Regression models for competing risks

Per Kragh Andersen

Section of Biostatistics, University of Copenhagen

DSBS Course
Survival Analysis in Clinical Trials
January 2018

Overview

- Definitions and example
- Models for cause-specific hazards
- Direct models for cumulative incidences
- Alternative models, summary

`www.biostat.ku.dk/~pka/SACT18-part1`

Definitions and example

Another example

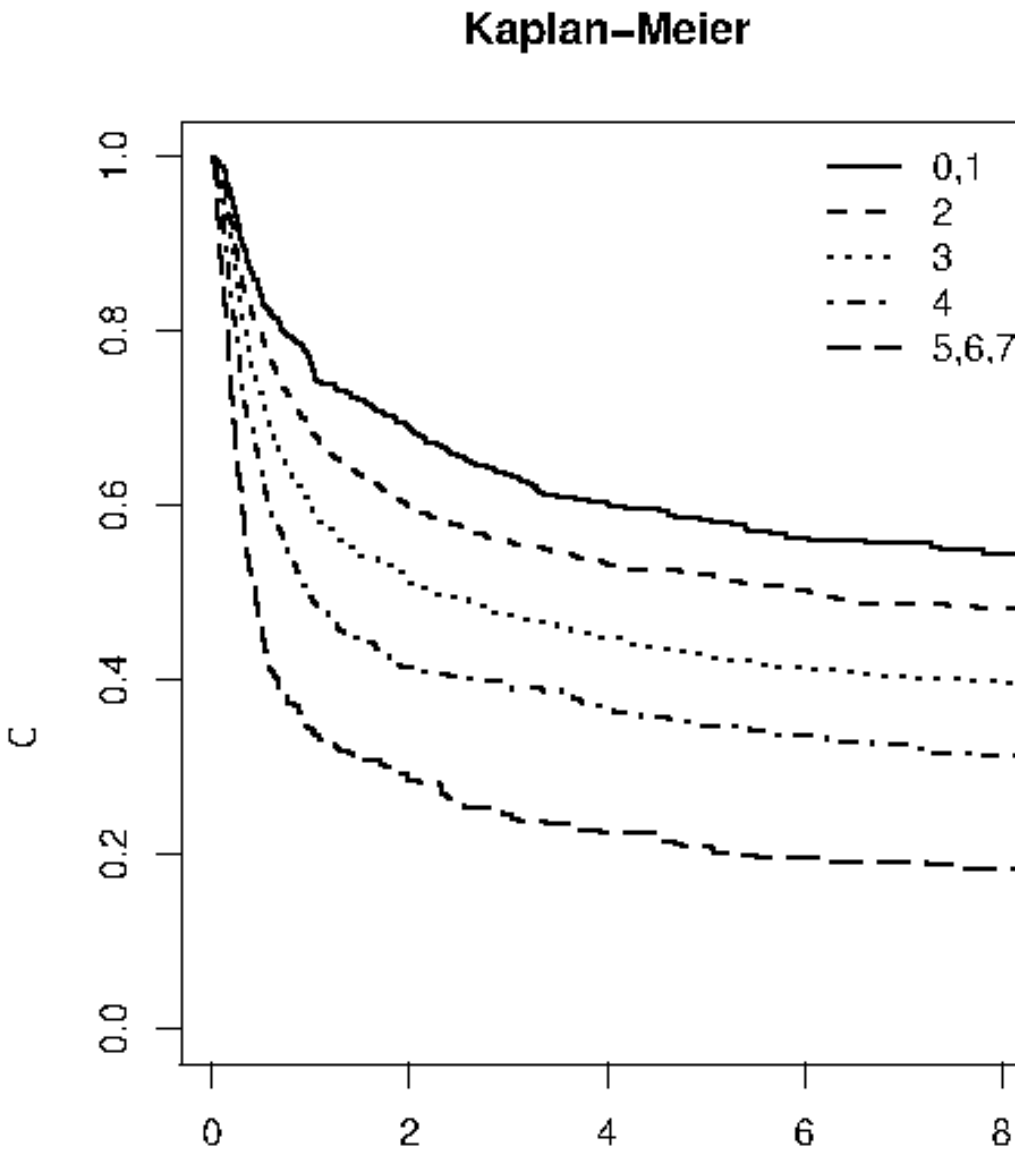
- Data from the European group for Blood and Marrow Transplantation (EBMT)
- All (3982) chronic myeloid leukemia (CML) patients with an allogeneic stem cell transplantation from an HLA-identical sibling or a matched unrelated donor during the years 1997–2000.
- Study effect of EBMT risk score with values 0–7, here grouped into five groups: 0, 1 ($n = 506$), 2 ($n = 1159$), 3 ($n = 1218$), 4 ($n = 745$), and 5, 6, 7 ($n = 354$).
- Points obtained from: donor type (2), stage (3), age (3: 20,40), female-to-male (2), time from diagnosis (2: 12 mo.)
- Failure from transplantation may either be due to relapse or to non-relapse mortality (NRM). Often these two endpoints are taken together to relapse-free survival (RFS), which is the time from transplantation to either relapse or death, whichever comes first.

Summary table

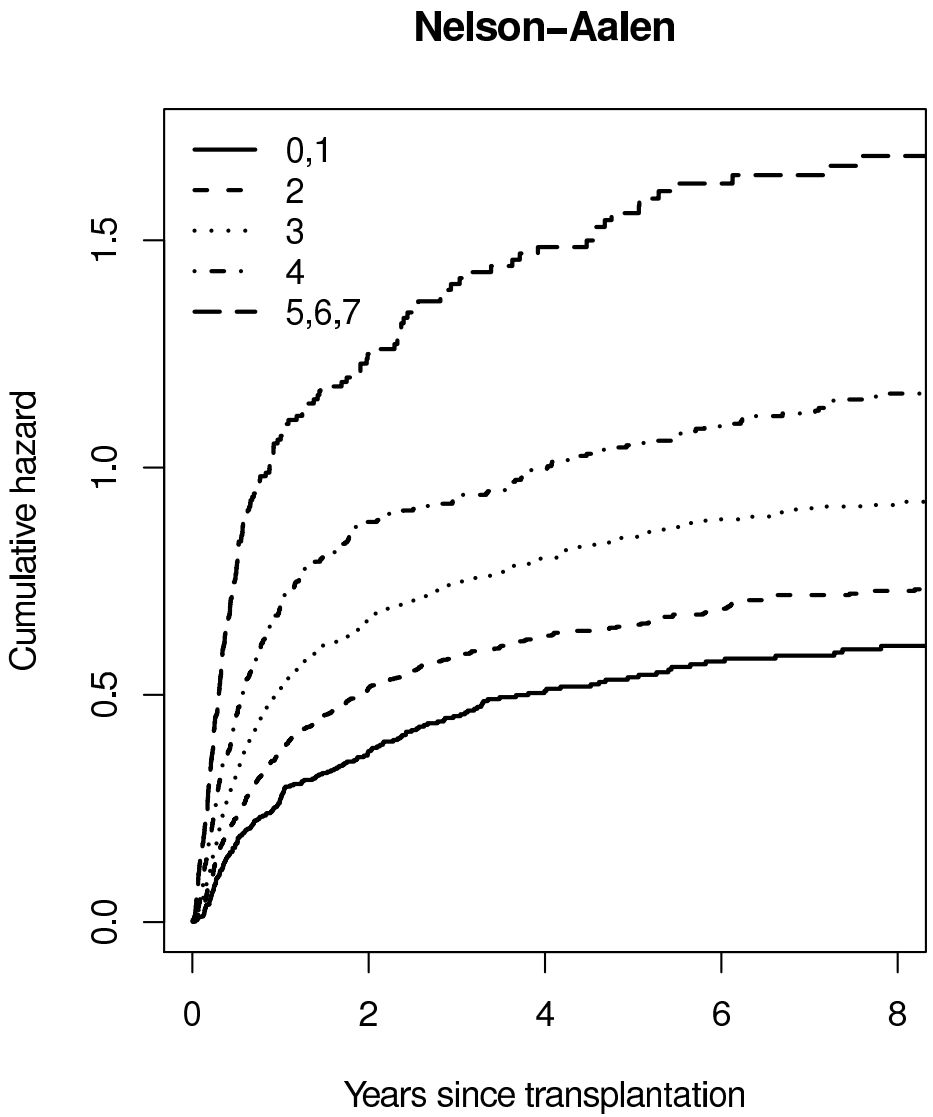
EBMT risk group	Relapse *n* (%)	NRM *n* (%)	Censored *n* (%)	Total *n* (%)
0,1	113 (22.3)	94 (18.6)	299 (59.1)	506 (100)
2	247 (21.3)	323 (27.9)	589 (50.8)	1159 (100)
3	292 (24.0)	404 (33.2)	522 (42.9)	1218 (100)
4	193 (25.9)	300 (40.3)	252 (33.8)	745 (100)
5,6,7	112 (31.6)	169 (47.7)	73 (20.6)	354 (100)

Next, we study RFS in relation to risk group.

Kaplan-Meier curves



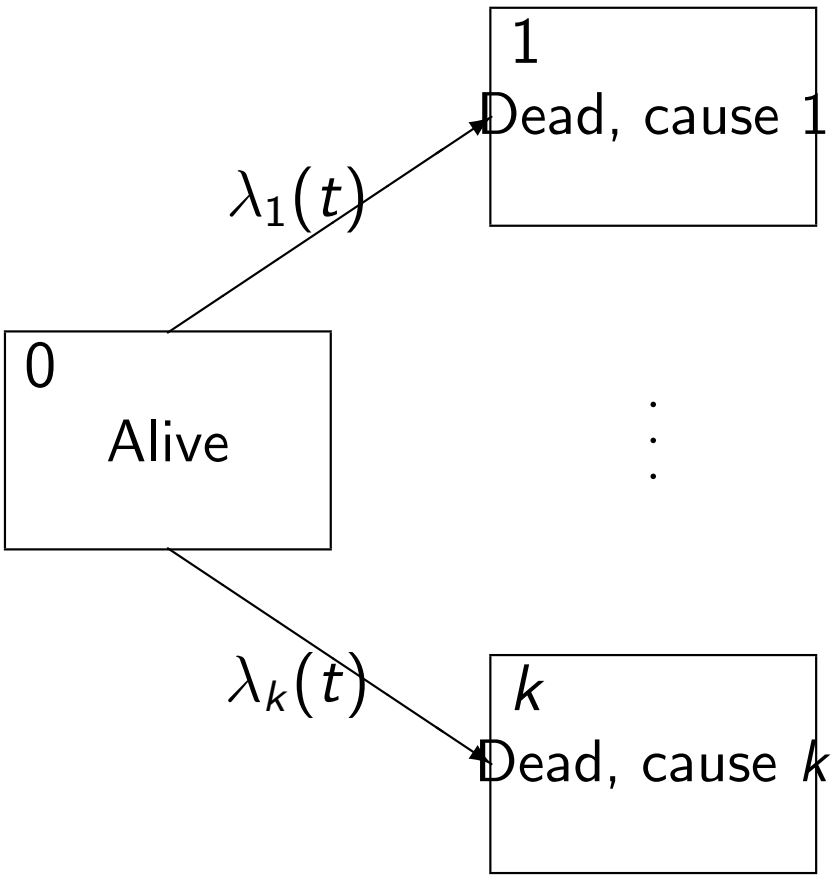
Nelson-Aalen curves



Doing it in SAS

```
/* Kaplan-Meier for risk groups: RFS */

PROC PHREG DATA=ebmt PLOT(OVERLAY=ROW)=SURV;
     MODEL days*dc(0)=;
     STRATA riskscore;
     BASELINE / METHOD=PL;
RUN;

/* Nelson-Aalen for risk groups: RFS */

PROC PHREG DATA=ebmt PLOT(OVERLAY=ROW)=CUMHAZ;
     MODEL days*dc(0)=;
     STRATA riskscore;
RUN;
```

The competing risks multi-state model



Basic parameters

Cause-specific hazards $j = 1, 2, \ldots$ ("transition intensities"):

$$\lambda_j(t) \approx P(\text{state } j \text{ time } t + dt \mid \text{state } 0 \text{ time } t)/dt.$$

State occupation probabilities:

1. Overall survival function:

$$\begin{aligned} S(t) &= P(\text{alive time } t) \\ &= \exp\Big(-\int_0^t \sum_j \lambda_j(u)du\Big). \end{aligned}$$

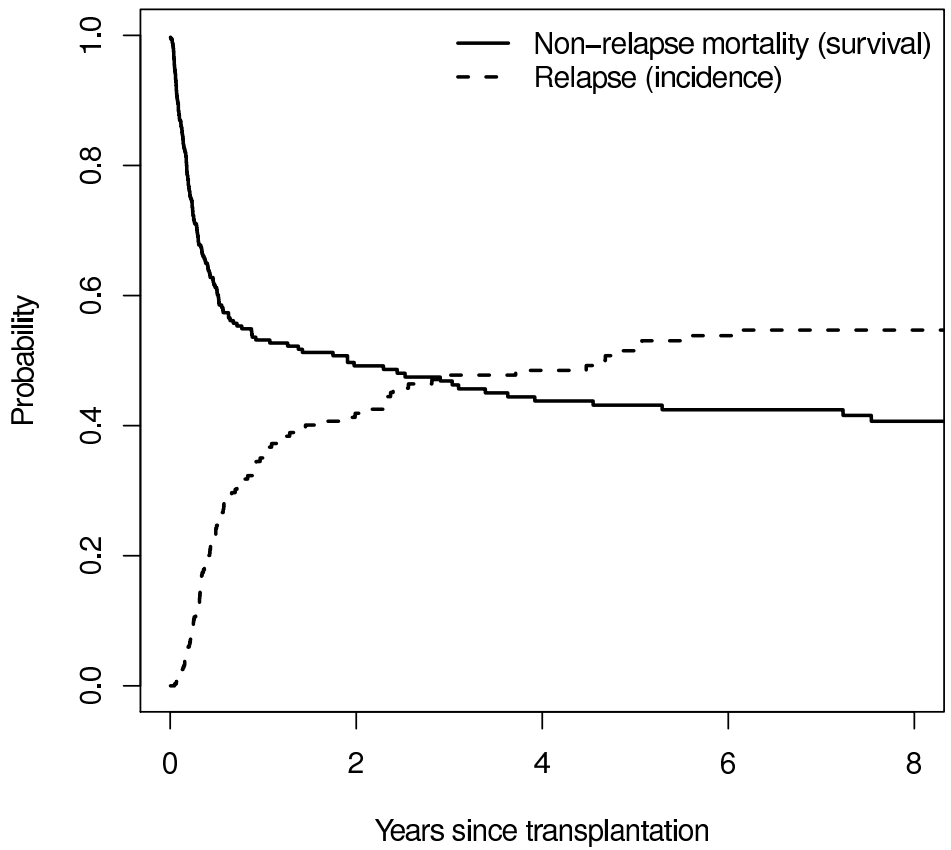
2. Cumulative incidences $j = 1, 2, \ldots$:

$$\begin{aligned} F_j(t) &= P(\text{dead from cause } j \text{ before time } t) \\ &= \int_0^t S(u)\lambda_j(u)du. \end{aligned}$$

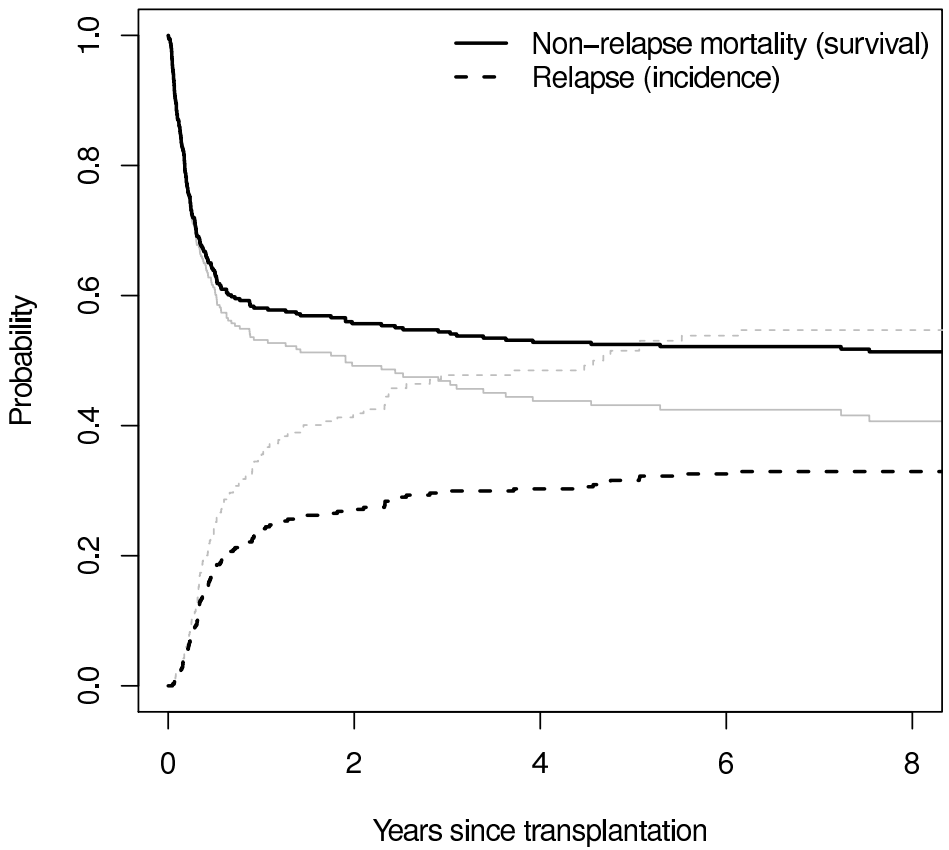
Cumulative incidence. vs. 1-KM

We look at risk group 5,6,7 and compare the 1-Kaplan-Meier estimates with the correct Aalen-Johansen estimates for relapse and for NRM.

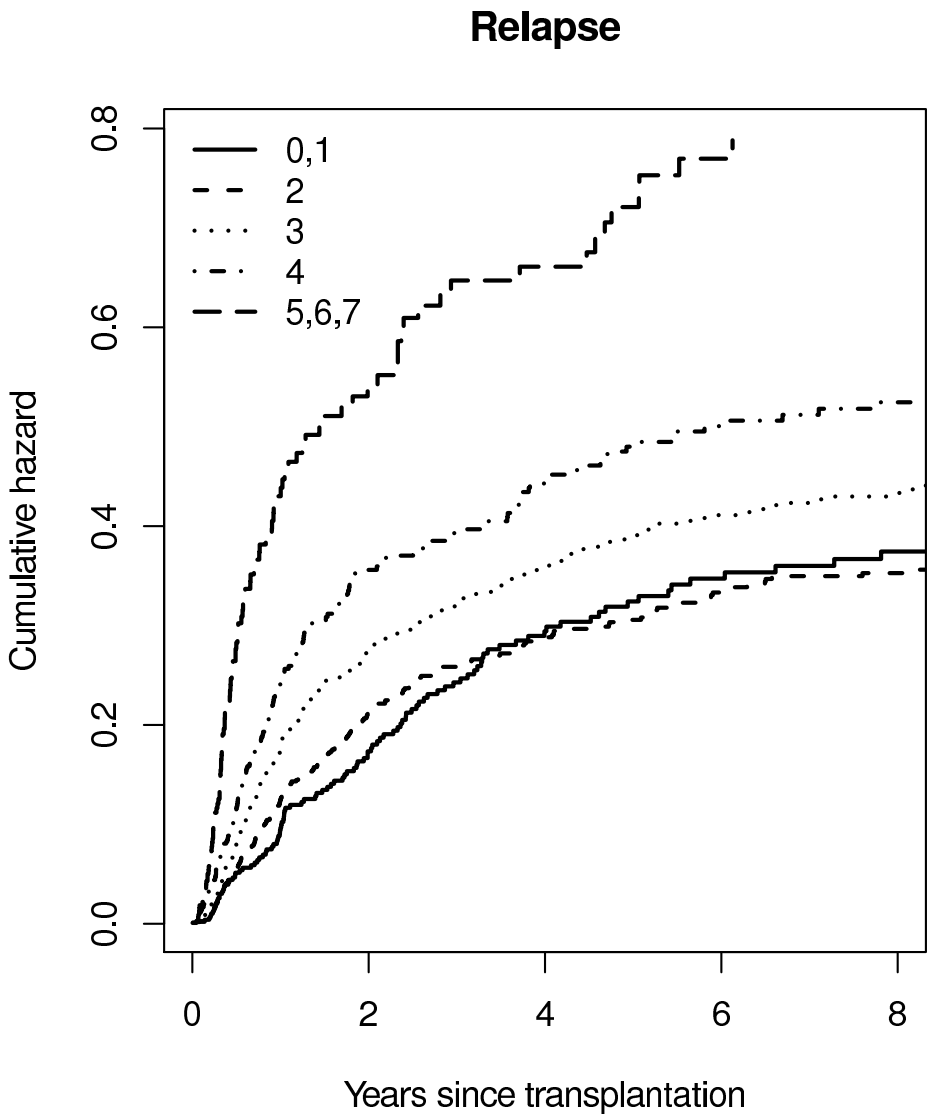
Kaplan-Meier curves



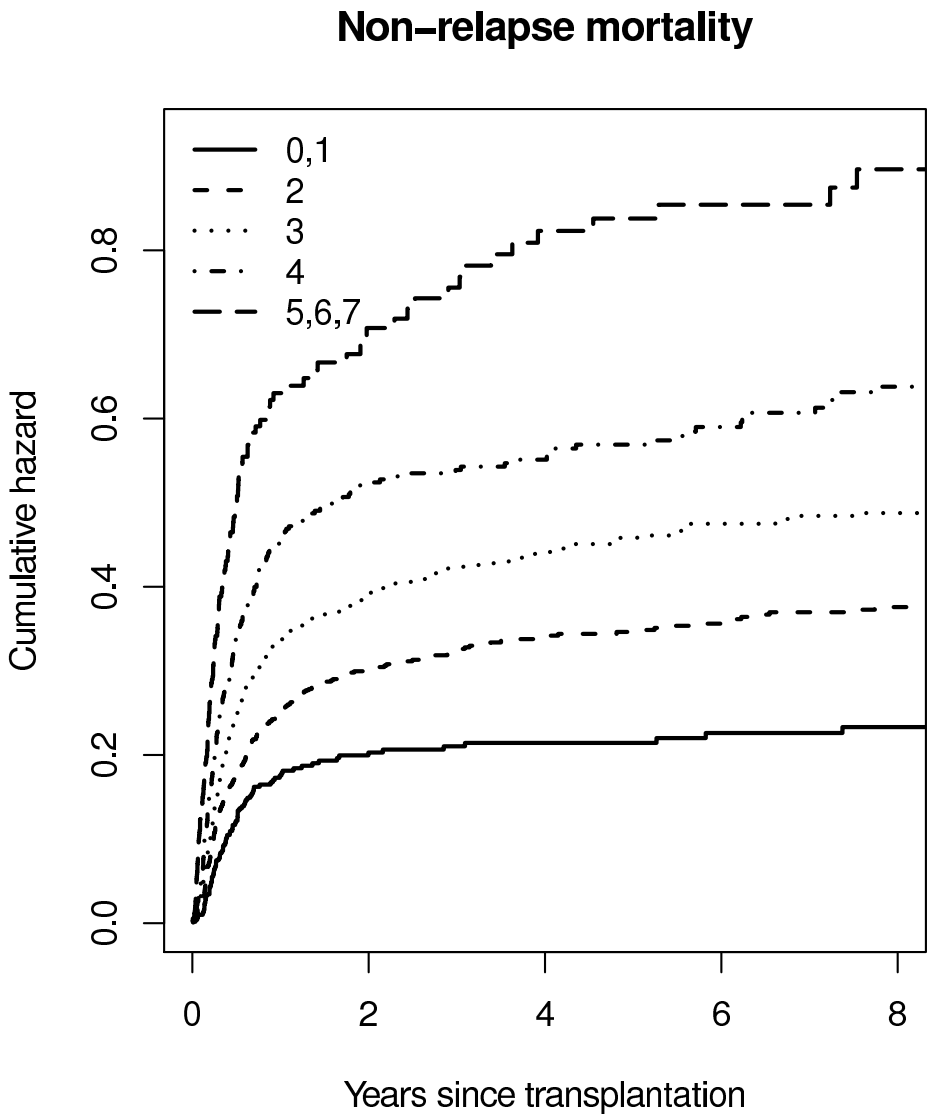
Cumulative incidence curves



Nelson-Aalen curves: relapse



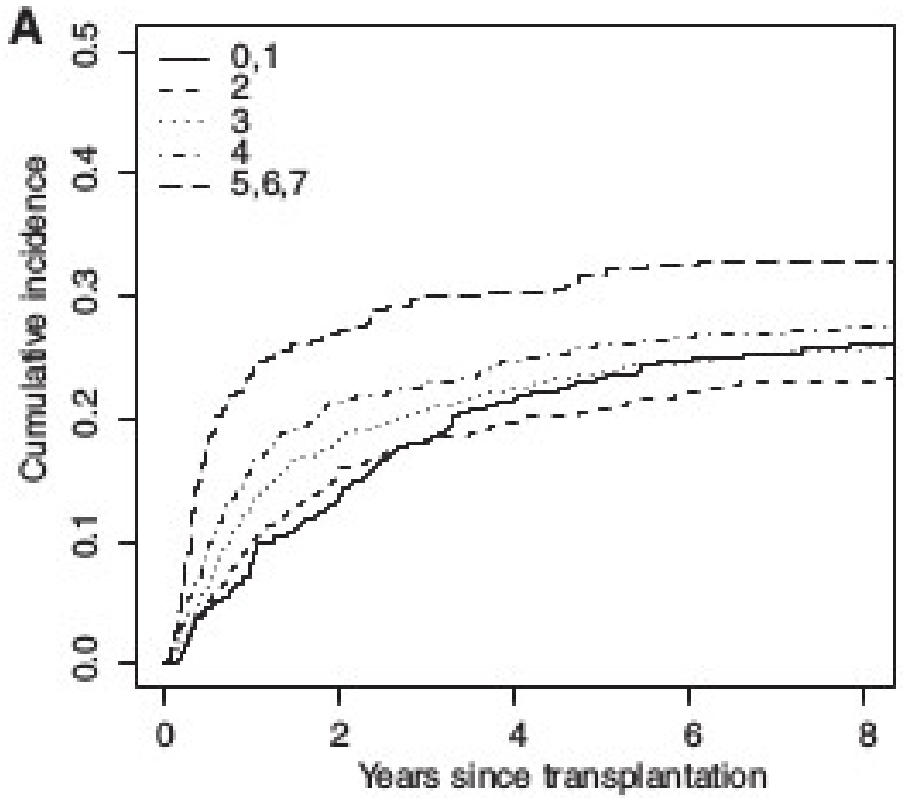
Nelson-Aalen curves: NRM



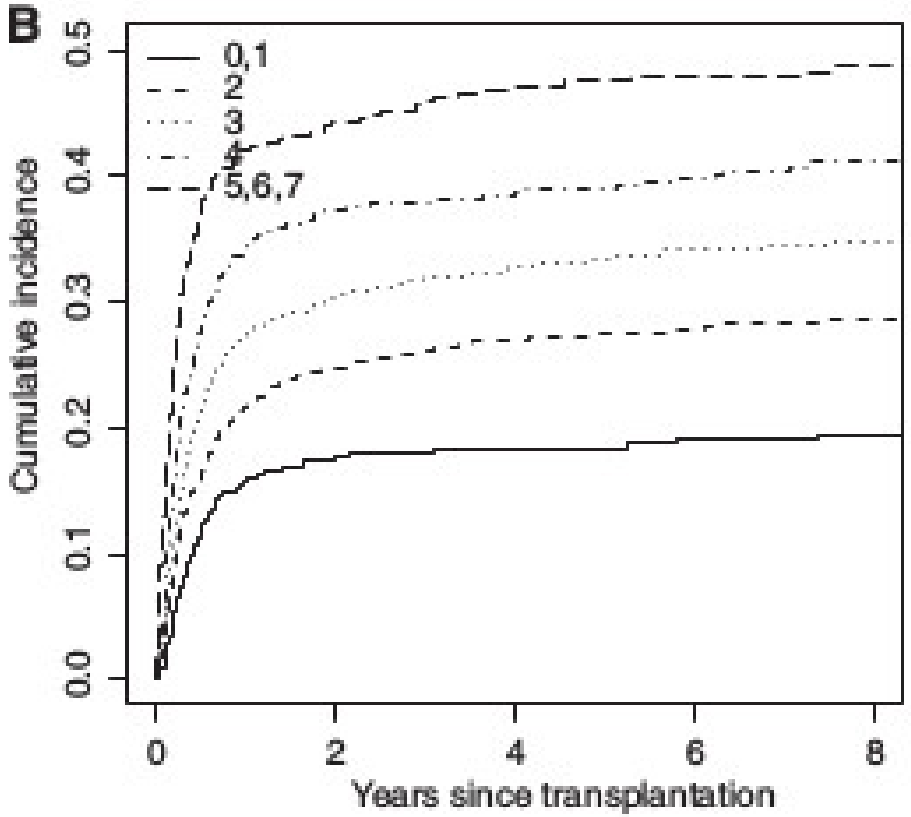
Doing it in SAS

```
/* Nelson-Aalen for risk groups: Relapse */

PROC PHREG DATA=ebmt PLOT(OVERLAY=ROW)=CUMHAZ;
     MODEL days*dc(0 2)=;
          STRATA riskscore;
RUN;

/* Nelson-Aalen for risk groups: NRM */

PROC PHREG DATA=ebmt PLOT(OVERLAY=ROW)=CUMHAZ;
     MODEL days*dc(0 1)=;
          STRATA riskscore;
RUN;
```

Aalen-Johansen curves: relapse



Aalen-Johansen curves: NRM



Doing it in SAS

```
/* Cumulative incidence for risk groups:
   Relapse */

PROC PHREG DATA=ebmt PLOT(OVERLAY=ROW)=CIF;
    MODEL days*dc(0)=/EVENTCODE=1;
        STRATA riskscore;
RUN;

/* Cumulative incidence for risk groups: NRM */

PROC PHREG DATA=ebmt PLOT(OVERLAY=ROW)=CIF;
    MODEL days*dc(0)=/EVENTCODE=2;
        STRATA riskscore;
RUN;
```

NB: `PROC PHREG` fits 'an empty Fine-Gray model' (more later) and reports the baseline estimate which is not quite Aalen-Johansen.

Models for cause-specific hazards

Likelihood

Data: $(\widetilde{T}_i, D_i)$, $\quad i = 1, ..., n$ where $D_i = j, j = 1, ..., k$ if observed failure from cause j, $D_i = 0$ if censored.
Likelihood:

$$
\begin{aligned}
L &= \prod_{i=1}^{n} S(\widetilde{T}_i) \prod_{j=1}^{k} (\lambda_j(\widetilde{T}_i))^{I(D_i=j)} \\
&= \prod_{i=1}^{n} \left(\exp(-\sum_{j=1}^{k} \Lambda_j(\widetilde{T}_i))\right) \prod_{j=1}^{k} (\lambda_j(\widetilde{T}_i))^{I(D_i=j)} \\
&= \prod_{j=1}^{k} \left(\prod_{i=1}^{n} \exp(-\Lambda_j(\widetilde{T}_i))(\lambda_j(\widetilde{T}_i))^{I(D_i=j)}\right).
\end{aligned}
$$

Inference for cause-specific hazards

Note:

- Product over causes, j,
- The jth factor is what we would get if only that cause were studied *and all other causes were right-censorings*
- This has nothing to do with "independence" of causes - it is solely a consequence of the definition of cause-specific hazards as hazards of exclusive events.
- It means that all standard hazard-based models for survival data apply when analyzing cause-specific hazards
  - non-parametric: estimate $\Lambda_j(t) = \int_0^t \lambda_j(u)\mathrm{d}u, j = 1, \ldots, k$ by Nelson-Aalen estimator, compare using, e.g. logrank tests
  - parametric models
  - Cox regression, (Poisson regression, Aalen model)

Cox models for cause-specific hazards

Model for cause j:

$$\lambda_j(t \mid X) = \lambda_{0j}(t) \exp(\beta_j^{\mathsf{T}} X),$$

that is, separate baseline hazards and separate regression coefficients for each cause.

It is technically possible to fit Cox models for cause-specific hazards with

- identical or proportional baselines for some causes
- regression coefficients that are shared between several causes

However, that is rarely relevant!

These features may be more relevant for other multi-state models than the competing risks model - more in Part II of the course.

Cox models for cause-specific hazards

Fit the model for one cause at a time and, technically, consider failures from other causes as censored observations. This provides the correct likelihood.

```
/* Cox model for risk groups: Relapse */

PROC PHREG DATA=ebmt;
     CLASS riskscore (REF="0");
     MODEL days*dc(0 2)=riskscore/RL;
RUN;

/* Cox model for risk groups: NRM */

PROC PHREG DATA=ebmt;
     CLASS riskscore (REF="0");
     MODEL days*dc(0 1)=riskscore/RL;
RUN;
```

EBMT: Cox models for cause-specific hazards

EMBT risk group	Relapse HR (95% ci)	NRM HR (95% ci)
0,1		
2	1.01 (0.81–1.27)	1.57 (1.25–1.97)
3	1.28 (1.03–1.59)	2.01 (1.61–2.52)
4	1.57 (1.25–1.99)	2.68 (2.12–3.37)
5,6,7	2.67 (2.06–3.47)	3.98 (3.09–5.13)

Same rate of relapse in group 2 as in group 0,1.

Estimation of cumulative incidences from hazards

Estimate $F_j(t \mid X)$ by plug-in:

$$\widehat{F}_j(t \mid X) = \int_0^t \widehat{S}(u- \mid X) d\widehat{\Lambda}_j(u \mid X).$$

Here,

$$\widehat{\Lambda}_j(u \mid X) = \widehat{\Lambda}_{j0}(u) \exp(\widehat{\beta}_{j1} X_1 + ... + \widehat{\beta}_{jp} X_p)$$

is the cumulative cause-j-hazard estimate from the Cox model and $\widehat{S}(u \mid X)$ the Cox model based estimator for the overall survival function, e.g.,

$$\widehat{S}(u \mid X) = \exp\left(- \sum_j \widehat{\Lambda}_j(u \mid X) \right),$$

or, preferably, the corresponding product-integral estimator.

Estimation of cumulative incidences from hazards

To do this in SAS, we need 'simultaneous access' to results from all (k) Cox models. One way of doing this is to use a 'data duplication trick'. For $k = 2$ causes:

1. create a data set with $2n$ records, and in both versions of the data, keep the `id` (if relevant) and `time` variables
2. in the first version, set `version=1;` and in the second set `version=2;`
3. for all covariates `cov` (NB:numerical), set `cov1=cov*(version=1);` in the first version and set `cov2=cov*(version=2);` in the second version
4. define the new failure indicator: `fail=(version=1)*(cause=1)+(version=2)*(cause=2);`
5. fit the Cox model

```
MODEL time*fail(0)=cov1 cov2/RL;
STRATA version;
```

and keep baseline hazards using a `BASELINE` statement.

SAS code

Data set `single` with `n` records and variables `time, cause, cov`:

```
DATA double; SET single single;
version=1+(_N_ > n);
fail=(version=1)*(cause=1)+(version=2)*(cause=2);
cov1=cov*(version=1); cov2=cov*(version=2);
RUN;
```

Note: `cov` should be numerical, i.e., categorical variables should be represented by dummies.
Models with *common effects of the covariate* may be obtained by replacing `cov1 cov2` in the `MODEL` statement by the original variable `cov`.

Estimation of cumulative incidences from hazards

A (non-user friendly!) `SAS MACRO` called `CUMINC` is available from:
`www.biostat.ku.dk/~pka`
for computing the cumulative incidences for given covariate patterns:

1. create a data set `pattern` with the desired covariate patterns
2. fit a stratified Cox model and keep the baseline hazard
3. invoke the macro

The macro creates a data set (by default called 'data') including the estimated cumulative incidence for each cause for the specified covariate patterns.

SAS code

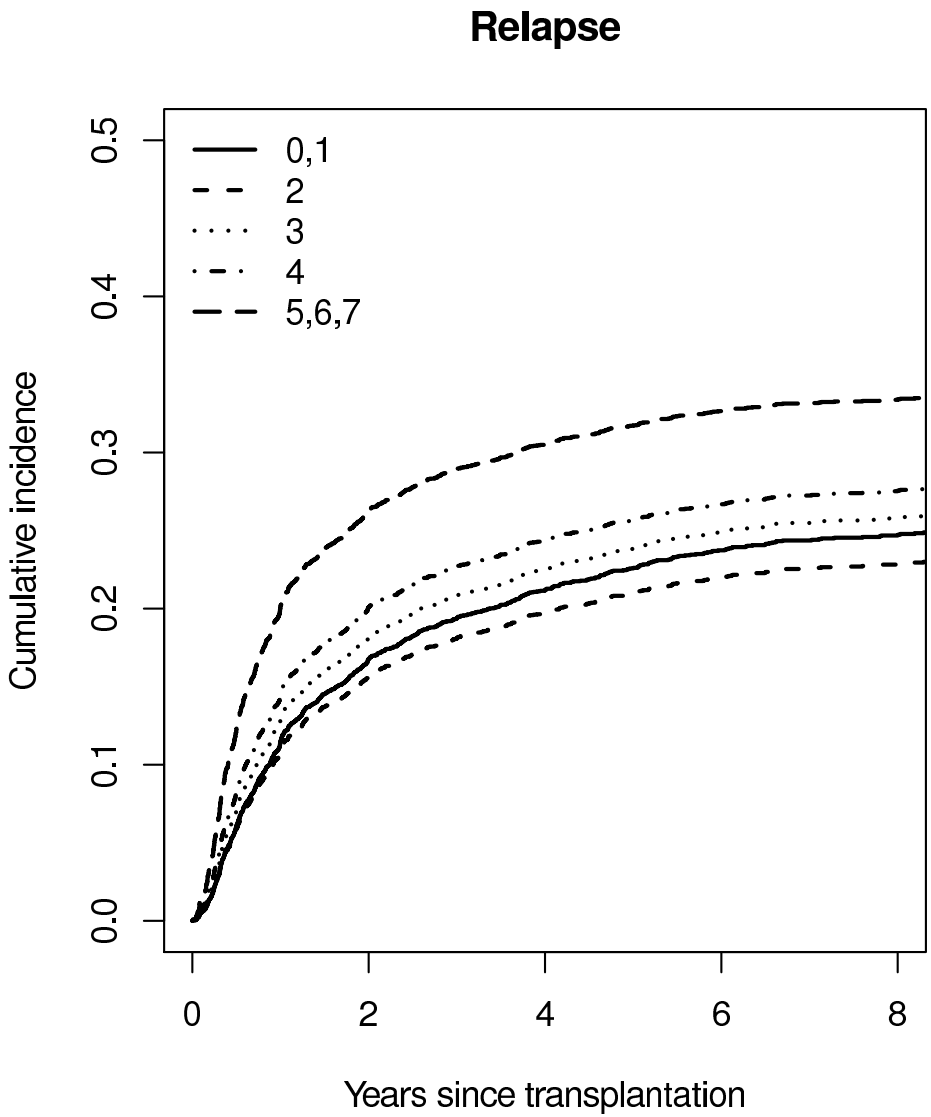
```
DATA pattern;
INPUT cov1 cov2;
DATALINES;
47 0
0 47
; /* For each pattern (here 1): 1 line for each
    cause */
RUN;

PROC PHREG DATA=double;
MODEL time*fail(0)=cov1 cov2/RL;
STRATA version;
BASELINE OUT=ciData COVARIATES=pattern
    SURVIVAL=pred / NOMEAN METHOD=CH;
RUN;
```

%CUMINC(ciData,version,time,pred);

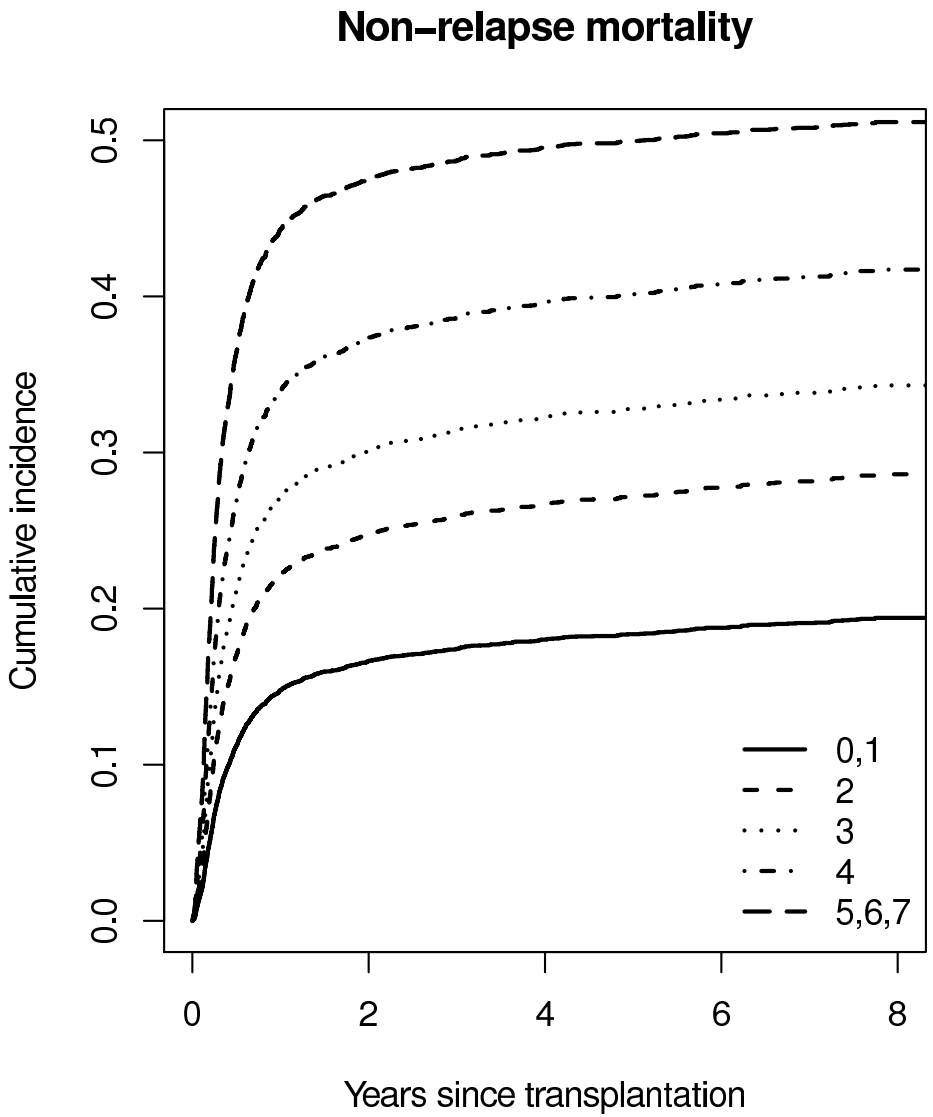
EBMT example

We now predict the cumulative incidences for relapse and NRM for each of the 5 EMBT risk groups based on Cox models for the two cause-specific hazards.
(Analyses were performed using the `mstate` package in `R` but could have been done using the `SAS MACRO` if the Cox models were fitted using dummy variables for the `riskscore`.)

EBMT example, relapse



EBMT example, NRM



Cumulative incidences from cause-specific Cox models

Important to notice:

- The Cox models impose a simple structure between covariates and *rates*.
- Due to the non-linear relationship between rates and risks in a competing risks model, this simple relationship does not carry over to the cumulative incidences.
- In particular, the way in which a covariate affects a rate can be different from the way in which it affects the corresponding risk: this will depend on how it affects the rates for the competing causes.
- EBMT example: group 2 vs. 0,1, relapse

Direct models for the cumulative incidence

Cumulative incidence regression models

The fact that plugging-in cause-specific hazard models does not provide parameters that in a simple way describe the relationship between covariates and cumulative incidences has led to the development of direct regression models for the cumulative incidences.

The most widely used such model is the *Fine-Gray* model. Recall from a Cox model for all-cause mortality that:

$$\log(-\log(1 - F(t \mid X))) = \log(\Lambda_0(t)) + \beta^{\mathsf{T}} X.$$

Fine & Gray (1999, *JASA*) studied the similar model for a cumulative incidence:

$$\log(-\log(1 - F_j(t \mid X))) = \log(\widetilde{\Lambda}_{0j}(t)) + \widetilde{\beta}_j^{\mathsf{T}} X.$$

The Fine-Gray model

This is a model for the hazard for the improper random variable

$$T_j^* = T \cdot I(D = j) + \infty \cdot I(D \neq j),$$

i.e. for

$$\tilde{\lambda}_j(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \log(1 - F_j(t)).$$

That is, the transformation which for all-cause mortality takes us from cumulative risk to hazard is used for a cumulative incidence in a competing risks model.

The resulting $\widetilde{\lambda}_j(t)$ is denoted the *sub-distribution hazard* and the Fine-Gray model is thus a proportional sub-distribution hazards model.

The Fine-Gray model

A problem is that, while the hazard function has the useful "rate" interpretation:

$$\lambda(t) \approx P(\text{die before } t + dt \mid \text{ alive } t)/dt, \quad dt \text{ small},$$

and so has the cause-specific hazard:

$$\lambda_1(t) \approx P(\text{die from cause 1 before } t + dt \mid \text{ alive } t)/dt, \quad dt \text{ small},$$

the sub-distribution hazard has *not*. Thus

$$\widetilde{\lambda}_1(t) \approx P(\text{die from cause 1 before } t + dt \mid$$
either alive at t *or* dead from a competing cause by $t)/dt, \quad dt$ small.

The Fine-Gray model

The model for the sub-distribution hazard is:

$$\widetilde{\lambda}_j(t \mid X) = \widetilde{\lambda}_{0j}(t) \exp(\widetilde{\beta}_1 X_1 + ... + \widetilde{\beta}_p X_p),$$

but, while a "sub-distribution hazard" sounds like a hazard, it is not! Therefore, the resulting parameters $\exp(\widetilde{\beta})$ in the Fine-Gray model have a rather indirect interpretation as "sub-distribution hazard ratios".
Sadly, `SAS` has chosen to implement the model in `PROC PHREG` which means that the output is identical to that from a standard Cox model.

Anyway, the model is being used quite a bit and it is, indeed, useful by giving parameters that directly link the cumulative incidence to covariates.

Math

With no censoring, Fine and Gray defined the cause j "risk set"

$$\widetilde{R}_j(t) = \{i : (T_i \geq t) \text{ or } (T_i \leq t, D_i \neq j)\}$$

and $\widetilde{\beta}_j$ is estimated by the partial likelihood score equation

$$U_j(\widetilde{\beta}_j) = \sum_i I(D_i = j)\Big(X_i - \frac{\sum_{m \in \widetilde{R}_j(T_i)} X_m \exp(\widetilde{\beta}_j^{\mathsf{T}} X_m)}{\sum_{m \in \widetilde{R}_j(T_i)} \exp(\widetilde{\beta}_j^{\mathsf{T}} X_m)}\Big) = 0$$

corresponding to replacing times of failures from causes other than j by $+\infty$.
With known (e.g., "administrative") censoring (at U_i), the cause j risk set is replaced by

$$\widetilde{R}_j(t) = \{i : (T_i \wedge U_i \geq t) \text{ or } (T_i \leq t, D_i \neq j, U_i \geq t)\}.$$

Math

- To identify this 'risk set', we need to know the time U of censoring for a subject who failed.
- With general censoring, an Inverse Probability of Censoring Weighted (IPCW) score equation is used and to use this, a model for censoring is needed.
- In the simplest case, one uses the 'Kaplan-Meier for censoring', that is, estimating $P(U > t)$. (In this analysis 'failures are censorings'.)
- If censoring depends on covariates then a model for $P(U > t \mid X)$ is needed for the weights, e.g. a Cox model.
- IPCW is a technique which is often used when analyzing *recurrent events* - more in Part II of the course.

The Fine-Gray model

The Fine-Gray model provides parameters describing the relationship between the covariates and the cause j risk. For example, for a binary covariate X_1 with an estimated regression coefficient $\widetilde{\beta}_1 > 0$ it follows that for all values, X_2^0, for the other covariates in the model we have that

$$\widehat{F}_j(t \mid X_1 = 1, X_2^0) > \widehat{F}_j(t \mid X_1 = 0, X_2^0).$$

The positive regression coefficient has the *qualitative* meaning that individuals with $X_1 = 1$ have a uniformly increased cause j cumulative incidence compared to those with $X_1 = 0$.
However, the resulting estimates $\exp(\widetilde{\beta}_j)$ are sub-distribution hazard ratios, so the *quantitative* meaning of the regression coefficient is not simple.
The model is related to the Gray (1988, *Ann. Statist.*) test for comparison of cumulative incidences.

Fine-Gray models for EBMT data

EMBT risk	Relapse			NRM		
group	$\widetilde{\beta}$	SD	exp($\widetilde{\beta}$) (95% ci)	$\widetilde{\beta}$	SD	exp($\widetilde{\beta}$) (95% ci)
0,1						
2	-0.068	0.111	0.93 (0.75–1.16)	0.443	0.116	1.56 (1.24–1.96)
3	0.072	0.108	1.07 (0.87–1.33)	0.661	0.114	1.94 (1.55–2.42)
4	0.161	0.117	1.17 (0.93–1.48)	0.906	0.118	2.48 (1.96–3.12)
5,6,7	0.439	0.135	1.55 (1.19–2.02)	1.185	0.131	3.27 (2.53–4.22)

Somewhat lower risk of relapse in group 2 than in group 0,1.

Other link functions than $\log(-\log)$) - more to come!

SAS code

```
/* Fine-Gray for risk groups: relapse */

PROC PHREG DATA=ebmt;
     CLASS riskscore (REF="0");
     MODEL days*dc(0)=riskscore/EVENTCODE=1;
RUN;

/* Fine-Gray for risk groups: NRM */

PROC PHREG DATA=ebmt;
     CLASS riskscore (REF="0");
     MODEL days*dc(0)=riskscore/EVENTCODE=2;
RUN;
```

Alternative models, summary

Other models for the cumulative incidence

We have studied two classes of competing risks regression models - those based on rates and those based on risks.
When interest focuses on a single time point τ, the average cause-j risk difference for Z at τ may be estimated based on either approach using direct standardization (the g-formula):

$$\frac{1}{n}\Big(\sum_i \widehat{F}_j(\tau \mid Z=1, X_i) - \sum_i \widehat{F}_j(\tau \mid Z=0, X_i)\Big),$$

where $\widehat{F}_j$ is predicted from the regression model. For this approach, it does not matter whether or not regression parameters have nice interpretations.

But what if we want parameters with a nice interpretation for $F_j(\tau \mid X)$ or for several time points $\tau_1, ..., \tau_m$?

Pseudo-observations

Let $\widehat{F}_j$ be the Aalen-Johansen estimator and $\widehat{F}_j^{(-i)}$ the same estimator applied to the data set (of size $n-1$) obtained by eliminating subject i.
Then the *pseudo-observation* for the (possibly incompletely observed) cause j failure indicator $I(T_i \leq \tau, D_i = j)$ is:

$$F_{ji}(\tau) = n \cdot \widehat{F}_j(\tau) - (n-1) \cdot \widehat{F}_j^{(-i)}(\tau).$$

This may be used as response variable for a generalized linear model

$$g(F_j(\tau \mid X)) = \alpha_0 + \alpha_1 X_1 + ... + \alpha_p X_p$$

and parameters may be estimated by solving the 'usual kind' of GEE.
Choosing g =cloglog we get a Fine-Gray model at τ but other links (log, logit, identity) may provide parameters with a more direct interpretation.
`SAS MACRO` is available for computing the pseudo-observations.

Summary

- In studies of all-cause mortality, risks (probabilities, cumulative incidences) can be computed from rates (hazards) and vice versa - in other words the two functions contain *equivalent* information
- In studies of events which will not eventually happen for every one in the population, this is no longer the case and death (and maybe other events) are *competing risks* which need to be addressed
- In such cases, the risk of a given cause depends on the rates for *all* competing causes
- Therefore, using '1-Kaplan-Meier for a single cause' as a risk estimator is (upward) biased
- The magnitude of the bias depends on the frequency of the competing events

Summary (ctd.)

- A rather simple, unbiased estimator for the risk exists - the 'Aalen-Johansen' estimator
- Effects of covariates on rates (cause-specific hazards) may be (qualitatively) different from their effects on the risks (cumulative incidences)
- Rates may be analysed using standard hazard based methods from survival analysis (Nelson-Aalen, Cox, Poisson, logrank, ...)
- Risks may be analysed by 'pluging-in' results from such hazard models or directly using, e.g. the Fine-Gray model
- Interpretation of coefficients from a Fine-Gray model is not appealing but other link functions may be used

Rates vs. risks - quotes

- Latouche A., Allignol A., Beyersmann J., Labopin M., Fine J.P.: A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *J. Clin. Epidemiol.* (2013) **66**, 648-653.
- Koller, M.T., Raatz, H., Steyerberg, E.W., Wolbers, M.: Competing risks and the clinical community: irrelevance or ignorance? *Stat. in Med.* (2012) **31**, 1089-1097. "etiology hypotheses are most naturally formulated in terms of cause-specific hazards ... absolute risk of events are the natural basis for prognosis"
- Andersen, P.K., Geskus, R.B., de Witte, T., Putter, H.: Competing risks in epidemiology: Possibilities and pitfalls. *Int. J. Epidemiol.* (2012) **41**, 861-870 quote Koller et al.