Choice of time-scale in the Cox model for epidemiologic cohort studies where entry has no direct biological relevance

Paul Dickman and Anna Johansson
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

October 28, 2005

Slides available at `http://www.pauldickman.com/teaching/`

Time-varying exposure in epidemiological cohort studies

- Study 1: We wish to examine the association between exposure to radioactive iodine and incidence of thyroid cancer among survivors of the Chornobyl accident. From a biological perspective it is important to consider
  - age at exposure
  - time since exposure
  - (attained age)
- Time since exposure is a 'time-varying' explanatory variable (the value changes with time) whereas age at exposure is fixed for each individual.
- Study 2: Invite women from the general population to participate in a cohort study; follow-up to assess the association between diet and incidence of breast cancer.
- From a biological perspective it is important to consider age at time of follow-up (attained age), a 'time-varying' explanatory variable.

- Time since entry is not of direct biological interest.
- The choice of variables to adjust for in a statistical model should be based, first and foremost, on biological and clinical considerations; we should only adjust for time-since entry if it has direct biological relevance.
- How do we, technically, adjust for the fact that a single exposure variable can assume multiple values for a single individual?
- One approach is to 'split' the person-time for each individual into bands, creating a data set containing multiple observations for each individual.
- This is what we do with Poisson regression; can adjust for two (but not three) time-varying explanatory variable.

- Estimates from the Cox model are always adjusted for one time-varying variable (the underlying time-scale) automatically.
- We get to adjust for one time-varying confounder 'for free'.
- It is therefore sensible to choose the most important time-varying confounder as the underlying time-scale.
- For many epidemiological cohort studies this is attained age.
- Can adjust for a second time-varying variable by splitting the data.

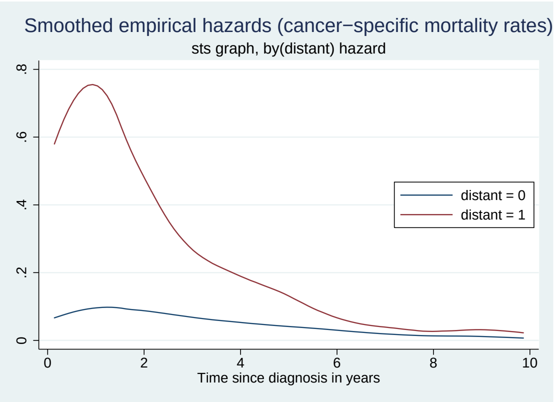
The Cox proportional hazards model

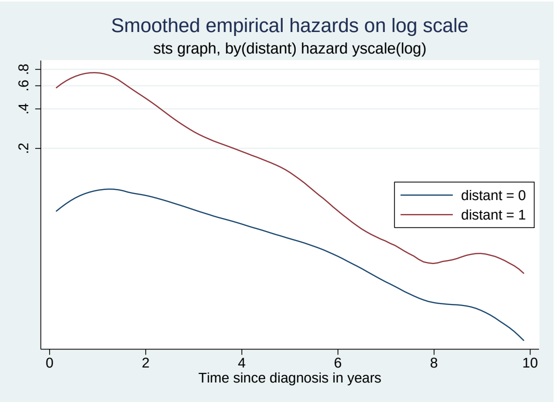
- The 'intercept' in the Cox model, the hazard (event rate) for individuals with all covariates z at the reference level, is an arbitrary function of time[1], often called the baseline hazard and denoted by $\lambda_0(t)$.
- The hazard at time t for individual with other covariate values is a multiple of the baseline
$$\lambda(t|z) = \lambda_0(t)\exp(\beta' z).$$
- Can extend the model to a 'stratified Cox model' which has separate baseline hazards for each level of some factor $j = 1, \ldots, J$
$$\lambda(t|j, z) = \lambda_{0j}(t)\exp(\beta' z).$$

[1] time t can be defined in many ways, e.g., attained age, time-on-study, calendar time, etc.

Example: survival of patients diagnosed with colon carcinoma

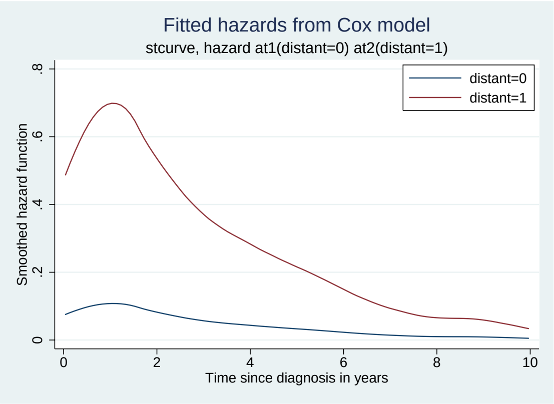
- Patients diagnosed with colon carcinoma in Finland 1984–95. Potential follow-up to end of 1995; censored after 10 years.
- Outcome is death due to colon carcinoma.
- Time-scale t is time-since-diagnosis in years.
- Interested in the effect of clinical stage at diagnosis (distant metastases vs no distant metastases).


Smoothed empirical hazards (cancer−specific mortality rates)
sts graph, by(distant) hazard

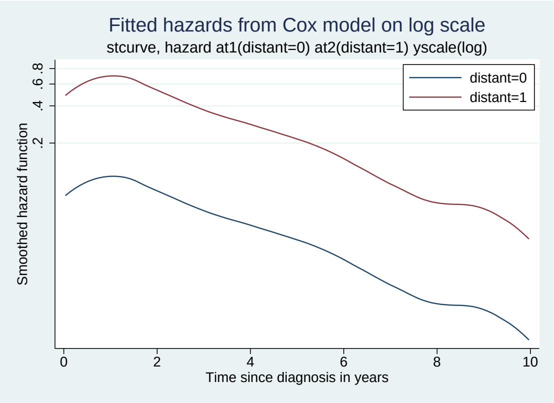

Smoothed empirical hazards on log scale
sts graph, by(distant) hazard yscale(log)

Fit a Cox model

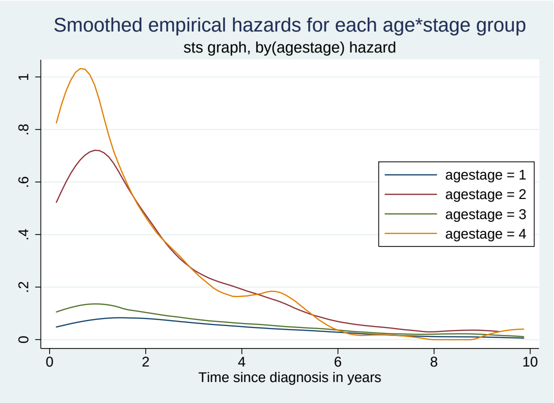
```
. stcox distant, basehc(base)

         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.25
             origin:  time dx

No. of subjects =        14648          Number of obs   =      14648
No. of failures =         7186
Time at risk    =  64134.28611
                                        LR chi2(1)      =    6164.83
Log likelihood  =   -62951.506          Prob > chi2     =     0.0000

--------------------------------------------------------------------
     _t | Haz. Ratio   Std. Err.    z    P>|z|   [95% Conf. Interval]
--------+-----------------------------------------------------------
distant |   7.190404   .1833347  77.37   0.000   6.839905    7.558863
--------------------------------------------------------------------
```


Fitted hazards from Cox model

stcurve, hazard at1(distant=0) at2(distant=1)


Fitted hazards from Cox model on log scale

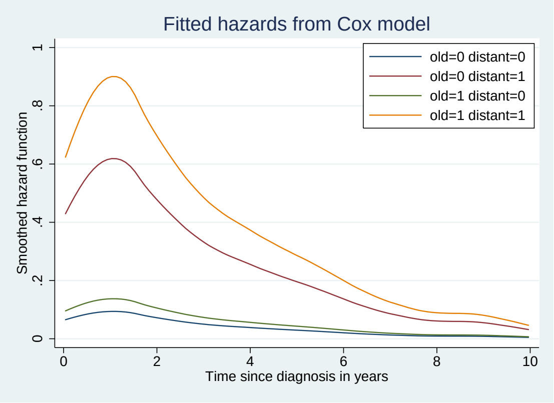
stcurve, hazard at1(distant=0) at2(distant=1) yscale(log)


Smoothed empirical hazards for each age*stage group

sts graph, by(agestage) hazard

Fit a Cox model adjusted for age at diagnosis

```
. stcox distant old, basehc(base)

         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.25
             origin:  time dx

No. of subjects =        14648         Number of obs   =      14648
No. of failures =         7186
Time at risk    =  64134.28611
                                       LR chi2(2)      =    6496.87
Log likelihood  =   -62785.488         Prob > chi2     =     0.0000

--------------------------------------------------------------------
     _t | Haz. Ratio   Std. Err.    z    P>|z|   [95% Conf. Interval]
--------+-----------------------------------------------------------
distant |   7.252431    .185139  77.61   0.000   6.898494    7.624528
    old |    1.57537   .0384735  18.61   0.000    1.50174    1.652611
--------------------------------------------------------------------
```


Fitted hazards from Cox model

Possible models

- Let a_0 be the age at entry and t the time-on-study.
- Using time-on-study as the time scale and adjusting for age at entry we have

$$\lambda(t|a_0, z) = \lambda_0(t)\exp(\xi a_0 + \gamma' z) \qquad \text{(Korn model 4).}$$

- Using attained age as the time scale we have

$$\lambda(a|a_0, z) = \lambda_0(a)\exp(\beta' z) \qquad \text{(Korn model 5).}$$

- Model (4) is appropriate for the cancer survival data but not for epidemiological cohort studies where time-on-study has no direct relevance.

- Nevertheless, this model is commonly applied in epidemiology.
- Model (5) is appropriate for epidemiological cohort studies (provided there are no cohort or period effects).
- Korn et al. [1] argue for the model with age as the time-scale and stratified on birth cohorts B_j

$$\lambda(a|b_0 \in B_j, z) = \lambda_{0j}(a)\exp(\beta' z) \qquad \text{(Korn model 3)}$$

that is, separate baseline hazards for each birth cohort.

- We will focus on a comparison of models (4) and (5), those most commonly applied in epidemiology.
- In particular, we will study conditions under which model (5) is correct but model (4) provides estimates without large bias.

Similarity of models (4) and (5)

- Assume model (5) is appropriate (hazard depends on attained age and there are no period or birth cohort effects).
- We also assume, for the moment, that the exposure of interest does not vary over time.
- Korn et al. suggested two conditions under which the γ's estimated from model (4) are similar to the β's estimated from model (5) is
  1. the baseline hazard $\lambda_0(a) = c\exp(\psi a)$ for some $c > 0$ and ψ; or
  2. the baseline ages, a_0, are independent of the covariates z.
- Thiébaut and Bénichou (2004) [2] performed simulations and observed bias even when the second condition was met.
- First condition can be written as $\ln[\lambda_0(a)] = \ln(c) + \psi a$;
  – we require the log hazard to be a linear function of (attained) age.

Example 1 from Korn et al.; condition 1 is satisfied

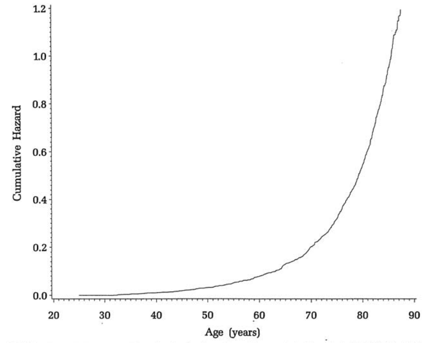


FIGURE 1. Cumulative hazard for mortality as a function of age (age ≥25 years) for women being followed in the NHANES I Epidemiologic Followup Study.

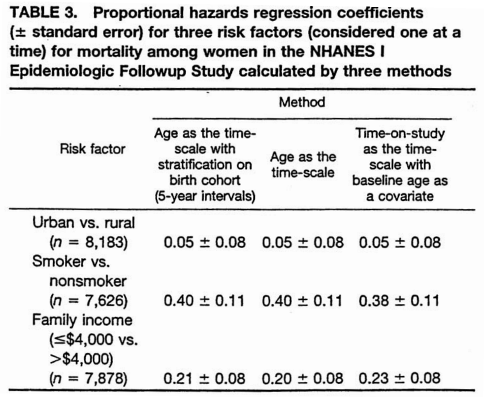


TABLE 3. Proportional hazards regression coefficients (± standard error) for three risk factors (considered one at a time) for mortality among women in the NHANES I Epidemiologic Followup Study calculated by three methods

Risk factor	Method		
	Age as the time-scale with stratification on birth cohort (5-year intervals)	Age as the time-scale	Time-on-study as the time-scale with baseline age as a covariate
Urban vs. rural (n = 8,183)	0.05 ± 0.08	0.05 ± 0.08	0.05 ± 0.08
Smoker vs. nonsmoker (n = 7,626)	0.40 ± 0.11	0.40 ± 0.11	0.38 ± 0.11
Family income (≤$4,000 vs. >$4,000) (n = 7,878)	0.21 ± 0.08	0.20 ± 0.08	0.23 ± 0.08

Example 2 from Korn et al.; condition 1 is not satisfied

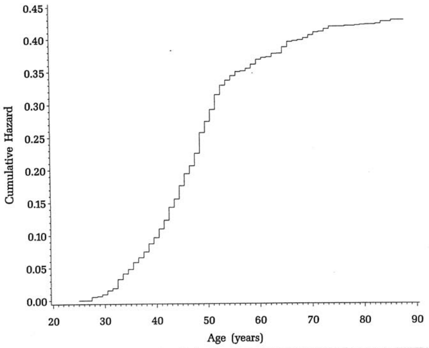


FIGURE 2. Cumulative hazard for ovary removal as a function of age (age ≥25 years) for women being followed in the NHANES I Epidemiologic Followup Study.

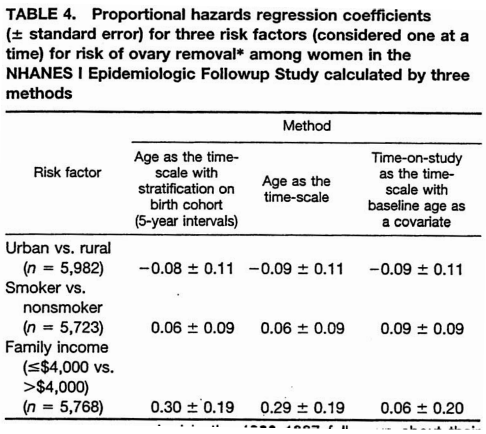


TABLE 4. Proportional hazards regression coefficients (± standard error) for three risk factors (considered one at a time) for risk of ovary removal* among women in the NHANES I Epidemiologic Followup Study calculated by three methods

Risk factor	Method		
	Age as the time-scale with stratification on birth cohort (5-year intervals)	Age as the time-scale	Time-on-study as the time-scale with baseline age as a covariate
Urban vs. rural (n = 5,982)	−0.08 ± 0.11	−0.09 ± 0.11	−0.09 ± 0.11
Smoker vs. nonsmoker (n = 5,723)	0.06 ± 0.09	0.06 ± 0.09	0.09 ± 0.09
Family income (≤$4,000 vs. >$4,000) (n = 5,768)	0.30 ± 0.19	0.29 ± 0.19	0.06 ± 0.20

Simulation study of Thiébaut and Bénichou (2004)

- Designed to simulate risk of breast cancer in the E3N cohort, 100k French women aged 40–65 years at recruitment (1989/90).
- Exposure of interest is menopausal status at recruitment (time-fixed) and menopausal status (time-varying).
- Table I: Covariate independent of age at entry
- Table II: Covariate dependent on age at entry; $\beta = 0$
- Table III: Covariate dependent on age at entry; $\beta = \ln(5)$

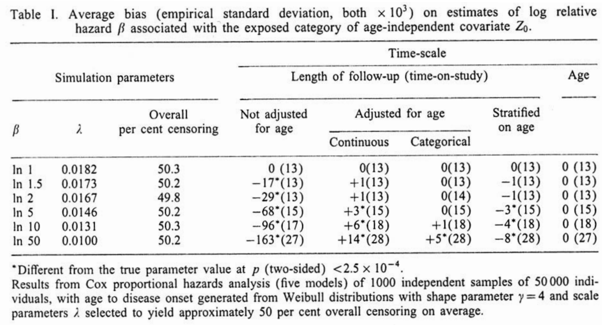


Table I. Average bias (empirical standard deviation, both $\times 10^3$) on estimates of log relative hazard β associated with the exposed category of age-independent covariate Z_0.

Simulation parameters			Time-scale				
			Length of follow-up (time-on-study)				Age
			Not adjusted for age	Adjusted for age		Stratified on age	
β	λ	Overall per cent censoring		Continuous	Categorical		
ln 1	0.0182	50.3	0 (13)	0(13)	0(13)	0(13)	0 (13)
ln 1.5	0.0173	50.2	−17*(13)	+1(13)	0(13)	−1(13)	0 (13)
ln 2	0.0167	49.8	−29*(13)	+1(13)	0(14)	−1(13)	0 (13)
ln 5	0.0146	50.2	−68*(15)	+3*(15)	0(15)	−3*(15)	0 (15)
ln 10	0.0131	50.3	−96*(17)	+6*(18)	+1(18)	−4*(18)	0 (18)
ln 50	0.0100	50.2	−163*(27)	+14*(28)	+5*(28)	−8*(28)	0 (27)

*Different from the true parameter value at p (two-sided) $<2.5 \times 10^{-4}$.
Results from Cox proportional hazards analysis (five models) of 1000 independent samples of 50 000 individuals, with age to disease onset generated from Weibull distributions with shape parameter $\gamma = 4$ and scale parameters λ selected to yield approximately 50 per cent overall censoring on average.

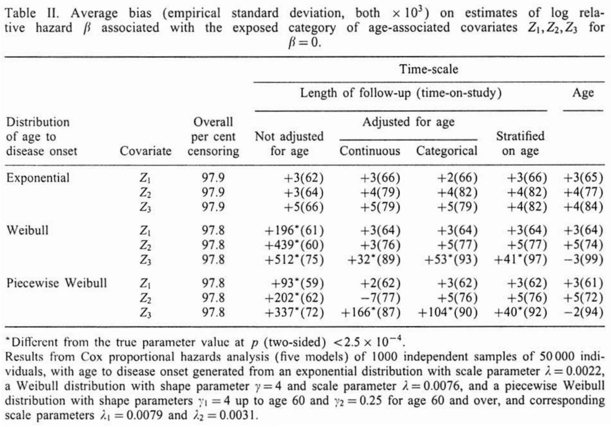


Table II. Average bias (empirical standard deviation, both $\times 10^3$) on estimates of log relative hazard β associated with the exposed category of age-associated covariates Z_1, Z_2, Z_3 for $\beta = 0$.

			Time-scale				
			Length of follow-up (time-on-study)				Age
Distribution of age to disease onset	Covariate	Overall per cent censoring	Not adjusted for age	Adjusted for age		Stratified on age	
				Continuous	Categorical		
Exponential	Z_1	97.9	+3(62)	+3(66)	+2(66)	+3(66)	+3(65)
	Z_2	97.9	+3(64)	+4(79)	+4(82)	+4(82)	+4(77)
	Z_3	97.9	+5(66)	+5(79)	+5(79)	+4(82)	+4(84)
Weibull	Z_1	97.8	+196*(61)	+3(64)	+3(64)	+3(64)	+3(64)
	Z_2	97.8	+439*(60)	+3(76)	+5(77)	+5(77)	+5(74)
	Z_3	97.8	+512*(75)	+32*(89)	+53*(93)	+41*(97)	−3(99)
Piecewise Weibull	Z_1	97.8	+93*(59)	+2(62)	+3(62)	+3(62)	+3(61)
	Z_2	97.8	+202*(62)	−7(77)	+5(76)	+5(76)	+5(72)
	Z_3	97.8	+337*(72)	+166*(87)	+104*(90)	+40*(92)	−2(94)

*Different from the true parameter value at p (two-sided) $<2.5 \times 10^{-4}$.
Results from Cox proportional hazards analysis (five models) of 1000 independent samples of 50 000 individuals, with age to disease onset generated from an exponential distribution with scale parameter $\lambda = 0.0022$, a Weibull distribution with shape parameter $\gamma = 4$ and scale parameter $\lambda = 0.0076$, and a piecewise Weibull distribution with shape parameters $\gamma_1 = 4$ up to age 60 and $\gamma_2 = 0.25$ for age 60 and over, and corresponding scale parameters $\lambda_1 = 0.0079$ and $\lambda_2 = 0.0031$.

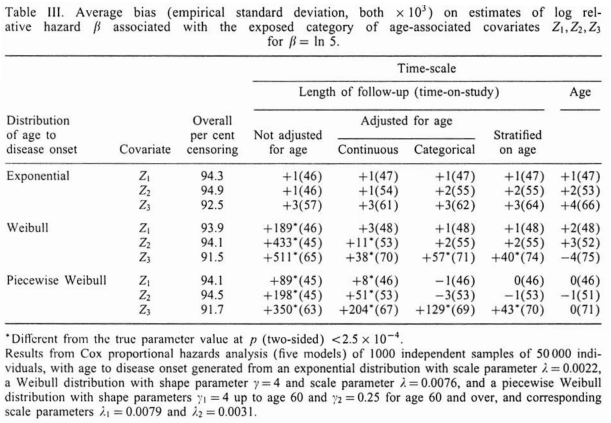


Table III. Average bias (empirical standard deviation, both $\times 10^3$) on estimates of log relative hazard β associated with the exposed category of age-associated covariates Z_1, Z_2, Z_3 for $\beta = \ln 5$.

			Time-scale				
			Length of follow-up (time-on-study)				Age
				Adjusted for age			
Distribution of age to disease onset	Covariate	Overall per cent censoring	Not adjusted for age	Continuous	Categorical	Stratified on age	
Exponential	Z_1	94.3	+1(46)	+1(47)	+1(47)	+1(47)	+1(47)
	Z_2	94.9	+1(46)	+1(54)	+2(55)	+2(55)	+2(53)
	Z_3	92.5	+3(57)	+3(61)	+3(62)	+3(64)	+4(66)
Weibull	Z_1	93.9	+189*(46)	+3(48)	+1(48)	+1(48)	+2(48)
	Z_2	94.1	+433*(45)	+11*(53)	+2(55)	+2(55)	+3(52)
	Z_3	91.5	+511*(65)	+38*(70)	+57*(71)	+40*(74)	−4(75)
Piecewise Weibull	Z_1	94.1	+89*(45)	+8*(46)	−1(46)	0(46)	0(46)
	Z_2	94.5	+198*(45)	+51*(53)	−3(53)	−1(53)	−1(51)
	Z_3	91.7	+350*(63)	+204*(67)	+129*(69)	+43*(70)	0(71)

*Different from the true parameter value at p (two-sided) $<2.5 \times 10^{-4}$.

Results from Cox proportional hazards analysis (five models) of 1000 independent samples of 50 000 individuals, with age to disease onset generated from an exponential distribution with scale parameter $\lambda = 0.0022$, a Weibull distribution with shape parameter $\gamma = 4$ and scale parameter $\lambda = 0.0076$, and a piecewise Weibull distribution with shape parameters $\gamma_1 = 4$ up to age 60 and $\gamma_2 = 0.25$ for age 60 and over, and corresponding scale parameters $\lambda_1 = 0.0079$ and $\lambda_2 = 0.0031$.

Mortality in relation to snus use; a cohort of Swedish men

- Randomly selected men (n=9976) aged 14–99 (at entry) living in Uppsala county 1973
- Participants were:
  - Invited to oral examination (at dentist)
  - Questionnaire on snus use (plus tobacco use & life-style factors)
  - Followed-up for cancer and death 1973-2002 via population registers
- 1427 (14%) were snus users; 8408 (84%) non-users at baseline
- If cumulative dose is the underlying exposure of interest and we model exposure (hours/day) at baseline as a fixed covariate then age-at-entry may approximate cumulative dose at entry and time-since entry may approximate cumulative dose during follow-up.

Results from fitting various models (outcome is death)

	time-on-study as time-scale			attained age as time-scale		
		adj. for age entry	adj. for att age		adj. for time-on-st	stratified age entry
	1	2 (4)	3	4 (5)	5	6
snus hrs/day						
0	1.0 (ref)	1.0 (ref)	1.0 (ref)	1.0 (ref)	1.0 (ref)	1.0 (ref)
1–6	1.1 (1.0-1.2)	1.0 (0.9-1.1)	1.1 (1.0-1.3)	1.0 (0.9-1.1)	1.1 (1.0-1.2)	1.0 (0.9-1.1)
7–15	1.5 (1.3-1.8)	1.0 (0.9-1.2)	1.2 (1.0-1.4)	1.0 (0.9-1.2)	1.1 (0.9-1.3)	1.0 (0.9-1.2)
16–24	2.5 (1.6-3.7)	1.3 (0.9-2.0)	1.8 (1.2-2.7)	1.3 (0.9-2.0)	1.6 (1.0-2.3)	1.3 (0.8-1.9)
time	1.5 sec	1.3 sec	14 sec	5 min	55 min	15 sec

- Which model is appropriate?
  - Model 1 is not appropriate but suggests the heavy users are older.
  - Age-specific mortality is exponential so we expect models 2 & 4 to be similar.
  - Model 6 extends 2 & 4 by controlling for cohort & period effects (suggests there are no such effects)
  - Models 3 & 5 both adjust for two time-scales. Why is time-on study a confounder given adjustment for attained age? Cumulative dose effect?

Categories used

- Age-at-entry: 0-19, 20-24, 25-29, 30-34, . . . , 90-94, 95+ yrs
- Attained age: 0-19, 20-24, 25-29, 30-34, . . . , 90-94, 95+ yrs
- Attained follow-up: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29 yrs
- Results are preliminary and not adjusted for potential confounders (smoking, alcohol, etc.). Such adjustment will be performed after data have been cleaned.

SAS processing times for model 4

obs	time (seconds)
100	0.3
1000	1.3
5000	17
10000	300
20000	?

References

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

[2] Thiébaut ACM, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med* 2004; **23**:3803–3820.