Survival Analysis

Lecture 3

Marta Fiocco[1,2] & Hein Putter [1]

(1) Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
(2) Mathematical Institute Leiden University

Outline

Truncation
Left truncated data

Right and left truncated data

Non-parametric estimation

Kaplan-Meier method

Nelson-Aalen

Informative censoring

- Truncated data arise when a variable is observable only over some portion of its range
- Truncation is a property of the underlying distribution from which the data arise
- The values of a random variable may be observable only when they are greater than a specific lower bound, only when the values exceed a threshold value
- All values of such a random variable that fall outside these bounds are never observable, and consequently their existence is not known to us
- This property of a statistical distribution is called truncation
- Truncation can be observed not only in lifetime data, but also in other types of data

- Left truncation is the most common form of truncation in lifetime data
- Called also late entry
- Example: all electronic goods are sold after they are tested for some pre-specified number of hours
- When we buy a TV or a refrigerator, their lifetimes have already exceeded the threshold value that the manufacturer decided to be the testing period, say 200 hours
- Some of the units may have failed during the testing period, but neither their lifetimes nor the number of such failed units are known to us. All we know is that the units sold have lifetime $X > 200$
- We observe only items (subjects) who live beyond the entry time

- random variable X left truncated at a point τ^L (i.e. realizations of X are observable only when they exceed τ^L
- pdf is

$$f_{LT}(x) = \frac{f(x)}{1 - F(\tau^L)}, \; x > \tau^L$$

- left truncated cdf

$$F_{LT}(x) = \frac{F(x) - F(\tau^L)}{1 - F(\tau^L)}, \; x > \tau^L$$

Outline

Truncation
Left truncated data
Right and left truncated data
Non-parametric estimation
Kaplan-Meier method
Nelson-Aalen
Informative censoring

- Though not common in lifetime data, right truncation occurs when the values of a random variable can be observed only when they are smaller than an upper bound (the right truncation point)
- all values greater than this upper bound are not observable
- in reliability data are collected in such a way that only items satisfying certain conditions are observed
- time to first failure are collected only for those items that fail over a given calendar time period $(0, T)$
- an item that enters service at calendar time u_i in $(0, T)$ and has failure time Y_i is observable iff $Y_i \leq \tau_i$, where $\tau_i = T - u_i$

- If the right truncation point is τ^R, then the pdf and cdf of a right truncated random variable are
- pdf is

$$f_{RT}(x) = \frac{f(x)}{F(\tau^R)}, \; x < \tau^R$$

- right truncated cdf

$$F_{RT}(x) = \frac{F(x)}{F(\tau^R)}, \; x < \tau^R$$

Marta Fiocco [1,2] & Hein Putter[1]

Censoring and truncation

- Though the censoring and truncation seem to be close as far as incompleteness of data is concerned, they are actually distinctly different
- Censoring is a property of the sample from a population, while truncation is a property of the population itself
- consider left truncation and left censoring: in left censoring, we know that some units failed before a specified time τ, and the number of such units is known to us

- All that we know about these failed units is that their lifetimes belong to the interval $[0, \tau]$ without any more specific knowledge
- In left truncation, we have no information at all on all these units that failed before the left truncation point, say τ
- We know that the existing units have exceeded the threshold τ, and some units may have failed before τ
- But we have no knowledge on the failed units at all, including the fact that if such units even existed

Outline

Truncation
Left truncated data
Right and left truncated data
Non-parametric estimation
Kaplan-Meier method
Nelson-Aalen
Informative censoring

Non-parametric estimation

- Previous week we have discussed parametric models for survival data
- But, especially in clinical research, non-parametric methods are much more often used
- Or semi-parametric, in regression models for survival data (Cox model)
- Topic of today: non-parametric methods
  - For the survival function (Kaplan-Meier estimator)
  - For the (cumulative) hazard function (Nelson-Aalen estimator)

Recall survival data

In the presence of censoring

- For an individual i, $i = 1, \ldots, n$, we define
  - x_i: event time
  - c_i: censoring time
- We *observe*
  - $t_i = \min(x_i, c_i)$: observed time
  - δ_i taking value
    - 1 if $x_i \leq c_i$ (event is observed)
    - 0 is $x_i > c_i$ (censored observation)

Outline

Truncation
Left truncated data
Right and left truncated data
Non-parametric estimation
Kaplan-Meier method
Nelson-Aalen
Informative censoring

Kaplan-Meier method

- Aim is to estimate the survival function of T

$$S(t) = P(T > t)$$

- The probability of being event-free at t or the probability that the survival time is greater than t
- Based on possibly censored survival data
- We need to account for the censoring somehow ...
- Assuming that censoring is unrelated to potential survival time

More notation

- Suppose that the events occur at D distinct times $t_1 < t_2 < \ldots < t_D$
- At time t_i there are d_i events (deaths)
- Let Y_i be the number of individuals who are *at risk* at time t_i
  - Individuals who are alive and in follow-up just before t_i (including individuals experiencing the event of interest at t_i)
- The quantity d_i / Y_i is an estimate of the probability that an individual who survives to just prior to time t_i experiences the event at time t_i
- An estimate of the *hazard* at t_i
- This is the basic quantity from which we will construct estimators of the survival function and the cumulative hazard rate

Kaplan-Meier method

- Example: Survival times of 20 patients (in days)
- 54, 76, 80, 121, 150, 177^+, 195, 221, 221, 257^+, 260^+, 310, 310^+, 390, 420^+, 503, 580^+, 670^+, 680, 685^+ (+ means: censored)
- In case of no censoring

$$\hat{S}(t) = \frac{\textit{no. of patients surviving beyond } t}{n}$$

- In the example: $\hat{S}(100) = \frac{17}{20}$
- The problem is observation 177^+

Kaplan-Meier method

- Divide time axis in as small as possible intervals, typically days
- Suppose censoring happens only at end of an interval
- $\hat{p}_i$: estimated conditional probability of surviving day i
- d_i: number of deaths at day i
- Y_i: number at risk at time t_i
- d_i/Y_i: estimate of the conditional probability that an individual who has survived to just prior to time t_i dies at time t_i
- $(1 - d_i/Y_i)$: estimate of the conditional probability that an individual who has survived to just prior to time t_i will survive t_i
- Multiply these estimates to obtain an estimator of the survival function

Kaplan-Meier method

In the example we have:

$$\hat{S}(200) = \hat{p}_1 \cdot \hat{p}_2 \cdot \hat{p}_3 \ldots \hat{p}_{200}$$

More precisely we get (other p's are 1):

$$\hat{S}(200) = \hat{p}_{54} \cdot \hat{p}_{76} \cdot \hat{p}_{80} \cdot \hat{p}_{121} \cdot \hat{p}_{150} \cdot \hat{p}_{195}$$

$$= \frac{19}{20} \times \frac{18}{19} \times \frac{17}{18} \times \frac{16}{17} \times \frac{15}{16} \times \frac{13}{14} \times = 0.696$$

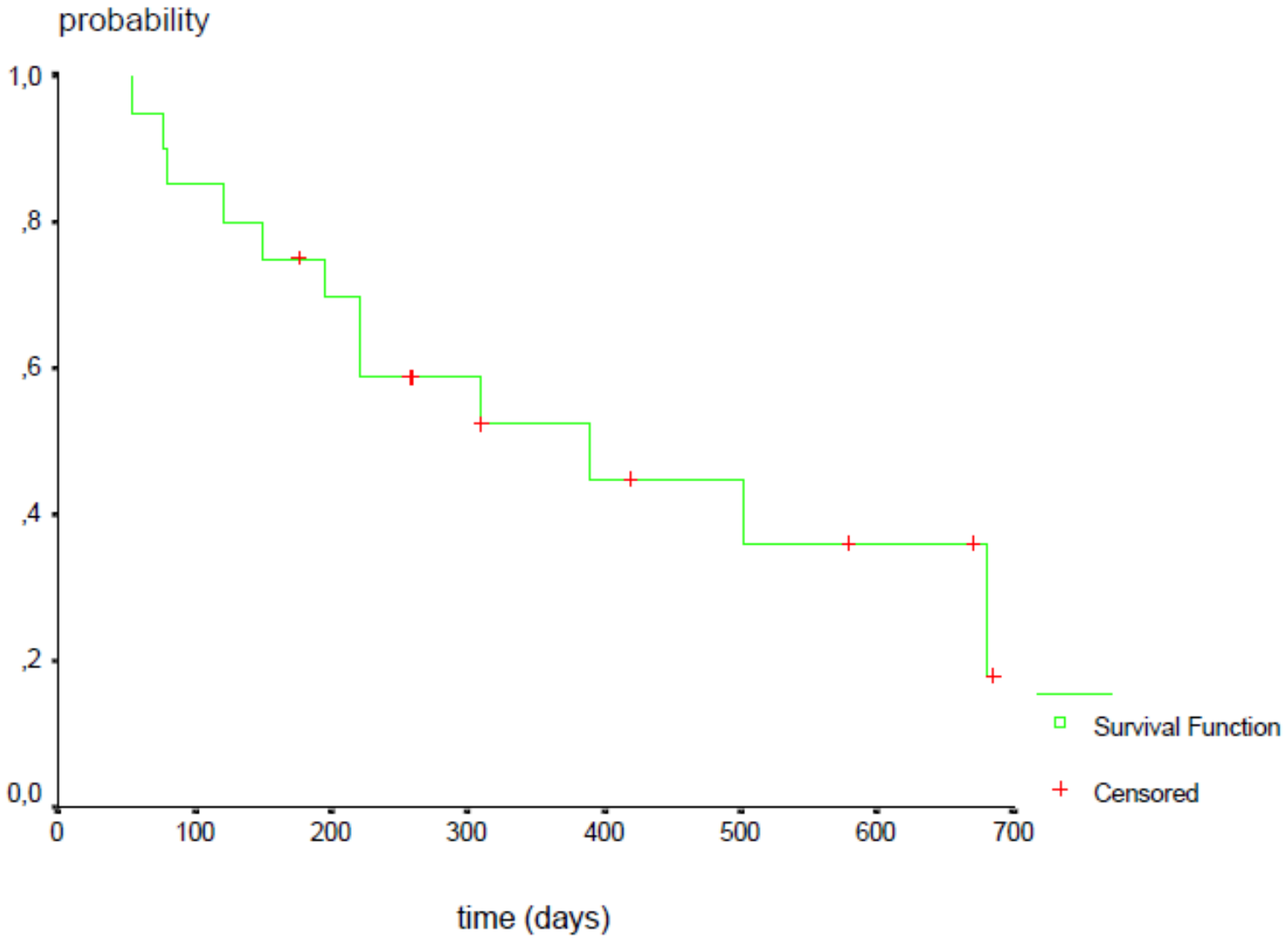
Example of Kaplan-Meier survival table

time of death	no of death	no under observation	death probability	survival probability	cumulative survival probability
54	1	20	1/20	19/20	19/20 = .95
76	1	19	1/19	18/19	18/19 x .95 =.90
80	1	18	1/18	17/18	17/18 x .90 =.85
121	1	17	1/17	16/17	16/17 x .85 =.80
150	1	16	1/16	15/16	15/16 x .80 =.75
195	1	14	1/14	13/14	13/14 x .75 =.696
221	2	13	2/13	11/13	11/13 x .696 =.589
310	1	9	1/9	8/9	8/9 x .589 = .523
390	1	7	1/7	6/7	6/7 x .523 = .449
503	1	5	1/5	4/5	4/5 x .449 = .359
680	1	2	1/2	1/2	1/2 x .359 = .180

177+

decrease when censored data

Estimate of the survival curve

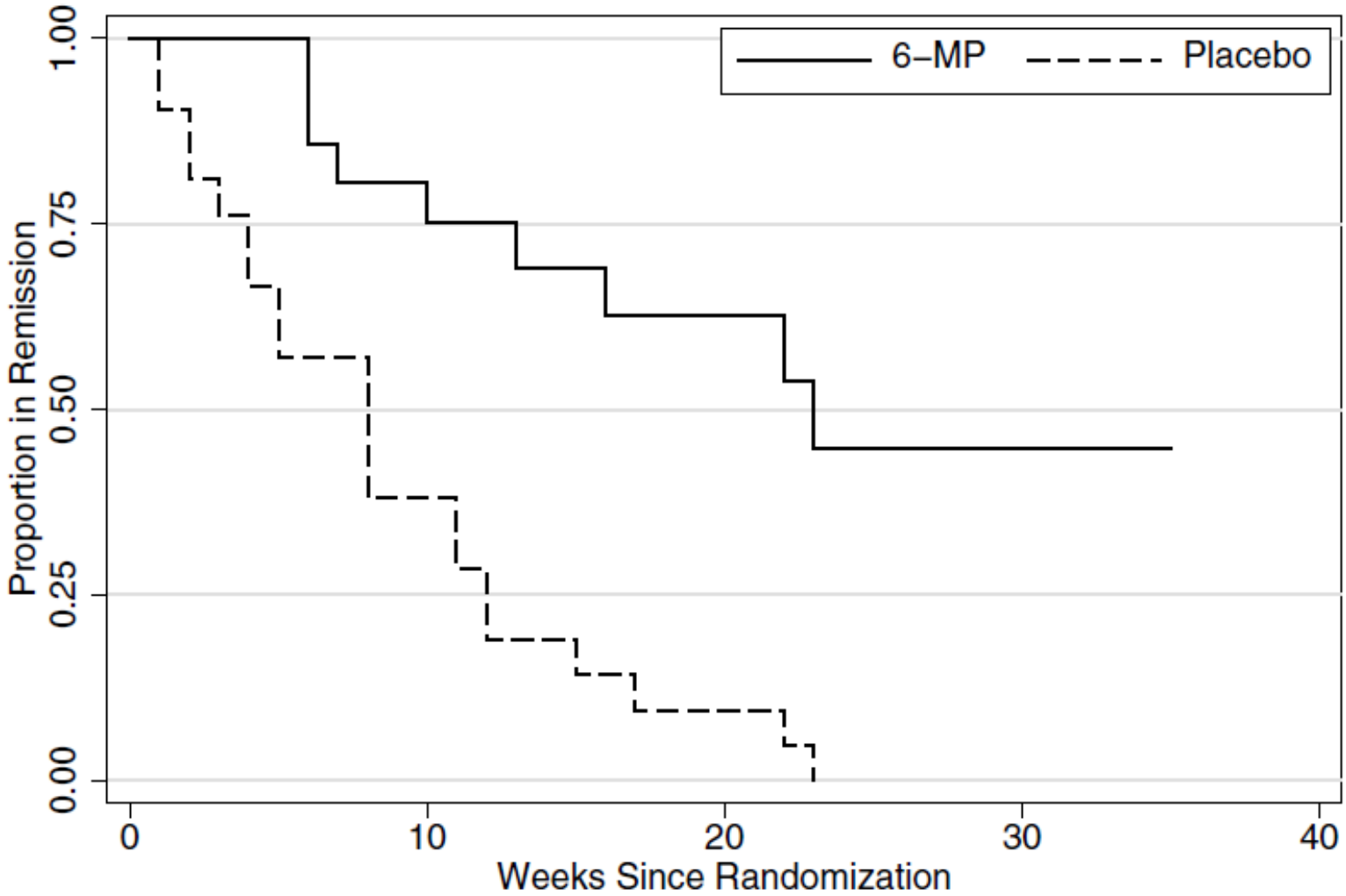


The product-limit estimator

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right)$$

- The product-limit estimator is a step function with jumps at the observed event times
- The size of these jumps depends not only on the number of events observed at each event time t_i, but also on the pattern of the censored observations prior to t_i

Survival curves by treatment for leukemia patients (Ex in Table 1.1)



Variance estimation

- Greenwood's formula

$$\widehat{V}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

- Simpler variance estimate (Aalen and Johansen 1978)

$$\widetilde{V}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i^2}$$

- This estimator and Greenwood's estimator tend to underestimate the true variance of the Kaplan- Meier estimator for small to moderate samples
- On average, Greenwood's estimator tends to come closest to the true variance and has a smaller variance except when Y_i is very small

Outline

Truncation
Left truncated data
Right and left truncated data
Non-parametric estimation
Kaplan-Meier method
Nelson-Aalen
Informative censoring

Estimation of the cumulative hazard function

- One way is to use the product-limit estimator to estimate the cumulative hazard function $H(t) = -ln[S(t)]$
- Estimator of the cumulative hazard $\widehat{H}(t) = -ln[\widehat{S}(t)]$
- Alternative estimator of $H(t)$ (Nelson-Aalen)

$$\widetilde{H}(t) = \sum_{i:t_i \leq t} \frac{d_i}{Y_i}$$

- Note that it uses $\frac{d_i}{Y_i}$, just like the Kaplan-Meier
- Estimated variance of the Nelson-Aalen estimator

$$\sigma^2_H(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2}$$

- Just like the Kaplan-Meier can be used to estimate the cumulative hazard, so the Nelson-Aalen can be used to estimate the survival function

$$\widetilde{S}(t) = \exp[-\widetilde{H}(t)]$$

- Construction of the Nelson-Aalen estimator and its estimated variance for the 6-MP group

Time t	$\tilde{H}(t) = \sum_{t_i \le t} \frac{d_i}{Y_i}$	$\sigma^2_H = \sum_{t_i \le t} \frac{d_i}{Y_i^2}$.	*Standard Error*
$0 \le t < 6$	0	0	0
$6 \le t < 7$	$\frac{3}{21} = 0.1428$	$\frac{3}{21^2} = 0.0068$	0.0825
$7 \le t < 10$	$0.1428 + \frac{1}{17} = 0.2017$	$0.0068 + \frac{1}{17^2} = 0.0103$	0.1015
$10 \le t < 13$	$0.2017 + \frac{1}{15} = 0.2683$	$0.0103 + \frac{1}{15^2} = 0.0147$	0.1212
$13 \le t < 16$	$0.2683 + \frac{1}{12} = 0.3517$	$0.0147 + \frac{1}{12^2} = 0.0217$	0.1473
$16 \le t < 22$	$0.3517 + \frac{1}{11} = 0.4426$	$0.0217 + \frac{1}{11^2} = 0.0299$	0.1729
$22 \le t < 23$	$0.4426 + \frac{1}{7} = 0.5854$	$0.0299 + \frac{1}{7^2} = 0.0503$	0.2243
$23 \le t < 35$	$0.5854 + \frac{1}{c} = 0.7521$	$0.0503 + \frac{1}{c^2} = 0.0781$	0.2795

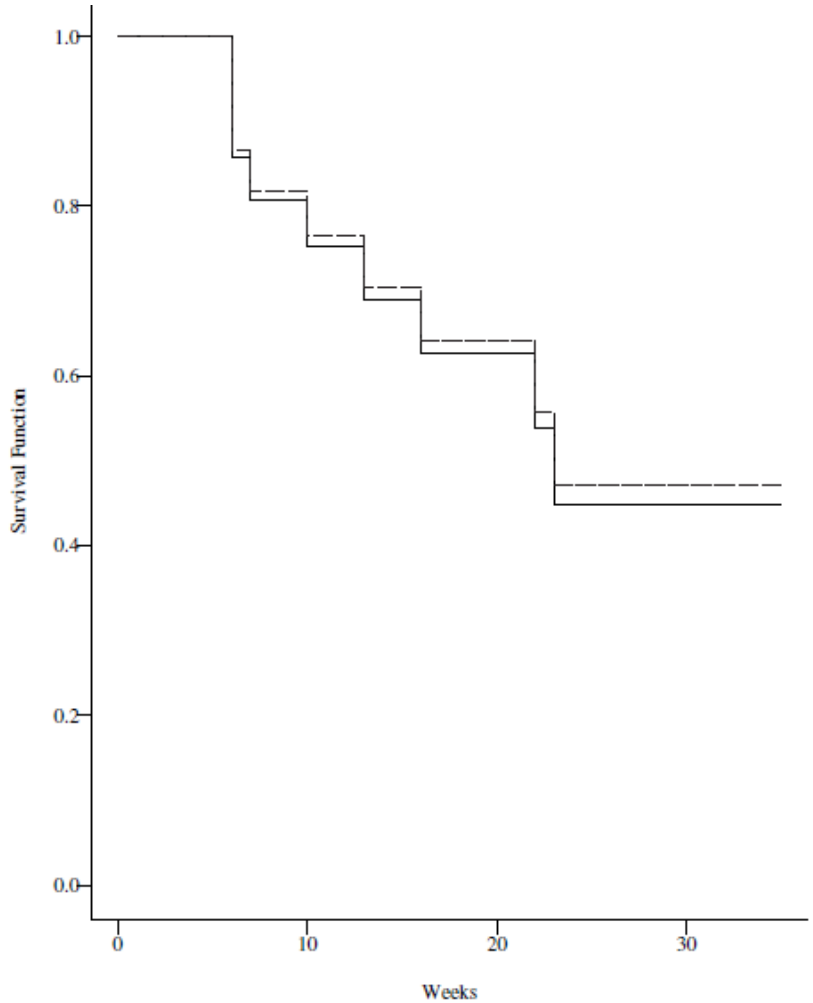


Figure 4.1A *Comparison of the Nelson–Aalen (------) and Product-Limit (————) estimates of the survival function for the 6-MP group.*

The Nelson-Aalen estimate

- It is an estimate of the integrated (cumulative) hazard function
- A plot of the cumulative hazard function is useful
  - To check distributional assumptions
  - To check proportional hazards assumption (later)

Outline

Truncation
Left truncated data
Right and left truncated data
Non-parametric estimation
Kaplan-Meier method
Nelson-Aalen
Informative censoring

Independent censoring

- Both Nelson-Aalen and Kaplan-Meier (product-limit) estimator are based on assumption of *noninformative censoring*
- This assumption is used when d_i/Y_i (based on those *excluding* patients censored before t_i) is used as estimate of $P(T = t_i|T \geq t_i)$
- It implies that knowledge of a censoring time for an individual gives no further information about this person's likelihood of survival at a future time had the individual continued on the study
- Most frequent example of violation: patients feeling ill drop out and are lost to follow-up (=> censored)

Examples of informative censoring

- Study on alcohol relapse: include two censoring mechanisms:
  - The first one was end of data collection when some individuals are still abstinent; this censoring mechanism is not informative
  - The second one concerns individuals who were lost to follow up after being abstinent for one year
  - Reasons for lost to follow up: moved out of town (or other random occurrences) then ***censoring is not informative***
  - Lost to follow-up because they started drinking again and stopped notifying investigators; then ***censoring is informative***

What to do in case of informative censoring

- In case of informative censoring the only way out is to jointly model the survival and censoring processes, which very often would imply knowledge that is not present
- These methods rely on correct estimation of the censoring mechanism, but will generally be less biased than methods ignoring informative censoring (e.g., Kaplan-Meier)