Modelling multiple timescales using flexible parametric survival models

Hannah Bower* Therese M-L. Andersson, Michael J. Crowther and Paul C. Lambert

*Department of Medical Epidemiology and Biostatistics

Karolinska Institutet, Sweden

Nordic and Baltic Stata Users Group meeting

1st September 2017

Motivation

- Defining the timescale(s) of interest is essential in any time-to-event analysis
- Different timescales could be important for different outcomes
    - For example, time since diagnosis when considering survival after a diagnosis of breast cancer
    - Or, attained age for the incidence of breast cancer
- There are occasions when several timescales are simultaneously of interest
    - Incidence of breast cancer: attained age & time since childbirth

Motivation

Suppose we have two timescales of interest. How are these commonly accounted for?

Motivation

Suppose we have two timescales of interest. How are these commonly accounted for?

One option:

- Select the most important timescale as the primary timescale
- Split the data on the second timescale and include several indicator variables in the model for this second timescale

Motivation

Suppose we have two timescales of interest. How are these commonly accounted for?

One option:

- Select the most important timescale as the primary timescale
- Split the data on the second timescale and include several indicator variables in the model for this second timescale
  - Splitting data and fitting models to split data can be computationally intensive
  - The effect of the second timescale is not continuous

Motivation

Suppose we have two timescales of interest. How are these commonly accounted for?

Another option:

- Select the most important timescale as the primary timescale
- Ignore the second timescale, or use some fixed time effect of the second timescale (e.g., age at diagnosis for attained age)

Motivation

Suppose we have two timescales of interest. How are these commonly accounted for?

Another option:

- Select the most important timescale as the primary timescale
- Ignore the second timescale, or use some fixed time effect of the second timescale (e.g., age at diagnosis for attained age)
    - Won't accurately account for the effect of the second timescale

Motivation

If we wanted to capture the effect of multiple timescales, how would we do it more accurately?

Motivation

If we wanted to capture the effect of multiple timescales, how would we do it more accurately?

- Time increases in the same way independent of the scale
- Thus, one timescale is a function of the other
  - Where is the origin of the timescale?

Motivation

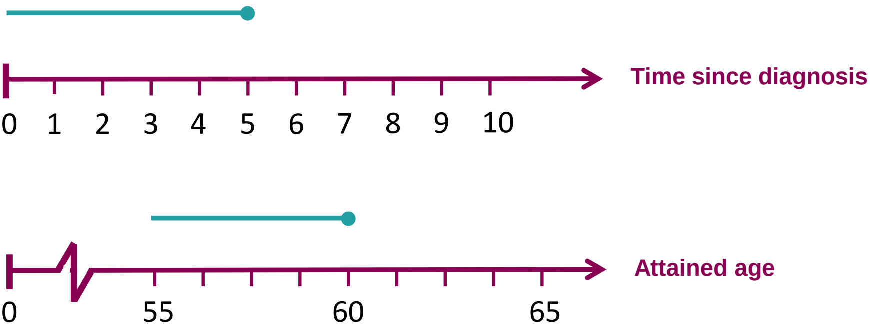
If we wanted to capture the effect of multiple timescales, how would we do it more accurately?

- Time increases in the same way independent of the scale
- Thus, one timescale is a function of the other
    - Where is the origin of the timescale?
- For example, consider time since diagnosis of a disease t_{diag} and attained age t_{age}

$$t_{age} = \text{age}_{diag} + t_{diag}$$

Motivation

- If t_{diag} = 5 & age_{diag}=55, t_{age} = 60



The `strcs` command

- Previously developed `strcs` to model the log hazard using flexible parametric survival models (FPSMs)
- FPSMs usually model the log cumulative hazard
- Initially `strcs` was developed to deal with problems when modelling multiple time-dependent effects
- We realised they could be used to model multiple timescales

Flexible parametric survival models

- Flexible parametric survival models (FPSMs) use restricted cubic splines (RCS) to model some form of the hazard function
- RCS are piecewise cubic polynomials joined together at points called knots
  - Continuous 1st, and 2nd derivatives at the knots, linear before first and after last knot
- RCS are able to capture complex hazard functions which standard parametric models may struggle to capture

FPSMs on the log hazard scale

- Non-proportional FPSM on the log hazard scale looks like:

$$\ln(\boldsymbol{h}(\boldsymbol{t};\boldsymbol{x})) = \underbrace{s(\ln(\boldsymbol{t});\gamma_{\boldsymbol{0}})}_{\text{spline function}} + \overbrace{\boldsymbol{x}\beta}^{\text{covariates}} + \underbrace{\sum_{k=1}^{D} s(\ln(\boldsymbol{t});\gamma_{\boldsymbol{k}})\boldsymbol{x}_{\boldsymbol{k}}}_{\text{time-dependent effects}}$$

Maximum likelihood estimation

Log-likelihood

$$\ln L_i = d_i \ln\{h(t_i)\} - H(t_i)$$

- d_i = event indicator
- $h(t_i)$ = hazard function
- $H(t_i)$ = cumulative hazard function

$$H(t_i) = \int_0^t h(u_i)du$$

Maximum likelihood estimation

Log-likelihood

$$\ln L_i = d_i \ln\{h(t_i)\} - H(t_i)$$

- **FPSMs on the log hazard scale**: numerical integration required to get cumulative hazard function

$$H(t_i) = \int_0^t h(u_i)du$$

The `stmt` command

- `stmt` is a Stata command which fits multiple timescales using FPSMs on the log hazard scale
- Is specifically designed to model multiple timescales and is an extension of `strcs`
- `stmt` uses Mata to numerically integrate the hazard function using Gaussian quadrature
- The first timescale is specified using the `stset` command
- Still being developed

`stmt` syntax

`stmt` *varlist*, [`time1`(*sub-options*) `time2`(*sub-options*)
`time3`(*sub-options*) ...]

Timescale-specific sub-options

- `df(#)` - degrees of freedom for effect of timescale
- `start`(*varname*) - starting value of second & third timescales
- `tvc`(*varlist*) - variables with time-dependent effects
- `logtoff` - create restricted cubic spline for untransformed time (default is log time scale)
- Plus other options & timescale-specific sub-options found in the `stpm2` and `strcs` commands

Example: Orchiectomy dataset

- Swedish prostate cancer patients (60 961 observations)
- Interested in risk of hip fracture after bilateral orchiectomy
- Timescales of interest:
  - Time since diagnosis of prostate cancer
  - Attained age
- Variable of interest is `orch`, indicator for orchiectomy

Example: Two timescales, proportional hazards

```
. stset dateexit, fail(frac = 1) enter(datecancer)
> origin(datecancer) scale(365.25)

. stmt orch, time1(df(3)) time2(start(agediag) df(5) logtoff)
```

Example: Two timescales, proportional hazards

```
. stset dateexit, fail(frac = 1) enter(datecancer)
> origin(datecancer) scale(365.25)

. stmt orch, time1(df(3)) time2(start(agediag) df(5) logtoff)
```

$$\ln(\boldsymbol{h}(\boldsymbol{t})) = \underbrace{s_{t1}(\ln(\boldsymbol{t});\gamma_{\boldsymbol{t1}})}_{\text{time since diagnosis}} + \overbrace{s_{t2}(\boldsymbol{t} + \text{age}_{diag};\gamma_{\boldsymbol{t2}})}^{\text{attained age}} + orch$$

Example: Two timescales, proportional hazards

```
. stmt orch, time1(df(3)) time2(start(agediag) df(5) logtoff)

Log likelihood =  -7464.385                       Number of obs     =     60,961
------------------------------------------------------------------------------
             | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        orch |   1.579357    .083613     8.63   0.000     1.423694     1.75204
-------------+----------------------------------------------------------------
rcs          |
     __t1_s1 |   .0129676    .025773     0.50   0.615    -.0375467    .0634818
     __t1_s2 |  -.0206878   .0251947    -0.82   0.412    -.0700686     .028693
     __t1_s3 |   .0235215   .0259144     0.91   0.364    -.0272698    .0743129
     __t2_s1 |   .6799227   .0332591    20.44   0.000     .6147361    .7451092
     __t2_s2 |  -.1234378   .0342275    -3.61   0.000    -.1905225   -.0563532
     __t2_s3 |   .0913521   .0296776     3.08   0.002     .0331852    .1495191
     __t2_s4 |   .0038328   .0248068     0.15   0.877    -.0447878    .0524533
     __t2_s5 |   .0180132   .0214929     0.84   0.402    -.0241121    .0601384
       _cons |   -5.17632   .0348153  -148.68   0.000    -5.244557   -5.108084
------------------------------------------------------------------------------
```

Example: Two timescales, proportional hazards

```
. stmt orch, time1(df(3)) time2(start(agediag) df(5) logtoff)

Log likelihood =  -7464.385                     Number of obs     =     60,961
------------------------------------------------------------------------------
             | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        orch |   1.579357    .083613     8.63   0.000     1.423694     1.75204
-------------+----------------------------------------------------------------
rcs          |
     __t1_s1 |   .0129676    .025773     0.50   0.615    -.0375467    .0634818
     __t1_s2 |  -.0206878   .0251947    -0.82   0.412    -.0700686     .028693
     __t1_s3 |   .0235215   .0259144     0.91   0.364    -.0272698    .0743129
     __t2_s1 |   .6799227   .0332591    20.44   0.000     .6147361    .7451092
     __t2_s2 |  -.1234378   .0342275    -3.61   0.000    -.1905225   -.0563532
     __t2_s3 |   .0913521   .0296776     3.08   0.002     .0331852    .1495191
     __t2_s4 |   .0038328   .0248068     0.15   0.877    -.0447878    .0524533
     __t2_s5 |   .0180132   .0214929     0.84   0.402    -.0241121    .0601384
       _cons |   -5.17632   .0348153  -148.68   0.000    -5.244557   -5.108084
------------------------------------------------------------------------------
```

Example: Two timescales, proportional hazards

```
. stmt orch, time1(df(3)) time2(start(agediag) df(5) logtoff)

Log likelihood =  -7464.385                        Number of obs     =     60,961
------------------------------------------------------------------------------
             | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        orch |   1.579357     .083613     8.63   0.000     1.423694     1.75204
-------------+----------------------------------------------------------------
rcs          |
     __t1_s1 |   .0129676     .025773     0.50   0.615    -.0375467     .0634818
     __t1_s2 |  -.0206878    .0251947    -0.82   0.412    -.0700686      .028693
     __t1_s3 |   .0235215    .0259144     0.91   0.364    -.0272698     .0743129
     __t2_s1 |   .6799227    .0332591    20.44   0.000     .6147361     .7451092
     __t2_s2 |  -.1234378    .0342275    -3.61   0.000    -.1905225    -.0563532
     __t2_s3 |   .0913521    .0296776     3.08   0.002     .0331852     .1495191
     __t2_s4 |   .0038328    .0248068     0.15   0.877    -.0447878     .0524533
     __t2_s5 |   .0180132    .0214929     0.84   0.402    -.0241121     .0601384
       _cons |   -5.17632    .0348153  -148.68   0.000    -5.244557    -5.108084
------------------------------------------------------------------------------
```

Example: Two timescales, non-proportional hazards

```
. stmt orch, time1(df(3))  ///
 > time2(start(agediag) df(5) logtoff tvc(orch) dftvc(3))
```

Example: Two timescales, non-proportional hazards

```
. stmt orch, time1(df(3))  ///
 > time2(start(agediag) df(5) logtoff tvc(orch) dftvc(3))
```

```
Log likelihood = -7454.3291                     Number of obs     =     60,961
------------------------------------------------------------------------------
             | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        orch |   1.770931   .1044573     9.69   0.000      1.57759    1.987968
-------------+----------------------------------------------------------------
rcs          |
     __t1_s1 |   .0142601   .0258053     0.55   0.581    -.0363173    .0648376
     __t1_s2 |  -.0196129   .0251721    -0.78   0.436    -.0689494    .0297235
     __t1_s3 |   .0268569   .0258941     1.04   0.300    -.0238946    .0776085
     __t2_s1 |   .7620801   .0410964    18.54   0.000     .6815326    .8426276
     __t2_s2 |  -.1308936   .0415365    -3.15   0.002    -.2123036   -.0494835
     __t2_s3 |   .1362839   .0345208     3.95   0.000     .0686243    .2039435
     __t2_s4 |   .0188686   .0258904     0.73   0.466    -.0318756    .0696129
     __t2_s5 |   .0165599   .0216135     0.77   0.444    -.0258018    .0589216
__t2_s_orch1 |  -.2428242   .0686272    -3.54   0.000    -.3773311   -.1083172
__t2_s_orch2 |  -.0150246   .0680762    -0.22   0.825    -.1484516    .1184023
__t2_s_orch3 |  -.1123459   .0509553    -2.20   0.027    -.2122165   -.0124754
       _cons |  -5.213729   .0370125  -140.86   0.000    -5.286272   -5.141186
------------------------------------------------------------------------------
```

Predictions

- We are in the process of writing a predict command to be used after `stmt`
- Interested in predicting
  - Hazard for different values of the timescales
  - Survival
  - Hazard ratio over time
  - Hazard differences
  - Others?

Predictions: current syntax

`predict` *newvar*, { `hazard` | `xb` } [`startt1`(#) `startt2`(#) `startt3`(#) `followup`(#) `n`(#) `at`(*varname* # ...) `zeros`]

Options

- `startt1`(#) - Prediction entry time for timescale 1
- `startt2`(#) - Prediction entry time for timescale 2 (etc. for timescale 3)
- `followup`(#) - Follow-up time for prediction
- `n`(#) - How many intervals are needed for predictions up to the follow-up
- `at`(*varname* #) - Predict at values of other variables in the model
- Others are to be included

Prediction example

```
. stmt orch, time1(df(3))  ///
 > time2(start(agediag) df(5) logtoff tvc(orch) dftvc(3))
```

Prediction example

```
. stmt orch, time1(df(3))  ///
 > time2(start(agediag) df(5) logtoff tvc(orch) dftvc(3))

. predict haz, hazard startt1(0) startt2(70) followup(3) ///
>  n(10) at(orch 1)
```

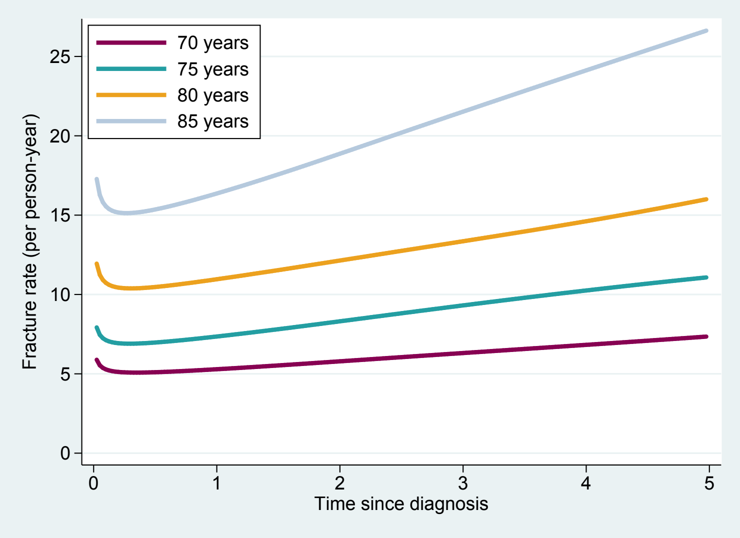
Prediction example

```
     +-----------------------------+
     | t1_haz   t2_haz         haz |
     |-----------------------------|
  1. |      0       70           . |
  2. |     .3     70.3   .00508109 |
  3. |     .6     70.6   .00512982 |
  4. |     .9     70.9    .0052459 |
  5. |    1.2     71.2   .00538255 |
     |-----------------------------|
  6. |    1.5     71.5   .00552947 |
  7. |    1.8     71.8   .00568337 |
  8. |    2.1     72.1   .00584082 |
  9. |    2.4     72.4    .0059984 |
 10. |    2.7     72.7   .00615495 |
     |-----------------------------|
 11. |      3       73   .00631038 |
     +-----------------------------+
```

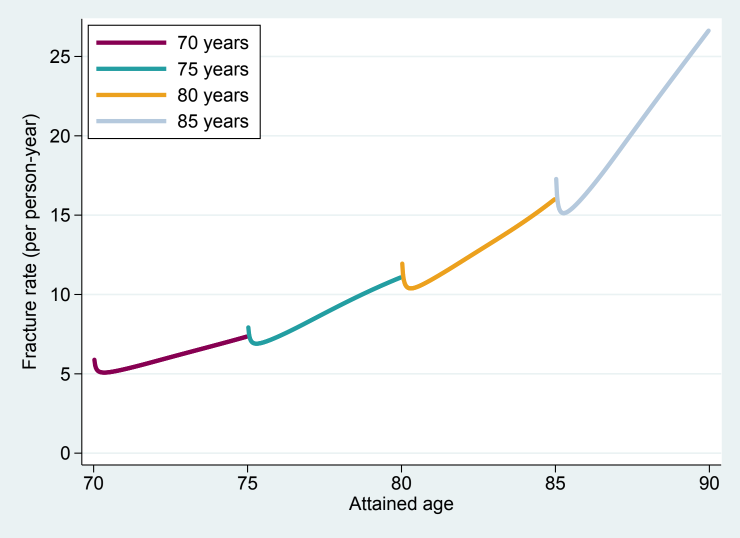
Prediction example

```
forvalues age = 70(5)85 {
      predict haz_`age', hazard startt1(0) startt2(`age') ///
          followup(5) n(200) at(orch 1)
}
```

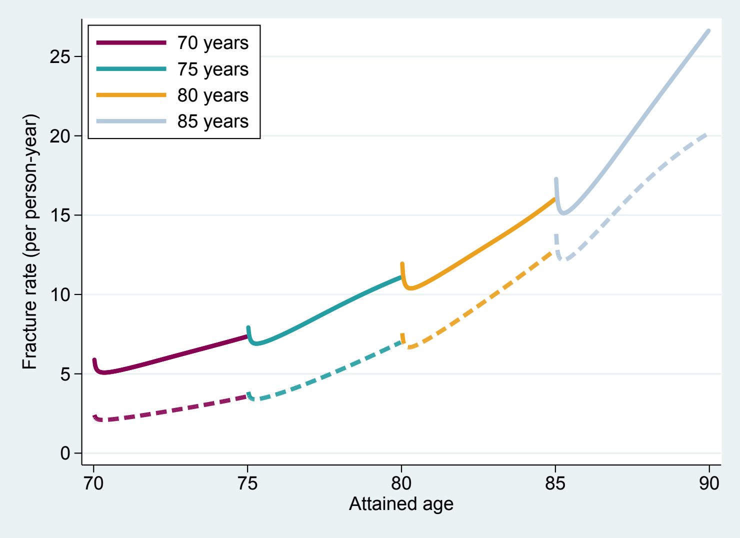
Prediction example



Prediction example



Prediction example



Ongoing work

- Interactions between the timescales
- Allow timescales for some individuals and not others
- More timescales?
- Predictions
- Suggestions?

Advantages and disadvantages

Disadvantages

- Numerical integration can be slow if you have large datasets
  - $N = 686$, model fits in $\approx$ 6 secs
  - $N = 60961$, model fits n $\approx$ 40 secs
  - $N = 423298$, model fits in $\approx$ 9 mins
  - A Poisson model with split data to model the second timescale will take a while to fit

Advantages

- Easy way for users to model multiple timescales & get predictions
- Models multiple timescales in a continuous way

References

[1] H. Bower, M. J. Crowther, and . P.C. Lambert.
strcs: A command for fitting flexible parametric survival models on the log-hazard scale.
The Stata Journal, 16:989–1012, 2016.

[2] P. Royston and M. K. B. Parmar.
Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects.
Statistics in Medicine, 21(15):2175–2197, Aug 2002.