Competing Risks

Terry Therneau

19 Sept 2023

1 Introduction

Figure 1 shows 4 multistate models. This chapter deals with the one in upper right, competing risks. Everyone starts in the same state, that there are multiple possible transitions from that state, but each subject only experiences one of them.

In R the coding for competing risks is quite simple. The final state is a factor variable whose first level corresponds to censored observations, and remaining levels are the possible endpoints. An id variable is required, even if a subject has only one line of data. Otherwise this looks exactly like the data set for an ordinary Cox model.

This is my own big picture view of this portion.

- Setup
  - Creating the multi-state data set, used for the Aalen-Johansen estimate and for the mulitstate hazards model, is essentially the same as other multistate cases. (Though it is one of the simpler cases.)
  - Check the data
  - Simple statistics include counts, rates, and AJ estimate. Start there.
- Key measures of importance are hazard ratios, $p(t)$ and mean time in state for each outcome. Do at least 2.
- An additive model for any one $\neq$ additive for another. A large coefficient in one $\neq$ a large coefficient in another.
  - Users want an additive model of course – which ones work? How do you fit them?
  - Population marginal means (PMM) are a strong alternative (PMM has many names).
- Don't forget to check assumptions.

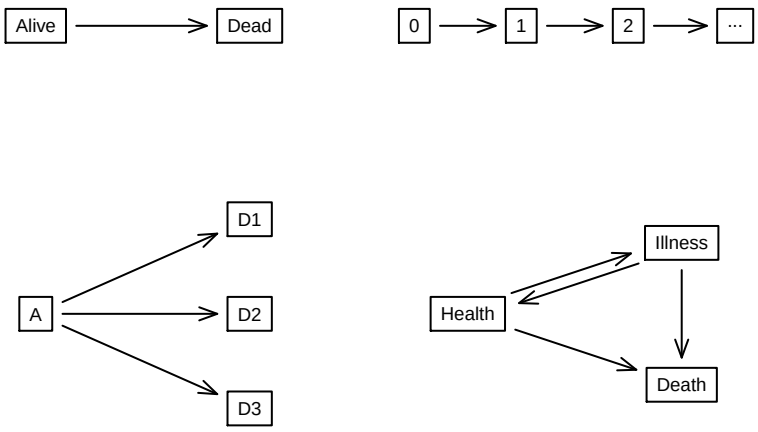

Figure 1: Four multistate models. The upper left panel depicts simple survival, the upper right panel depicts sequential events, the lower left panel illustrates competing risks, and the lower right panel shows a multistate illness-death model.

2 Aalen-Johansen estimator

Let λ_{jk} be the empirical hazard estimator, the number of observed transitions from state j to k at time t divided by the number who were at risk of such a transition.

$$\lambda_{jk}(t) = \frac{dN_{jk}(t)}{Y_j(t)}$$
$$N_{jk}(t) = \sum_i N_{ijk}(t)$$
$$Y_j(t) = \sum_i Y_{ij}(t)$$

Let A be a matrix of the estimates, i.e., with $A_{jk} = \lambda_{jk}$ for all $j \neq k$. (It would be natural to use Λ for the matrix, but that symbol has been widely used for the cumulative hazard, so it is already taken.) The diagonal of A is then filled in such that each row sums to 0. Let $H = I + A$ be the similar matrix but with rows that each sum to 1. Each row of H describes what happened at time t: the diagonal contains the fraction of observations in each state who "stayed home" at time t and the off diagonal the fractions who moved to another state. For any state j with no members at time t the convention is to set that row of A to zero.

The Aalen-Johansen estimate of the probability in state is

$$\hat{p}(t) = p(0) \prod_{s \leq t} H(s)$$

The starting vector $p(0)$ is very often (1, 0, 0, ...), e.g., everyone starts in the enrolled state.

A special case of the AJ estimate is when there are only two states and one transition, e.g., alive $\longrightarrow$ dead. Let $d(t)$ and $n(t)$ be the number of deaths and the number at risk at each time. Then

$$H(t) = \begin{pmatrix} \frac{n(t)-d(t)}{n(t)} & \frac{d(t)}{n(t)} \\ 0 & 1 \end{pmatrix}$$

The second row of the matrix encodes the fact that death is an absorbing state. Simple matrix multiplication shows that

$$p_1(t) = \prod_{s \le t} \frac{n(t) - d(t)}{n(t)} \qquad p_2(t) = 1 - p_1(t)$$

which we immediately recognize as the Kapan-Meier. That is, the KM is a special case of the AG.

In the survival package, we have

```
> sfit1 <- survfit(Surv(time, event) ~ group, data=mydata, id=ptnum,
                   istate= cstate)
> sfit2 <- survfit(Surv(time1, time2, event) ~ group, data=mydata, id=ptnum,
                   istate= cstate)
```

- The second form is used if there is delayed entry. Use with a time dependent `group` is invalid.
- The KM is computed if the `event` variable is 0/1 or TRUE/FALSE, and the AJ if it is a factor. In the second case the first level of the factor is required to encode "no event at this time" for the subject, e.g. censoring. The labels of the factor levels are the user's choice.
- For the AG the `id` statement is mandatory, it names a variable which labels which rows of the data set belong to which subject.
- The `istate` variable is optional, if missing the code assumes that all subjects start in the same state, which will be labeled "(s0)" in the output. Not everyone need start in the same state.

An alternate to the AJ is the exponential product

$$\hat{p}(t) = p(0) \prod_{s \le t} e^{A(s)}$$

3 Free light chain

The `flchain` data set in the survival package is a stratified sample containing 1/2 of the subjects from a study of the relationship between serum free light chain (FLC) and mortality. The original sample contains samples on approximately 2/3 of the residents of Olmsted County aged 50 or greater. Plasma cells each create a unique immunoglobulin that is constructed from a light chain

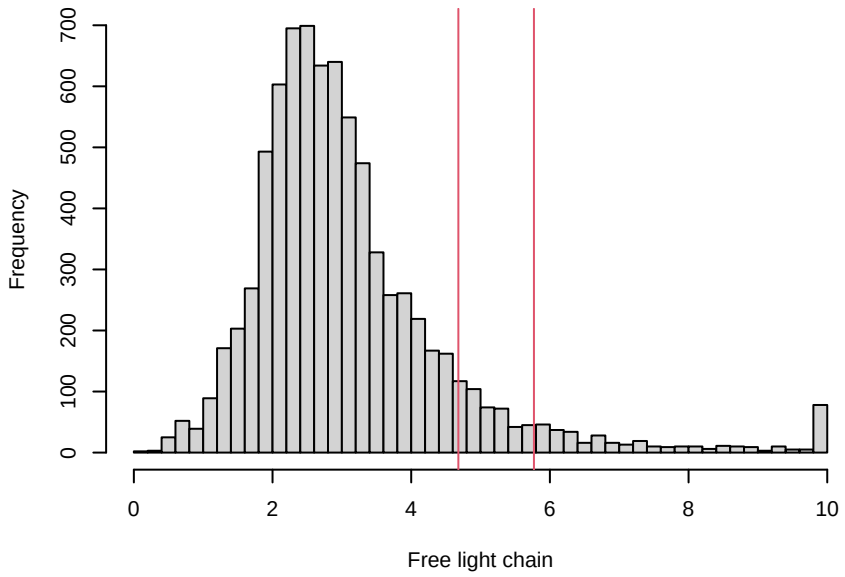

Figure 2: Free light chain quantiity for the 7874 study subjects, along with the 90th and 95th percentiles.

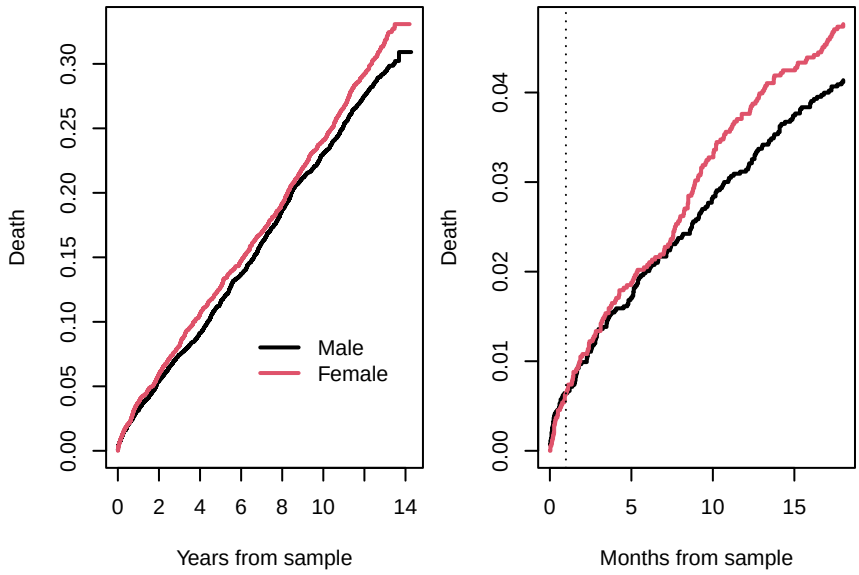

Figure 3: Overall survival for the FLC study, along with zoom in on the first 18 months. The dotted line is at 30 days.

and heavy chain portion. For unclear reasons, the cellular machinery produces an excess of light chain, which is excreted into the blood and cleared by the kidneys.

Figure 3 shows the overall survival for the FLC study, along with a blowup of the first 1.5 years. We see an increased mortality for the first month; it is about 8 months before the usual male/female divergence. The entry criteria for the study was any Olmsted County resident age 50 or older, not yet sampled, who had blood work done that results in excess sera. Many of these people are getting simple checkups, but a subset will of course be very ill. One can argue that the group of interest, with respect to a useful prediction of future mortality, are those with > 30 days of follow-up.

The next lines show the distribution of cause of death, using the chapter heading of the ICD codes. For this example, we will collapse all but the top 2 into a single category of "other".

```
                     dtype
chapter               censor cardiac cancer other
  Blood                    0       0      0     4
  Circulatory              0     722      0     0
  Congenital               0       0      0     3
  Digestive                0       0      0    62
  Endocrine                0       0      0    48
  External Causes          0       0      0    66
  Genitourinary            0       0      0    42
  Ill Defined              0       0      0    38
  Infectious               0       0      0    30
  Injury and Poisoning     0       0      0    14
  Mental                   0       0      0   144
  Musculoskeletal          0       0      0    14
  Neoplasms                0       0    556     0
  Nervous                  0       0      0   129
  Respiratory              0       0      0   242
  Skin                     0       0      0     4
  <NA>                  5685       0      0     0

Call:
survcheck(formula = Surv(futime, dtype) ~ 1, data = flc2, id = id)

Unique identifiers       Observations        Transitions
              7803               7803               2118

Transitions table:
         to
from      cardiac cancer other (censored)
  (s0)        722    556   840       5685
  cardiac       0      0     0          0
  cancer        0      0     0          0
  other         0      0     0          0
```

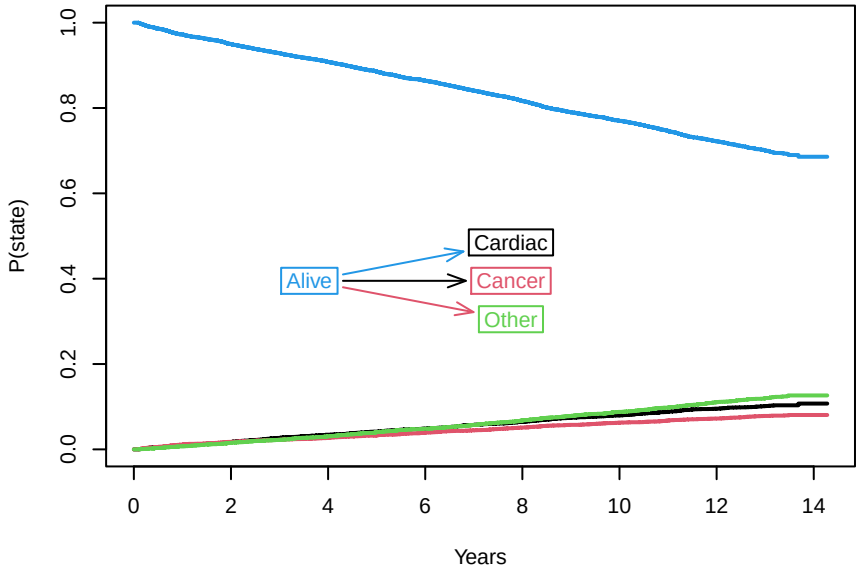

Figure 4: Aalen-Johansen estimates of probability in state, for the 4 states.

```
Number of subjects with 0, 1, ... transitions to each state:
         count
state        0    1
  cardiac 7081  722
  cancer  7247  556
  other   6963  840
  (any)   5685 2118
```

Figures 4 and 5 show the results of an Aalen-Johansen estimate $p(t)$. The result, at each time point, is a vector of the estimated probability of being in each state. By definition the sum over states has to be 1, $\sum p_k(t) = 1$. It is thus not necessary to show all 4 curves and it is common to omit the least interesting of them as in figure **??**.

One useful estimate is the sojourn time or restricted mean time in state (RMST). For a two state alive:dead model (simple survival) a more common label for time in the alive state is restricted mean survival time (RMST). For competing risks, the sojourn time in the non-initial state is sometimes called the years of life lost (YLL), in this case .34 of the first 10 years after the FLC assay is lost to cancer.

```
> print(aj1, rmean=10, digits=2)
Call: survfit(formula = Surv(years, dtype) ~ 1, data = flc2, id = id,
    influence = TRUE)

          n nevent rmean se(rmean)*
(s0)   7803      0  8.84      0.029
```

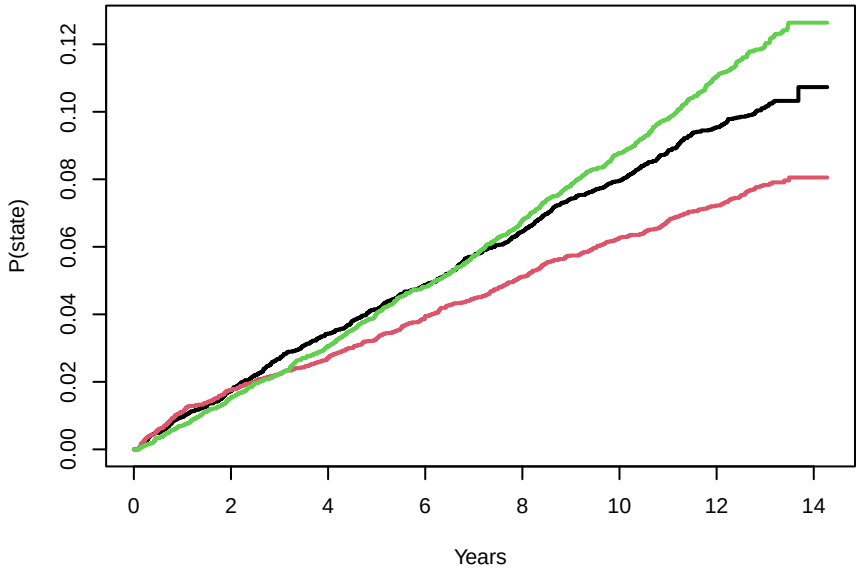

Figure 5: Aalen-Johansen estimates, omitting the alive state.

```
cardiac 7803    722  0.41      0.019
cancer  7803    556  0.34      0.017
other   7803    840  0.41      0.018
   *restricted mean time in state (max time = 10 )
```

When there are multiple covariates the set of curves can get a bit large, for example create curves by sex, flc10, and quartiles of age.

```
> agegrp <- cut(flc2$age, c(0, 55, 63, 72, 100))
> flcmany <- survfit(Surv(years, dtype) ~ agegrp + sex + flc10, flc2, id=id)
> dim(flcmany)
strata states
    16      4
> print(flcmany[,2], digits=2, rmean=10)
Call: survfit(formula = Surv(years, dtype) ~ agegrp + sex + flc10,
    data = flc2, id = id)

                                          n nevent rmean*
agegrp=(0,55], sex=F, flc10=0, cardiac    997      9  0.029
agegrp=(0,55], sex=F, flc10=1, cardiac     40      2  0.445
agegrp=(0,55], sex=M, flc10=0, cardiac    913     23  0.089
agegrp=(0,55], sex=M, flc10=1, cardiac     32      4  0.412
agegrp=(55,63], sex=F, flc10=0, cardiac  1060     27  0.086
agegrp=(55,63], sex=F, flc10=1, cardiac    58      2  0.181
```

```
agegrp=(55,63], sex=M, flc10=0, cardiac   988     39  0.150
agegrp=(55,63], sex=M, flc10=1, cardiac    55      4  0.313
agegrp=(63,72], sex=F, flc10=0, cardiac   948     41  0.125
agegrp=(63,72], sex=F, flc10=1, cardiac    75     14  0.686
agegrp=(63,72], sex=M, flc10=0, cardiac   769     71  0.370
agegrp=(63,72], sex=M, flc10=1, cardiac   101     22  1.197
agegrp=(72,100], sex=F, flc10=0, cardiac  933    222  1.047
agegrp=(72,100], sex=F, flc10=1, cardiac  200     70  2.232
agegrp=(72,100], sex=M, flc10=0, cardiac  452    103  0.879
agegrp=(72,100], sex=M, flc10=1, cardiac  182     69  2.284
   *restricted mean time in state (max time = 10 )
```

The multi-state hazard model is simple to fit using `coxph`. The result is a set of coefficients for each transition.

```
> cfit <- coxph(Surv(years, dtype) ~ age + sex + log(creat) + flc10,
              flc2, id=id)
> print(cfit, digits=2)
Call:
coxph(formula = Surv(years, dtype) ~ age + sex + log(creat) +
    flc10, data = flc2, id = id)


1:2            coef exp(coef) se(coef) robust se    z      p
  age        0.1203    1.1279   0.0045    0.0047 25.5 <2e-16
  sexM       0.1843    1.2024   0.0866    0.0885  2.1   0.04
  log(creat) 0.9687    2.6346   0.1436    0.1687  5.7  9e-09
  flc10      0.6595    1.9338   0.0996    0.1021  6.5  1e-10


1:3            coef exp(coef) se(coef) robust se    z      p
  age        0.0534    1.0549   0.0046    0.0047 11.3 <2e-16
  sexM       0.3128    1.3673   0.1011    0.1036  3.0  0.003
  log(creat) 0.2204    1.2465   0.2132    0.2286  1.0  0.335
  flc10      0.6003    1.8226   0.1316    0.1328  4.5  6e-06


1:4            coef exp(coef) se(coef) robust se    z      p
  age        0.1223    1.1301   0.0042    0.0044 28.1 <2e-16
  sexM       0.2479    1.2814   0.0828    0.0879  2.8  0.005
  log(creat) 0.1366    1.1464   0.1740    0.2295  0.6  0.552
  flc10      0.6855    1.9847   0.0961    0.1011  6.8  1e-11

 States:  1= (s0), 2= cardiac, 3= cancer, 4= other
```

```
Likelihood ratio test=2573  on 12 df, p=<2e-16
n= 6456, number of events= 1912
   (1347 observations deleted due to missingness)
```

Notice that the coefficients we get for cardiac are exactly what come from a simple Cox model using cardiac death as an end point with all other outcomes treated as censored. Hazards can be computed one at a time. However, as shown in figure 6, absolute risk (survival) can not be computed in this way, using the Kaplan-Meier. Doing so is a common and fundamental mistake.

```
> ctest <- coxph(Surv(years, dtype=="cardiac") ~ age + sex + log(creat) + flc10,
                 flc2)
> print(ctest, digits=2)
Call:
coxph(formula = Surv(years, dtype == "cardiac") ~ age + sex +
    log(creat) + flc10, data = flc2)

             coef exp(coef) se(coef)    z      p
age        0.1203    1.1279   0.0045 26.8 <2e-16
sexM       0.1843    1.2024   0.0866  2.1   0.03
log(creat) 0.9687    2.6346   0.1436  6.7  2e-11
flc10      0.6595    1.9338   0.0996  6.6  4e-11

Likelihood ratio test=1166  on 4 df, p=<2e-16
n= 6456, number of events= 654
   (1347 observations deleted due to missingness)
```

Here is a test for proportional hazards. We see a smattering of small p, of which age:other is the strongest. Figure 7 shows the associated plots for the two smallest.

```
                  chisq df       p
age_1:2         0.75245  1 0.38570
sex_1:2         0.00821  1 0.92778
log(creat)_1:2  3.37786  1 0.06608
flc10_1:2       1.55692  1 0.21212
age_1:3         4.09692  1 0.04296
sex_1:3         1.51234  1 0.21878
log(creat)_1:3  4.46535  1 0.03459
flc10_1:3       1.71869  1 0.18986
age_1:4        11.57518  1 0.00067
sex_1:4         2.02730  1 0.15449
log(creat)_1:4  4.31147  1 0.03786
flc10_1:4       3.03470  1 0.08150
GLOBAL         41.31455 12 4.3e-05
```

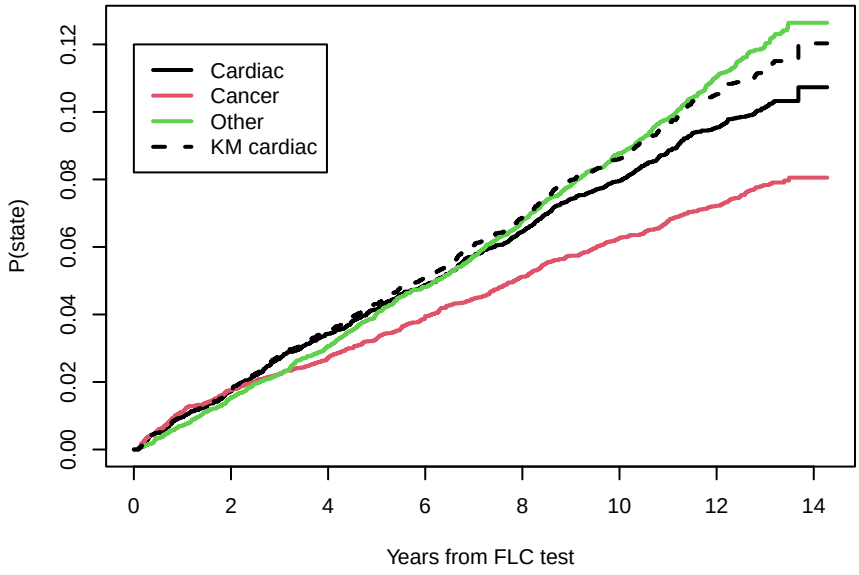

Figure 6: Aalen-Johansen estimates for the FLC data (solid) along with a KM for the cardiac endpoint (dashed).

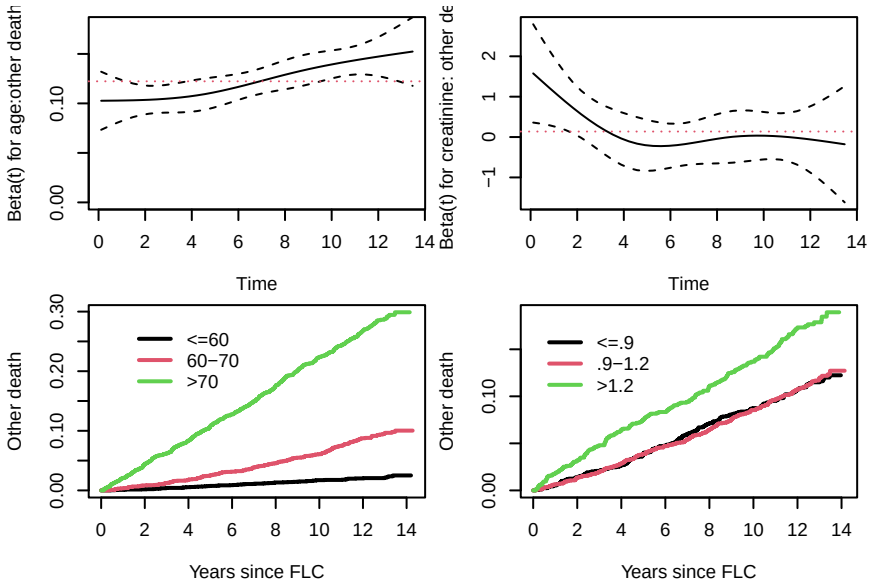

Figure 7: Estimates of time-dependent $\beta(t)$ for two covariates, other death endpoint. Dotted line is the Cox model coefficient.

3.1 Predicted survival

Predictions from a single state or multi-state PH model are available for any given values of the X variables. The challenge, of course, is which X values to use as a good summary. Here is an example

```
> dummy <- expand.grid(age= c(55, 63, 72), sex=c("F", "M"), creat=1, flc10=0:1)
> psurv <- survfit(cfit, newdata=dummy)
> dim(psurv)
  data states
    12      4
> print(psurv[,2], rmean=10)
Call: survfit(formula = cfit, newdata = dummy)

               n nevent     rmean*
1, cardiac  6456    654 0.05174764
2, cardiac  6456    654 0.13289781
3, cardiac  6456    654 0.37276681
4, cardiac  6456    654 0.06182471
5, cardiac  6456    654 0.15792080
6, cardiac  6456    654 0.43709231
7, cardiac  6456    654 0.09832515
8, cardiac  6456    654 0.24821007
9, cardiac  6456    654 0.66573158
10, cardiac 6456    654 0.11683571
11, cardiac 6456    654 0.29200945
12, cardiac 6456    654 0.76484738
   *restricted mean time in state (max time = 10 )


  age sex RMTS:low FLC RMTS:high FLC     delta
1  55   F   0.05174764    0.09832515 0.0465775
2  63   F   0.13289781    0.24821007 0.1153123
3  72   F   0.37276681    0.66573158 0.2929648
4  55   M   0.06182471    0.11683571 0.0550110
5  63   M   0.15792080    0.29200945 0.1340887
6  72   M   0.43709231    0.76484738 0.3277551
```

This has created 48 curves, 12 for each state. To look at the effect of high FLC on cardiac death, we might show this as 6 pairs, using line types and colors.

Can we get a single number summary for the YLL due to FLC?

3.2 The Aalen-Johansen estimate, again

At any time point t create the transition matrix H. The first row of H is the dispostion of all those in state 1 at time $t - 0$, the second row all those in state 2, etc. As an example, for the flc2 data set above at 1874 days we have (6710, 330, 263, 322) subeject currently in the entry,

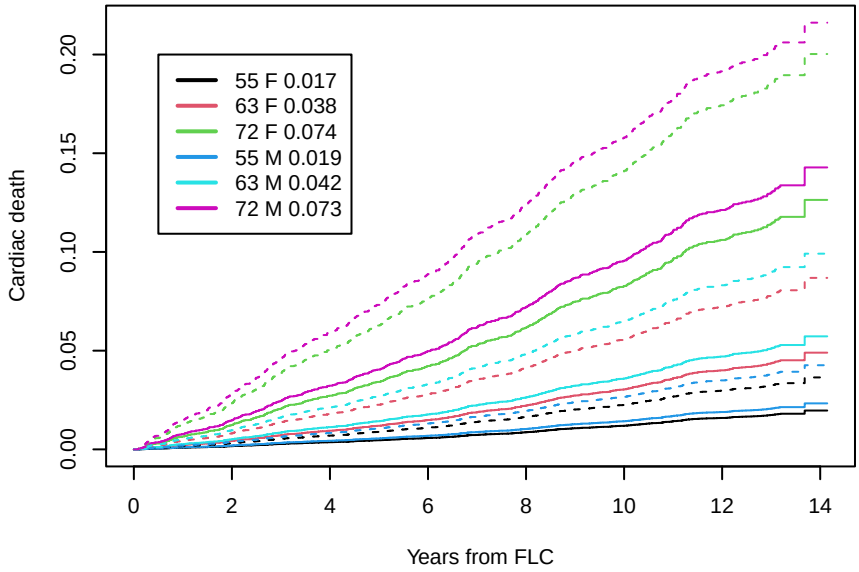

Figure 8: Curves for FLC in the lower 90% (solid) versus the upper 10% (dashed), for 6 combinations of age and sex (colors).

cardiac death, cancer death, and other death states (182 have been censored before day 1874), There were 2 cardiac deaths and 2 other deaths on that day, leading to

$$H(1874) = \begin{pmatrix} 6706/6710 & 2/6710 & 0 & 2/6710 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The AJ estimate at time t is

$$\hat{p}(t) = p(0) \prod_{s/let} H(s)$$

For simplicity label the states as 0= entry, 1= cardiac death, 2= cancer death and 3 = other death. Then $p(0) = (1, 0, 0, 0)$, and it is fairly easy to show that

$$\begin{aligned} p_0(t) &= S(t) \\ p_1(t) &= \int_0^t \lambda_{01}(z) S(z-) dz \\ p_2(t) &= \int_0^t \lambda_{02}(z) S(z-) dz \\ p_3(t) &= \int_0^t \lambda_{03}(z) S(z-) dz \end{aligned}$$

where $S(t)$ is the usual Kapan-Meir for death of any type and λ_{jk} are element of the H matrix above. Many will recognize this as the "cumulative incidence" (CI) formula. That is, the CI is a special case of the Aalen-Johansen.

When you use the KM, incorrectly, one essentially replaces the first row by (6708/6710, 2/6710, 0, 0).

Let $H = I + A$, i.e., the diagonal of A is set so that each row sums to zero. An alternate estimate is

$$p(t) = p(0) \prod_{s \le t} e^{A(s)}$$

where exp is the matrix exponential, defined as

$$\exp(A) = I + A + A^2/2! + A^3/3! + \dots$$

Many textbooks will contain a formula where the Kaplan-Meier is replaced by the exponential of the cumulative hazard, $S^\dagger(t) = \exp(\sum_{s/let} A_{00}(t)$, usually written in terms of λ. This leads to an estimate which does not satisfy $\sum_k p_k(t) = 1$.

4 Death and dementia in the MCSA

(A copy of this data set is not provided).

Mayo Clinic Study of Aging

- 6258 subjects
- 726 dementia, 1990 deaths, 1/2 the dementias occur after active participation
- Taken from the MCSA, an age/sex stratified random sample from Olmsted County, Minnesota
- Covariates
  - APOE e4 allele: risk factor for amyloidosis
  - CMC score: 0-7, count of morbidities
  - Education

Figure 9 shows the death and dementia rates as a function of age. The dementia rate is not loglinear!

5 The special role of age

When working with human subjects data, as the authors do, age and sex appear as covariates in nearly every model. Figure 10 shows overall US death rates as a function of age and sex. It displays the remarkable fact that, over the interval from age 40 to 100, rates closely align to the very simple model log(rate) = a + b*age + c* male. The same is true for other populations than the US total. The great majority of our clinical research falls into this age range; most studies will span only a portion of it, in which case the model fits even better.

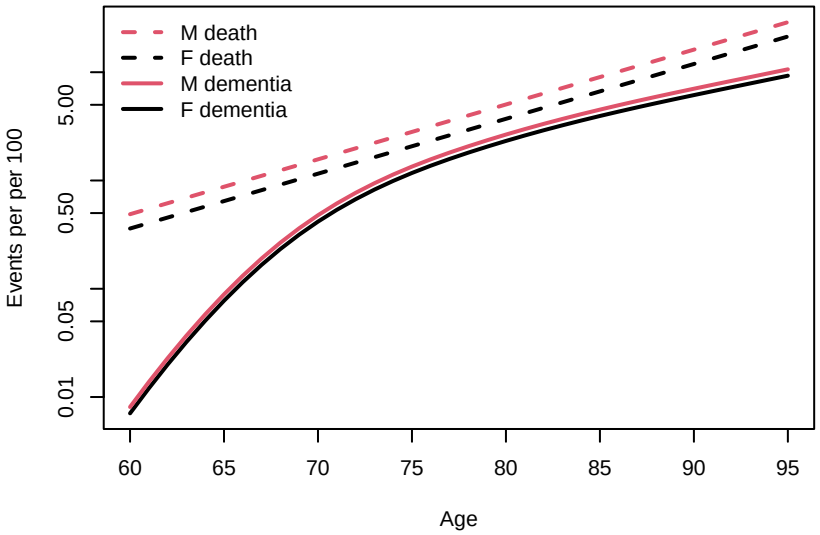

Figure 9: Overall rates of death and dementia, by age, in the MCSA.

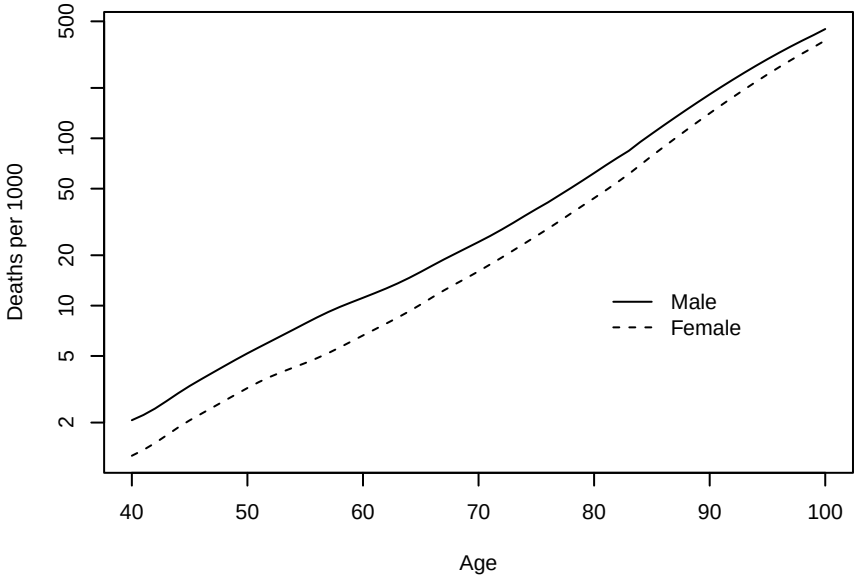

Figure 10: Death rates per 1000, United States 2010.

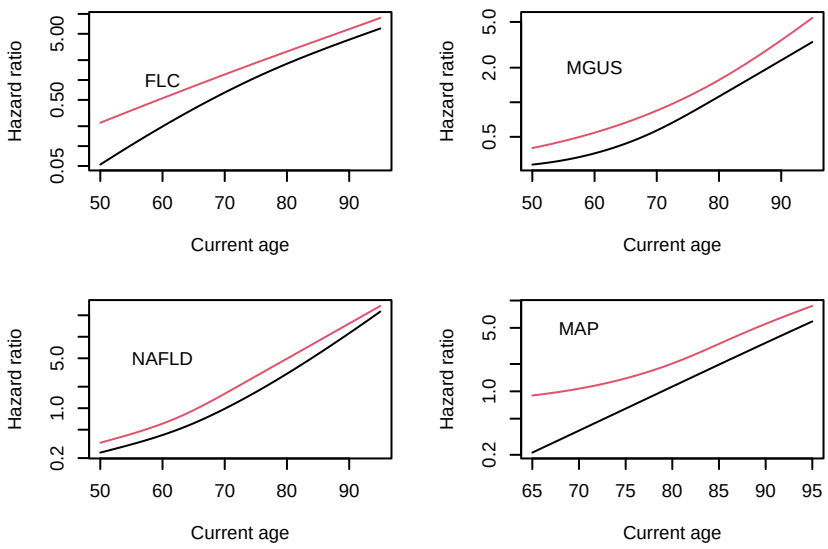

Figure 11: Predicted hazard ratios from Cox model fits with current age and sex as covariates; free light chain (FLC), monoclonal gammopathy of undetermined significance (MGUS), non-alcoholic fatty liver disease (NAFLD) and the minority and aging project (MAP).

What is perhaps more surprising is how often this same relationship holds, at least approximately, in research studies. Figure 11 shows simple age and sex fits for four different studies. In each Cox model fit we used current age rather than enrollment age as the (time dependent) covariate; death is the endpoint. The FLC study solicited Olmsted County subjects over the age of 50, with a blood draw as part of their current visit, for permission to run an extra test on the obtained sample (at no cost). Participation was very high, and so we might expect to see a reprise of the overall US pattern. The MGUS study followed subjects for whom a particular (rather rare) test had been ordered, so represent a select subset. The NAFLD study contains post-diagnosis follow-up for all subjects diagnosed with fatty liver disease along with age and sex matched controls. The Rush University Memory and Aging project enrolled community subjects for a study of long term cognitive changes.

The models all used time-dependent age (current age) $a(t)$ as a covariate. If subject i has an event at time t and s is the 0/1 sex variable, and $a(t)$ is linear, then the Cox hazard ratio satisfies

$$\begin{aligned}
\frac{e^{\eta_i}}{\sum_j Y_j(t) e^{\eta_j}} &= \frac{e^{\beta_1 a_i(t) + \beta_2 s_i}}{\sum_j Y_j(t) e^{\beta_i a_j(t) + \beta_2 s_j}} \\
&= \frac{e^{\beta_1 (t + a_i(0)) + \beta_2 s_i}}{\sum_j Y_j(t) e^{\beta_i (t + a_j(0)) + \beta_2 s_j}} \\
&= \frac{e^{\beta_1 a_i(0) + \beta_2 s_i}}{\sum_j Y_j(t) e^{\beta_i a_j(0) + \beta_2 s_j}}
\end{aligned}$$

where $a_i(0)$ is the age at enrollment for subject i. That is, if the age effect is linear, then Cox models using the age at enrollment (time fixed covariate) and using current age (time dependent) are identical. The former model is much easier to fit, since a time-dependent age covariate does not need to be created. As well, the code to create predicted survival curves from the fitted model is substantially easier.

For this reason, nearly every published analysis uses baseline age as a time fixed covariate; but checks for whether the necessary linearity assumption is justified are extremely rare. The fact that linearity of age will often be close to true is, in our opinion, one of the reasons for the remarkable success of the Cox model. Quoting GEP Box: "Assumptions, whether implied or clearly stated, are never exactly true. So the question you need to ask is not 'Is the model true?' (it never is) but 'Is the model good enough for this particular application?' " Linear age + additive sex will quite often be "good enough".

However, a caution is in order, particularly for studies that span a large age range, either through a broad enrollment window or long follow-up. Notable in all the cases above is how very *large* the age effect is. Most studies target effects of 1.3 fold or greater, and are excited to find them; the age effects above can span a 100 fold change in risk. Indeed, the sex effects in each of the 4 cases are large, 1.5 – 2.2 fold, but look visually small in comparison. A benign time dependent covariate that is associated with age may appear highly significant, simply by acting as a surrogate for any non-linearity of age, "subject has turned gray" for instance.

As a simple and perhaps silly example add the variable "has retired", coded as the time depedent covariate of 1 if age > 65 and 0 otherwise, and another variable RMD with a cutoff of age 72, an indicator of whether the subject will have begun to take RMD withdrawals from their pension plan.

```
> test <- coxph(Surv(futime, status) ~ age + male + I(age > 65), nafld1,
               subset= (futime > 7))
> # and as a time dependent
> temp <- subset(nafld1, ,c(id, age, male))
> ndata <- tmerge(temp, nafld1, id=id, death= event(futime/365.25, status))
> temp <- data.frame(id= ndata$id, rmd65 = 65- ndata$age, rmd72 = 72-ndata$age)
> ndata <- tmerge(ndata, temp, id=id, retire= tdc(rmd65), rmd= tdc(rmd72))
> coxph(Surv(tstart, tstop, death) ~ age + male + retire, ndata)
Call:
coxph(formula = Surv(tstart, tstop, death) ~ age + male + retire,
    data = ndata)

            coef exp(coef)  se(coef)      z        p
age     0.107397  1.113376  0.003465 30.995  < 2e-16
male    0.373637  1.453009  0.054320  6.878 6.05e-12
retire -0.325160  0.722412  0.102315 -3.178  0.00148

Likelihood ratio test=2309  on 3 df, p=< 2.2e-16
n= 21623, number of events= 1364
> coxph(Surv(tstart, tstop, death) ~ age + male + rmd, ndata)
Call:
coxph(formula = Surv(tstart, tstop, death) ~ age + male + rmd,
```

```
    data = ndata)

          coef exp(coef)  se(coef)      z        p
age   0.100913  1.106180  0.003825 26.381  < 2e-16
male  0.373525  1.452848  0.054322  6.876 6.15e-12
rmd  -0.060566  0.941231  0.096097 -0.630    0.529

Likelihood ratio test=2300  on 3 df, p=< 2.2e-16
n= 21623, number of events= 1364
> # time-dependent is far from significant
```

We see that addition of age65 has a very small p, but 72 not at all. The upshot is that a variable like "Medicare insurance" could appear to be very significant when it is only a bystander to the non-linearity. (A closer look at the lower left panel in figure 11 hints that the age 65 indicator may simply be closer to the bend.)

Using age scale One way to deal with a possible non-liner age effect is to use age as the underlying time scale for the proportional hazards model. There are two advantages: it perfectly adjusts for age via matching, and the interpretation of the resulting coefficients is in many cases more natural. Coefficients will reflect the relative hazard of any subject as compared to others of the same age, rather than comparing to others at the same time from enrollment. For a study such as NAFLD, the time since enrollment has little biological meaning for a matched control: it is the age they were randomly selected. A possible downside to using age as the time scale is that there will no longer be an age coefficient in the model, i.e., no p-value. However, for long term processes such as these everyone already knows that higher age is a risk factor for death, there is no utility to "proving" it once again.

One aspect that is often overlooked is symmetry, the model using time-since-entry as the underlying scale used age as a covariate; conversely, models using age a covariate should consider time since enrollment as a covariate. As a example look at the MGUS and MAP data sets. For MGUS we have divided the time on study into 3 epochs of 0–6, 6–12 and $>$ 12 months of followup, using the last year as reference, while MAP is divided as 1, 5 and 10+ years. Results are shown in Table 1. In the MGUS data, death rates in the first 6 months after enrollment are 2.6 fold higher than in later follow-up, while in the MAP study death rates in the first year are $< 1/2$ of those at later times. For MGUS the underlying test, serum protein electrophoresis, is normally ordered in the diagnostic work-up for patients with serious illness. The results we see are due to a selection effect: many of these patients are facing imminent mortality. In fact, in a follow-up study (not shown) we found a similar mortality excess for all patients who had the test ordered, not just those with a positive result.

The MAP study displays a more common result, one that we see with most studies that involve voluntary patient recruitment. Essentially, patients who are in extremis are much less likely to enroll in a long term monitoring study. Comparing two 80 year old subjects, for instance, one of whom was enrolled 5 years ago and the other enrolled last week, the second of these is unlikey to die in the next few months — if they had such serious illness, and were aware of it, the subject would not have enrolled. The opposite can happen in a chronic disease that waxes and wanes; subjects currently experiencing difficultly may have a greater motivation to enter a

```
      male 0-1y 1-2y 2-5y 5-10y
MGUS  1.45 2.58 1.26
MAP   1.64 0.42 0.55 0.75
FLC   1.52 1.96 1.36 1.26  1.18
NAFLD 1.44 0.96 0.97 0.97  1.01
```

Table 1: Hazard ratios for models on age scale, with age and time from enrollment as covariates. (To do: format as a real table.) The time intervals for MGUS are 0-6m, and 6-12m, with > 12m as the reference, for MAP that are 0-1y, 1-5y and 5-10y, with > 10 as reference; column labels are correct for FLC and the NAFLD control subjects.

	Male sex	Hgb
Time since entry scale, adjusted for age	0.49 (0.07)	-0.15 (0.02)
Age scale, adjusted for time since entry	0.48 (0.07)	-0.13 (0.02)
Time since entry, unadjusted for age	0.43 (0.07)	-0.20 (0.02)
Age scale, unadjusted for time	0.50 (0.07)	-0.15 (0.02)

Table 2: Three different models for the MGUS data.

research study.

The last row of the table applies the same criteria to all subjects who were chosen as controls for the NAFLD study. This did not involve a for cause visit, nor explicit per-study consent, and we here see no effect of the time since random selection.

In spite of my exhortation that you should always look, the time from enrollment bias can often have little if any effect on coefficients of interest. Both treated and control will have the same bias applied. See the mfit3a and mfit3b fits above, for instance.

Absolute risk curves Another aspect of using age scale is that the survival curves, derived from a Cox model, will now be on age scale. Predicted curves are, as always for a Cox model, for one or more hypothetical subjects with specified covariate values. In addition, we need to specify a starting age, e.g., the predicted curve for a male, alive at age 60, for future ages.

How do we deal with the time-dependent covariate, time since enrollment? Our usual choice is to fix this at the reference level of 10+ years (or whatever was chosen); that is, to show the curve for a hypothetical subject who was chosen at random from the population, without selection effects due to disease or consent, since that matches application of the results to the population at large.

Table 2 shows coefficients for the MGUS data using the three possible models. A spline fit (not shown) suggests that the hemoglobin effect is approximately linear. The coefficients for sex and hgb are nearly identical for three of the four fits, something we have seen replicated in other cases but not all. Formally, it is most correct to adjust the age scale fit for entry, since the effect is highly significant, but because that variable is well balanced with respect to other covariate the impact on coefficients is small. (Most of the effect is in the first epoch, when all are still present). Age is a much larger effect, and not as well balanced.

Figure 12 shows predicted survival curves, on the left for enrollment scale and on the right using age scale. The right panel shows curves with and without adjustment for the selection

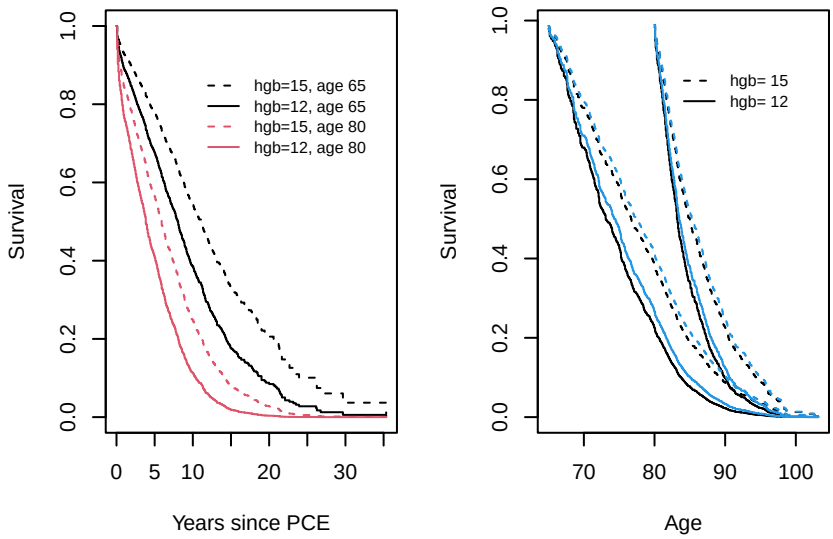

Figure 12: Predicted survival curves for a male with hgb of 12 and 15, starting at age 65 or 80. The left panel is from a Cox model using time since enrollment scale, the right panel from a model using age scale. In the right panel the second color indicates curves that have been adjusted for the initial selection effect.

effect, where without = a model without time from entry as a covariate, and with = predicted curves from a model that does include time from entry, predictions for a subject that did not experience that selection effect.