

Survival models with health monitoring

Peter McCullagh

Department of Statistics
University of Chicago

Rutgers University Statistics Symposium
Statistics for the Century of Data
May 2, 2014

Outline

Survival studies

Survival processes

Temporal realignment

Examples

Distribution theory

Study plan for a survival study

Continuous recruitment of patients

At recruitment on date d :

- determine patient eligibility

- measure covariates x_i

- measure baseline health variables $Y_i(0)$

- assign treatment arm by randomization $i \mapsto a_i$

At annual or semi-annual check-ups:

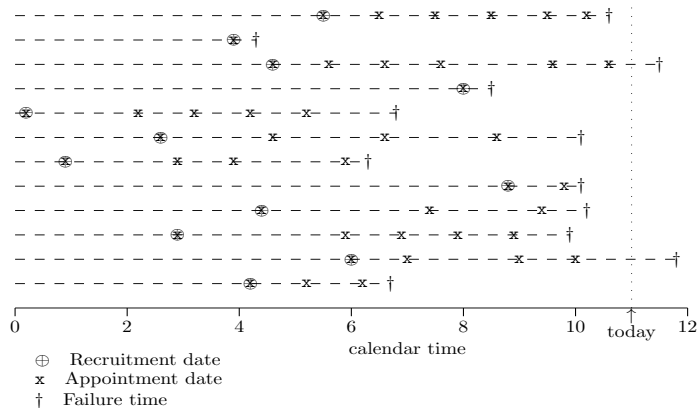
- record date $d + t$

- measure health status or quality-of-life $Y_i(t)$

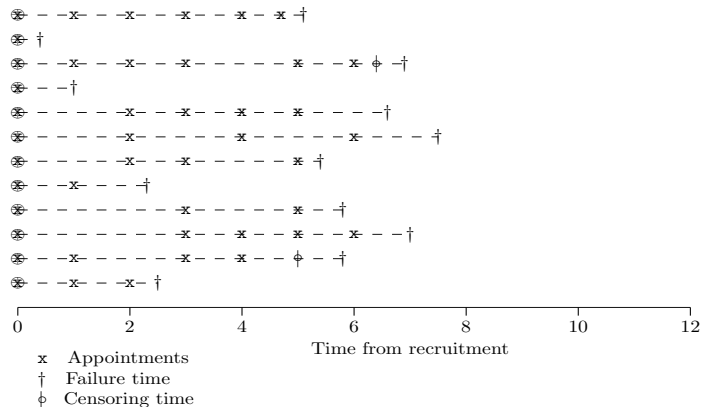
Record death if it occurs, and obtain date

After xx years, analyze data

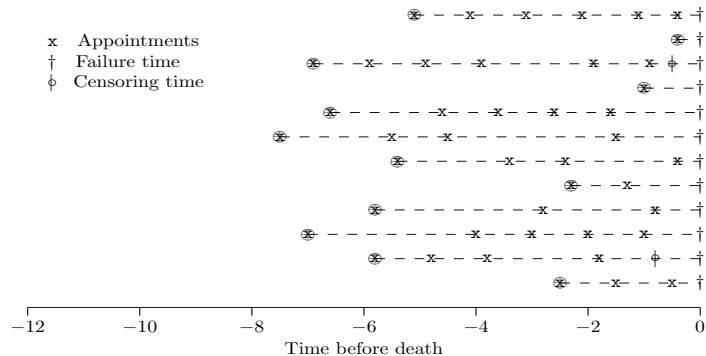
Diaries for 12 patients in calendar time



Diaries for 12 patients aligned by recruitment



Diaries for 12 patients aligned by death



Data structure for patient i

One record: $(d_i, x_i, a_i), (T_i, \mathbf{t}_i, Y_i[\mathbf{t}_i])$

Recruitment date d_i (calendar time);

Covariates x_i (age, sex,...)

Treatment arm a_i : control, active1,...;

Failure time T_i (at date $d_i + T_i$)

Appointment schedule: $\mathbf{t}_i \subset [0, T_i)$: $\mathbf{t} = (0, t_1, t_2, \dots)$
(relative to recruitment date)

Health values $Y_i[\mathbf{t}] = (Y_i(0), Y_i(t_1), \dots)$

Censoring time if failure time unavailable

Implications of exchangeability

Discussion confined to subset of patients $\{i : x_i = x_0\}$

\mathcal{R} : space of medical records (one patient, one time)

(Exch): Response values are distributed exchangeably.

(i) T_1, T_2, \dots exchangeable (real-valued, $T_i \in \mathbb{R}^+$)

(ii) $\mathbf{t}_1, \mathbf{t}_2, \dots$ exchangeable (random subsets of \mathbb{R}^+)

(iii) $Y_1(0), Y_2(0), \dots$ exch baseline health-values; \mathcal{R} -valued

(iv) $Y_1(t), Y_2(t), \dots$ exch at time $t = 20!$

(v) $Y_1(\cdot), Y_2(\cdot), \dots$ exch \mathcal{R} -valued processes

(vi) $(T_1, Y_1(\cdot)), (T_2, Y_2(\cdot)), \dots$ jointly exch

(vii) $Y_1(T_1 - s), Y_2(T_2 - s), \dots$ exch at $s > 0$

Inevitable dependencies

$Y_i(t)$ health status of patient i at time $t \geq 0$

T_i failure time: (RIP: $d_i + T_i$)

$\mathbf{t}_i \subset [0, T_i)$: appointment schedule

Robust health and longevity: $Y(0)$ and T

$Y(0)$ and $\mathbf{t} \subset [0, T)$

$Y(0)$ and $\#\mathbf{t}$

Implications for annual check-ups:

$T_1 = 3.5$; $\mathbf{t}_1 = (0, \dots, 2)$; $Y_1[\mathbf{t}_1] = (Y_1(0), \dots, Y_1(2)) \in \mathcal{R}^3$

$T_2 = 9.7$; $\mathbf{t}_2 = (0, \dots, 5)$; $Y_2[\mathbf{t}_2] = (Y_2(0), \dots, Y_2(5)) \in \mathcal{R}^6$

$(T_1, Y_1(\cdot)) \sim (T_2, Y_2(\cdot))$ exchangeably distributed

Given $\#\mathbf{t}_1 = 1$, $\#\mathbf{t}_2 = 6$, $Y_1(0)$ and $Y_2(0)$ not exch

\implies Sequence length and sequence values not independent

\mathcal{R} -valued survival process

\mathcal{R} state space: health values for one appointment

e.g. $\mathcal{R} = \{0, 1\}$ (dead/alive);

$\mathcal{R} = \mathbb{R}$ (pulse rate), $Y(t) \in \mathcal{R}$

Absorbing state b :

$Y(t) = b$ implies $Y(t') = b$ for $t' \geq t$.

$Y(0) \neq b$: alive at recruitment!

$T = \sup\{t: Y(t) \neq b\} < \infty$: no immortals

Focus on time-reversed process (revival process)

$Z_i(s) = Y_i(T_i - s)$ at time s prior to failure

$Z_i(s) \neq b$ for $s > 0$

$Z_i(T_i) = Y_i(0)$ baseline value at recruitment

Two statistical assumptions

Survival process: $Y_i(t)$

Survival time: $T_i = \sup\{t: Y_i(t) \neq b\}$ (finite)

Revival process: $Z_i(s) = Y_i(T_i - s)$

Transformation: $Y(\cdot) \mapsto (T, Z(\cdot))$

Appointments: $\mathbf{t}_i \subset [0, T_i)$, $\mathbf{s}_i = T_i - \mathbf{t}_i \subset (0, T_i]$

Sampling assumption: (appointment dates)

$\mathbf{t}^{(k)} \subset \mathbf{t}$ subset of initial k appointments

$\mathbf{t} \perp\!\!\!\perp Y \mid (\mathbf{t}^{(k)}, Y[\mathbf{t}^{(k)}, T])$

Rationale for temporal realignment:

Temporal patterns: forward time versus reverse time

Evidence?

Distributional and likelihood factorizations

Density of $(T, \mathbf{t}^{(k)}, Y[\mathbf{t}^{(k)}])$ at $(t, \mathbf{t}^{(k)}, y)$:

$$\begin{aligned} & f(t) \times \prod_{j < k} p(t_j, y_j | \mathcal{H}(t_{j-1}), T = t) \\ = & f(t) \times \prod_{j < k} p(y_j | t_j, \mathcal{H}(t_{j-1}), T = t) \times \prod_{j < k} p(t_j | \mathcal{H}(t_{j-1}), T = t) \\ = & f(t) \times g_k(y; t - \mathbf{t}^{(k)} | T = t) \times \prod_j p(t_j | \mathcal{H}(t_{j-1}), T = t), \\ = & \text{surv dens} \times \text{revival density given } T \times \text{appt process} \end{aligned}$$

Bear in mind that $p(Y(t_k) = b) > 0!!$

State space and observation space

State space \mathcal{R} for health process $Y_i(\cdot)$:

$Y_i(t)$ = pulse rate of i at time t ; $\mathcal{R} = \mathbb{R}$

$Y_i(t)$ = (blood pressure, pulse, CD4 count); $\mathcal{R} = \mathbb{R}^3$

Observation space for one patient:

for value at one time $\mathcal{S} = \mathcal{R}$

for values at two times $\mathcal{S} = \mathcal{R}^2$

In general, for an indefinite number of visits:

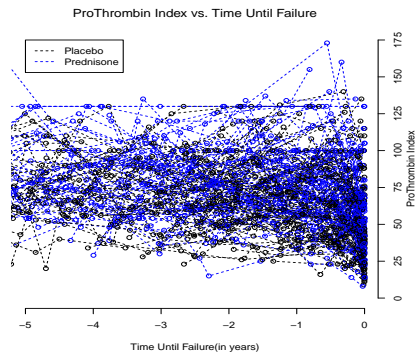
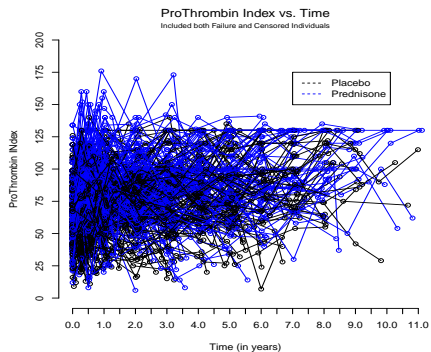
$$\mathcal{S} = \bigcup_{j=0}^{\infty} \mathcal{R}^j$$

Observation space for n patients:

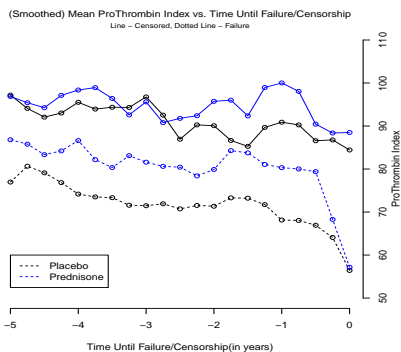
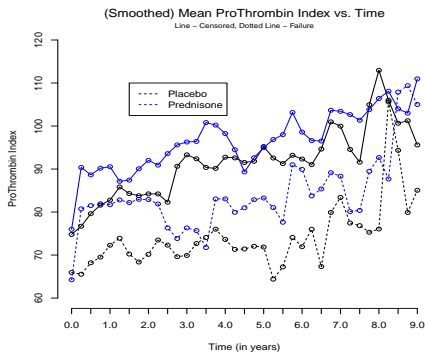
$$\mathcal{S} \times \mathcal{S} \times \cdots \times \mathcal{S} = \mathcal{S}^n$$

Distribution exchangeable on \mathcal{S}^n with respect to permutation of patients

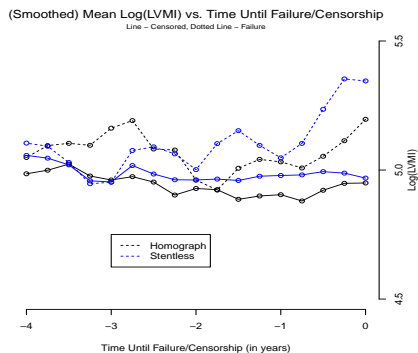
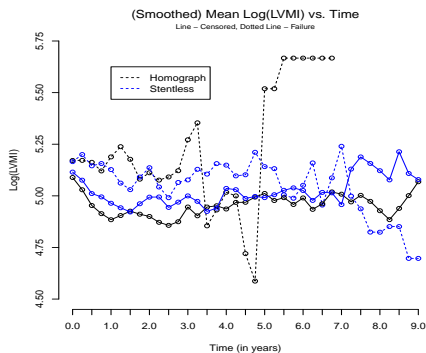
Cirrhosis, prednizone and prothrombin plots



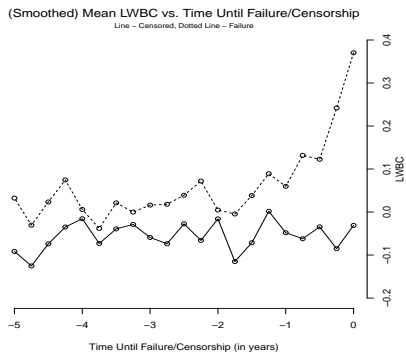
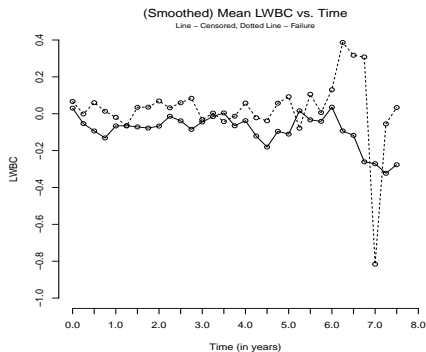
Cirrhosis, prednizone and prothrombin mean plots



Heart disease: Log LVMI temporal plots



Chronic Myeloid leukaemia: Log WBC count



Survival process distribution

$Z(s) \equiv Y(T - s)$: health status at time s before failure

$Z(s) \in \mathcal{R}$ space of health records

Finite-dimensional restriction: $Z[\mathbf{s}] = (Z(s_1), \dots, Z(s_k)) \in \mathcal{R}^k$

Joint density at $y \in \mathcal{R}^k$ is $g_k(y; \mathbf{s})$

Joint density of observation $(T, \mathbf{t}, Y[\mathbf{t}])$ at (t, \mathbf{t}, y)

$$f(t) \times p(\mathbf{t} | T = t) \times g_{\#\mathbf{t}}(y; t - \mathbf{t})$$

on the space

$$\mathbb{R}^+ \times \bigcup_{k=0}^{\infty} \mathcal{R}_+^k$$

where \mathcal{R}_+ also includes the appointment date.

Statistical aspects: parametric models

Joint density function: Likelihood function

$$f(t; \beta) \times p(\mathbf{t} | T = t) \times g_{\# \mathbf{t}}(y; t - \mathbf{t}, \theta | T)$$

β : governing the survival distribution

θ : governing the revival process

$\mathbf{t} \perp\!\!\!\perp Y(\cdot) | T; \quad Z(\cdot) \perp\!\!\!\perp T$

Predictive survival distribution given $Y[\mathbf{t}^{(k)}] = y$

$$p(t | Y[\mathbf{t}^{(k)}] = y) \propto f(t; \beta) \times g(y, t - \mathbf{t}^{(k)}; \theta)$$

NOT including $p(\mathbf{t}^{(k)} | T = t)$ because ...

Should include $p((\mathbf{t}^{(k)}, \dots) | T = t)$

Illustration by simulation

$$T \sim \exp(\mu = 5),$$

$$E(Z(s)) = 10 + 10s/(10 + s)$$

$$\text{cov}(Z(s), Z(s')) = 1 + \delta_{ss'} + \exp(-|s - s'|)$$

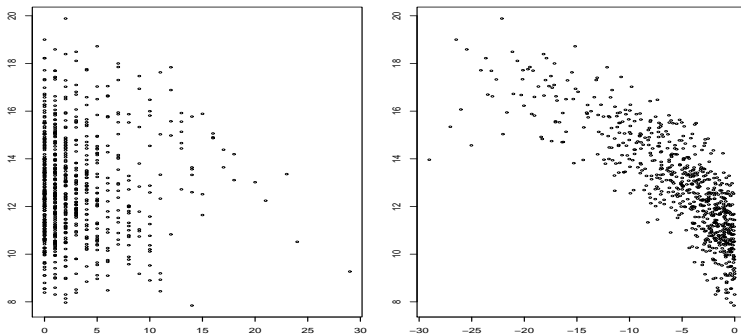


Figure : Simulated health status sequences aligned by recruitment time (left) and the same sequences aligned by failure time (right)

Randomization and treatment effects

Primary analysis: survival distribution as function of treatment
Standard methods and assumptions

Secondary analysis: prothrombin trajectory versus treatment

Response: $Y_i(t)$: blood coagulation index of i at time $t \geq 0$

At recruitment: $a_i(0) = \text{NULL} \neq \text{Control}$

Post-randomization: $a_i(t) = \text{prednizone or control for } t > 0$

\Rightarrow at least three levels

In reverse time:

$Z_i(s) = Y_i(T_i - s)$: same value, different time.

$\bar{a}_i(s) = a_i(T_i - s)$ treatment level at revival time s

$\bar{a}_i(T_i) = \text{NULL}$ (pre-randomization)

Case study: cirrhosis and prothrombin sequence

Study design:

Cirrhosis diagnosed by biopsy: Copenhagen: 1962–1969

Randomized treatment: Prednizone or Control

Blood coagulation index Y measured every ~ 6 months

Baseline variables: sex, age,... (not available in data)

Short versus long sequences:

	Record length							
	1–2	3–4	4–5	7–8	9–10	11–12	13–14	15–16
mean	55.2	64.6	70.3	72.2	79.2	80.5	71.9	54.0
sd	20.7	23.0	19.5	18.2	20.9	19.8	22.2	5.3
#patients	56	69	65	45	33	13	8	3

Cirrhosis case study (contd)

Table 1: Average prothrombin levels indexed by T and t .

Survival time (T)	Time t after recruitment (yrs)								
	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8+
0-1	58.0								
1-2	72.5	66.4							
2-3	72.6	73.2	66.0						
3-4	69.8	71.2	68.5	54.2					
4-5	68.5	75.7	72.5	74.6	57.7				
5-6	70.5	77.3	73.5	57.1	64.5	60.9			
6-7	81.8	73.6	81.1	80.6	79.4	75.5	75.8		
7-8	84.4	88.8	88.1	92.1	85.2	81.2	84.3	88.1	
8+	77.3	73.6	87.0	74.1	92.0	80.3	89.2	79.4	84.7

ANOVA for prothrombin averages

Table 2: ANOVA decomposition for Table 1

u/v	$\ P_u Y\ ^2 - \ P_v Y\ ^2$	d.f.	M.S.	
$(R + C + D)/(R + C)$	544.4	7	77.8	diag
$(R + C + D)/(R + D)$	238.9	7	34.1	cols
$(R + C + D)/(C + D)$	818.8	7	117.0	rows
$RC/(R + C + D)$	498.3	21	23.7	resid

R, C, D : rows, cols, diag=reverse time (9-level factors)

Prothrombin mean square variation associated with
survival time T (rows) 117.0: $F = 4.9$
time remaining to failure: (diag) 77.8: $F = 3.3$
time following recruitment: (cols) 34.1: $F = 1.4$
residual: 23.7

⇒ negligible prothrombin variation associated with time measured from recruitment

Cirrhosis revival model

Basic no-frills revival model for prothrombin level:

$$Z_i(\mathbf{s}) = Y_i(T_i - \mathbf{s}) = \mu(\mathbf{s}, T_i) + \eta_i(\mathbf{s}) + \epsilon_i + \epsilon'_{i\mathbf{s}}$$

η zero-mean cts GP; ϵ, ϵ' iid GPs

$$E(Z_i(\mathbf{s}) | T) = \alpha + \tau_{\bar{a}_i(\mathbf{s})} + \beta_0 T_i + \beta_1 \mathbf{s} + \beta_2 \log(\mathbf{s} + \delta)$$

$$\text{cov}(Z_i(\mathbf{s}), Z_j(\mathbf{s}') | T) = \sigma_1^2 \delta_{ij} K(\mathbf{s}, \mathbf{s}') + \sigma_2^2 \delta_{ij} + \sigma_3^2 \delta_{ij} \delta_{\mathbf{s}\mathbf{s}'}$$

Key points:

responses for distinct patients are indep (and i.d.)

temporal trend in mean level $\beta_1 \mathbf{s} + \beta_2 \log(\mathbf{s} + \delta)$

$\delta = 1$ day

treatment has *three* levels: *null, control, prednizone*

individual-specific additive random effects ϵ_i iid, const in time

temporal autocorrelation: $K(\mathbf{s}, \mathbf{s}') = \exp(-|\mathbf{s} - \mathbf{s}'|/\lambda)$

(AR1 with $\lambda = 1.67$ years)

Fitted coefficients in basic revival model

Fitted coefficients using uncensored cases only

Table 3: Regression coefficients in a revival model

Covariate	Coef	S.E.	Ratio
Null treatment	0.00	—	
Control	2.41	1.43	1.7
Prednizone	13.56	1.47	9.2
Survival (T)	1.75	0.47	3.7
Revival (s)	-2.12	0.47	-4.5
$\log(s + \delta)$	4.66	0.41	11.4

Note: huge signal associated with revival time scale s

REML estimated variance components:

AR1: 211.4

Individual 209.2

Residual 179.7

Model checking

Q1: *Is the treatment effect for long-term survivors the same as the treatment effect for short-term survivors?*

Interaction: *Treat.T* included in mean model

$2 * LR = 0.83$ on 2 d.f.

(not entirely obvious or trivial calculation)

Q2: *Does the treatment effect depend on revival time s ?*

Interaction 1: *Treat.s*: $2 \times LR = 3.90$ on 2 df

Interaction 2: *Treat.log(s)*: $2 \times LR = 8.86$ on 2 df

Q3: *Is the temporal mean model adequate?*

Extra covariance term: $-|\log(s) - \log(s')|$: $2 \times LR = 1.2$

Q4: *Is alignment by failure time adequate?:*

Remedy: include alignment-by-recruitment covariance term:

Extra covariance term: $-|t - t'|$: $2 \times LR = 2.38$ on 1 df

Censored versus uncensored patients

Q5: *Are the values estimated from censored patients compatible with those from uncensored patients?*

Table 5: Estimated regression coefficients in a revival model

Covariate	Uncensored		Censored		Std Ratio
	Coef	S.E.	Coef	S.E.	
Null treatment	0.00	—	0.00	—	
Control	2.41	1.43	4.11	1.82	0.73
Prednizone	13.56	1.47	11.49	1.74	-0.90
Survival (T)	1.75	0.47	2.78	0.37	1.71
Revival (s)	-2.12	0.47	-2.46	0.56	-0.47
$\log(s + \delta)$	4.66	0.41	1.41	4.02	-0.80

Values for censored patients computed by Walter Dempsey

Survival prognosis for patient u

Partial prothrombin sequence and survival prognosis

\mathbf{t}_u (days)	0	126	226	392	770	1127	1631	1855
$Y_u[\mathbf{t}_u]$	49	93	122	120	110	100	72	59

Revival model is a specification for the joint distribution

$$p(T_u = t) \times p(Y[t - \mathbf{t}_u] = y | T_u = t).$$

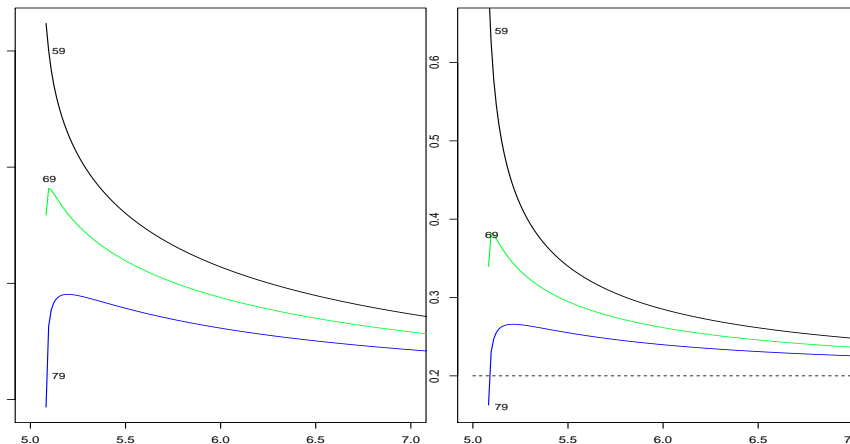
The second factor is Gaussian $N(\mu, \Sigma)$

$$\log(p(y | T_u = t)) = \text{const} - (y - \mu)' \Sigma^{-1} (y - \mu) / 2$$

where μ is linear in $t - \mathbf{t}_u$ and $\log(t - \mathbf{t}_u + \delta)$.

Survival prognosis for patient 402 (contd)

log modification factors for the predictive survival density (left panel) and hazard functions (right panel).



Three versions of the record for patient 402:
Final prothrombin value is 59, 69 or 79.

References,...

- Andersen, P.K., Hansen, L.H. and Keiding, N. (1991) Assessing the influence of reversible disease indicators on survival. *Statistics in Medicine* **10**, 1061-1067.
- Cox, D.R. (1972) Regression models and life tables (with discussion). *J. Roy. Statist. Soc. B* **34**, 187-220.
- Clifford, D. and McCullagh, P. (2006) The regress function. *R News*, **6** 6-10.
- Dempsey, W. (2014) galton.uchicago.edu/~wdempsey/research.html
- Diggle, P.J., Farewell, D. and Henderson, R. (2007) Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal (with discussion). *Applied Statistics* **56**, 499-550.
- Diggle, P.J., Sousa, I. and Chetwynd, A. (2008) Joint modeling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine* **27**, 2981-2998.
- Henderson, R., Diggle, P. and Dobson, A. (2000) Joint modeling of longitudinal measurements and event time data. *Biostatistics* **1**, 465-480.
- Kurland, B.F., Johnson, L.L., Egleston, B.L. and Diehr, P. (2009) Longitudinal data with follow-up truncated by death: match the analysis method to the research aims. *Statistical Science* **24**, 211-222.
- Rosthøj, Keiding, N. and Schmiegelow, N. (2012) Estimation of dynamic treatment strategies for maintenance therapy of children with acute lymphoblastic leukaemia: an application of history-adjusted marginal structural models. *Statistics in Medicine* **31**, 470-488.
- van Houwelingen, H.C. and Putter, H. (2012) *Dynamic Prediction in Clinical Survival Analysis*. Monographs on Statistics and Applied Probability 123; CRC Press.