Estimating stroke-free life expectancy using a multi-state model

Ardo van den Hout[1], Riccardo Marioni[2]
Fiona Matthews[1]

[1]*MRC Biostatistics Unit, Cambridge, UK*

[2]*Department of Public Health and Primary Care, Cambridge University, UK*

REVES, Paris, May 2011

MRC Biostatistics Unit

Estimating stroke-free life expectancy using a multi-state model

Outline:

Research interest & data

Multi-state model:

time-dependent hazards & missing states

Application

Life expectancies

Conclusion

Research interests

- History of stroke in the older population:
  - Risk factors for stroke: sex and education
  - Total residual life expectancy (LE) = stroke-free LE + LE with a history of stroke
- The wider scope is the study of ageing

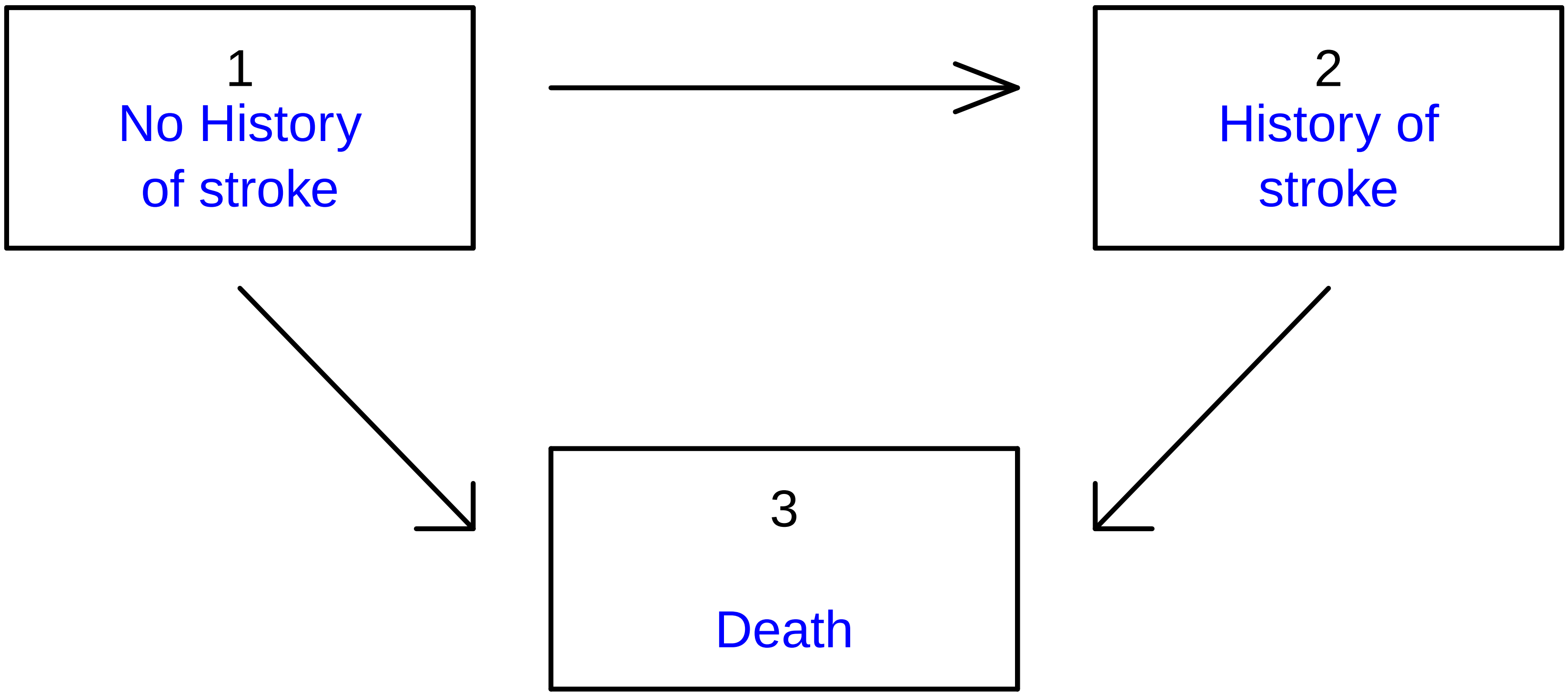
Data

- Longitudinal data available from the MRC Cognitive Function and Ageing Study (CFAS, `www.cfas.ac.uk`)

- No (reliable) data on exact time of stroke

- History of stroke = one or more strokes in the past

- Three-state model for history of stroke:



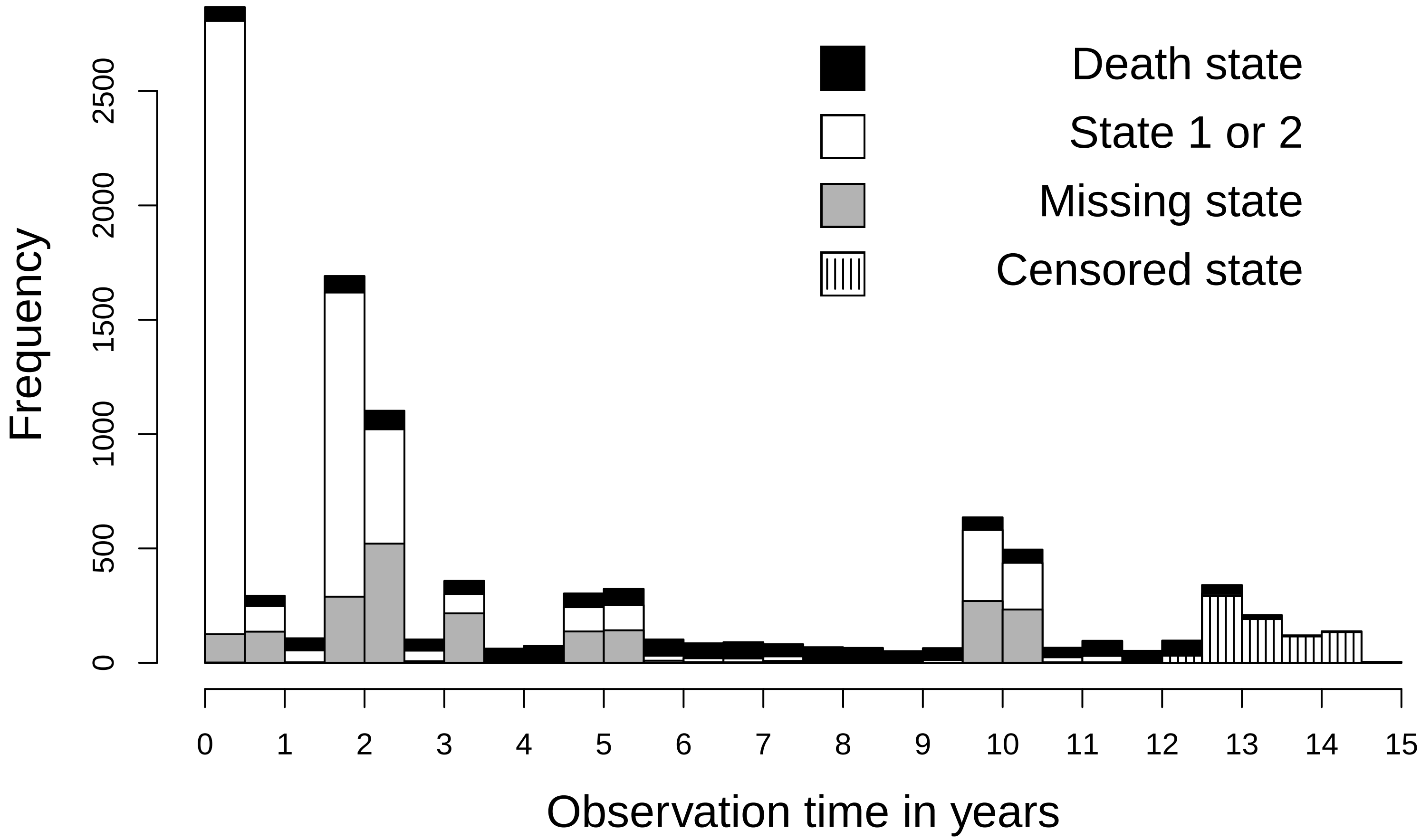
Censoring

- Occurrence of stroke is process in continuous time
- Pre-scheduled interviews: transitions between living states are interval censored
- In CFAS: death times are known
- Right-censoring at end of 14 years of follow-up

Missing data

- Pre-scheduled interviews
- A missed interview: missing data
- Missing values for state can be seen as panel data
  But...
  - Info on time-dependent covariates is also missing
  - If reason for missing an interview is related to the process under investigation, then ignoring this can lead to bias

- Panel data CFAS. For subset ($N = 2321$) in application:



Continuous-time model with interval-censoring

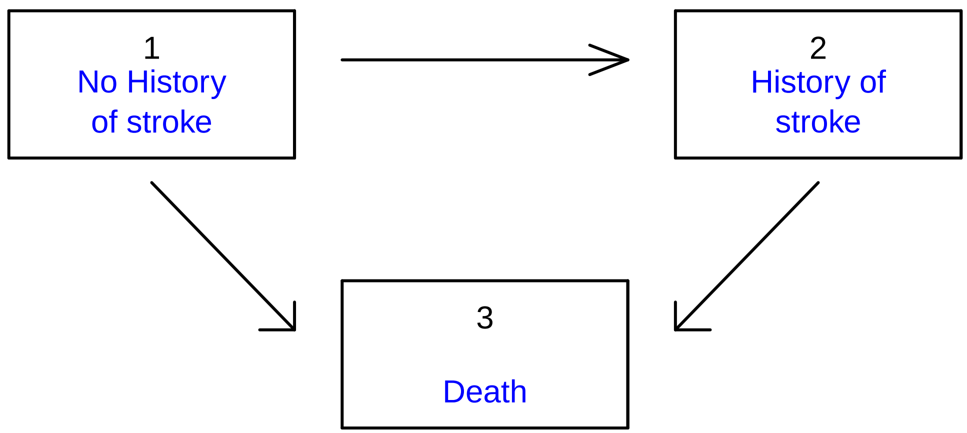
- Model fitted with `msm` in R (Jackson, 2011)

- Details of MLE will be skipped (Kalbfleisch & Lawless, 1985)

- Two aspects will be discussed briefly:
  - Time-dependent hazards
  - Missing values for states

Time-dependent hazards

- Three-state illness-death model for stroke:



- Transition-specific hazards: $q_{12}(t), q_{13}(t), q_{23}(t) > 0$

- Log-linear model: $\log[q_{rs}(t)] = \boldsymbol{\beta}_{rs}^{\top}\boldsymbol{z}(t)$

- Age as time scale

- Loglinear model in application:

  $\log[q_{rs}(\texttt{age})] = \beta_{rs.0} + \beta_{rs.1}\texttt{age} + \beta_{rs.2}\texttt{ybirth} + \beta_{rs.3}\texttt{sex} + \beta_{rs.4}\texttt{educ}$

- Or, equivalently with $t = \texttt{age}$

$$q_{rs}(t) \;=\; \lambda_{rs} \exp[\gamma_{rs} t] \exp[\boldsymbol{\alpha}_{rs}^{\top} \boldsymbol{z}]$$

- Age as time scale
  - Piecewise-constant hazards in likelihood
  - Example: for observation times t_1, t_2, t_3, t_4 hazards are assumed to be constant within $(t_1, t_2]$, $(t_2, t_3]$, $(t_3, t_4]$

Missing states. Basic idea

- Example: likelihood contribution for times t_1, t_2, t_3 with missing state at t_2

$$\mathbb{P}(X_{t_1}, X_{t_3}) \;=\; \sum_{x=1,2} \mathbb{P}(X_{t_1}, X_{t_2} = x, X_{t_3})$$

  Sum over all possible states at time t_2.

- Adding missing states improves piecewise-constant approximation!

Application: data

- Subset of CFAS: data from Newcastle
  State table:

	To				
From	1	2	3	Missing	Right-censored
1	2942	105	837	855	382
2	0	304	176	60	43
Missing	24	8	542	1200	341

- $N = 837 + 176 + 542 + 382 + 43 + 341 = 2321$

- 1441 women, 880 men

- Age at baseline:

< 70	$(70, 75]$	$(75, 80]$	$(80, 85]$	> 85
761	559	541	318	142

- Number of records per individual for living states:

1	2	3	4	5	6	7	8	9
566	788	629	151	99	54	21	12	1

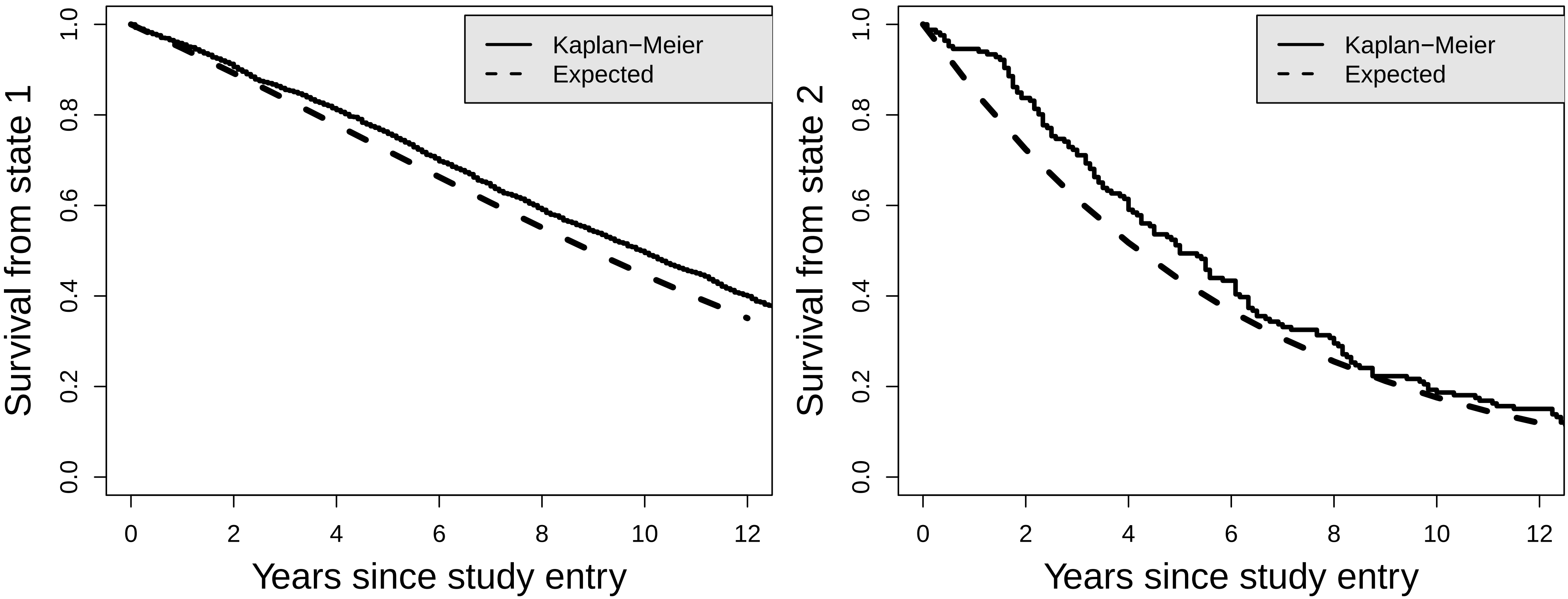
Fitted model

- Equation for transition intensities:

$$\log[q_{rs}(\mathrm{age})] = \beta_{rs.0} + \beta_{rs.1}\texttt{age} + \beta_{rs.2}\texttt{ybirth} + \beta_{rs.3}\texttt{sex} + \beta_{rs.4}\texttt{educ}$$

age		**ybirth**		**sex** (men ≡ 1)	
$\beta_{12.1}$	0.11 (0.05)	$\beta_{12.2}$	0.03 (0.05)	$\beta_{12.3}$	0.40 (0.20)
$\beta_{13.1}$	0.09 (0.01)	$\beta_{13.2}$	$<$ 0.01 (0.01)	$\beta_{13.3}$	0.36 (0.08)
$\beta_{23.1}$	0.05 (0.02)	$\beta_{23.2}$	-0.01 (0.02)	$\beta_{23.3}$	0.43 (0.13)
educ (10 or more yrs of educ ≡ 1)					
$\beta_{12.4}$	-0.02 (0.23)	$\beta_{13.4}$	-0.27 (0.10)	$\beta_{23.4}$	0.16 (0.16)

- Model validation is not straightforward
  - Varying observation times
  - Interval censoring
  - Missing data
- Heuristic check: assess predicted survival


Kaplan−Meier
Expected
Survival from state 1
Years since study entry
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
Kaplan−Meier
Expected
Survival from state 2
Years since study entry
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12

Life expectancies

- Residual life expectancy (LE) at a given age t_0

- Alive/death survival: LE is the expectation of the remaining years spent alive (U):

$$\begin{aligned}\mathbb{E}(U|t_0, \mathcal{Z}) &= \int_0^\infty u f(u|t_0, \mathcal{Z}) du = \int_0^\infty S(u|t_0, \mathcal{Z}) du \\ &= \int_0^\infty \mathbb{P}(X_{t_0+u} = 1|t_0, \mathcal{Z}) du\end{aligned}$$

- Multi-state survival: LE in state s given state r at t_0:

$$e_{rs}(t_0) = \int_0^\infty \mathbb{P}(X_{t+t_0} = s|\ X_{t_0} = r, \mathcal{Z}) dt$$

- LE in state s given state r is *occupancy time* for $T = \infty$ (Kulkarni, 2011)

- Marginal LE given by

$$e_{\bullet s}(t_0) = \sum_{r=\text{living state}} \mathbb{P}(X_{t_0} = r|\mathcal{Z}) e_{rs}(t_0)$$

  Needed: distribution of the living states at age t_0

- Total LE at age t_0 is

$$e(t_0) = \sum_{s=\text{living state}} e_{\bullet s}(t_0)$$

- Logistic regression model for distribution of states:

$$\mathbb{P}(\text{State 2 at baseline}) = \frac{\exp[\mu]}{1 + \exp[\mu]}$$

$$\mu = \alpha_0 + \alpha_1 \texttt{age}$$

- Multinomial regression for model with > 2 living states

Software for computation of life expectancies

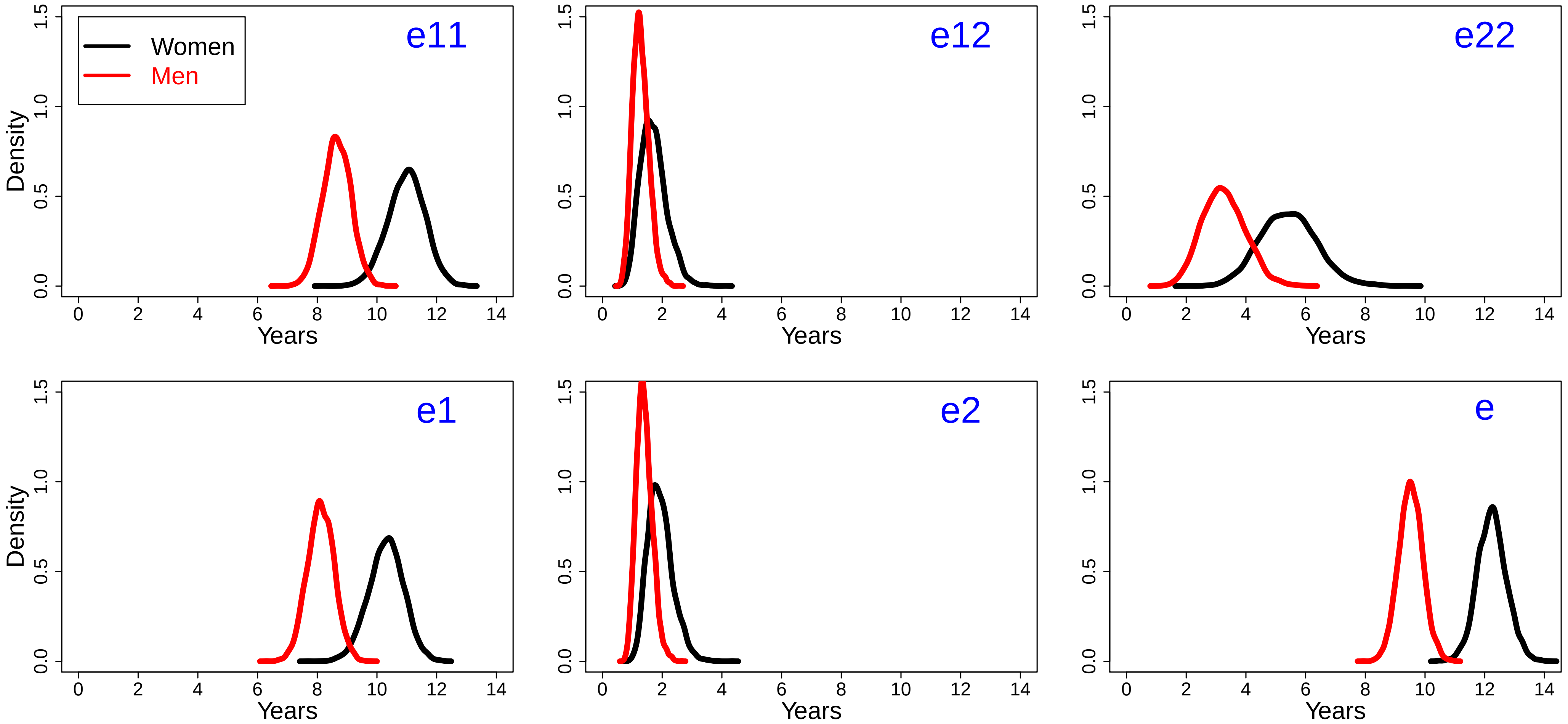
- Fit multi-state model with `msm` with time-dependent `age`

- Estimate and investigate LEs with additional `R` code:
  - Functions for computation, summarising and plotting
  - MLE simulation used to compute uncertainty
  - Currently implemented for 3-state and 4-state models

Example of output in application:

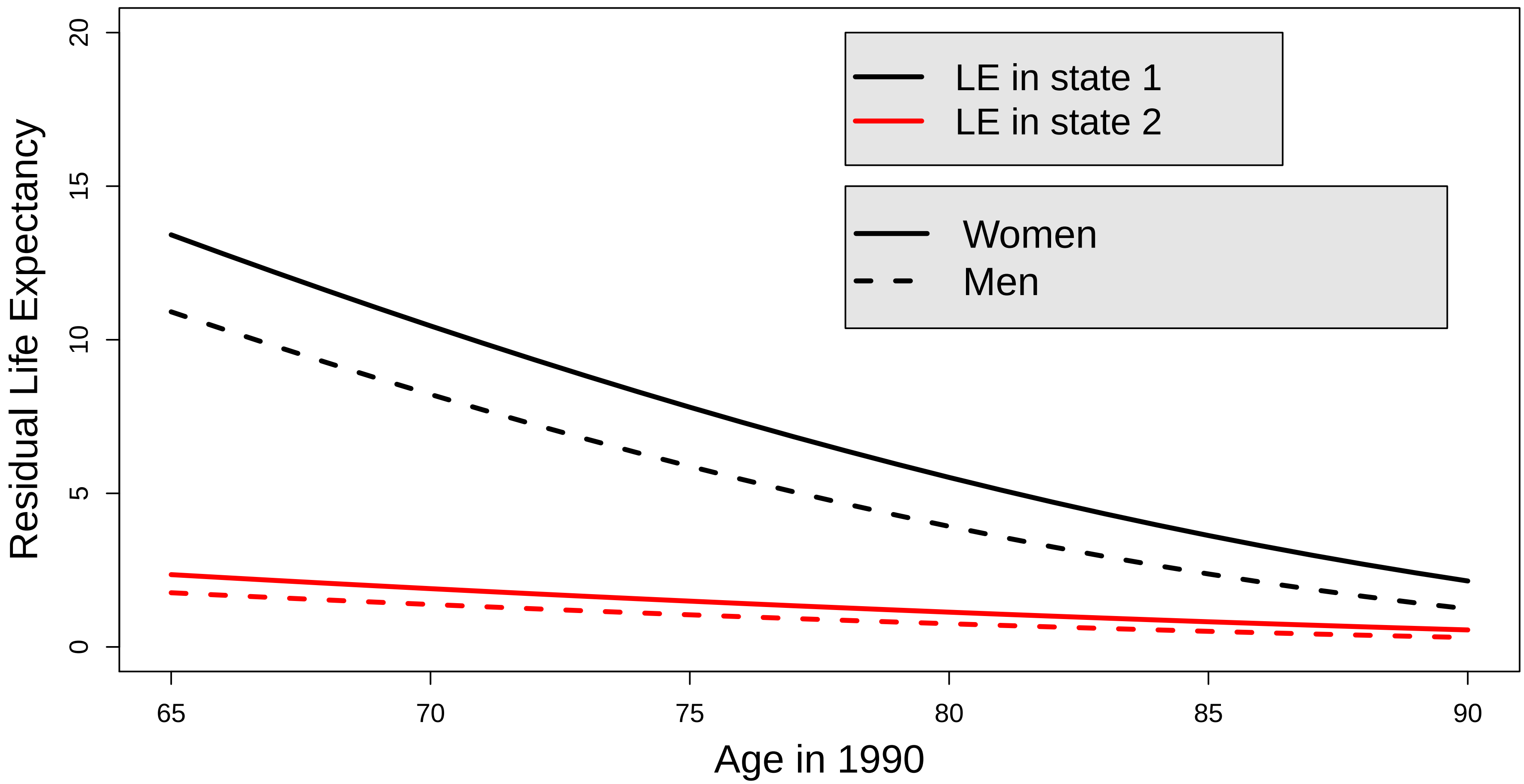
```
 age   ybrth  sex  educ
 70    20     0    1
Using simulation with  1000 replications

Point estimates, and mean, SEs, quantiles from simulation:
      pnt    mn   se 0.025q 0.975q
e11 11.14 10.97 0.60   9.82  12.18
e12  1.67  1.72 0.45   0.96   2.70
e21  0.00  0.00 0.00   0.00   0.00
e22  5.40  5.46 0.92   3.76   7.26
e1  10.45 10.30 0.58   9.18  11.43
e2   1.90  1.95 0.43   1.23   2.89
e   12.35 12.25 0.48  11.35  13.25
```



Women
Men
e11
e12
e22
e1
e2
e
Density
Density
Years
Years
Years
Years
Years
Years
0.0
0.5
1.0
1.5
0
2
4
6
8
10
12
14

For $\geq$ 10 years of education:



Conclusion (I)

- Continuous-time illness-death model for stroke. Flexibility w.r.t. censoring and inclusion of covariates

- Missing values for state can be taken into account

- LEs using model parameters. Complete info on uncertainty. Alternative to multi-state life tables methods

- Computations with available software

Conclusion (II)

- Topics in other work
  - Non-ignorable missing data
  - Misclassification of states
  - Piecewise-constant hazards with a grid independent of observations

- Future work
  - Functional form of regression equation
  - Model validation