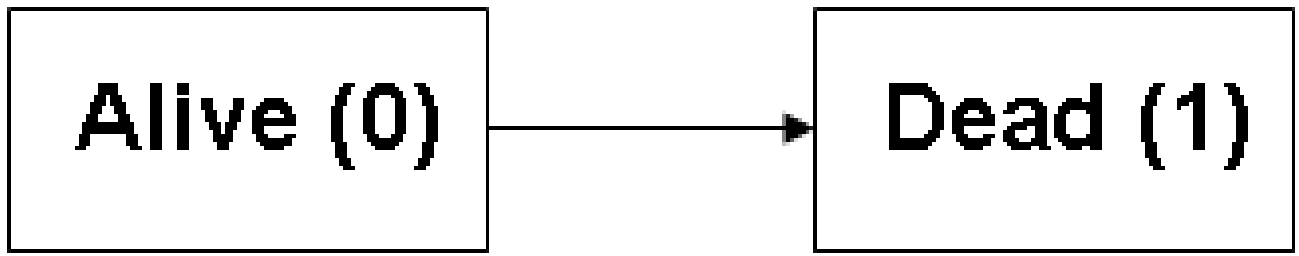
1. Event history analysis - Introduction

September 01 - 25, 2015

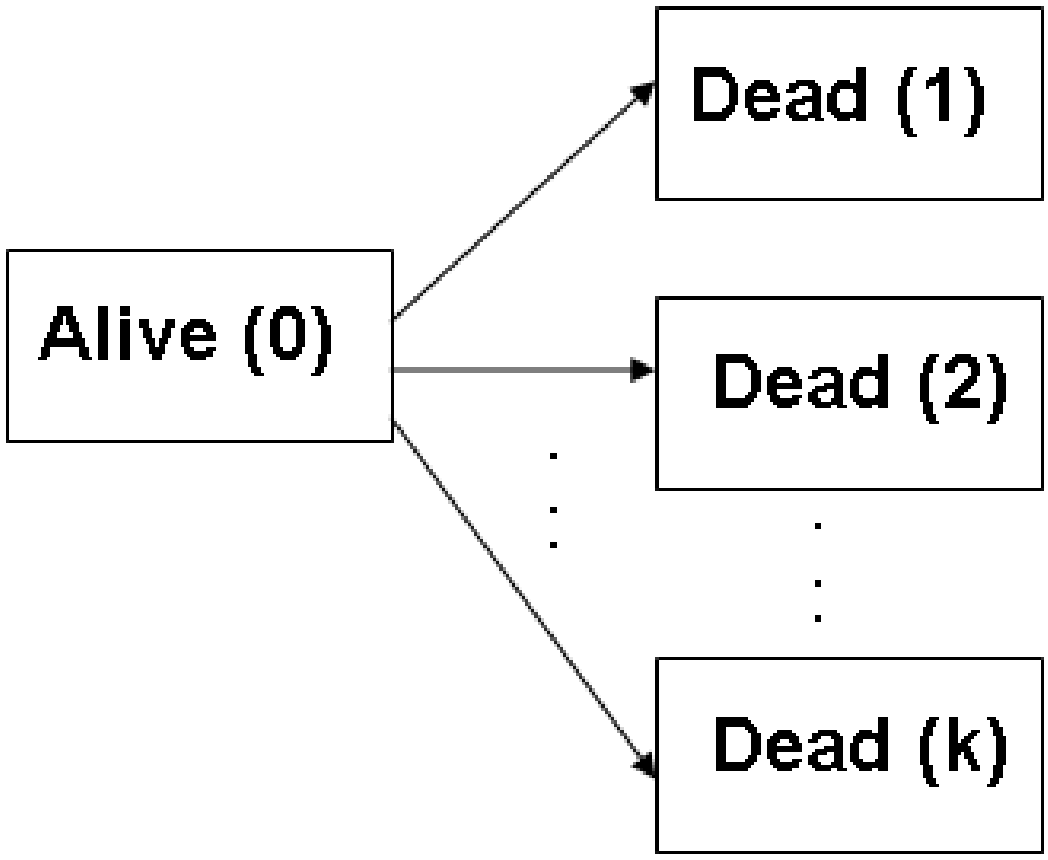
Course: Event history analysis

- A statistical methodology used in settings where one is interested in the occurrence of events.
- Examples: death, marriage, birth of a child, on-set of a disease
- Classical survival analysis: focuses on a single event for each individual and describes its occurrence using survival curves and hazard rates. Interest is mainly in understanding its causes or establishing risk factors.
- Event histories are generated by connecting together several events for an individual as they occur over time.

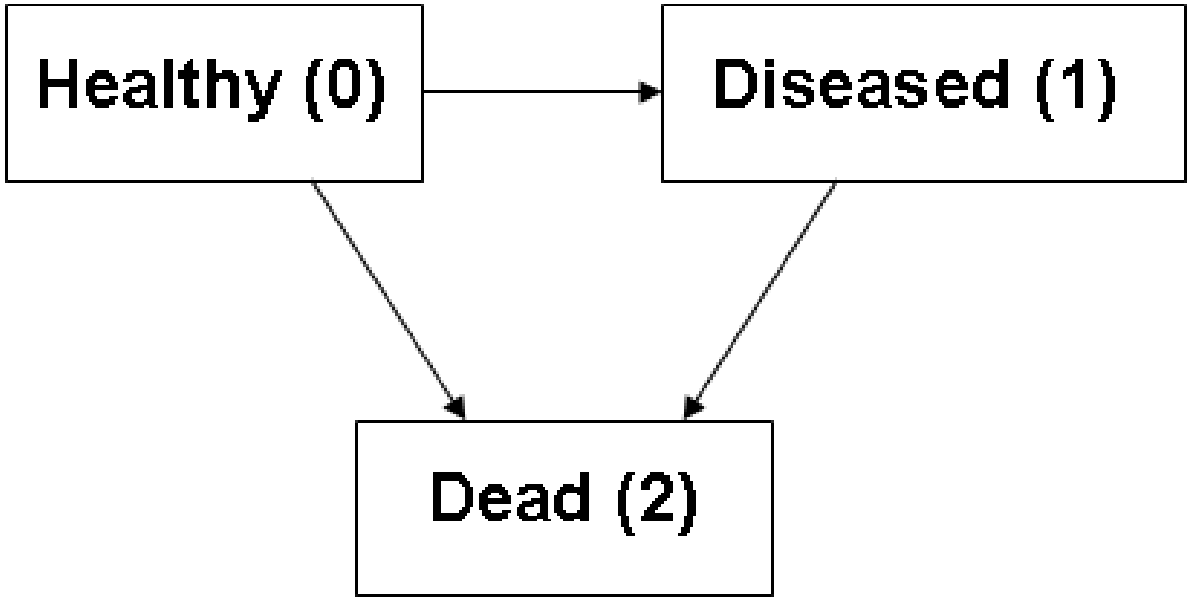
Example 1: Survival model



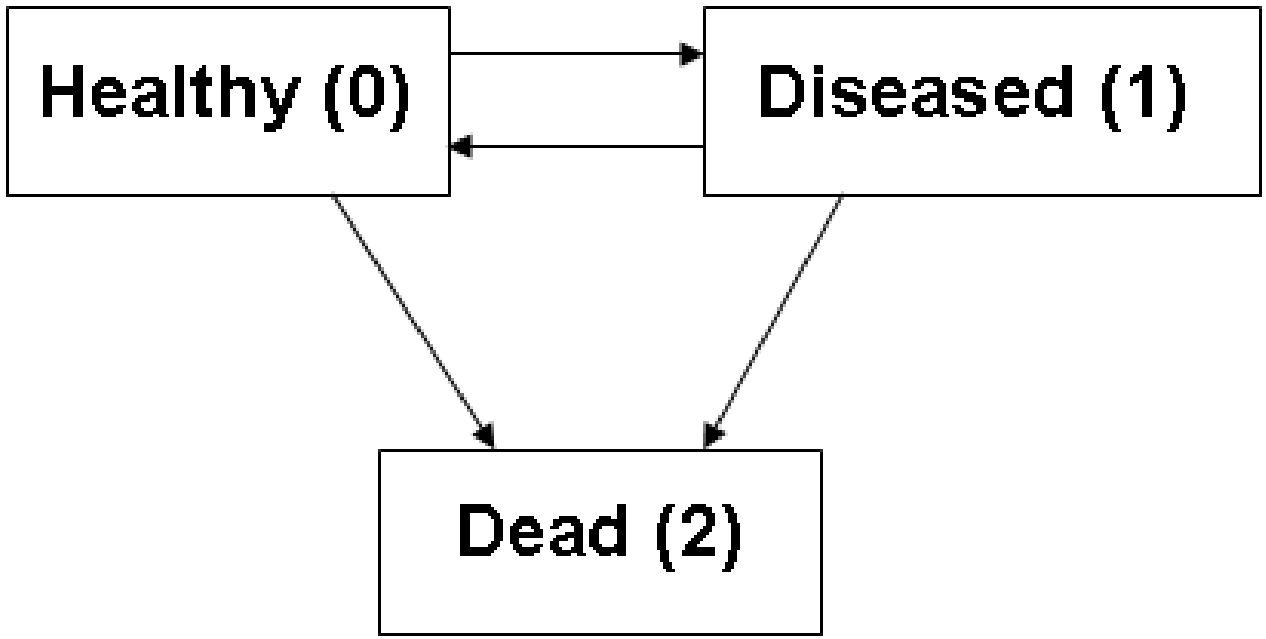
Example 2: Competing risks model



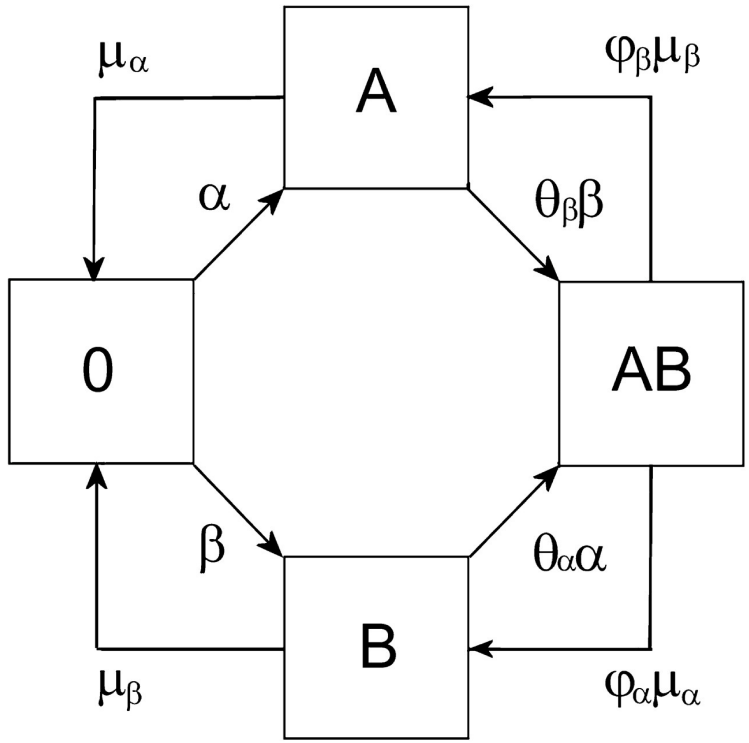
Example 3: Progressive illness-death



Example 4: Illness-death model



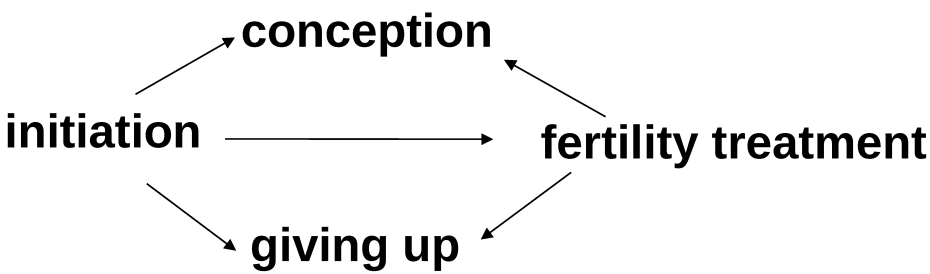
Example 5: Dynamics of infection



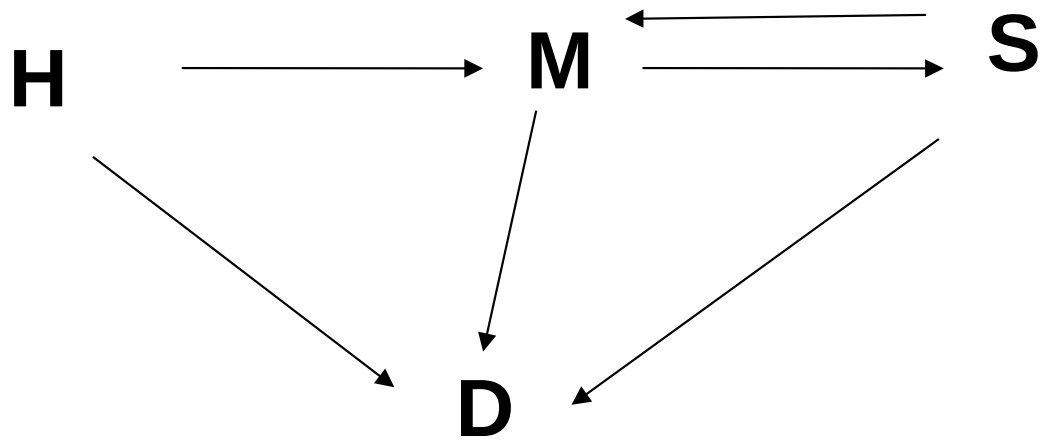
Auranen K et al. Am. J. Epidemiol. 2009;171:169-176

American Journal of
EPIDEMIOLOGY

Example 6: Multi-state model for time to pregnancy



Example 7: Event of marriage as a multi-state model



H = unmarried/single, M = married/cohabiting/union,
S = separated/widowed/divorced, D = dead

Multi-states and times

- Definitions of states?
- Which is an absorbing state (death)?
- How to enter a state?
- How to leave a state?
- Which transitions are possible?
- What are the transition times?

Event history analysis (1)

- Analysis or study of a cohort (group of individuals) each moving among a finite number of states.
- Analysis of life histories of individuals
- Exact transition times in continuous time form the modelling basis of the phenomenon.
- Applications in epidemiology, engineering, economics ....

Event history analysis (2)

Why study survival and event history analysis as a special methodology?

- One has to wait for the event to occur and at the end of the study, the event of interest might not have occurred for all.
- Often these times are only incompletely observed and are refer to as censored survival times.
- (Right) censoring -
  - migration and no coverage in the registers
  - death due to other causes before the end of the follow-up
  - fixed study period and hence, alive at the end of the follow-up
- (Left) truncation - unavailability of an individual for observation due to for example, death
- Data are often mixture of complete and incomplete observations.

Hypothetical clinical study

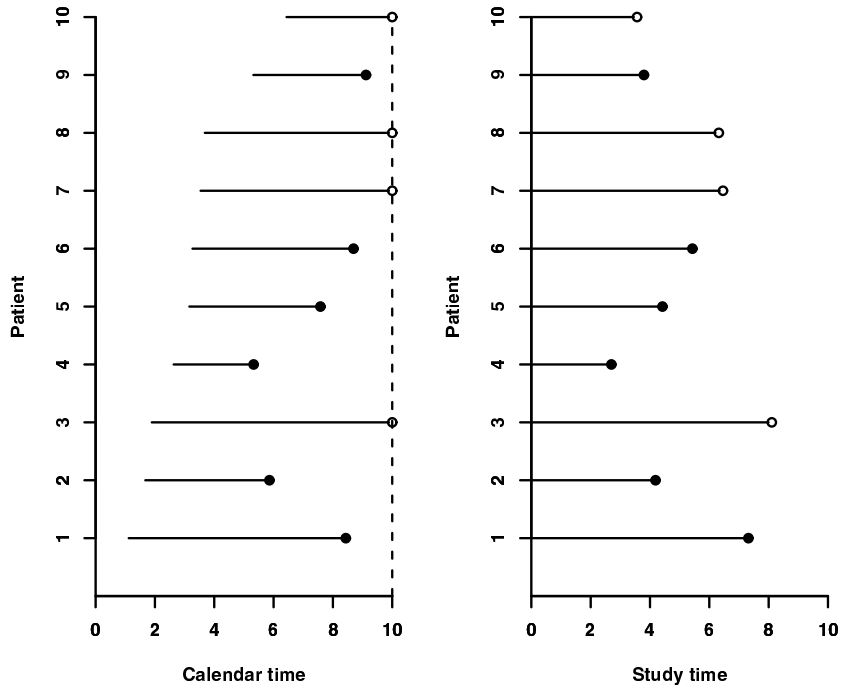


Fig. 1.1 *Patient follow up in a hypothetical clinical study with 10 patients. Left panel shows the actual calendar time. Right panel shows the same observations in the study time scale, where time 0 for each individual is his or her entry into the study. A filled circle indicates occurrence of the event, while an open circle indicates censoring. In the left panel the dotted vertical line indicates the closing date of the study.*

Emphasis on time

When do we start the clock? What should be the *study* time scale?

Many possible different time scales that may be used:

Starting point	Time scale
Birth	Age
Any fixed date	Calendar time
First exposure	Time exposed
Entry into the study	Time in study
Disease onset	Time since onset
Start of treatment	Time on treatment

Risk set at time t

The set of individuals for which the event of interest has not happened before a given time t, and who have not been censored before time t

Data: $\{7.32, 4.19, 8.11^*, 2.70, 4.42, 5.43, 6.46^*, 6.32^*, 3.80, 3.50^*\}$

In the example, at time 0 there were 10 individuals, at time 8 only 1 individual is left and at the end of the study no individuals are left.

Concept (1)

Two basic concepts: survival function and hazard rate

- Time (failure time): nonnegative random variable mostly (absolutely) continuous (X)
- Censoring time: period elapsed in which the event of interest has not occurred (C)
- Observed failure time: $T = min(X, C)$
- Censoring indicator: an indicator whether the event of interest has occurred or not ($d = 1$, if event observed and 0, otherwise)
- Survival (type) Data: (T, d) (incomplete data since failure time are missing for $d = 0$)

Concept (2)

NOTE: The censoring causes removal of a subject from observation, but after censoring (that is time C) the subject is still at risk of failure - a subject does not cease to run the risk of failure simply because he/she has ceased to participate in a follow-up study. The interpretation of censoring due to competing causes is different.

Concept (3)

- Survival function $S(t) = P(X > t), t > 0$, non-increasing right-continuous function of t with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$.
- Hazard rate, failure rate: $(\lambda(t))$ given that an individual has survived up to time t, the probability of dying in a short interval after t

$$\lambda(t) \quad = \quad \lim_{h\to 0} \frac{1}{h} P(t \leq X < t + h \mid X \geq t)$$

- Cumulative hazard rate: hazard rate integrated over an interval $[0, t)$

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Mathematical connections

Two fundamental mathematical conncetions between the survival function and the hazard rate.

- $$\Lambda'(t) = \lambda(t) \quad = \quad -\frac{S'(t)}{S(t)} = -\frac{d \log S(t)}{dt}$$

- $$S(t) = \exp\{-\Lambda(t)\} = \exp\{-\int_0^t \lambda(u)du\}$$

- Note: Any nonnegative function $\lambda(t)$ that satisfies $\int_0^t \lambda(u)du < \infty$ for some $t > 0$ and $\int_0^\infty \lambda(u)du = \infty$ can be the hazard function of a continuous random variable.

Expected residual life

Other representation of failure time distribution is the expected residual life at time t

$$\begin{aligned} r(t) &= E(X - t \mid X \geq t] \\ &= \frac{\int_t^\infty (u-t) f(u) du}{S(t)}. \end{aligned}$$

This uniquely determines a continuous survival function with finite mean

$$S(t) = \frac{r(0)}{r(t)} \exp\{-\int_0^t \frac{du}{r(u)}\}.$$

Concept (6)

T is a discrete random variable taking values $a_1 < a_2 \ldots$ with probabilities $f(a_1), f(a_2), \ldots$. The discrete hazard function $(\lambda_i, i = 1, 2, \ldots)$ uniquely determines the distribution of T.

$$\begin{aligned} \lambda_i &= P(T = a_i \mid T \geq a_i) = \frac{f(a_i)}{1 - \sum_{j=1}^{i-1} f(a_j)} \\ S(t) &= \prod_{j; a_j \leq t} (1 - \lambda_j) \\ f(a_i) &= \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j) \end{aligned}$$

Concept (7)

T has discrete and continuous components with continuous component $\lambda_c(t)$ and discrete components $(\lambda_i, i = 1, 2, \ldots)$ at the discrete times $a_1 < a_2 < \ldots$.

$$
\begin{aligned}
S(t) &= \exp\{-\int_0^t \lambda_c(u)du\} \prod_{j;a_j \le t} (1-\lambda_j) \\
\Lambda(t) &= \int_0^t \lambda_c(u)du + \sum_{j;a_j \le t} \lambda_j \\
d\Lambda(t) &= \lambda_i, \quad t = a_i, \\
&= \lambda_c(t)dt, \ \textit{otherwise}
\end{aligned}
$$

Concept (8)

In general,

$$S(t) = \mathcal{P}_0^t[1 - d\Lambda(u)]$$

where the product integral $\mathcal{P}$ is defined by

$$\mathcal{P}_0^t[1 - d\Lambda(u)] = \lim \prod_{k=1}^{r} \{1 - [\Lambda(u_k) - \Lambda(U_{k-1}]\},$$

where $0 = u_0 < u_1 \ldots < u_r = t$ and the limit is taken as $r \to \infty$ and $\max(u_i - u_{i-1}) \to 0$.

Inference

Statistical inference questions:

- How to estimate the survival function and hazard rate?
- Are there parametric distributions describing the data?
- How to use the data collected during the baseline survey to estimate the effects on the survival? (survival regression models)
- Alternatives to parametric models - semiparametric and nonparametric approaches to the survival data analysis
- General models - estimation of transition probabilities and intensities

Formulation

As described above - standard statistical models in terms of distributions

Counting processes and their intensities

Event history analysis (3)

- Event history data needs dynamic methods (follow individuals over time).
- Covariate values may change over time.
- Covariates may depend on the history at each follow-up time point.

Historical development of EHA (1)

- *Aalen* - concepts from demography and life table analysis were studied using continuous-time martingale theory, stochastic integration, and counting process theory.
  Reference: O. O. Aalen (1975). Statistical inference for a family of counting processes. PhD thesis, University of California, Berkley.
- *Kaplan and Meier* - interpreted the classical life table analysis statistically and and gave a widely used Kaplan-Meier estimator for the survival function in 1958.
  Reference: Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. JASA, 53, 457-481, 562-563.

Historical development of EHA (2)

- *Nelson* - presented a hazard plot for incomplete failure time data and it was independently studied by Aalen in 1972 in his master's thesis.
- *Gehan (1965), Mantel (1966), Efron (1967), Breslow (1970)* - nonparametric rank tests for censored data
- *Breslow and Crowley (1974)* - weak convergence of the Nelson-Aalen and Kaplan-Meier estimators.
- *Aalen and Johansen (1978)* - introduced the product-integral as the canonical transformation from hazard to distribution function.

Historical development of EHA (3)

- *Gill (1979-80)* - estimation and two-sample tests in the classical censored survival data problem. (PhD thesis)
- *Cox (1972)* - assumed the death rate/intensity to be a product of an unspecified function of time common to all individuals and a known function of a linear combination of covariates wih coefficients to be estimated. He also introduced partial likelihood for the estimation.
- *Prentice and Kalbfleisch (1978-80)* - authority on censoring and likelihood and on the hazard function approach to models for several types of failure.
- Arjas and Haara (1984, 1989) - mathematical rigorisation of censoring patterns which was further developed by Andersen (1988).

Discussion articles

1. Niels Keiding. Event history analysis. Annual Review of Statistics and its Applications 2014; 1: 333-360.
2. Per Kragh Andersen, Steen Z Abildstrom, and Susanne Rosthøj. Competing risks as a multi-state model. Statistical Methods in MedicalResearch 2002; 11: 203-215.
3. Kate Bull and David J. Spiegelhalter. Tutorial in biostatistics survival analysis in observational studies. Statistics in medicine, vol. 16, 1041-1074 (1997).
4. Niels Keiding, Oluf K. Højbjerg Hansen and Ditte Nørbø Sørensen. The current duration approach to estimating time to pregnancy. Scandinavian J of Statistics 2012; 185-204.

Reference books

1. Aalen, O, Borgan, Gjessing, H. *Survival and Event History Analysis*. Springer-Verlag, (2008).
2. Broström G. *Event History Analysis with R*. CRC Press, Taylor & Francis Group (2012).
3. Andersen, P. K., Borgan, O, Gill, R. D., Keiding, N. *Statistical Models Based on Counting Processes*. Springer-Verlag, (1993).
4. Clayton D and Hills M. *Statistical Methods in Epidemiology*. Oxford University Press, New York (2002).
5. Kalbfleisch, J. D., and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: Wiley, (2002).
6. Cook, R.J. and Lawless, J. F. *Statistical analysis of recurrent events*. Springer-Verlag, (2007).