts for calendar effects);  
proc phreg data=stend;  
  model (stage2,endage)*evca3(0) =  
    alcyr1 alcyr2 alcyr3 alcyr4 alcyr5  
    ageatq nonwhite dt_kcal famhx bmic2 bmic3  
    menoc2 menoc3 b2m2 b2m3 b3m2 b3m3  
    menar12 fftp0 fftp2529 fftp30 spmp3yr  
    ertle5 ertgt5 /rl;  
  strata birthc;  
  title Using age as the time scale, stratified by birth cohort;  
run;
```

```
* 5. Time-on-study, allowing for different age effects for younger and older ages;  
proc phreg data=stend;  
  model pmths*evca3(0) =  
    alcyr1 alcyr2 alcyr3 alcyr4 alcyr5  
    age150 agege50  
    nonwhite dt_kcal famhx bmic2 bmic3  
    menoc2 menoc3 b2m2 b2m3 b3m2 b3m3  
    menar12 fftp0 fftp2529 fftp30 spmp3yr  
    ertle5 ertgt5 /rl;  
  title Time-on-study, allowing for diff age effects for younger & older ages;  
run;
```

Table 1. Alcohol consumption in the year before baseline in the CTS. Hazard rate ratio (95% confidence interval)^a.

Method ^c	non-drinkers	grams/day of alcohol					p-value for trend
		<5	5-9	10-14	15-19	≥20	
	[449] ^b	[269]	[227]	[197]	[125]	[161]	
#1 Time	1.00	1.06 (0.91-1.24)	1.05 (0.90-1.23)	1.14 (0.97-1.35)	1.04 (0.85-1.27)	1.34 (1.12-1.61)	0.009
#2 Age	1.00	1.03 (0.89-1.20)	1.03 (0.88-1.21)	1.11 (0.94-1.31)	1.00 (0.82-1.22)	1.27 (1.06-1.53)	0.044
#3 Time	1.00	1.06 (0.91-1.24)	1.05 (0.90-1.23)	1.14 (0.97-1.35)	1.04 (0.85-1.27)	1.34 (1.12-1.61)	0.009
#4 Age	1.00	1.03 (0.89-1.20)	1.03 (0.88-1.21)	1.11 (0.94-1.31)	1.00 (0.82-1.22)	1.27 (1.06-1.53)	0.045
#5 Time	1.00	1.04 (0.89-1.21)	1.03 (0.88-1.21)	1.11 (0.94-1.32)	1.01 (0.82-1.23)	1.28 (1.07-1.54)	0.035

^a Adjusted for age, race, daily caloric intake, family history of breast cancer, age at menarche, nulliparity/age at first full-term pregnancy, physical activity, estrogen therapy use/duration, and an interaction term for body mass index and menopausal status.
^b [Number of cases].

^c Methods:

1. Time-on-study as the time-scale, with age at baseline included as a covariate in the model.
2. Age as the time-scale, i.e. subjects enter and exit at their age.
3. Time-on-study as the time-scale, with current age as a time-dependent covariate in the model.
4. Age as the time-scale, stratifying the model by birth cohort (5-year intervals).
5. Time-on-study as the time-scale, allowing for different age effects for younger and older ages.

Table 2. Percent consuming ≥ 20 grams/day of alcohol, by age at baseline.

Age	Total N	≥ 20 g/d of alcohol	
		n	%
20-29	3,365	144	4.3
30-39	12,789	598	4.7
40-49	25,065	1,651	6.6
50-59	21,916	2,054	9.4
60-69	13,753	1,642	11.9
70-79	7,536	699	9.3
80-84	1,458	101	6.9