

Vertical modeling: a "pattern mixture" approach for competing risks data

Alina Nicolaie, Hans van Houwelingen, Hein Putter

Leiden University Medical Center, The Netherlands

The 30th ISCB Conference,
Prague, Czech Republic, 23-27 August 2009

Outline

Data

Modeling

- The competing risks model

- Notation

Estimation

- The key concepts

Vertical modeling

- Intuition

- No covariates

- Covariates

Discussion

Description of the data

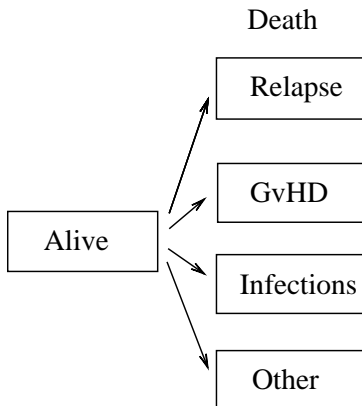
- ▶ data from the EBMT registry on 8966 patients with allogeneic hematopoietic stem cell transplantation (HSCT)
- ▶ time is measured in years from transplantation
- ▶ Number of events for each cause of death

Event	Alive	Relapse	GvHD	Infections	Other
Number	5656	1098	834	454	924

- ▶ covariate disease subtype: "AML"(39%), "ALL"(21%), "CML"(40%)
- ▶ objective: to study whether patterns of causes of death change over time



Our competing risks model





Notation

- ▶ Set T =the time to death, $T \geq 0$
 D =the cause of death, $D \in \{1, \dots, J\}$
 C =the right-censoring time, $C \geq 0$
 Z =the vector of covariates
- ▶ observe $(\min(T, C), \Delta = \mathbf{1}\{T < C\} \cdot D, Z)$
- ▶ assume independent censoring (T and C are independent, given Z)
- ▶ the aim: to estimate the **cumulative incidence** function

$$P(T, D) = P(T \leq t, D = j) =: F_j(t) \quad (1)$$

How to estimate this quantity?

- ▶ several possibilities:
 - ▶ Cause-specific hazards approach (1978)
 - ▶ Larson and Dinse's model (1985)
 - ▶ Fine and Gray's model (1999)
 - ▶ Vertical modeling

The key concepts

- ▶ Cause-specific hazards: **the cause-specific hazard** $\lambda_j(t)$, $j = 1, \dots, J$

$$\lambda_j(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{P}(t \leq T < t + \Delta t, D = j | T \geq t)}{\Delta t}$$

- ▶ Larson and Dinse's model: the joint distribution of (T, D) expressed as

$$P(T, D) = P(T \leq t | D = j)P(D = j)$$

- ▶ Fine and Gray's model: **the subdistribution hazard** α_j , $j = 1, \dots, J$

$$\alpha_j(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{P}(t \leq T < t + \Delta t, D = j | T \geq t \cup (T \leq t \cap D \neq j))}{\Delta t}$$

Vertical modeling: first, intuition...

- ▶ when analyzing the dynamic of failures
 - ▶ instead of being concerned with the cause-specific rates...
 - ▶ you may want to:
 1. analyze the rate at which a failure occurs, irrespective of its type(=an overall "view")
 2. determine the probability, given a failure occurred, that it is of a specific type(=a conditional probability)

Now, formulas...

- ▶ the joint distribution of (T, D) expressed as

$$P(T = t, D = j) = P(D = j | T = t)P(T = t) \quad (2)$$

- ▶ the "natural" decomposition

Two key concepts

- ▶ define **the total hazard** $\lambda_{\bullet}(t)$, $t \geq 0$, by

$$\lambda_{\bullet}(t) := \sum_{j=1}^J \lambda_j(t)$$

$$\lambda_{\bullet}(t) = \lim_{\Delta t \rightarrow 0} \frac{\text{Prob}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (3)$$

- ▶ define **the relative cause-specific hazard**

$$\pi_j(t) := \frac{\lambda_j(t)}{\lambda_{\bullet}(t)}, \quad t \geq 0.$$

Issues about the relative hazard

- ▶ essentially, it can be any function with values in $[0, 1]$
- ▶ it describes a **local time** behaviour

$$\pi_j(t) = \text{Prob}(D = j | T = t)$$



$$\sum_{j=1}^J \pi_j(t) = 1$$

- ▶ f_{\bullet} the density corresponding to the distribution of T

$$F_j(t) = \int_0^t \pi_j(u) f_{\bullet}(u) du$$

How to model?

$$P(T = t, D = j) = P(D = j|T = t)P(T = t)$$

- ▶ the driving force for $P(T)$ is $\lambda_{\bullet}(t)$
- ▶ the driving forces for $P(D|T)$ are $(\pi_j(t))_{j \in J}$

How to estimate them? Firstly, without covariates

- ▶ for the failure time:
 - ▶ all failures are considered as event, irrespective of the cause of failure
 - ▶ the most obvious choice is the nonparametric Kaplan-Meier estimator
- ▶ for the cause of failure: we need a model for the relative hazard, but...

The model for the relative hazard

- ▶ estimate π_j "model free" via

$$\hat{\pi}_j(t_k) = \frac{\# \text{ patients failing from cause } j \text{ at time } t_k}{\# \text{ patients failing at time } t_k} \quad (4)$$

- ▶ **problem**: in the absence of ties

$$\hat{\pi}_j(t_k) = 1 \text{ and } \hat{\pi}_l(t_k) = 0 \text{ for all } l \neq j$$

...is multinomial logistic regression

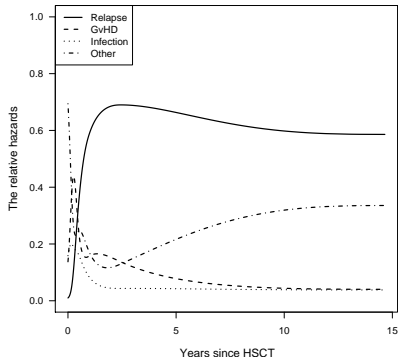
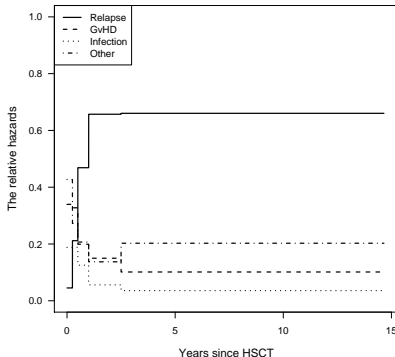
- ▶ the natural choice for the relative hazard is multinomial regression model

$$\pi_j(t) = \frac{\exp(\beta_j^\top \mathbf{B}(t))}{\sum_{l=1}^J \exp(\beta_l^\top \mathbf{B}(t))} \quad (5)$$

where $\beta_j = (\beta_{j1}, \dots, \beta_{jp})$ are regression coefficients, $B(t)$ time functions

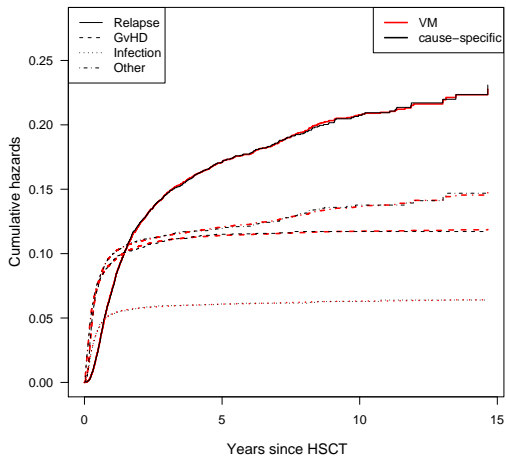
- ▶ intuition through $B(t) = (I_{[0,0.25)}(t), I_{[0.25,0.5)}(t), I_{[0.5,1)}(t), I_{[1,2.5)}(t), I_{[2.5,\infty)}(t))$
- ▶ smoothing through $\mathbf{B}(t) = (B_1(t), B_2(t), \dots, B_4(t))$, where B_i are cubic spline

The piece-wise constant and smoothed relative hazards

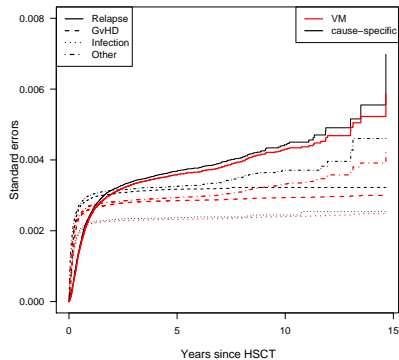
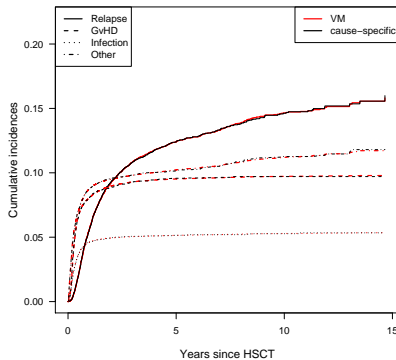




The cumulative hazards



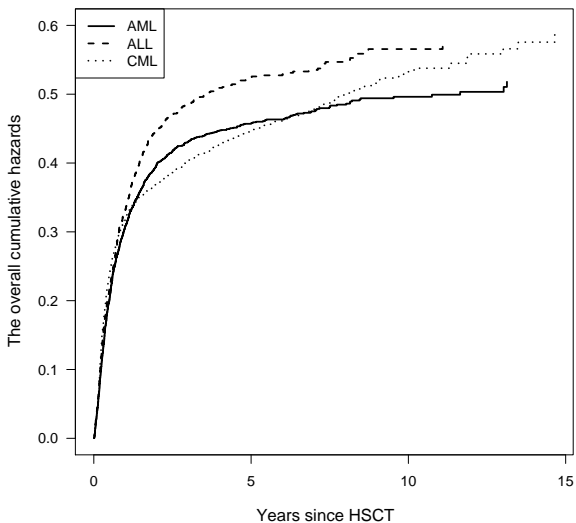
The cumulative incidences and their standard errors



Modeling with covariates

- ▶ for the failure time
 - ▶ we can estimate different Kaplan-Meier curves (for different values of the covariate, when categorical)
 - ▶ or, we can use a proportional hazards model

The overall cumulative hazard (AML, ALL, CML)



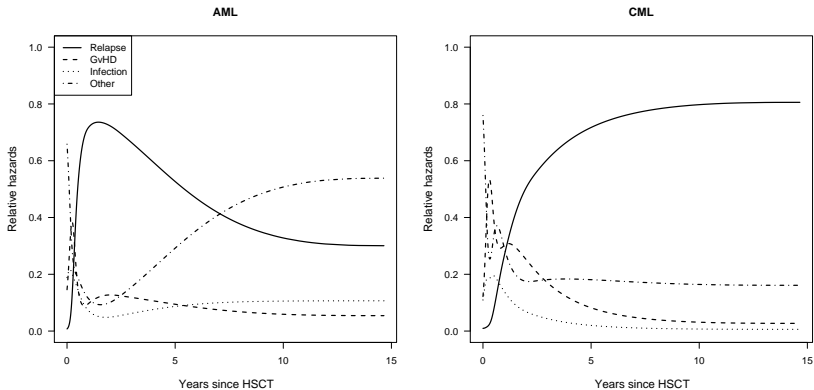
Modeling with covariates

- ▶ for the cause of failure: a multinomial regression model with covariates and a pre-specified function of time (and, perhaps, their interactions as predictors)
- ▶ here we chosen

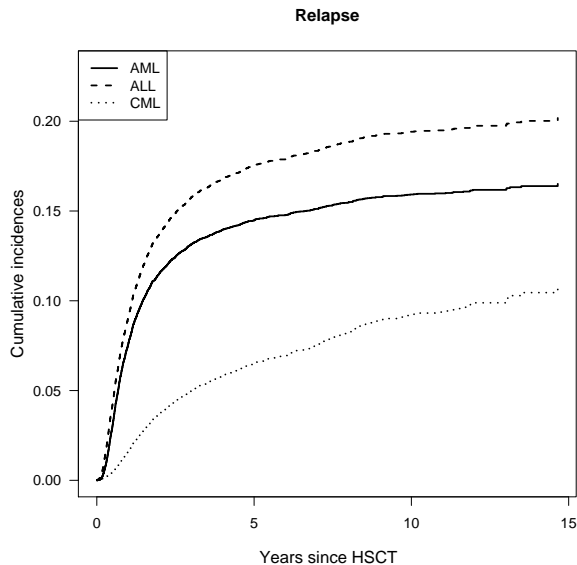
$$\pi_j(t) = \frac{\exp(\beta_j^\top \mathbf{B}(t) * Z)}{\sum_{l=1}^J \exp(\beta_l^\top \mathbf{B}(t) * Z)}, j = 1, \dots, J.$$

where * stands for interaction and Z =disease subtype

The smoothed relative hazards



The cumulative incidence



Discussion

- ▶ our model leads to a gain in efficiency in estimation, by providing more accurate information on cumulative incidences
- ▶ does not assume proportionality of cause-specific hazards or of subdistribution hazards
- ▶ no need to censor, for a specific cause of death in turn, failures due to other causes
- ▶ missing causes of death influence only the estimates of the relative hazards, but not the total hazard
- ▶ the **vm** function soon available in the **mstate** package

References



Larson, G. and Dinse, G.

A mixture model for the regression analysis of competing risks data.
Applied Statistics, (3)34, 1985, 201-211 , 1985.



Prentice, R. L. and Kalbfleisch, J. D.

The analysis of failure times in the presence of competing risks.
Biometrics, 34, 541-554 , 1978.



Putter, H. and Fiocco, M. and Geskus, R. B.

Tutorial in biostatistics: Competing risks and multi-state models.
Statist Med, 26, 2389-2430 , 2007.



de Wreede, L., Fiocco, M., Putter, H.

The mstate package for estimation and prediction in non- and semi-parametric multi-state models.
Submitted, 2009.



Nicolaie, M. A., van Houwelingen, H., Putter, H.

Vertical modeling: a pattern mixture approach to competing risks data.
Submitted, 2009.