Analysis of competing risks data and simulation of data following predefined subdistribution hazards

Bernhard Haller

Institut für Medizinische Statistik und Epidemiologie
Technische Universität München

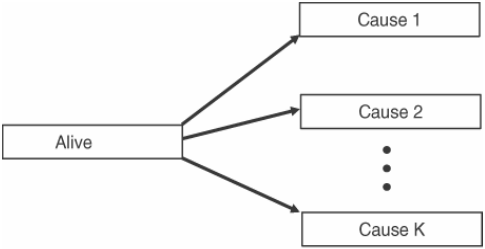
27.05.2013
Research Seminar

Table of contents

Table of contents

Introduction

- Time-to-event analysis
- Subjects can fail from one out of K mutually exclusive types of event
- Often relevant in clinical studies:
  Primary endpoint: Time to cancer-specific death / cardiac death / ...
- Special methods have to be conducted



Basic quantities

- Basic quantities:
  - Cumulative incidence function (CIF):
    $$F_k(t) = P(T \leq t, D = k) = \int_0^t \lambda_k(s) S(s-) \mathrm{d}s$$
  - Cause-specific hazard (CSH):
    $$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, K = k | T \geq t)}{\Delta t}$$
  - Overall survivor function $S(t)$:
    $$S(t) = \exp\big(-\sum_{l=1}^{K} \Lambda_l(t) \big)$$
- Different methods available for the analysis of competing risks data
- The "naïve" Kaplan-Meier estimator gives a biased estimate for the probability of an event of type k up to time t

Table of contents

Cause-specific hazards regression

Regression based on the cause-specific hazards (Prentice et al. 1978):

- Focus on cause-specific hazard rates using e.g. a Cox-type regression model:

$$\lambda_k(t|\mathbf{X}) = \lambda_{k,0}(t) exp(\boldsymbol{\beta}_k^\top \mathbf{X})$$

- Can be performed using standard Cox-regression software treating competing events as censored observations
- Estimated CIFs depend on CSHs for all event types

$$F_k(t|\mathbf{X}) = \int_0^t \lambda_k(s|\mathbf{X}) \, \exp\Big(- \sum_{l=1}^{K} \Lambda_l(t|\mathbf{X}) \mathrm{d}s \Big)$$

- CSHs completely determine the competing risks process
- Higher CSH for an event k does not necessarily translate into a higher event probability for k

The subdistribution hazard

- Introduced by Gray (1988)
- Aim of the subdistribution hazard:
  A "hazard function" that is directly linked to the CIF in the presence of competing risks
- Definition of the subdistribution hazard (SDH) for event k:

$$\gamma_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k | T \geq t \cup \{T < t, D \neq k\})}{\Delta t}.$$

- Subjects failing from an event $D \neq k$ remain in the risk set (until their potential censoring time).
- For the SDH (as known from standard survival analysis):

$$F_k(t|\mathbf{X}) = 1 - \exp\big(-\Gamma_k(t|\mathbf{X})\big)$$

- Competing events are considered implicitly in the adapted risk set

Subdistribution hazard regession

Introduced by Fine & Gray (1999)

- Focusing on the SDH for the event of interest
- Individuals failing from an event $D \neq k$ remain in the risk set until their potential censoring time
- Censoring time distribution is estimated from the censored observations
- Competing events are weighted using the inverse probability of censoring weighting (IPCW) approach
- A Cox-type regression model was proposed for the SDH:

$$\gamma_k(t|\mathbf{X}) = \gamma_{k;0}(t)\ \exp(\beta_k^{*\top}\mathbf{X})$$

- Proportionality assumption often questionable in practice.

The mixture model approach

Alternative regression approach - introduced by Larson and Dinse (1985)

- $P(T, D) = P(D)\, P(T|D)$
- Proposed: Logistic regression for type of event, parametric model for conditional event time distributions
  - Larson & Dinse (1985): piecewise-exponential
  - Lau et al. (2011): generalized gamma distribution
- Likelihood contribution of subject i:

$$L_i = [\pi_i f_1(t_i)]^{I(d_i=1)} \times [(1-\pi_i) f_2(t_i)]^{I(d_i=2)}$$
$$\times [\pi_i S_1(t_i) + (1-\pi_i) S_2(t_i)]^{I(d_i=0)}$$

  with $f_k(t)$ and $S_k(t)$ denoting quantities of the cond. event time distributions
- Numerical maximization to determine ML-estimates

Table of contents

Simulation following predefined cause-specific hazards

Beyersmann et al. (2009), example for two possible event types:

- Define CSHs depending on covariates: $\lambda_1(t|\mathbf{X})$, $\lambda_2(t|\mathbf{X})$
- Determine the overall hazard rate: $\lambda(t|\mathbf{X}) = \lambda_1(t|\mathbf{X}) + \lambda_2(t|\mathbf{X})$.
- Generate an event time t_i for subject i with hazard rate $\lambda(t|\mathbf{x_i})$.
- Determine the event type d_i by running a Bernoulli experiment
  - $P(D_i = 1) = \lambda_1(t|\mathbf{x_i}) \,/\, \big(\lambda_1(t|\mathbf{x_i}) + \lambda_2(t|\mathbf{x_i})\big)$
  - $P(D_i = 2) = \lambda_2(t|\mathbf{x_i}) \,/\, \big(\lambda_1(t|\mathbf{x_i}) + \lambda_2(t|\mathbf{x_i})\big)$
- Draw possible censoring times from a censoring time distribution and determine event time and status accordingly.

Used in several research articles for investigation of competing risks methods.

Simulation following predefined subdistribution hazards

- Different approaches focusing on the SDHs were introduced
- Simulation mainly conducted using unit exponential mixture distributions
- Simulation using flexible prespecified subdistribution hazards not possible

Idea by Beyersmann et al. (2009):

- Use relationship between CSH and SDH:

$$\lambda_1(t|\mathbf{X}) = \gamma_1(t|\mathbf{X})\left(1 + \frac{F_2(t|\mathbf{X})}{S(t|\mathbf{X})}\right) \tag{1}$$

- Specify SDH for event of interest and one CSH
- Calculate other CSH following (1)
- Simulate event times using the CSHs to obtain data following the prespecified SDHs

Simulation following predefined subdistribution hazards

Problems:

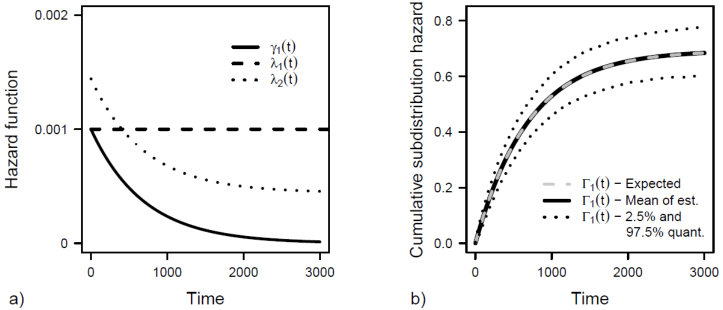
- Certain constraints on different quantities
  - All hazard functions have to be non-negative for all time points $t > 0$.
  - $\lim_{t \to \infty} F_k(t|\mathbf{X}) < 1 \quad \Leftrightarrow \quad \lim_{t \to \infty} \gamma_k(t|\mathbf{X}) = 0$
  - $\lim_{t \to 0} \lambda_k(t|\mathbf{X}) = \lim_{t \to 0} \gamma_k(t|\mathbf{X})$
- Simulation following time-varying CSHs needed

Simulation following predefined subdistribution hazards

Example:

- $\gamma_1(t) = 0.001 \ \exp\left(-\frac{0.001\,t}{\ln(2)}\right)$
- $\lambda_1(t) = 0.001$
- $\lambda_2(t|\mathbf{X}) = \gamma_1(t|\mathbf{X}) - \lambda_1(t|\mathbf{X}) - \frac{\mathrm{d}}{\mathrm{d}t} \ \ln\left(\frac{\gamma_1(t|\mathbf{X})}{\lambda_1(t|\mathbf{X})}\right)$

$$= 0.001 \ \exp\left(-\frac{0.001\,t}{\ln(2)}\right) - 0.001 + \frac{0.001}{\ln(2)}$$



Two methods for data generation (I)

Aim: Generate event times with hazard rate $\lambda(t|\mathbf{X}) = \lambda_1(t|\mathbf{X}) + \lambda_2(t|\mathbf{X})$

- Inversion method (see e.g. Bender et. al (2005))
- $U = exp\big(-\Lambda(t|\mathbf{X})\big) \qquad \Leftrightarrow \qquad T = \Lambda^{-1}(-ln\, U)$
  - $U \sim U[0,1]$
  - $\Lambda^{-1}(z)$ is the inverse function of the cumulative overall hazard function
- In general, numerical procedures for
  - the cumulative overall hazard function
  - the solution of $U = exp\big(-\Lambda(t|\mathbf{X})\big)$
- Can become very time-consuming

Two methods for data generation (II)

Based on binomial algorithm (by Sylvestre & Abrahamowicz (2008))

- Event time generation for discrete timepoints
- Start with subject $i = 1$
- Begin at time $t_j = 1$
- Prob. for any event at time t_j for subject i: $p(t_j|\mathbf{x_i}) = \lambda_1(t_j|\mathbf{x_i}) + \lambda_2(t_j|\mathbf{x_i})$
- Perform Bernoulli experiment to determine whether i failed at t_j
- Event at t_j
  - Determine type of event
  - Continue for subject i+1
- No event at t_j:
  - Continue for timepoint $t_j + 1$

The Binomial approach

- Investigated for different scenarios
  - One group
  - Two groups, constant SD hazard ratio
  - Two groups, time-varying SD hazard ratio
  - One quantitative covariate
  - Multiple covariates (SDH regression model)
- Established methods were used to analyse the generated data
- Good behaviour of the data generating process
- Can lead to bindings in event times
- Amount of binding can be controlled by choice of hazard functions
- Published in:
  Haller B, Ulm K (2013) Flexible simulation of competing risks data following prespecified subdistribution hazards. Journal of Statistical Computation and Simulation. doi:10.1080/00949655.2013.793345.

Table of contents

The mixture model - notation (following Lau et. al, 2011)

- $P(T, D|\mathbf{X}) = P(D|\mathbf{X})\ P(T|D, \mathbf{X})$
- Probability for an event of type k: $\pi_k(\mathbf{X}) = P(D = k|\mathbf{X})$
- Density function of the conditional event-time distribution: $f_k(t|D = k, \mathbf{X}))$
- Cum. density fct. of the cond. event-time distribution: $F_k(t|D = k, \mathbf{X}))$
- Survivor function of the conditional event-time distribution: $S_k(t|D = k, \mathbf{X})$

- Subdensity function: $f_k^*(t|\mathbf{X}) = f_k(t|D = k, \mathbf{X})\,\pi_k(\mathbf{X})$
- Subdistribution function: $F_k^*(t|\mathbf{X}) = F_k(t|D = k, \mathbf{X})\,\pi_k(\mathbf{X})$

The mixture model - relationships

Following Lau et al. (2011)

- Cause-specific hazard function: $\lambda_k(t|\mathbf{X})$
- Subdistribution hazard function: $\gamma_k(t|\mathbf{X})$
- Overall survival function: $S(t|\mathbf{X})$

$$S(t|\mathbf{X}) = exp\big(- \sum_{l=1}^{K} \Lambda_l(t|\mathbf{X})\big) = \sum_{l=1}^{K} \pi_k(\mathbf{X})\, S_k(t|D=k,\mathbf{X}))$$

- Cumulative incidence function: $F_k^*(t|\mathbf{X})$

$$F_k^*(t|\mathbf{X}) = \int_0^t \lambda_k(s|\mathbf{X})\ \exp\Big(- \sum_{l=1}^{K} \Lambda_l(t|\mathbf{X})\mathrm{d}s\Big)$$

$$F_k^*(t|\mathbf{X}) = 1 - \exp\big(- \Gamma_k(t|\mathbf{X})\big)$$

$$F_k^*(t|\mathbf{X}) = F_k(t|D=k,\mathbf{X})\,\pi_k(\mathbf{X})$$

Estimating CSH and SDH from a mixture model

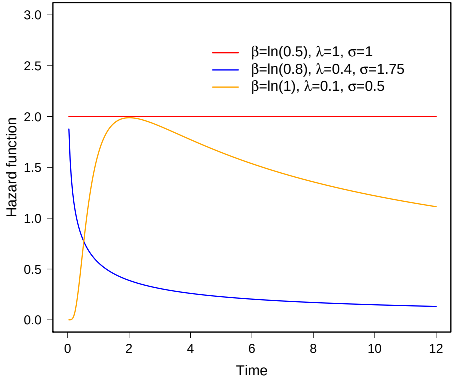
- CSH: $\lambda_k(t|\mathbf{X}) = f_k^*(t|\mathbf{X})/S(t|\mathbf{X})$
- SDH: $\gamma_k(t|\mathbf{X}) = f_k^*(t|\mathbf{X})/\Big(1 - F_k^*(t|\mathbf{X})\Big)$

Proposal by Lau et al:

- Use a generalized gamma distribution for conditional event times
- Allows flexible estimation of CSHs and SDHs

$$f(t) = \frac{|\lambda|}{\sigma t \Gamma(\lambda^{-2})} \left(\lambda^{-2}(e^{-\beta}t)^{\lambda/\sigma}\right)^{\lambda^{-2}} \exp\left(-\lambda^{-2}(e^{-\beta}t)^{\lambda/\sigma}\right)$$

The generalized gamma distribution



Problems / issues:

- Numerical instabilities
- Weighting of extreme observations
- Are all relevant forms covered?

Alternative approach: penalized B-splines

- Define set of basis functions $B_k(t)$, e.g. cubic splines
- The hazard function can be modelled using $B_k(t)$, e.g. (Rosenberg, 1995):

$$h(t|\boldsymbol{\theta}) = \sum_{k=-3}^{K} B_k(t) \exp(\boldsymbol{\theta})$$

- Roughness of the estimated hazard function will be penalized using
  - a smoothing parameter λ
  - a matrix D, e.g. second order differences

$$\mathbf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -2 & 1 & 0 \\ 0 & \dots & \dots & 0 & 1 & -2 & 1 \end{pmatrix}$$

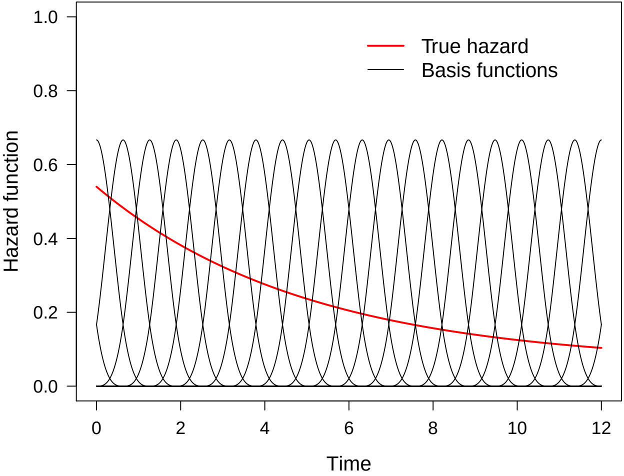
- Maximize penalized Log-Likelihood:

$$l_{pen} = l(\boldsymbol{\theta}; t, d, B) - \tfrac{1}{2}\,\lambda\,\boldsymbol{\theta}^\top \mathbf{D_k}^\top \mathbf{D_k}\,\boldsymbol{\theta}$$

- Find ML estimates by numerical optimization

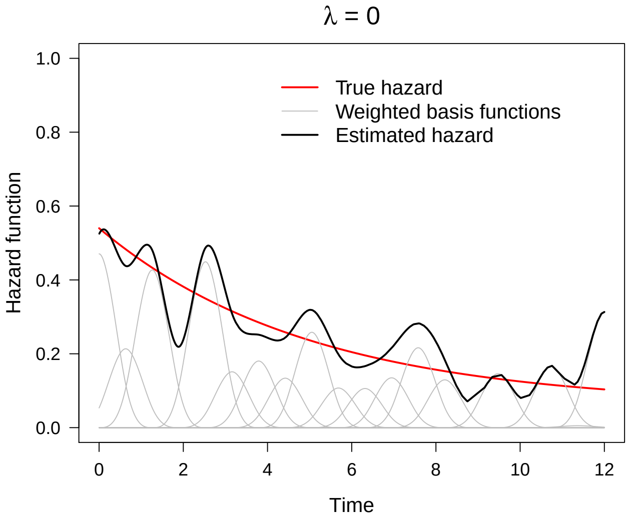
Possible approach

Illustration of P-spline approach for one possible endpoint



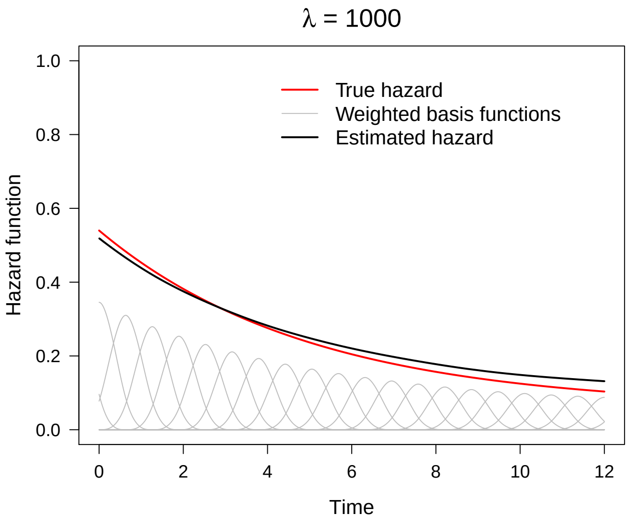
P-spline approach

Simulated example:



P-spline approach

Simulated example:



Adapting the P-Spline approach for the mixture model

- Different approaches available (hazard function, cumulative hazard function)
- Penalty matrix has to be adapted
- Implementation (*R*)
- Numerical maximization to find ML-estimates
  - Stable results?
  - Computation time?
- Confidence interval estimation for HRs (bootstrap)

Outlook

- Comparison of generalized gamma and P-spline approach
  - Real data examples::
    Data used by Lau et al. (2011) are available in *R*
  - Simulated data
    - Data generated according to mixture model approach
    - Predefined CSHs
    - Predefined SDHs
- Investigate roles of
  - Smoothing parameter
  - Number and placing of knots
  - Penalisation
  - Amount of censoring
  - ...
- Estimating "average hazard ratios" in adequate situations

References (I)

- Bender R, Augustin T, Blettner M (2005) Generating survival times to simulate Cox proportional hazards models. Stat Med 24: 1713-1723.
- Beyersmann J, Latouche A, BuchholzA, SchumacherM (2009) Simulating competing risks data in survival analysis. Stat Med 28(6):956–971.
- Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 94:496–509.
- Gray R (1988) A class of k-sample tests for comparing the cumulative incidence function in the presence of a competing risk. Ann Stat 16:1141–1154.
- Larson MG, Dinse GE (1985) A mixture model for the regression analysis of competing risks data. J R Stat Soc Ser C 34:201–211.

References (II)

- Lau B, Cole S, Gange S (2011) Parametric mixture models to evaluate and summarize hazard ratios in the presence of competing risks with time-dependent hazards and delayed entry. Stat Med 30:654–665.
- Prentice R, Kalbfleisch J, Peterson A, Flournoy N, Farewell V, Breslow N (1978) The analysis of failure times in the presence of competing risks. Biometrics 34:541–554.
- Rosenberg PS (1995) Hazard function estimation using B-Splines. Biometrics 51: 874-887.
- Sylvestre MP, Abrahamowicz M (2008) Comparison of algorithms to generate event times conditional on time-dependent covariates. Stat Med, 27(14):2618–2634.