A review of recurrent events methods with applications to a Danish dataset on atrial fibrillation

Olivier Bouaziz

Université Paris Descartes, MAP5

15 février 2016

Atrial fibrillation database

- Dataset on atrial fibrillation (AF) collected by cardiologist doctors : Jakob Schrøder, Ulrik Dixen, Per Lav Madsen, Christian Jøns, Bue Agner, Dana Li, Louise Bjørnager.
- Statistical analysis done by Torben Martinussen and myself.
- Patients with atrial fibrillation occasionnally experience episodes of rapid and irregular heart rate which may lead to hospitalization to the cardiology ward.
- Data collected from January 2009 until March 2014 of 175 patients who were enrolled, having either paroxysmal or persistent AF.

Atrial fibrillation database

The data :

- We know the exact dates of hospitalization.
- The event of interest is the time since study entry until hospitalization due to AF attack or cardioversion.
- Censoring : patients are censored at the end of the study.
- Terminal events : patients can move to permanent AF status or they can die.
- Covariates : type of AF (persistent, paroxysmal), gender, age at inclusion, alcohol consomption, tobacco use, hypertension, heart failure, heart valve disease, ischemic heart disease, hyperthyroidism, diabetes mellitus, COPD, kidney disease.

Atrial fibrillation database

Goal of the study :

- Find the risk factors for the occurence of further recurrent events and evaluate the effect of each risk factor through a regression model.
- Knowing the history of a patient predict the odds of getting a new recurrent event in the future.

1. Modelling the hazard rate and basic estimation techniques

2. Analysis of the atrial fibrillation dataset

3. Extended models

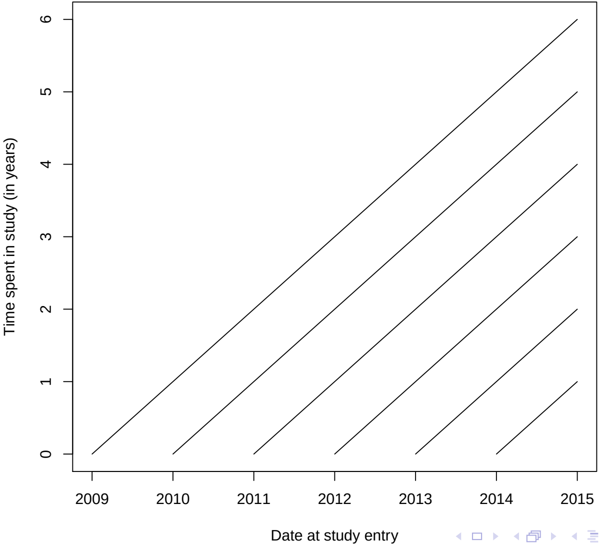
Contents

1. Modelling the hazard rate and basic estimation techniques

2. Analysis of the atrial fibrillation dataset

3. Extended models

Censoring effect : the Lexis diagram



Notations and framework

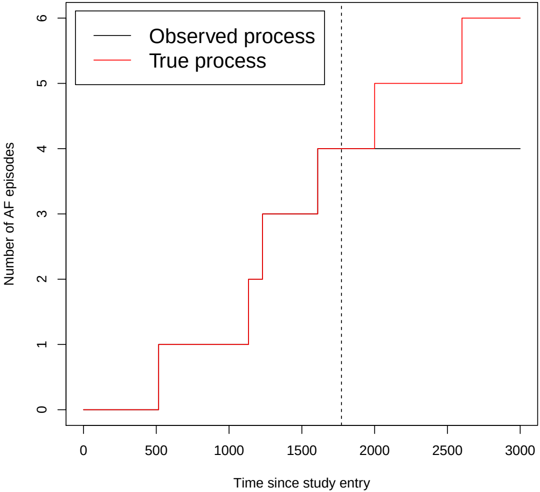
The time $t \geq 0$ represents time since study entry. Define for $i = 1, \ldots, n$:

- The process of interest : $\tilde{N}_i(t)$, $t \geq 0$, represents number of hospital admissions due to AF attacks since study entry.
- The time to terminal event : D_i
- The time to censoring : C_i
- The external covariate vector : $X_i(t) = (X_i^1(t), \ldots, X_i^p(t))^T$.

- We observe :
$$\begin{cases} X_i(t) \\ T_i = D_i \wedge C_i \\ N_i(t) = \tilde{N}_i(t \wedge T_i), i = 1, \ldots, n. \end{cases}$$

Counting process of interest v.s. observed counting process

start	end	event
0	516	TRUE
516	1134	TRUE
1134	1230	TRUE
1230	1609	TRUE
1609	1772	FALSE



The Andersen-Gill model with a terminal event

Let $\tilde{Y}(t) = I(D \geq t)$ represent the non-observed at-risk process. To account for a terminal event, we introduce the following model :

$$\mathbb{E}[d\tilde{N}(t)|X(t), \tilde{Y}(t)] = \tilde{Y}(t)\lambda(t|X(t))dt$$

which is equivalent to the formula :

$$\lambda(t|X(t)) = \lim_{\Delta t \to 0} \frac{\mathbb{P}[\Delta\tilde{N}(t) = 1|D \geq t, X(t)]}{\Delta t}$$

The independent censoring assumption

Let $Y(t) = I(D \wedge C \geq t)$ be the observed at-risk process.
We assume that

$$\mathbb{E}[d\tilde{N}(t)|X(t), \tilde{Y}(t)] = \mathbb{E}[d\tilde{N}(t)|X(t), Y(t)]$$

This condition is fulfilled if, for instance,

$$C \perp\!\!\!\perp (d\tilde{N}(t), D) \,|\, X(t).$$

Under this assumption, we have :

$$\mathbb{E}[dN(t)|X(t), Y(t)] = Y(t)\lambda(t|X)dt$$

where $N(t)$ and $Y(t)$ are observed processes !

Estimation of the Cox regression parameter

Let $t_{(1)} < t_{(2)} < \cdots < t_{(H)}$ denote the H unique observed ordered event times. In the Cox model,

$$\lambda(t|X(t)) = \lambda_0(t)\exp(\theta_0^T X(t)),$$

and the regression parameter is estimated through the Cox partial likelihood :

$$\hat{\theta} = \text{argmax}_\theta \prod_{h=1}^{H}\prod_{i=1}^{n}\left(\frac{\exp(\theta^T X_i(t_h))}{\sum_{j=1}^{n} Y_j(t_h)\exp(\theta^T X_j(t_h))}\right)^{dN_i(t_h)}.$$

This estimator is asymptotically normal and the asymptotic variance can be estimated via the robust variance estimator.

Nonparametric estimation of the cumulative mean function

Under the independent censoring assumption,

$$\mathbb{E}[dN(t)] = \mathbb{E}[d\tilde{N}(t)I(C \geq t)] = \mathbb{E}[d\tilde{N}(t)]\mathbb{P}[C \geq t].$$

Then

$$\mathbb{E}[\tilde{N}(t)] = \int_0^t \frac{\mathbb{E}[dN(u)]}{\mathbb{P}[C \geq u]} \left(= \int_0^t \frac{\mathbb{P}[D \geq u]\mathbb{E}[dN(u)]}{\mathbb{P}[T \geq u]} \right)$$

and the cumulative mean function is estimated by :

$$\widehat{\mathbb{E}[\tilde{N}(t)]} = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{n\hat{S}_C(u)} \left(= \int_0^t \frac{\sum_{i=1}^n \hat{S}_D(u)dN_i(u)}{\sum_{j=1}^n Y_j(u)} \right)$$

where $\hat{S}_C(u)$ (resp. $\hat{S}_D(u)$) is the Kaplan-Meier estimator of C (resp. D).

Summary

- Work directly on the observed recurrent process by modelling its hazard rate conditionally on being at risk.
- The main assumption is that C is independent of the recurrent events occurrences and of the terminal event (conditionally on the covariates).
- The terminal event is treated as a competing risk situation. Special care should be taken when dealing with marginal features such as expected number of recurrent events.
- Regression models such as Cox can deal with external time varying covariates.
- Almost always use the robust variance estimator when dealing with recurrent events. This is done through the **coxph** function with the **cluster** option.

Contents

1. Modelling the hazard rate and basic estimation techniques

2. Analysis of the atrial fibrillation dataset

3. Extended models

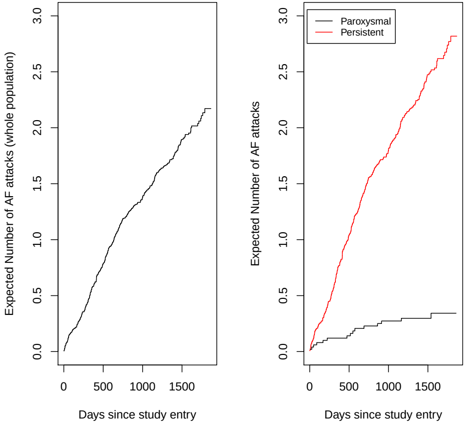
Some descriptive statistics

Variable	Levels	Value
AF type	paroxysmal	50(28.6)
	persistent	125(71.4)
gender	male	125(71.4)
	female	50(28.6)
age	median iqr	63.0{52.5, 68.0}
alcohol	0 − 5	93(56.4)
	5+	72(43.6)
	missing	10
tobacco	never	88(53.3)
	ex smoker	46(27.9)
	smoking	31(18.8)
	missing	10
hypertension	yes	82(46.9)
	no	93(53.1)
heart failure	yes	14(8.0)
	no	161(92.0)
heart valv dis	yes	12(6.9)
	no	163(93.1)
isch heart dis	yes	23(13.1)
	no	152(86.9)
diabetes	no	151(86.3)
	yes	24(13.7)
copd	yes	11(6.3)
	no	164(93.7)

- Number of patients : 175.
- 45 terminal events.
- 130 censored patients.
- Total number of observed recurrent events : 326.

AF nb	Freq
0	90
1	30
2	11
3	9
4	9
5	3
6	3
7	9
8	4
9	2
11	2
12	1
14	1
17	1

Estimation of the cumulative mean function



- One AF attack after 635 days on average.
- Two AF attacks after 1 613 days on average.
- One AF attack after 485 days on average for persistent patients.
- Two AF attacks after 1 146 days on average for persistent patients.

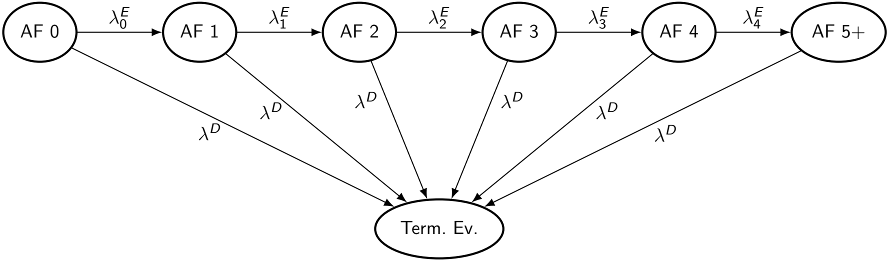
Cox analysis for the Atrial Fibrillation database

	Hazard ratio	2.5 %	97.5 %	se	robust se	p-value
AF type (persistent)	8.46	4.83	14.79	0.27	0.29	0.0000
gender (female)	1.00	0.63	1.58	0.13	0.23	0.9968
age	0.98	0.97	1.00	0.01	0.01	0.0392
hypertension (no)	0.83	0.51	1.36	0.12	0.25	0.4621
heart fail. (no)	0.96	0.44	2.10	0.26	0.40	0.9153
heart valv. dis. (no)	0.78	0.48	1.28	0.20	0.25	0.3235
isch. heart dis. (no)	1.22	0.45	3.28	0.23	0.51	0.6997
diabetes (yes)	0.25	0.09	0.73	0.27	0.54	0.0113
copd (no)	0.94	0.50	1.75	0.26	0.32	0.8412
alcohol (5+)	0.62	0.38	1.01	0.12	0.25	0.0545

- se : squared error obtained from the Poisson model.
- robust se : robust squared error obtained without assuming independent increments.

Multistate model

Cook R. J. and Lawless J. F. (2007) :



Goal : knowing the current state for a patient,

- Estimate the probability of experiencing further recurrent events.
- Estimate the probability of experiencing a terminal event.
- Estimate the probability of experiencing either further recurrent events or a terminal event.

Multistate model

Suppose for instance that patient i has experienced 1 recurrent event so far. We want to compute the probability that this patient experience further AF episodes in the future.

- $N_{i12}(t)$ indicates whether a transition from 1 to 2 recurrent events occurred for subject i over $[0, t]$.
- $N_i^D(t)$ indicates whether a transition from 1 recurrent event to terminal event occurred for subject i over $[0, t]$.

$$\mathbb{E}[dN_{i12}^E(t)|Y_{i1}(t), X] = Y_{i1}(t)\lambda_1^E(t|X)dt$$
$$\mathbb{E}[dN_i^D(t)|Y_{i1}(t), X] = Y_{i1}(t)\lambda^D(t|X)dt$$

where $Y_{i1}(t) = I(0 < E_{i1} < t < D, t < E_{i2})$ and $E_{i1}, E_{i2}, \dots$ represent the successive recurrent events for individual i.

Multistate model

- At time s patient i had already experienced one recurrent event (and only one) and was still alive and not in permanent AF state.
- The probability for this patient to experience a new recurrent event in the time interval $[s, t]$ is

Théorème

$$\mathbb{E}[N^E_{i12}(t) - N^E_{i12}(s) | 0 < E_{i1} < s < D, s < E_{i2}, X]$$
$$= \int_s^t \exp\{-\int_s^u \lambda^E_1(v|X)dv\} \exp\{-\int_s^u \lambda^D(v|X)dv\} \lambda^E_1(u|X)du.$$

Multistate model

- At time s patient i had already experienced one recurrent event (and only one) and was still alive and not in permanent AF state.
- The probability for this patient to experience a new recurrent event or a terminal event in the time interval $[s, t]$ is

Théorème

$$\mathbb{P}[D > t, E_{i2} > t | 0 < E_{i1} < s < D, s < E_{i2}, X] \\ = \exp\{-\int_s^t \lambda_1^E(u|X)du\} \exp\{-\int_s^t \lambda^D(u|X)du\}.$$

Multistate model

- At time s patient i had already experienced one recurrent event (and only one) and was still alive and not in permanent AF state.
- The probability for this patient to experience a terminal event in the time interval $[s, t]$ is

Théorème

$$\mathbb{P}[D > t | 0 < E_{i1} < s < D, s < E_{i2}, X]$$
$$= \exp\{-\int_s^t \lambda^D(u|X)du\}.$$

Modelization of the transition intensities

- The effect of the covariates is the same for each transition intensity.
- The transition intensities are assumed proportional with respect to the time to each other :

$$\lambda_s^E(t|X) = \lambda_0(t)\exp(\theta_0^T X + \beta_s), \quad \text{with } \beta_0 = 0.$$

In particular, we have, for a given covariate value x :

$$\frac{\lambda_s^E(t|X=x)}{\lambda_{s'}^E(t|X=x)} = \exp(\beta_s - \beta_{s'}).$$

Cox analysis for the AF episodes

Multistate model v.s. standard model

	Hazard ratio	2.5 %	97.5 %	robust se	p-value	p*
AF type (persistent)	4.46	2.90	6.87	0.22	0.0000	0.0000
gender (female)	1.12	0.83	1.50	0.15	0.4646	0.9968
age	0.99	0.97	1.00	0.01	0.0498	0.0392
hypertension (no)	0.84	0.60	1.19	0.18	0.3329	0.4621
heart fail. (no)	0.90	0.52	1.54	0.27	0.6942	0.9153
heart valv. dis. (no)	1.12	0.82	1.54	0.16	0.4790	0.3235
isch. heart dis. (no)	0.93	0.50	1.73	0.32	0.8171	0.6997
diabetes (yes)	0.49	0.21	1.14	0.43	0.0985	0.0113
copd (no)	0.96	0.66	1.38	0.19	0.8150	0.8412
alcohol (5+)	0.80	0.56	1.15	0.19	0.2368	0.0545
AF 1	2.44	1.64	3.63	0.20	0.0000	
AF 2	5.95	3.65	9.72	0.25	0.0000	
AF 3	4.86	2.77	8.52	0.29	0.0000	
AF 4	6.50	3.62	11.67	0.30	0.0000	
AF 5+	8.07	4.88	13.35	0.26	0.0000	

- The p* are the p-values obtained from the previous model, without adjustement with respect to the number of previous AF episodes.

Cox analysis for the terminal event

	Hazard ratio	2.5 %	97.5 %	se	p-value
AF type (persistent)	1.01	0.50	2.05	0.36	0.9741
gender (female)	1.10	0.55	2.20	0.35	0.7867
age	1.05	1.02	1.08	0.01	0.0017
hypertension (no)	0.71	0.37	1.36	0.33	0.2994
heart.fail (no)	0.55	0.21	1.46	0.49	0.2303
heart.valv.dis (no)	1.50	0.33	6.81	0.77	0.6024
isch.heart.dis (no)	1.22	0.51	2.94	0.45	0.6524
diabetes (yes)	1.00	0.42	2.36	0.44	0.9932
copd (no)	0.46	0.16	1.29	0.53	0.1407
alcohol (5+)	1.36	0.72	2.55	0.32	0.3394

Final models for prediction

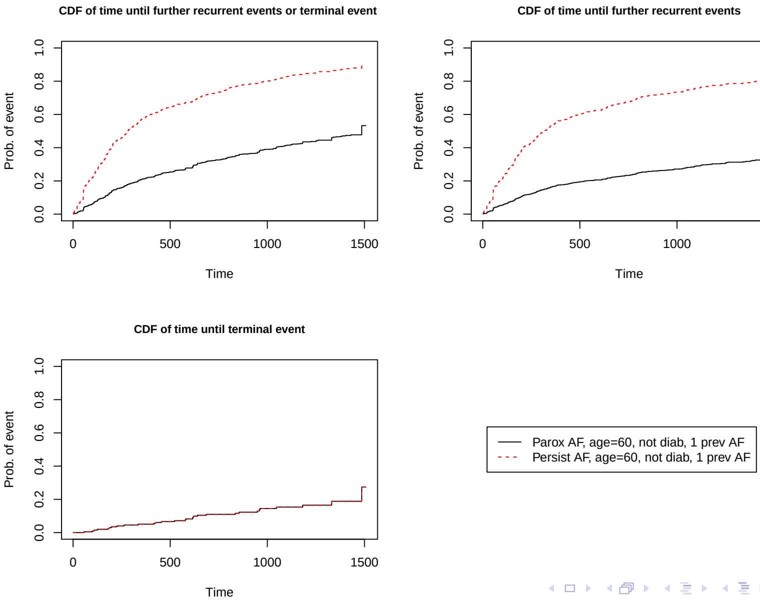
	Hazard ratio	2.5 %	97.5 %	robust se	p-value
AF type (persistent)	4.33	2.71	6.92	0.24	0.0000
age	0.99	0.98	1.00	0.01	0.0878
diabetes (yes)	0.53	0.23	1.26	0.44	0.1498
AF 1	2.55	1.73	3.76	0.20	0.0000
AF 2	6.05	3.89	9.42	0.23	0.0000
AF 3	5.07	3.03	8.49	0.26	0.0000
AF 4	7.05	3.96	12.55	0.29	0.0000
AF 5+	8.17	5.27	12.68	0.22	0.0000

For the recurrent event process

	Hazard ratio	2.5 %	97.5 %	se	p-value
AF type (persistent)	1.00	0.51	1.93	0.34	0.99
age	1.05	1.03	1.08	0.01	0.00
diabetes (yes)	1.17	0.52	2.66	0.42	0.71

For the terminal event

Prediction curves for $s = 240$ (8 months)

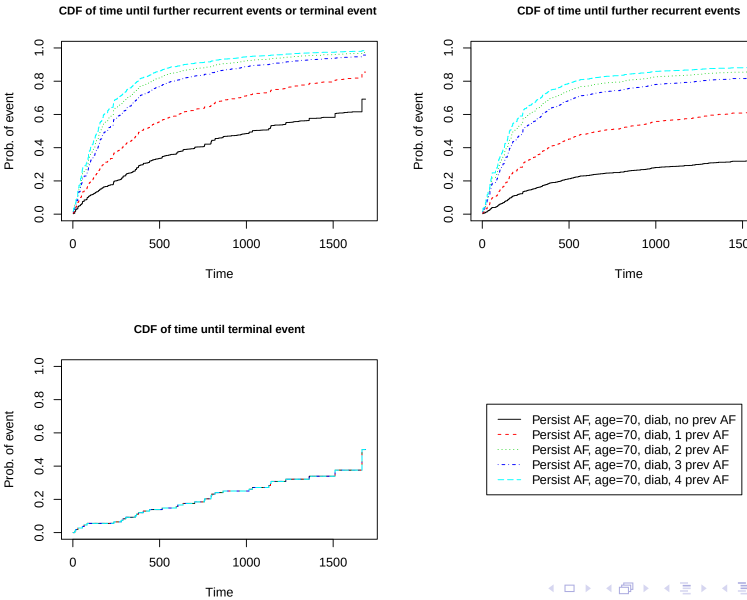


CDF of time until further recurrent events or terminal event

CDF of time until further recurrent events

CDF of time until terminal event

Prediction curves for $s = 180$ (3 months)



Contents

1. Modelling the hazard rate and basic estimation techniques
2. Analysis of the atrial fibrillation dataset
3. Extended models

Modelization of the transition intensities : the PWP model

Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981) proposed a model where :

- The effect of the covariates changes for each transition intensity,
- The baseline changes for each transition intensity.

Let $s = 0, 1, \ldots, B-1$ represent all states in the multistate model.

$$\lambda_s^E(t|X) = \lambda_0(t, s) \exp(\theta_0(s)^T X).$$

This model is overparametrized : with $B = 5$ and 11 covariates, $11 \times 5 = 55$ parameters to estimate !

The Cox partial likelihood in the PWP model

Let $t_{(1)} < t_{(2)} < \cdots < t_{(H)}$ denote the H unique observed ordered event times. The regression parameter is estimated through the stratified Cox partial likelihood :

$$L_n(\theta) = \prod_{h=1}^{H} \prod_{i=1}^{n} \prod_{s=0}^{B-1} \left(\frac{\exp(\theta(s)^T X_i(t_h))}{\sum_{j=1}^{n} Y_j^s(t_h) \exp(\theta(s)^T X_j(t_h))} \right)^{Y_i^s(t_h) dN_i(t_h)},$$

where $Y_j^s(t) = I(T_j \geq t, N_j(t-) = s)$ is the observed at-risk process in strata s.

$$\hat{\theta} = \text{argmax}_\theta L_n(\theta).$$

A penalized version of the PWP model

Bouaziz, O. and Guilloux, A (2015) proposed to penalise the log-likelihood.

Let $\theta_0 = (\theta_0^1(1), \dots, \theta_0^1(B), \dots, \theta_0^p(1), \dots, \theta_0^p(B))^\top$.

- Constrain the total-variation of the $\theta^j(s)$ to be “small”

$$\hat{\theta}_{TV} = \underset{\theta \in \mathbb{R}^{p \times B}}{\operatorname{argmin}} \left\{ -\log(L_n(\theta)) + \frac{\lambda_n}{n} \sum_{j=1}^{p} \sum_{s=1}^{B-1} \left| \theta^j(s) - \theta^j(s-1) \right| \right\}.$$

- If $\lambda_n = 0$, $\hat{\theta}_{TV}$ is the PWP estimator.
- If $\lambda_n / n = \infty$, $\hat{\theta}_{TV}$ is the classical estimator from the Andersen-Gill model (same $\hat{\theta}^j$ for all s).

Implementation of the penalized estimator

- The estimator can be rewritten as a Lasso estimator.
- Implementation through the **coxnet** function (R package **glmnet**).
- Regularization parameter λ_n chosen via 5-fold cross-validation, see Simon et al. (2011) or van Houwelingen et al. (2006).
- Programs available on my webpage !

Results from the event-specific penalized estimator

Hazard ratios for each covariate with respect to previous recurrent events experienced by a patient
($s = 0$: no recurrent events so far, $s = 1$: 1 recurrent event so far $\dots$).

	Hazard ratio	Hazard ratio*
AF type (persistent), all s	4.06	4.33
age, all s	1.00	0.99
diabetes (yes), $s = 0$	0.24	0.53
diabetes (yes), $s = 1, 2, 3, 4$	0.77	0.53

* HR from the previous model used for prediction

- AF type and diabetes seem to have the same effect with respect to previous number of recurrent events experienced by a patient.
- The data indicate a protective effect of diabetes for the first recurrent event. Then hazard ratio is close to 1.

A simple model with interaction to sum-up

- diab0 : interaction term of diabetes and $N(t-) = 0$ (no recurrent event so far).
- diab1 : interaction term of diabetes and $N(t-) \geq 1$ (one or more recurrent events so far).

	Hazard ratio	2.5 %	97.5 %	robust se	p-value
AF type (persistent)	4.41	2.74	7.09	0.24	0.0000
age	0.99	0.98	1.00	0.00	0.1221
diab0 (yes)	0.24	0.09	0.59	0.47	0.0021
diab1 (yes)	1.23	0.83	1.81	0.20	0.3015
AF 1	2.26	1.53	3.32	0.20	0.0000
AF 2	5.58	3.64	8.56	0.22	0.0000
AF 3	4.57	2.76	7.54	0.26	0.0000
AF 4	6.44	3.68	11.25	0.28	0.0000
AF 5+	7.43	4.86	11.37	0.22	0.0000

- Be careful with the interpretation of the diabetes effect : only 15 observed recurrent events for the diabetics !

Extension : frailty models

- A model that accounts for individual heterogeneity through a random effect $b_i \sim \mathcal{N}(0, \sigma^2)$:

$$\mathbb{E}[d\tilde{N}_i(t)|X_i(t), b_i, \tilde{Y}_i(t)] = \tilde{Y}_i(t)\lambda_0(t)\exp(\theta_0^T X_i(t) + b_i)dt$$

- Estimates obtained from the integrated partial likelihood, see Ripatti and Palmgren (2000).
- Implemented through the **coxme** package :

	HR	HR*	HR**	p-value	p-value*	p-value**
AF type (persistent)	7.48	8.68	4.33	0.0000	0.0000	0.0000
age	0.98	0.99	0.99	0.0470	0.0379	0.0878
diabetes (yes)	0.26	0.27	0.53	0.0021	0.0189	0.1498

* Without frailty
** Without frailty but stratified w.r.t. to previous number of recurrent events
Estimated variance of the frailty : $\hat{\sigma}^2 = 1.28$.

Extension : joint modeling of the terminal event and recurrent event process

- N_i^E : recurrent event process for individual i.
- N_i^D : terminal event process for individual i.
- $\tilde{Y}_i(t) = I(D_i \geq t)$: at risk-process for both processes of interest.
- b_i : shared random effect for both processes of interest.

$$\begin{cases} \mathbb{E}[dN_i^E(t)|\tilde{Y}_i(t), X] = \tilde{Y}_i(t)\lambda_0^E(t)\exp(\theta_0^T X_i(t) + b_i)dt \\ \mathbb{E}[dN_i^D(t)|\tilde{Y}_i(t), X] = \tilde{Y}_i(t)\lambda_0^D(t)\exp(\beta_0^T X_i'(t) + {b_i}^{\alpha})dt \end{cases}$$

Zeng, D. and Lin, D. Y. (2009) : matlab codes.
Rondeau, V., Mazroui, Y. and Gonzalez, J. R. (2012) : R package **frailtypack**.

Bibliography

[1] Per K Andersen and R. D. Gill. Cox's regression model for counting processes : a large sample study. *The Annals of Statistics*, 10(4) :1100–1120, 1982.

[2] O. Bouaziz and A. Guilloux. A penalized algorithm for event-specific rate models for recurrent events. *Biostatistics*, 16(2) :281–294, 2015.

[3] R. J. Cook and J. Lawless. *The statistical analysis of recurrent events*. Springer, 2007.

[4] D. Y. Lin, L. J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Society B*, 62(4) :711–730, 2000.

[5] R. L. Prentice, B. J. Williams, and A. V. Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2) :373–379, 1981.