STK4080/9080 SURVIVAL AND EVENT HISTORY ANALYSIS

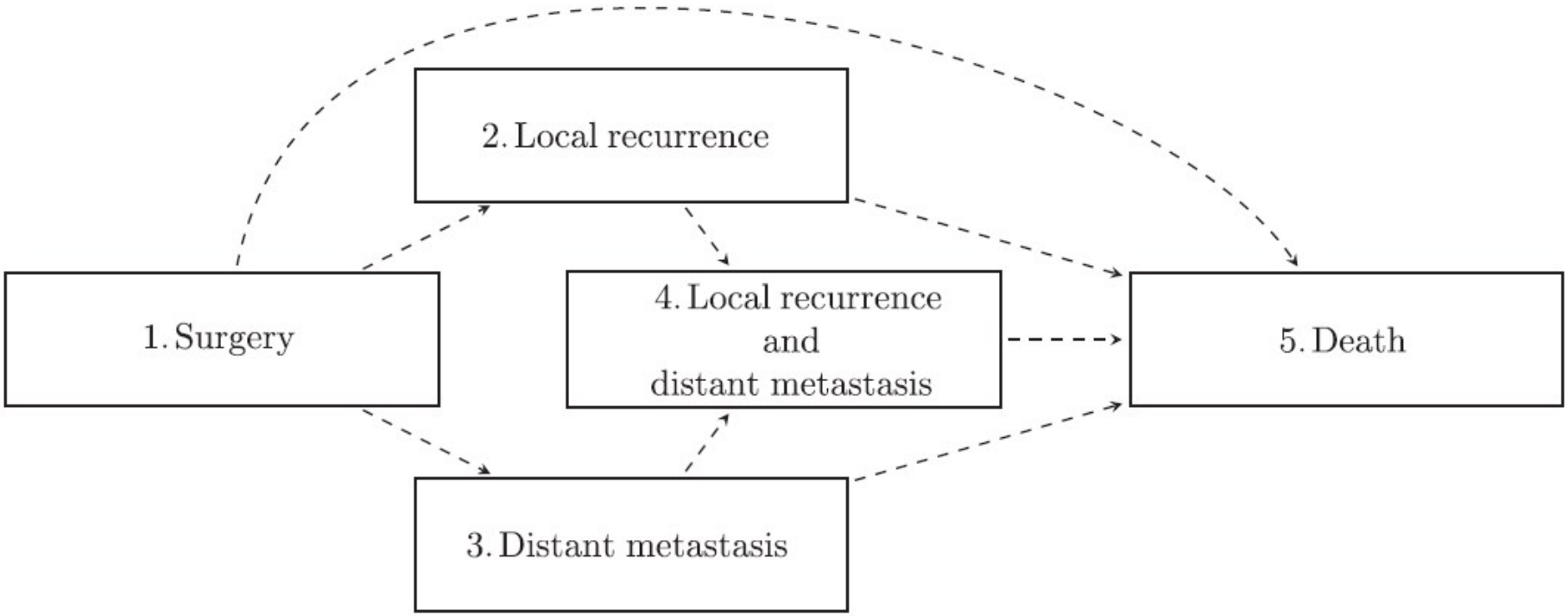
Slides 9: Multistate models

Bo Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim

https://www.ntnu.edu/employees/bo.lindqvist
bo.lindqvist@ntnu.no
boli@math.uio.no

University of Oslo, Spring 2021

Example: A multistate model for breast cancer



From:

Putter, H., Fiocco, M., & Geskus, R. B. (2007). **Tutorial in biostatistics: competing risks and multi-state models.** *Statistics in Medicine,* 26(11), 2389-2430.

Multistate models

We will consider stochastic processes $X(t)$ which move within finite state spaces $\mathcal{S} = \{0, 1, \ldots, k\}$; e.g.,

We will in particular consider

- The competing risks model
- The illness-death (e.g. healthy-illness-death)) model
- The general case (Aalen-Johansen estimator)

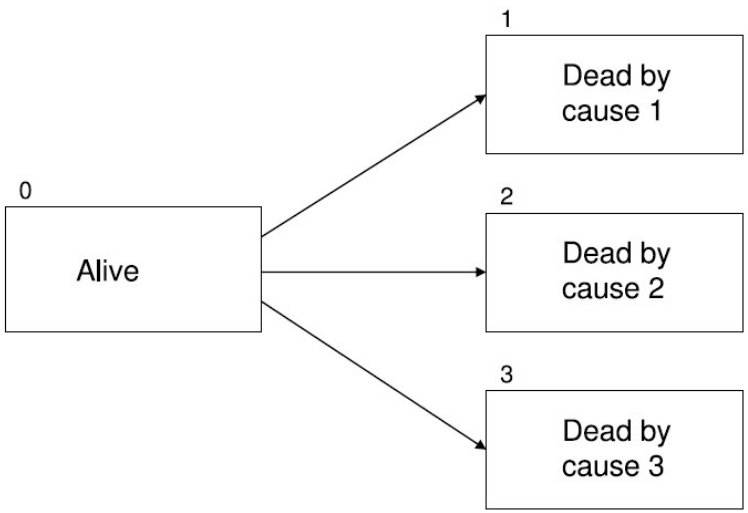


Fig. 3.5 *A model for competing risks with $k = 3$.*

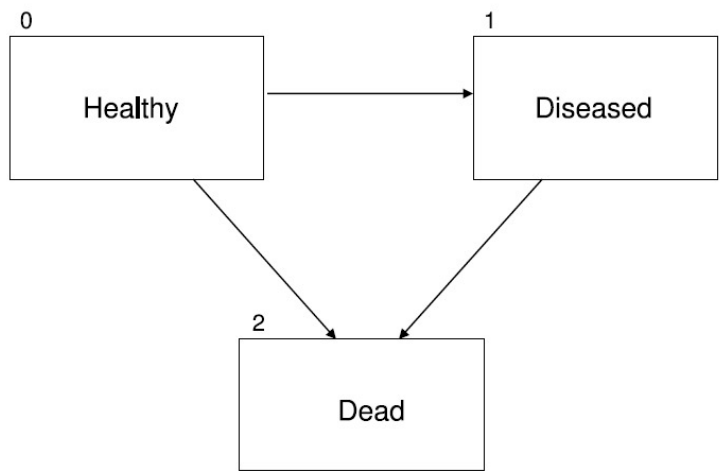


Fig. 3.19 *An illness-death model without recovery.*

Competing risks

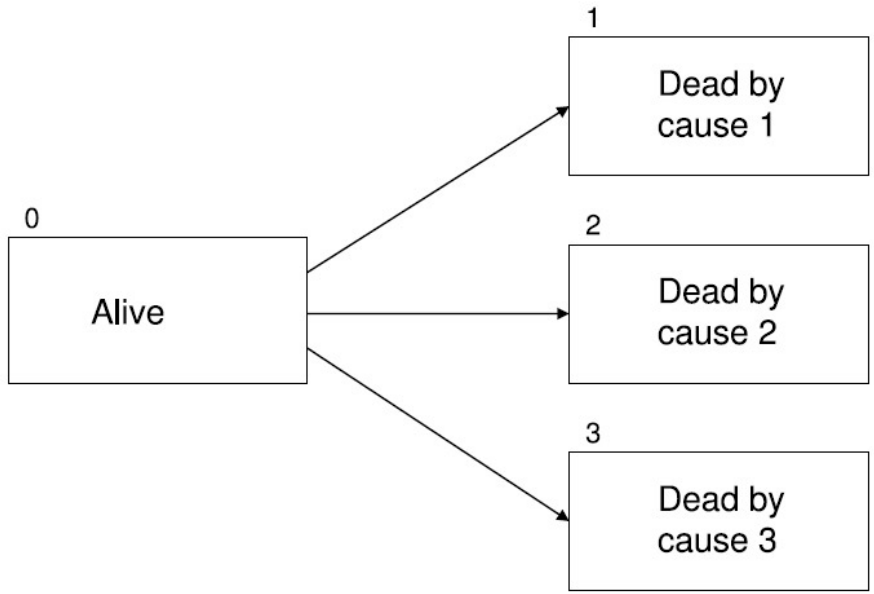


Fig. 3.5 *A model for competing risks with* $k = 3$.

Suppose that for each individual we observe

- the time to the event of interest ("failure"), T
- the cause of failure, $H \in \{1, 2, \ldots, k\}$

The pair (T, H) is the observation in a case of competing risks.

Examples

- Classical competing risks: Individuals subjected to multiple causes of death, for example
  - David Bernoulli (1760): How to disentangle the risk of dying from smallpox from other causes.
- Demography and actuarial sciences
  - "Multiple-decrement analysis"
- Cancer research:
  - Age at onset of cancer and cancer type
  - Disease relapse vs death in remission
- Reliability: Breakdown of a mechanical component, with several possible root causes for failure:
  - vibration
  - corrosion
  - etc.

Latent failure time approach to competing risks

Suppose the k causes are represented by potential failure times $T_1, \ldots, T_k$

One observes only:

- The smallest time, $T = \min_h T_h$
- Its index $H = \arg\min_h T_h$ (assumed *unique*)

Important issues:

- The marginal distributions of the T_j are often of primary interest, but are *non-identifiable* in general by observation of (T, H) only (even if we have an infinite number of observations of (T, H))
- Additional, but non-testable, assumptions may lead to identifiability (for example, independence of the T_j).
- **In biostatistics, one usually avoids the latent failure time approach and restricts attention to the pair (T, H).**
- Independent censoring at time C can easily be included, leading to the observation of the time $\tilde{T} = \min(T, C)$ together with H, where $H = 0$ means censoring.

Model specification: **The cumulative incidence function**

- The joint distribution of the observed pair (T, H) is given by the **cumulative incidence function**

$$F_h(t) = P(T \leq t, H = h) \qquad \text{for } t > 0,\ h = 1, \ldots k$$

- The marginal distributions of T and H are hence given by

$$\begin{aligned} F(t) &= P(T \leq t) = \sum_{h=1}^{k} F_h(t) \qquad \text{for } t > 0 \\ \pi_h &= P(H = h) = F_h(\infty) \qquad \text{for } h = 1, \ldots k \end{aligned}$$

Specification by **cause-specific hazard rates**

- The distribution of (T, H) can alternatively be specified by the **cause-specific hazard functions**:

$$\alpha_h(t) = \lim_{\Delta t \to 0} \frac{P\left(t < T \leq t + \Delta t, H = h \middle| T > t\right)}{\Delta t} = \frac{f_h(t)}{S(t)}$$

  where $f_h(t) = F_h'(t)$ is the derivative of the cumulative incidence function and $S(t) = 1 - F(t)$ is the survival function of T.
- It follows that $f_h(t) = S(t)\alpha_h(t)$ and hence we get, by integration, the sometimes useful formula for the cumulative incidence function,

$$F_h(t) = \int_0^t S(s)\alpha_h(s)ds$$

General notation of Chapter 3.4 in ABG

We consider stochastic processes $X(t)$ with state space $\mathcal{S} = \{0, 1, \ldots, k\}$, assumed to satisfy the Markov assumption:

$$P\left(X(t) = h | X(s) = g\right) = P\left(X(t) = h | X(s) = g, \mathcal{F}_s\right), \quad s < t, \ g, h \in \mathcal{S}$$

We can then define *transition probabilities*

$$P_{gh}(s, t) = P\left(X(t) = h | X(s) = g\right), \ s < t, \ g, h \in \mathcal{S}$$

and *transition intensities*

$$\alpha_{gh}(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(X(t + \Delta t) = h | X(t-) = g) \ \text{ for } \ g \neq h$$

Competing risks (3.4.1 in ABG)

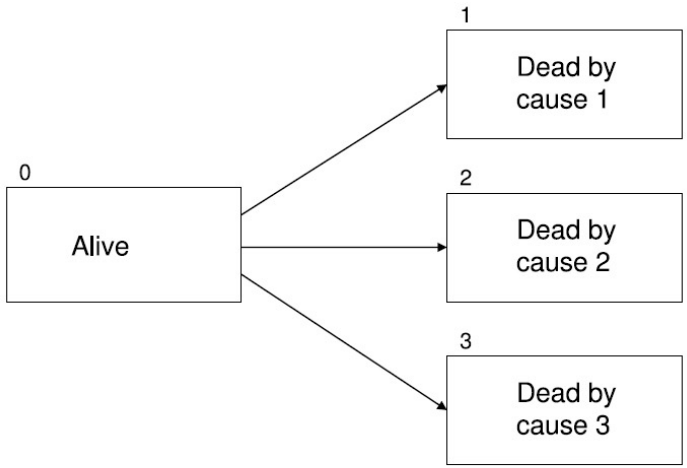


Fig. 3.5 *A model for competing risks with $k = 3$.*

Let $X(t)$ be the state at time t. The process then starts at $X(0) = 0$ and jumps to one (and only one) of the states ("causes") $\{1, 2, \ldots, k\}$.

Then if we let T be the time when the process jumps from state 0, and let H be the new state entered (i.e. the absorbing state), then we have

$$P_{00}(0,t) = P(T > t) = S(t) \text{ (the survival function)}$$
$$P_{0h}(0,t) = P(T \leq t,\ H = h) = F_h(t) \text{ (the cumulative incidence function}$$

Competing risks (3.4.1 in ABG)

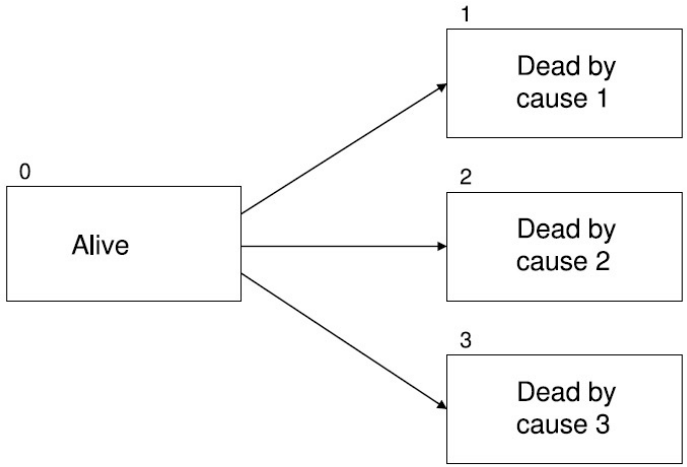


Fig. 3.5 *A model for competing risks with* $k = 3$.

The *cause-specific hazard* $\alpha_h(t)$ of cause h, as defined earlier, then is the transition intensity

$$\alpha_{0h}(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(X(t + \Delta t) = h | X(t-) = 0) \text{ for } h = 1, 2, \ldots, k$$

which in the counting process framework can be written

$$\alpha_{0h}(t)dt = P(\text{die from cause } h \text{ in } [t, t + dt) | \text{ alive at } t-)$$

Estimation of the cumulative cause-specific hazard

Assume that we have a sample of n individuals, where each individual is followed from an entry time to death or censoring.

- Let $T_1 < T_2 < \ldots$ be the times when deaths from any cause are observed,
- let $N_{0h}(t)$ be the process counting the number of individuals who are observed to die from cause h (i.e., make a transition from state 0 to state h) in the interval $[0, t]$,
- let $N_{0\bullet}(t) = \sum_{h=1}^{k} N_{0h}(t)$ for the total number of deaths in $[0, t]$, and let $Y_0(t)$ denote the number of individuals at risk (i.e., in state 0) just prior to time t.

Then the *cumulative cause-specific hazard function* for cause h, $A_{0h}(t) = \int_0^t \alpha_{0h}(s)ds$, is estimated by the Nelson-Aalen estimator

$$\hat{A}_{0h}(t) = \int_0^t \frac{dN_{0h}(s)}{Y_0(s)} = \sum_{T_j \le t} \frac{\Delta N_{0h}(T_j)}{Y_0(T_j)}$$

Notation used here: $\Delta N_{0h}(T_j) = \#\text{transitions } 0 \to h \text{ at } T_j$ (usually 0 or 1)

Estimation of the cumulative incidence function

The relations:

$$(\textit{Survival function})\ S(t) = P_{00}(0,t) \quad = \quad \exp\left(-\int_0^t \sum_{h=1}^{k} \alpha_{0h}(u)du\right)$$

$$(\textit{Cumulative incidence})\ F_h(t) = P_{0h}(0,t) \quad = \quad \int_0^t P_{00}(0,u-)\alpha_{0h}(u)du$$

suggest the estimators

$$\hat{P}_{00}(0,t) \quad = \quad \prod_{T_j \le t}\left(1 - \frac{\Delta N_{0\bullet}(T_j)}{Y_0(T_j)}\right), \quad \text{where } N_{0\bullet} = \sum_{h=1}^{k} N_{0h}(t)$$

$$\hat{P}_{0h}(0,t) \quad = \quad \sum_{T_j \le t} \hat{P}_{00}(0, T_{j-1})\frac{\Delta N_{0h}(T_j)}{Y_0(T_j)}$$

- $\hat{P}_{00}(0,t)$ *is exactly the KM-estimator for* $S(t)$
- $\alpha_{0h}(u)$ *is estimated by the increment of the NA-estimator for* $A_{0h}(t)$.

Estimated cumulative incidence function

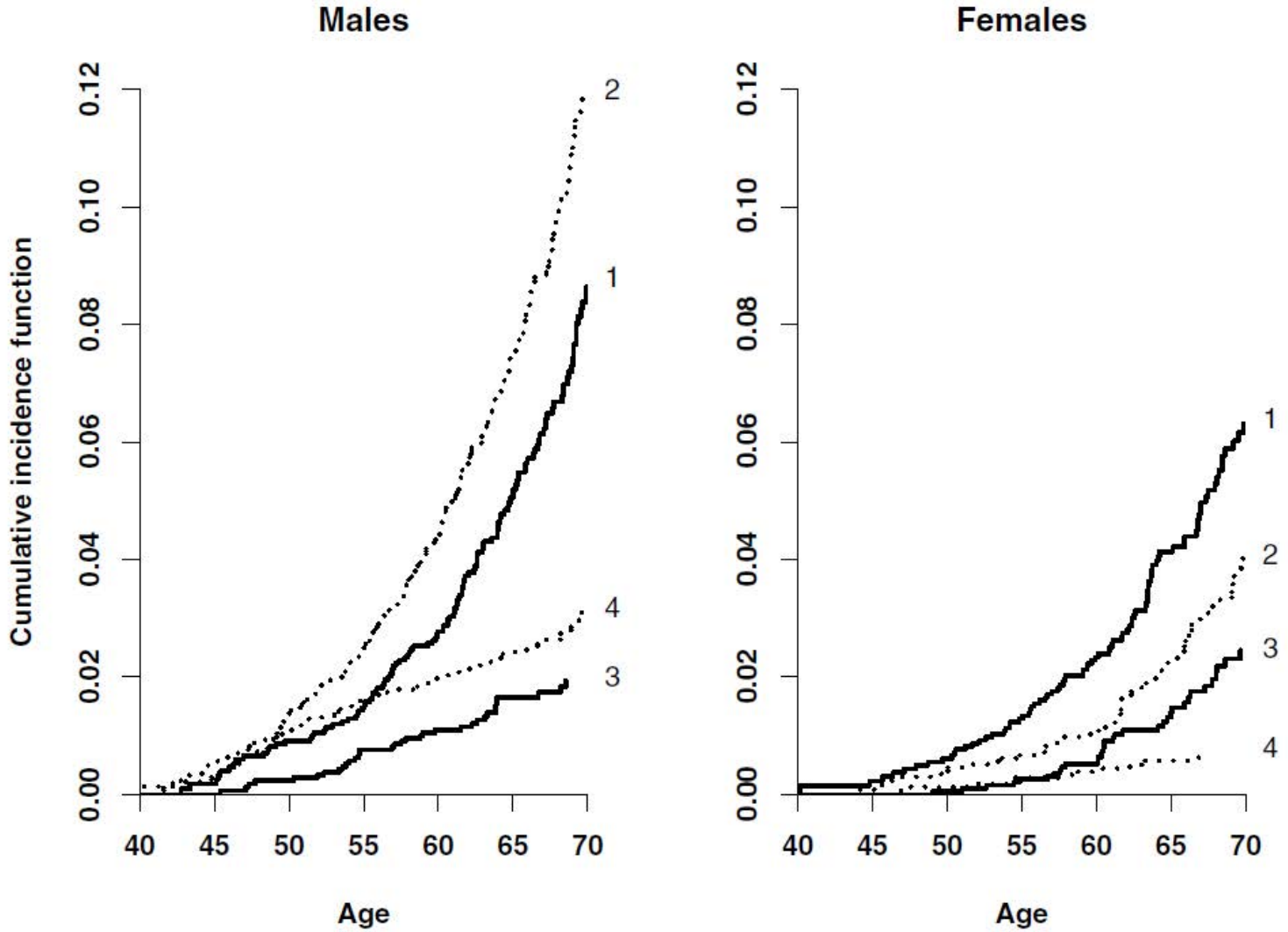


Fig. 3.17 *Empirical cumulative incidence functions for four causes of death among middle-aged Norwegian males (left) and females (right). 1) Cancer; 2) cardiovascular disease including sudden death; 3) other medical causes; 4) alcohol abuse, chronic liver disease, and accidents and violence.*

Simulated competing risks data

There are $n = 1000$ individuals, with latent failure times:

- $T_1 \sim \text{Weibull}(k = 2, b = 1)$, $\quad P(T_1 > t) = \exp\{-t^2\}$
- $T_2 \sim \text{Weibull}(k = 0.5, b = 1)$, $\quad P(T_2 > t) = \exp\{-t^{0.5}\}$
- $C \sim U[0, 2]$, $\quad$ (censoring times)

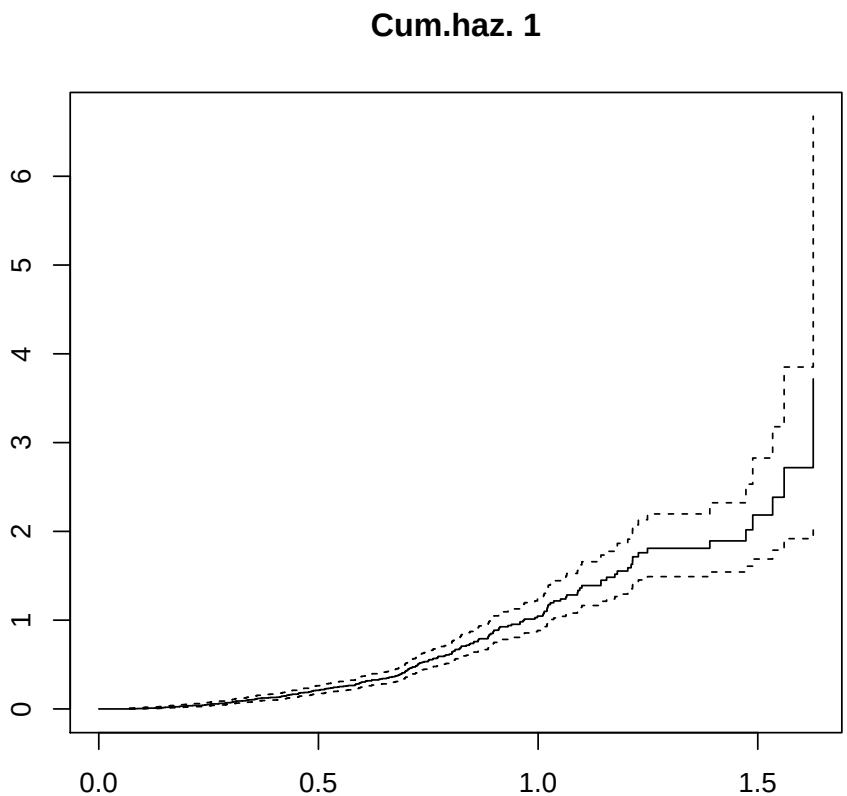
The observations are hence $(\tilde{T}, H)$ where

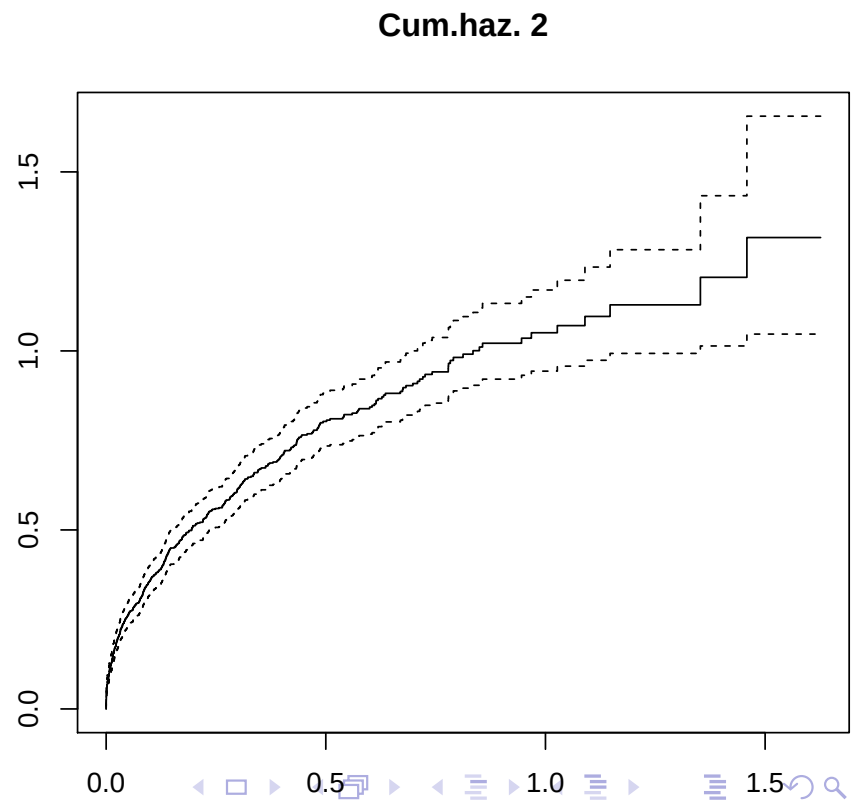
$$\tilde{T} = \min(T_1, T_2, C), \quad H = 1, 2, 0 \text{ according to whether } \tilde{T} = T_1, T_2, C$$

```
#R-CODE:
n=1000
time1=rweibull(n,2)
time2=rweibull(n,0.5)
censtime=runif(n)*2
obstime=pmin(time1,time2,censtime)
h = 1*(obstime==time1)+2*(obstime==time2)
```

Estimation of cumulative cause-specific hazard

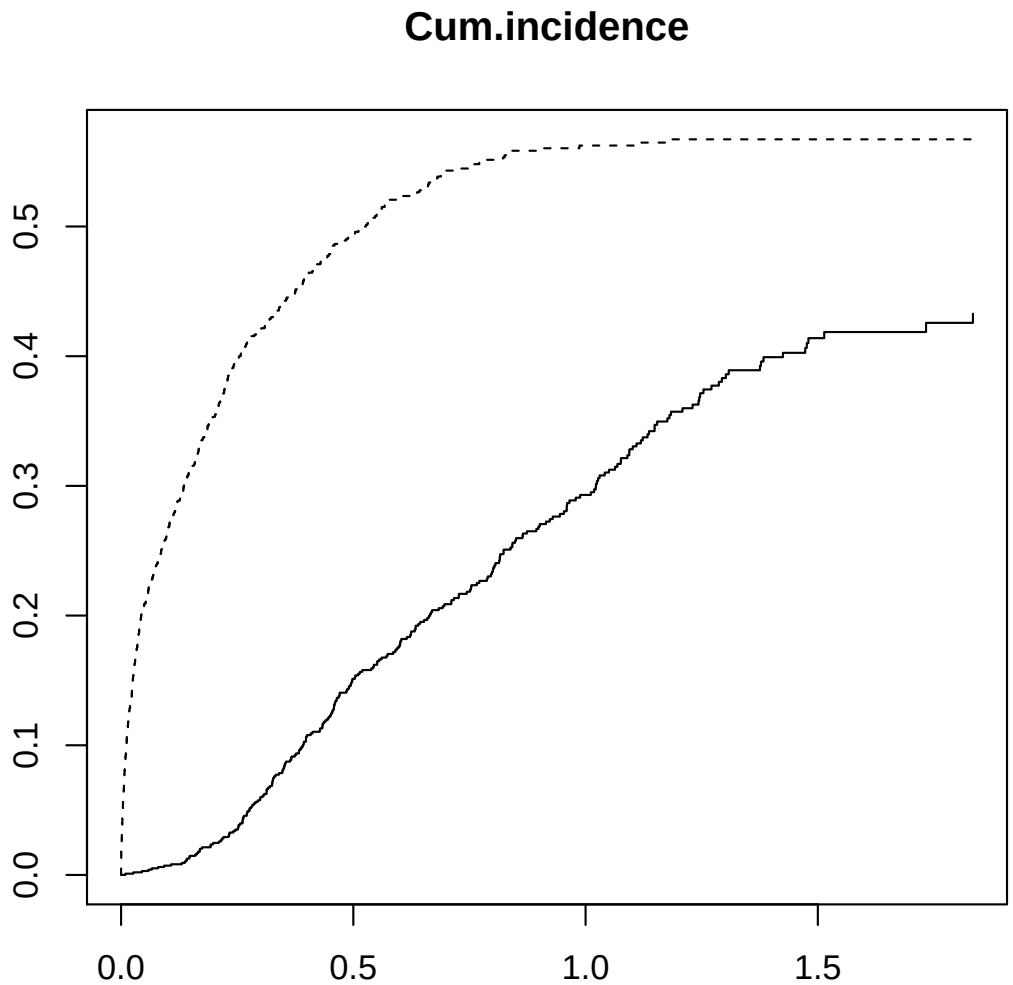
```
library(survival)
ch1 = survfit(Surv(obstime,h==1)~1,type="fh2")
ch2 = survfit(Surv(obstime,h==2)~1,type="fh2")
par(mfrow=c(1,2))
plot(ch1,fun="cumhaz",mark.time=F,main="Cum.haz.1")
plot(ch2,fun="cumhaz",mark.time=F,main="Cum.haz.2")
```





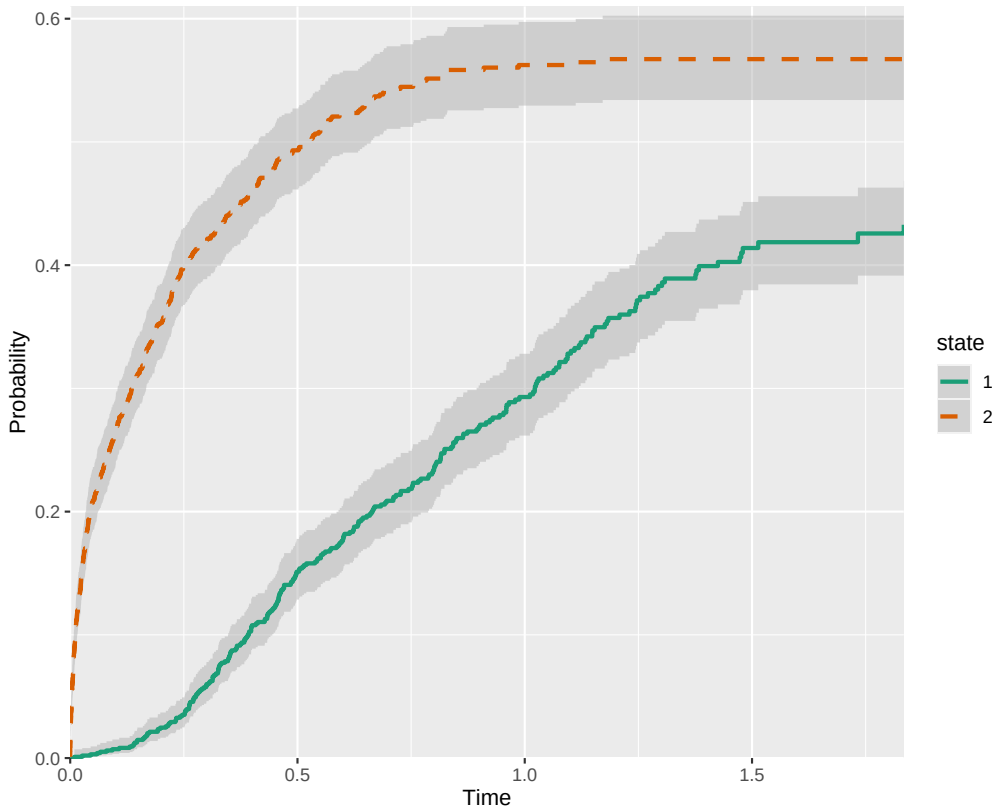
Estimation of cumulative incidence functions

```
#Estimation in the 'survival' package
ci.surv = survfit(Surv(obstime,h,type="mstate")~1)
plot(ci.surv,lty=1:2,main="Cum.incidence")
```



Estimation of cumulative incidence functions

```
#Estimation in the 'mstate' package
library(mstate)
library(ggplot2)
ci.sim = Cuminc(time=obstime,status=h)
plot(x = ci.sim,use.ggplot = TRUE,conf.type = "log",lty =
1:2,conf.int = 0.95)
```



The illness-death model without recovery

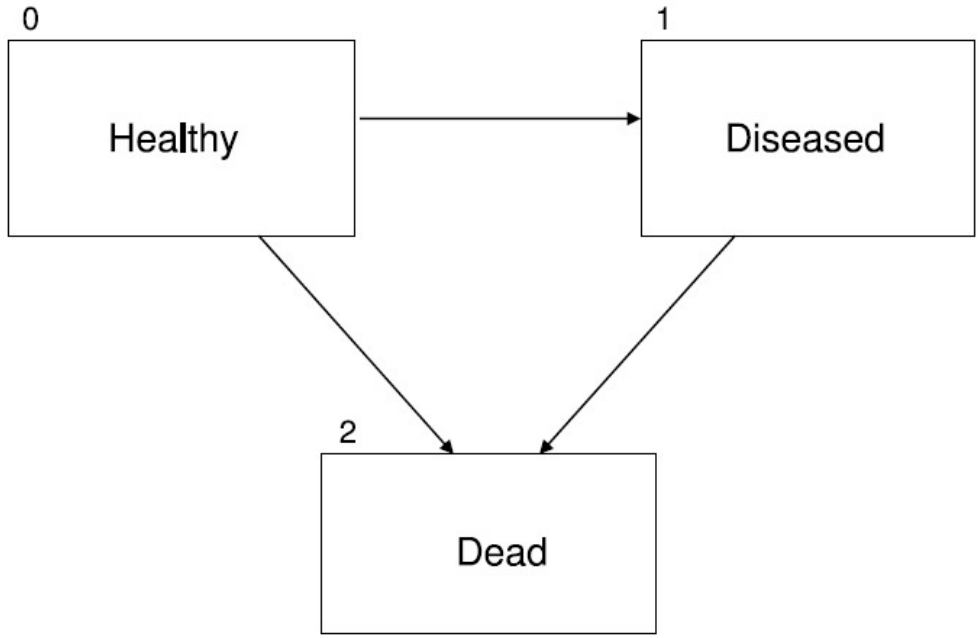


Fig. 3.19 *An illness-death model without recovery.*

The transition intensities $\alpha_{01}(t)$, $\alpha_{02}(t)$ and $\alpha_{12}(t)$ give the instantaneous probability of transition from one state to another (where arrows show the possible transitions).

Transition probabilities

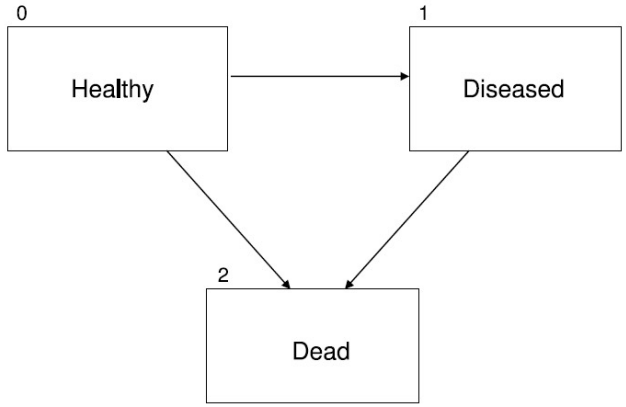


Fig. 3.19 *An illness-death model without recovery.*

$$P_{gh}(s,t) = P(\text{in state } h \text{ at time } t \mid \text{in state } g \text{ at time } s)$$

Then one may show

$$\begin{aligned} P_{00}(s,t) &= \exp\left\{-\int_s^t [\alpha_{01}(u)+\alpha_{02}(u)]du\right\} \\ P_{11}(s,t) &= \exp\left\{-\int_s^t \alpha_{12}(u)du\right\} \\ P_{01}(s,t) &= \int_s^t P_{00}(s,u-)\alpha_{01}(u)P_{11}(u,t)du \end{aligned}$$

Transition probabilities

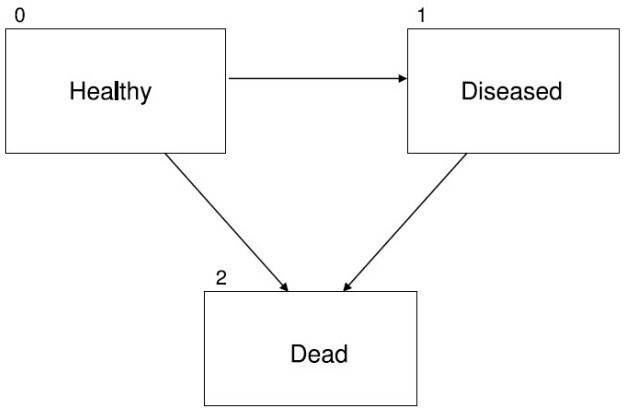


Fig. 3.19 *An illness-death model without recovery.*

$$P_{gh}(s,t) = P(\text{in state } h \text{ at time } t \mid \text{in state } g \text{ at time } s)$$

Further, one may show

$$\begin{aligned} P_{02}(s,t) &= \int_s^t P_{00}(s,u-)\alpha_{02}(u)du + \int_s^t P_{01}(s,u-)\alpha_{12}(u)du \\ P_{12}(s,t) &= 1 - P_{11}(s,t) \\ P_{22}(s,t) &= 1 \end{aligned}$$

Estimation of cumulative transition intensities

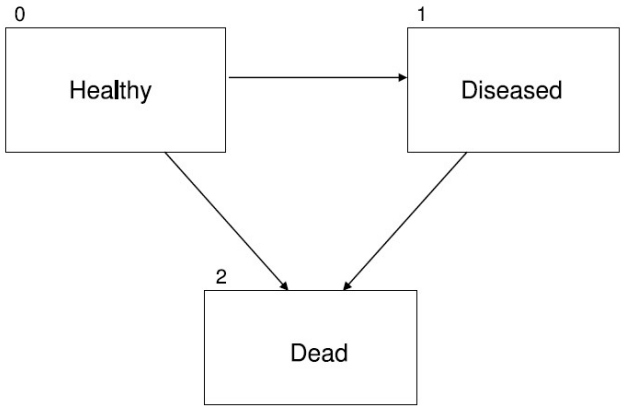
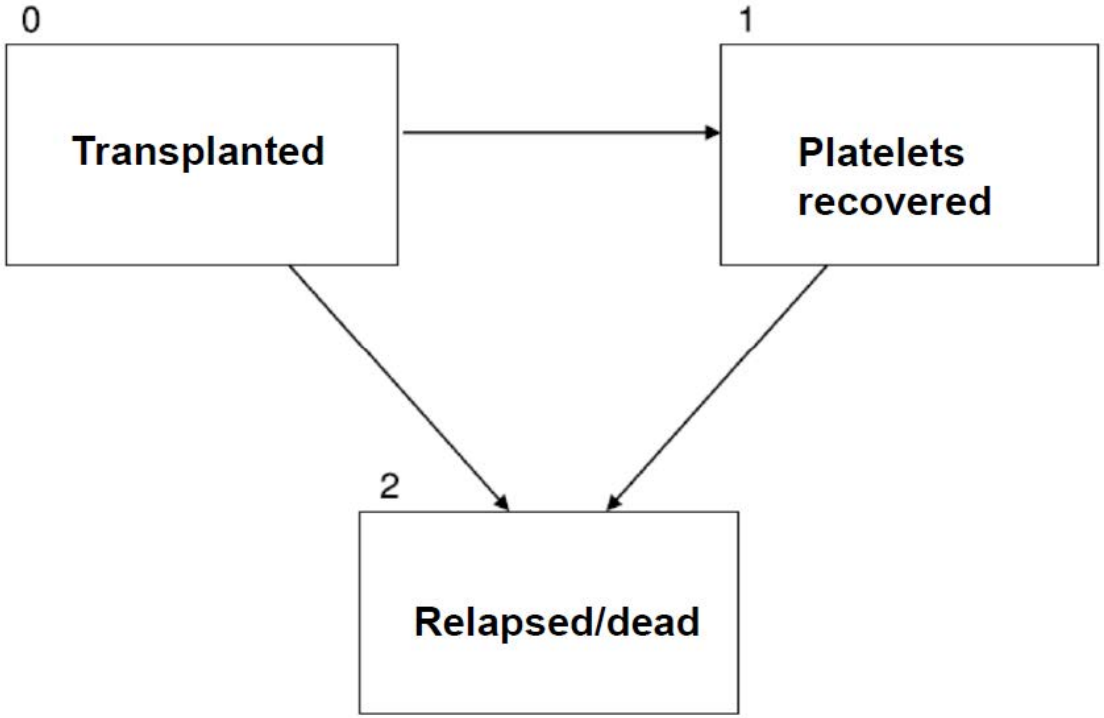


Fig. 3.19 *An illness-death model without recovery.*

Based on a sample from a population, we let $N_{gh}(t)$, for $(g,h) = (0,1),(0,2),(1,2)$, count the number of observed transitions from state g to state h in $[0,t]$, and let $Y_g(t)$ be the number of individuals in state g just prior to time t.

The intensity process of $N_{gh}(t)$ takes the multiplicative form $\lambda_{gh}(t) = \alpha_{gh}(t)Y_g(t)$, so we may use the Nelson-Aalen estimator $\hat{A}_{gh}(t)$ to estimate $A_{gh}(t) = \int_0^t \alpha_{gh}(u)du$.

Example 3.16: Bone marrow transplantation



- Platelet recovery, relapse and death for bone marrow transplant patients.
- 137 patients with acute leukemia have had a bone marrow transplantation.
- The time of the events "platelet recovery" and "death/relapse" are recorded

Example 3.16: Bone marrow transplantation.
Estimation of cumulative transition intensities

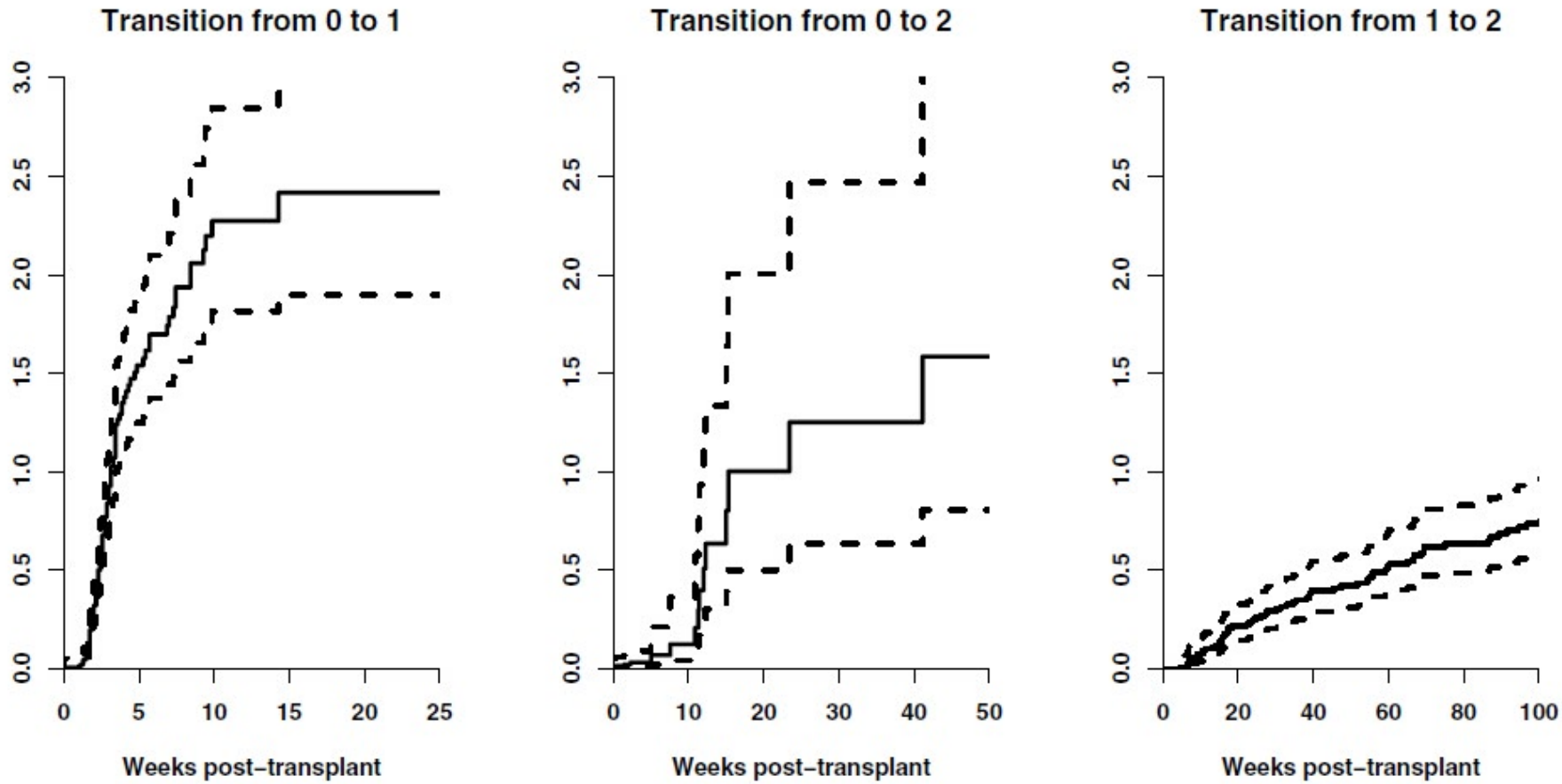


Fig. 3.20 *Nelson-Aalen estimates with log-transformed 95% confidence intervals of the cumulative transition intensities for the bone marrow transplant patients. The states are 0: "transplanted," 1: "platelet recovered," and 2: "relapsed or dead." Note that the time scale is not the same for the three estimates.*

Estimation of transition probabilities

$$\textit{Recall:} \quad P_{00}(s,t) = \exp\left\{-\int_s^t [\alpha_{01}(u)+\alpha_{02}(u)]du\right\}$$

$$P_{11}(s,t) = \exp\left\{-\int_s^t \alpha_{12}(u)du\right\}$$

$$P_{01}(s,t) = \int_s^t P_{00}(s,u)\alpha_{01}(u)P_{11}(u,t)du$$

This suggests the estimators:

$$\hat{P}_{00}(s,t) = \prod_{s<T_j\le t}\left(1-\frac{\Delta N_{0\bullet}(T_j)}{Y_0(T_j)}\right), \quad \text{where } N_{0\bullet}(t)=N_{01}(t)+N_{02}(t)$$

$$\hat{P}_{11}(s,t) = \prod_{s<T_j\le t}\left(1-\frac{\Delta N_{12}(T_j)}{Y_1(T_j)}\right)$$

$$\hat{P}_{01}(s,t) = \sum_{s<T_j\le t}\hat{P}_{00}(s,T_{j-1})d\hat{A}_{01}(T_j)\hat{P}_{11}(T_j,t)$$

Example 3.16: Bone marrow transplantation.
Estimation of transition probabilities $P_{01}(0, t)$

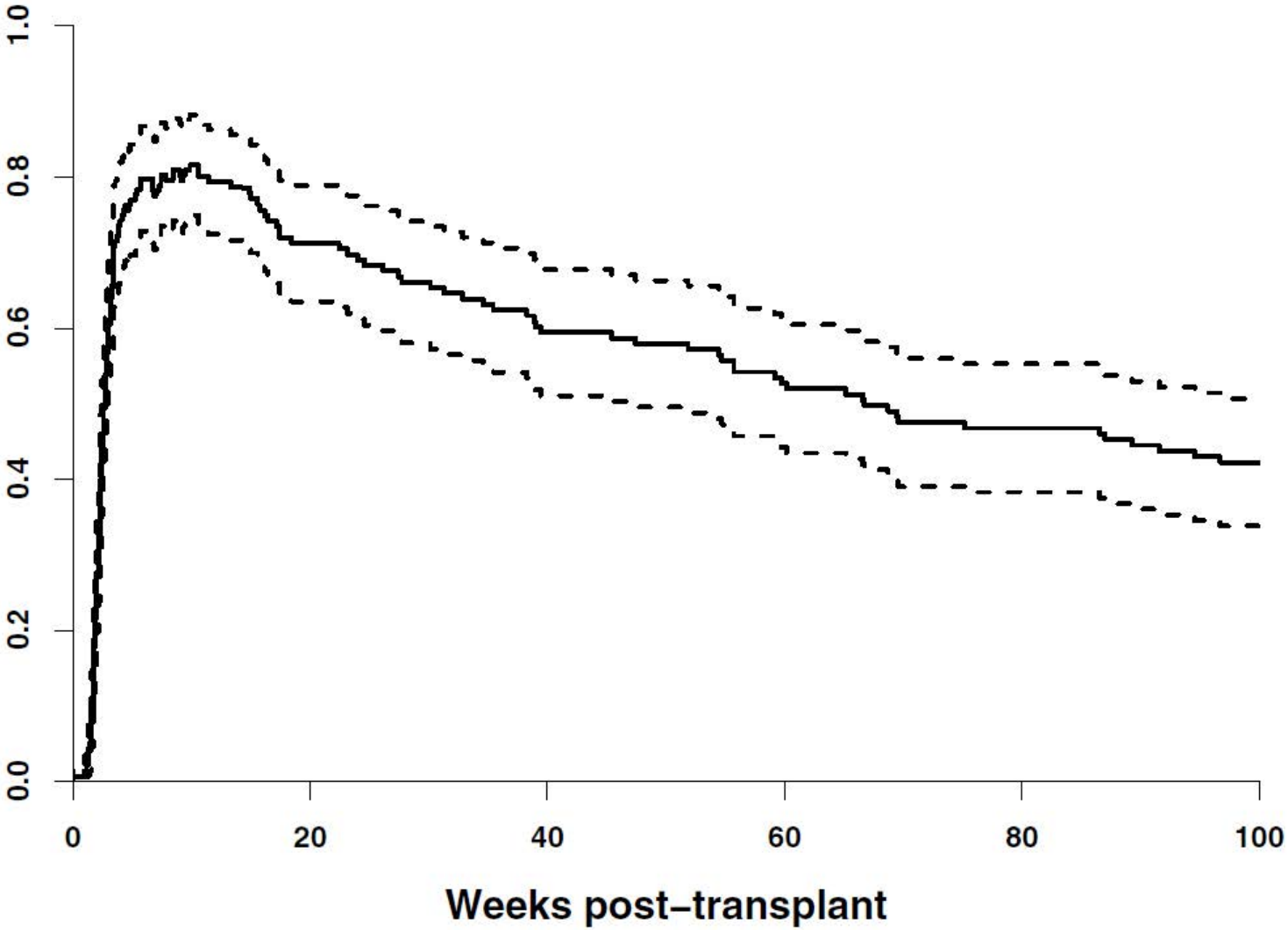


Fig. 3.21 *Empirical probability of being in response function* $\widehat{P}_{01}(0,t)$ *for the bone marrow transplant patients.*

A general Markov chain

Consider now a Markov chain $X(t)$ with state space $\mathcal{S} = \{0, 1, \ldots, k\}$ with transition *probabilities*

$$P_{gh}(s,t) = P\left(X(t) = h \mid X(s) = g\right), \; s < t, \; g, h \in \mathcal{S}$$

and transition *intensities*

$$\alpha_{gh}(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(X(t + \Delta t) = h \mid X(t-) = g) \;\text{ for }\; g \neq h$$

Consider the transition probability matrix

$$\mathbf{P}(s,t) = \begin{pmatrix} P_{00}(st) & P_{01}(s,t) & \cdots & P_{0k}(s,t) \\ P_{10}(st) & P_{11}(s,t) & \cdots & P_{1k}(s,t) \\ \cdots & \cdots & \cdots & \cdots \\ P_{k0}(st) & P_{k1}(s,t) & \cdots & P_{kk}(s,t) \end{pmatrix}$$

The key is to represent this matrix in terms of the transition intensities.

The Aalen-Johansen estimator

We have

$$\begin{aligned} P_{gh}(u, u+du) &= \alpha_{gh}(u)du; \quad g \neq h \\ P_{gg}(u, u+du) &= 1 - \sum_{h \neq g} \alpha_{gh}(u)du \end{aligned}$$

It can be deduced from this that the matrix $\mathbf{P}(s,t)$ can be represented as a matrix-valued product-integral

$$\mathbf{P}(s,t) = \prod_{s \leq u \leq t} (\mathbf{I} + d\mathbf{A}(u)) \qquad (*)$$

where $\mathbf{A}(t) = (A_{gh}(t))$ with $A_{gh}(t) = \int_0^t \alpha_{gh}(u)du$ being the *cumulative* transition intensities.

The $A_{gh}(t)$ are estimated by "ordinary" Nelson-Aalen estimators. and the $\mathbf{P}(s,t)$ are then estimated by plugging in the Nelson-Aalen estimates in (*). This corresponds to what we did in the special cases, and leads to the so-called *Aalen-Johansen* estimator.