Generic solution of the heterogeneity-induced competing risk problem in survival analysis

Ton Coolen

Institute for Mathematical and Molecular Biomedicine, King's College London
London Institute for Mathematical Sciences

- Survival analysis and competing risks
- Individual versus cohort level risk analysis
- Modelling heterogeneity-induced competing risks
- Applications: synthetic data and prostate cancer data

Generic solution of the heterogeneity-induced competing risk problem in survival analysis

Ton Coolen

Institute for Mathematical and Molecular Biomedicine, King's College London
London Institute for Mathematical Sciences

- Survival analysis and competing risks
- Individual versus cohort level risk analysis
- Modelling heterogeneity-induced competing risks
- Applications: synthetic data and prostate cancer data

Generic solution of the heterogeneity-induced competing risk problem in survival analysis

Ton Coolen

Institute for Mathematical and Molecular Biomedicine, King's College London
London Institute for Mathematical Sciences

- Survival analysis and competing risks
- Individual versus cohort level risk analysis
- Modelling heterogeneity-induced competing risks
- Applications: synthetic data and prostate cancer data

Generic solution of the heterogeneity-induced competing risk problem in survival analysis

Ton Coolen

Institute for Mathematical and Molecular Biomedicine, King's College London
London Institute for Mathematical Sciences

- Survival analysis and competing risks
- Individual versus cohort level risk analysis
- Modelling heterogeneity-induced competing risks
- Applications: synthetic data and prostate cancer data

Generic solution of the heterogeneity-induced competing risk problem in survival analysis

Ton Coolen

Institute for Mathematical and Molecular Biomedicine, King's College London
London Institute for Mathematical Sciences

- Survival analysis and competing risks
- Individual versus cohort level risk analysis
- Modelling heterogeneity-induced competing risks
- Applications: synthetic data and prostate cancer data

Survival analysis and competing risks

- N individuals, subject to R 'hazards' or 'risks'
  e.g. cancer recurrence, other death, end of trial ...
- If one event happens, others can no longer be observed
- Data, $i = 1 \ldots N$:

$$\begin{aligned} \mathbf{z}_i = (z_1^i, \ldots, z_p^i) : &\qquad \textit{values of } p \textit{ covariates} \\ t_i \geq 0 : &\qquad \textit{time of first event} \\ r_i \in \{1, \ldots, R\} : &\qquad \textit{type of first event} \end{aligned}$$

Question:

- Extract regularities that connect covariates to risks

Survival analysis and competing risks

- N individuals, subject to R 'hazards' or 'risks'
  e.g. cancer recurrence, other death, end of trial ...
- If one event happens, others can no longer be observed
- Data, $i = 1 \ldots N$:

$$
\begin{aligned}
\mathbf{z}_i = (z_1^i, \ldots, z_p^i): &\quad \textit{values of } p \textit{ covariates} \\
t_i \geq 0: &\quad \textit{time of first event} \\
r_i \in \{1, \ldots, R\}: &\quad \textit{type of first event}
\end{aligned}
$$

Question:

- Extract regularities that connect covariates to risks

competing risk problem

- If event times of risks correlated: informative censoring
  primary hazard rate contaminated by non-primary risks

$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \mathcal{P}(t_1|\mathbf{z})\mathcal{P}(t_2,\ldots,t_R|\mathbf{z})$$

- What would be primary risk survival function
  if all other risks were disabled?

  nontrivial ...

  – disabling non-primary risks affects also primary hazard rate
  – Tsiatis: without further assumptions one cannot infer
  risk correlations from survival data

- Most methods assume risk independence
  so non-primary risks don't affect primary hazard rate
  (Cox, KM, fraily models ...)

$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \mathcal{P}(t_1|\mathbf{z})\mathcal{P}(t_2,\ldots,t_R|\mathbf{z})$$

competing risk problem

- If event times of risks correlated: informative censoring
  primary hazard rate contaminated by non-primary risks

$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \mathcal{P}(t_1|\mathbf{z})\mathcal{P}(t_2,\ldots,t_R|\mathbf{z})$$

- What would be primary risk survival function
  if all other risks were disabled?

  nontrivial ...

  – disabling non-primary risks affects also primary hazard rate
  – Tsiatis: without further assumptions one cannot infer
  risk correlations from survival data

- Most methods assume risk independence
  so non-primary risks don't affect primary hazard rate
  (Cox, KM, fraily models ...)

$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \mathcal{P}(t_1|\mathbf{z})\mathcal{P}(t_2,\ldots,t_R|\mathbf{z})$$

competing risk problem

- If event times of risks correlated: informative censoring
  primary hazard rate contaminated by non-primary risks

$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \mathcal{P}(t_1|\mathbf{z})\mathcal{P}(t_2,\ldots,t_R|\mathbf{z})$$

- What would be primary risk survival function
  if all other risks were disabled?

  nontrivial ...

  – disabling non-primary risks affects also primary hazard rate
  – Tsiatis: without further assumptions one cannot infer
  risk correlations from survival data

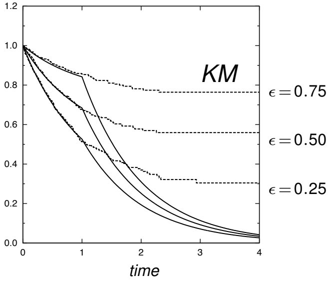
- Most methods assume risk independence
  so non-primary risks don't affect primary hazard rate
  (Cox, KM, fraily models ...)

$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \mathcal{P}(t_1|\mathbf{z})\mathcal{P}(t_2,\ldots,t_R|\mathbf{z})$$

a serious problem?

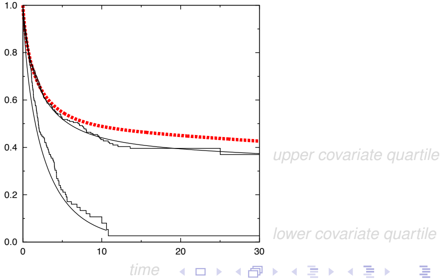
some illustrations ...

$$\mathcal{P}(t_1,t_2) = \epsilon\, e^{-t_2}\delta(t_1-t_2-1) + (1-\epsilon)e^{-t_1-t_2}$$



KM & Cox-Breslow estimators

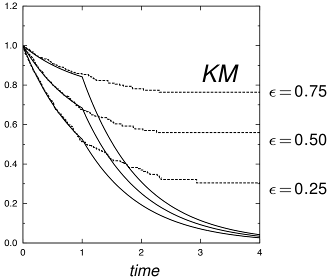
true survival curves



a serious problem?

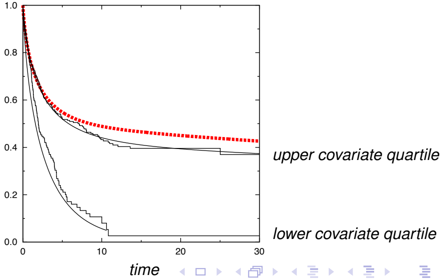
some illustrations ...

$$\mathcal{P}(t_1,t_2) = \epsilon\, e^{-t_2}\delta(t_1-t_2-1) + (1-\epsilon)e^{-t_1-t_2}$$



KM & Cox-Breslow estimators

true survival curves



Possible causes of informative censoring

Say we have 1000 people in a cohort
two risks, hazard rates h_A and h_B

- homogeneous cohort:
  all *individuals* have (h_A, h_B)

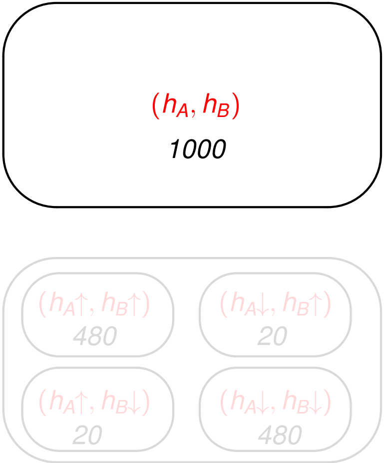
- heterogeneous cohort,
  four subgroups:

Possible causes of informative censoring

Say we have 1000 people in a cohort
two risks, hazard rates h_A and h_B

- homogeneous cohort:
  all *individuals* have (h_A, h_B)

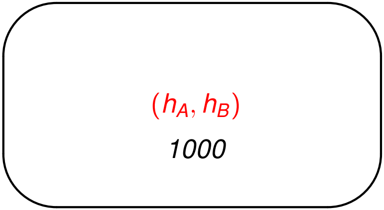
- heterogeneous cohort,
  four subgroups:



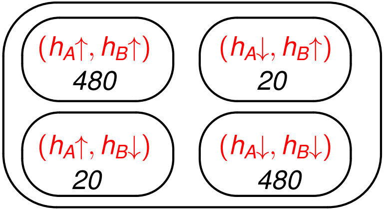
Possible causes of informative censoring

Say we have 1000 people in a cohort
two risks, hazard rates h_A and h_B

- homogeneous cohort:
  all *individuals* have (h_A, h_B)



- heterogeneous cohort,
  four subgroups:



to make progress:

model all risks and their relations

at individual and cohort level

individual versus cohort level risk description

	cohort:	individual i:
event time statistics:	$\mathcal{P}(t_1,\ldots,t_R)$	$\mathcal{P}_i(t_1,\ldots,t_R)$
cause-specific hazard rates:	$h_r(t)$	$h_r^i(t)$
cause-specific survival functions:	$S_r(t)$	$S_r^i(t)$

links:

$$\mathcal{P}(t_1,\ldots,t_R)=\frac{1}{N}\sum_{i=1}^N \mathcal{P}_i(t_1,\ldots,t_R) \qquad S_r(t)=\frac{1}{N}\sum_{i=1}^N S_r^i(t)$$

$$h_r(t)=\frac{\sum_{i=1}^N h_r^i(t)\mathrm{e}^{-\sum_{r'=1}^R\int_0^t \mathrm{d}s\ h_{r'}^i(s)}}{\sum_{i=1}^N \mathrm{e}^{-\sum_{r'=1}^R\int_0^t \mathrm{d}s\ h_{r'}^i(s)}}$$

to make progress:

model all risks and their relations
at individual and cohort level

individual versus cohort level
risk description

	cohort:	individual i:
event time statistics:	$\mathcal{P}(t_1,\ldots,t_R)$	$\mathcal{P}_i(t_1,\ldots,t_R)$
cause-specific hazard rates:	$h_r(t)$	$h_r^i(t)$
cause-specific survival functions:	$S_r(t)$	$S_r^i(t)$

links:

$$\mathcal{P}(t_1,\ldots,t_R)=\frac{1}{N}\sum_{i=1}^N \mathcal{P}_i(t_1,\ldots,t_R) \qquad S_r(t)=\frac{1}{N}\sum_{i=1}^N S_r^i(t)$$

$$h_r(t)=\frac{\sum_{i=1}^N h_r^i(t)\mathrm{e}^{-\sum_{r'=1}^R\int_0^t \mathrm{d}s\ h_{r'}^i(s)}}{\sum_{i=1}^N \mathrm{e}^{-\sum_{r'=1}^R\int_0^t \mathrm{d}s\ h_{r'}^i(s)}}$$

to make progress:

model all risks and their relations

at individual and cohort level

individual versus cohort level
risk description

	cohort:	individual i:
event time statistics:	$\mathcal{P}(t_1,\ldots,t_R)$	$\mathcal{P}_i(t_1,\ldots,t_R)$
cause-specific hazard rates:	$h_r(t)$	$h_r^i(t)$
cause-specific survival functions:	$S_r(t)$	$S_r^i(t)$

links:

$$\mathcal{P}(t_1,\ldots,t_R)=\frac{1}{N}\sum_{i=1}^{N}\mathcal{P}_i(t_1,\ldots,t_R) \qquad S_r(t)=\frac{1}{N}\sum_{i=1}^{N}S_r^i(t)$$

$$h_r(t)=\frac{\sum_{i=1}^{N}h_r^i(t)\mathrm{e}^{-\sum_{r'=1}^{R}\int_0^t \mathrm{d}s\; h_{r'}^i(s)}}{\sum_{i=1}^{N}\mathrm{e}^{-\sum_{r'=1}^{R}\int_0^t \mathrm{d}s\; h_{r'}^i(s)}}$$

Complexity levels of cohorts

level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 2 : heterogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$

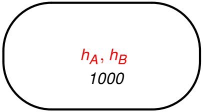
level 4 : individual *and* cohort level competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) \neq \prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$

Complexity levels of cohorts

level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})=\prod_r \mathcal{P}(t_r|\mathbf{z})$$



level 2 : heterogeneous cohort, no competing risks

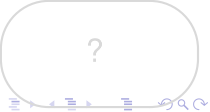
$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})=\prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})\neq\prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 4 : individual *and* cohort level competing risks

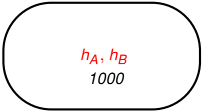
$$\mathcal{P}_i(t_1,\ldots,t_R)\neq\prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})\neq\prod_r \mathcal{P}(t_r|\mathbf{z})$$



Complexity levels of cohorts

level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})=\prod_r \mathcal{P}(t_r|\mathbf{z})$$



level 2 : heterogeneous cohort, no competing risks

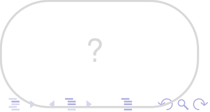
$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})=\prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})\neq\prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 4 : individual *and* cohort level competing risks

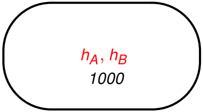
$$\mathcal{P}_i(t_1,\ldots,t_R)\neq\prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})\neq\prod_r \mathcal{P}(t_r|\mathbf{z})$$



Complexity levels of cohorts

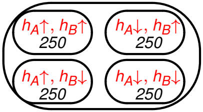
level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$



level 2 : heterogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$

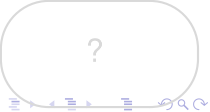


level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 4 : individual *and* cohort level competing risks

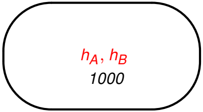
$$\mathcal{P}_i(t_1,\ldots,t_R) \neq \prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$



Complexity levels of cohorts

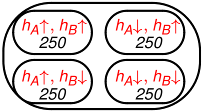
level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})=\prod_r \mathcal{P}(t_r|\mathbf{z})$$



level 2 : heterogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})=\prod_r \mathcal{P}(t_r|\mathbf{z})$$

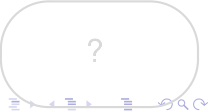


level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R)=\prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})\neq\prod_r \mathcal{P}(t_r|\mathbf{z})$$

level 4 : individual *and* cohort level competing risks

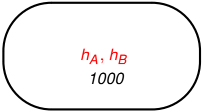
$$\mathcal{P}_i(t_1,\ldots,t_R)\neq\prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z})\neq\prod_r \mathcal{P}(t_r|\mathbf{z})$$



Complexity levels of cohorts

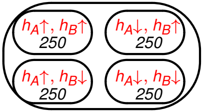
level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$



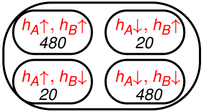
level 2 : heterogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$



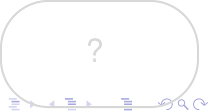
level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$



level 4 : individual *and* cohort level competing risks

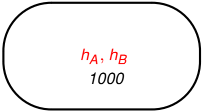
$$\mathcal{P}_i(t_1,\ldots,t_R) \neq \prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$



Complexity levels of cohorts

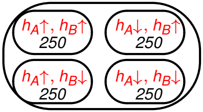
level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$



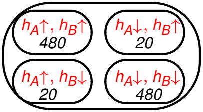
level 2 : heterogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$



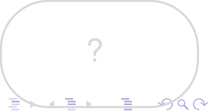
level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$



level 4 : individual *and* cohort level competing risks

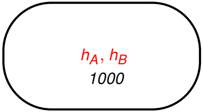
$$\mathcal{P}_i(t_1,\ldots,t_R) \neq \prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$



Complexity levels of cohorts

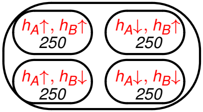
level 1 : homogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \overline{\mathcal{P}}(t_r|\mathbf{z}_i)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$



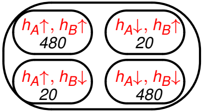
level 2 : heterogeneous cohort, no competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) = \prod_r \mathcal{P}(t_r|\mathbf{z})$$



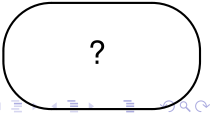
level 3 : heterogeneity-induced competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) = \prod_r \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$



level 4 : individual *and* cohort level competing risks

$$\mathcal{P}_i(t_1,\ldots,t_R) \neq \prod_{r=1} \mathcal{P}_i(t_r)$$
$$\mathcal{P}(t_1,\ldots,t_R|\mathbf{z}) \neq \prod_r \mathcal{P}(t_r|\mathbf{z})$$



Heterogeneity-induced competing risks

Natural description

covariate-conditioned joint *distribution*
of all cause-specific hazard rates:

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] = \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} \prod_r \delta_{\rm F}[h_r - h_r^i]}{\sum_{i,\mathbf{z}_i=\mathbf{z}} 1}$$

$h_r^i = \{h_r^i(t)\}$
risk r hazard rate
of individual i

Disabling non-primary risks:

$h_r^i \to 0$ *for all $r > 1$*

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] \to \mathcal{W}[h_1|\mathbf{z}] \prod_{r>1} \delta_{\rm F}[h_r] \qquad \mathcal{W}[h_1|\mathbf{z}] = \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} \delta_{\rm F}[h_1 - h_1^i]}{\sum_{i,\mathbf{z}_i=\mathbf{z}} 1}$$

Data log-likelihood:

$$\mathcal{L}(D|\mathcal{W}) = \sum_{i=1}^{N} \log \int \{{\rm d}h_1 \ldots {\rm d}h_R\}\, \mathcal{W}[h_1,\ldots,h_R|\mathbf{z}_i]\, h_{r_i}(t_i) {\rm e}^{-\sum_{r=1}^{R} \int_0^{t_i} {\rm d}s\, h_r(s)}$$

Heterogeneity-induced competing risks

Natural description

covariate-conditioned joint *distribution*
of all cause-specific hazard rates:

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] = \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} \prod_r \delta_{\rm F}[h_r - h_r^i]}{\sum_{i,\mathbf{z}_i=\mathbf{z}} 1}$$

$h_r^i = \{h_r^i(t)\}$
risk r hazard rate of individual i

Disabling non-primary risks:

$h_r^i \to 0$ *for all* $r > 1$

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] \;\to\; \mathcal{W}[h_1|\mathbf{z}] \prod_{r>1} \delta_{\rm F}[h_r] \qquad \mathcal{W}[h_1|\mathbf{z}] = \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} \delta_{\rm F}[h_1 - h_1^i]}{\sum_{i,\mathbf{z}_i=\mathbf{z}} 1}$$

Data log-likelihood:

$$\mathcal{L}(D|\mathcal{W}) = \sum_{i=1}^{N} \log \int \{{\rm d}h_1 \ldots {\rm d}h_R\}\, \mathcal{W}[h_1,\ldots,h_R|\mathbf{z}_i]\; h_{r_i}(t_i) {\rm e}^{-\sum_{r=1}^R \int_0^{t_i} {\rm d}s\; h_r(s)}$$

Heterogeneity-induced competing risks

Natural description

covariate-conditioned joint *distribution*
of all cause-specific hazard rates:

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] = \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} \prod_r \delta_{\rm F}[h_r - h_r^i]}{\sum_{i,\mathbf{z}_i=\mathbf{z}} 1}$$

$h_r^i = \{h_r^i(t)\}$
risk r hazard rate
of individual i

Disabling non-primary risks:

$h_r^i \to 0$ *for all* $r > 1$

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] \;\to\; \mathcal{W}[h_1|\mathbf{z}] \prod_{r>1} \delta_{\rm F}[h_r] \qquad \mathcal{W}[h_1|\mathbf{z}] = \frac{\sum_{i,\mathbf{z}_i=\mathbf{z}} \delta_{\rm F}[h_1 - h_1^i]}{\sum_{i,\mathbf{z}_i=\mathbf{z}} 1}$$

Data log-likelihood:

$$\mathcal{L}(D|\mathcal{W}) = \sum_{i=1}^{N} \log \int \{{\rm d}h_1 \ldots {\rm d}h_R\}\, \mathcal{W}[h_1,\ldots,h_R|\mathbf{z}_i]\; h_{r_i}(t_i) {\rm e}^{-\sum_{r=1}^{R} \int_0^{t_i} {\rm d}s\; h_r(s)}$$

Decontamination formulae

'crude' cause-specific quantities:

$$S_r(t|\mathbf{z}) = \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s|\mathbf{z})}$$

$$h_r(t|\mathbf{z}) = \frac{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ h_r(t)\mathrm{e}^{-\sum_{r'}\int_0^t \mathrm{d}s\ h_{r'}(s)}}{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ \mathrm{e}^{-\sum_{r'}\int_0^t \mathrm{d}s\ h_{r'}(s)}}$$

decontaminated:

$$\tilde{S}_r(t|\mathbf{z}) = \int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s)}$$

$$\tilde{h}_r(t|\mathbf{z}) = \frac{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ h_r(t)\mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s)}}{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s)}}$$

Decontamination formulae

'crude' cause-specific quantities:

$$S_r(t|\mathbf{z}) = \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s|\mathbf{z})}$$

$$h_r(t|\mathbf{z}) = \frac{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ h_r(t)\mathrm{e}^{-\sum_{r'} \int_0^t \mathrm{d}s\ h_{r'}(s)}}{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ \mathrm{e}^{-\sum_{r'} \int_0^t \mathrm{d}s\ h_{r'}(s)}}$$

decontaminated:

$$\tilde{S}_r(t|\mathbf{z}) = \int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s)}$$

$$\tilde{h}_r(t|\mathbf{z}) = \frac{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ h_r(t)\mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s)}}{\int\{\mathrm{d}h_1 \ldots \mathrm{d}h_R\}\ \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]\ \mathrm{e}^{-\int_0^t \mathrm{d}s\ h_r(s)}}$$

Parametrisations of $\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}]$

proportional hazards at level of individuals

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] = \int \mathrm{d}\boldsymbol{\beta}_1\ldots\mathrm{d}\boldsymbol{\beta}_R \int \{\mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_R\}\, \mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R;\lambda_1,\ldots\lambda_R) \times \prod_r \delta_{\mathrm{F}}\Big[h_r - \lambda_r \mathrm{e}^{\beta_r^0+\sum_{\mu=1}^p \beta_r^\mu z_\mu}\Big]$$

includes as special cases:
Cox regression, frailty models, random effect models, ...

- e.g. latent class heterogeneity:

$$\mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R;\lambda_1,\ldots,\lambda_R) = \mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R)\prod_{r=1}^R \delta_{\mathrm{F}}[\lambda_r - \hat{\lambda}_r]$$

$$\mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R) = \sum_{\ell=1}^L w_\ell \prod_{r=1}^R \delta(\boldsymbol{\beta}_r - \hat{\boldsymbol{\beta}}_r^\ell)$$

$$\hat{\boldsymbol{\beta}}_r^\ell = (\hat{\beta}_r^{\ell 0},\ldots,\hat{\beta}_r^{\ell p})$$

Parametrisations of $\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}]$

proportional hazards at level of individuals

$$\mathcal{W}[h_1,\ldots,h_R|\mathbf{z}] = \int \mathrm{d}\boldsymbol{\beta}_1\ldots\mathrm{d}\boldsymbol{\beta}_R \int \{\mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_R\}\, \mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R;\lambda_1,\ldots\lambda_R) \times \prod_r \delta_{\mathrm{F}}\Big[h_r - \lambda_r \mathrm{e}^{\beta_r^0+\sum_{\mu=1}^p \beta_r^\mu z_\mu}\Big]$$

includes as special cases:
Cox regression, frailty models, random effect models, ...

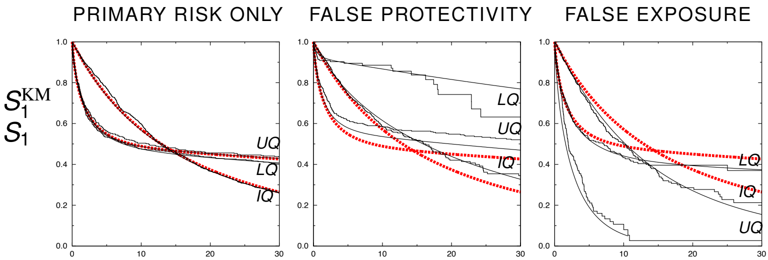
- e.g. latent class heterogeneity:

$$\mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R;\lambda_1,\ldots,\lambda_R) = \mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R)\prod_{r=1}^R \delta_{\mathrm{F}}[\lambda_r-\hat{\lambda}_r]$$

$$\mathcal{M}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_R) = \sum_{\ell=1}^L w_\ell \prod_{r=1}^R \delta(\boldsymbol{\beta}_r-\hat{\boldsymbol{\beta}}_r^\ell)$$

$$\hat{\boldsymbol{\beta}}_r^\ell = (\hat{\beta}_r^{\ell 0},\ldots,\hat{\beta}_r^{\ell p})$$

Applications – synthetic data

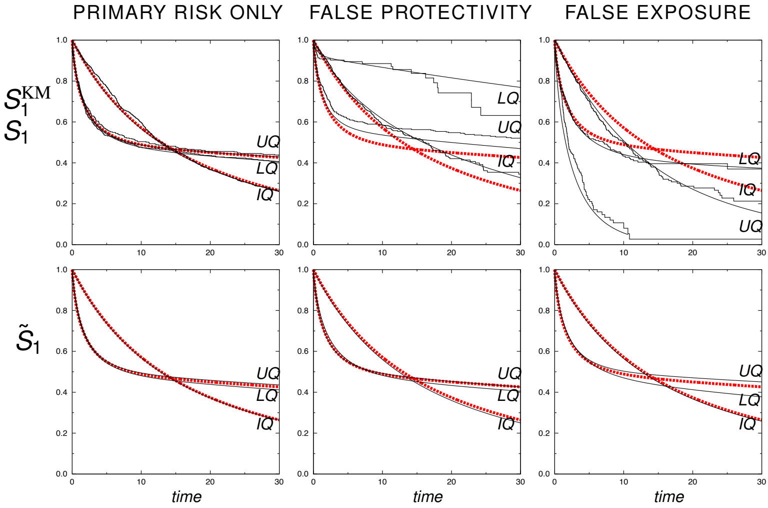


S_1^{KM} : *Kaplan-Meier*

S_1 : *crude survival curve*

red dashed: true survival curves

Applications – synthetic data



S_1^{KM} : *Kaplan-Meier*

S_1 : *crude survival curve*

red dashed: true survival curves

$\tilde{S}_1$: *decontaminated curves*

retrospective class identification

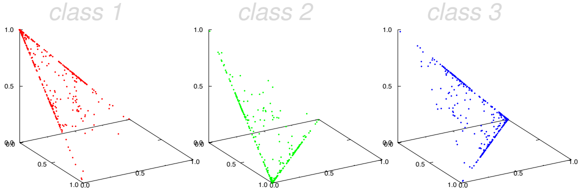
$$P(\ell|t,r,\mathbf{z}) = \frac{w_\ell\, e^{\hat{\boldsymbol{\beta}}_r^\ell\cdot\mathbf{z}-\sum_{r'=1}^R \exp(\hat{\boldsymbol{\beta}}_{r'}^\ell\cdot\mathbf{z})\int_0^t ds\;\hat{\lambda}_{r'}(s)}}{\sum_{\ell'=1}^L w_{\ell'}\, e^{\hat{\boldsymbol{\beta}}_r^{\ell'}\cdot\mathbf{z}-\sum_{r'=1}^R \exp(\hat{\boldsymbol{\beta}}_{r'}^{\ell'}\cdot\mathbf{z})\int_0^t ds\;\hat{\lambda}_{r'}(s)}}$$

Data:

3 classes,
$w_1 = w_2 = w_3 = \frac{1}{3}$
2 competing risks

$\beta_1^1 = (0.5, 0.5, 0.5) + (2, 0, 2)$
$\beta_1^2 = (0.5, 0.5, 0.5) + (-2, -2, 0)$
$\beta_1^3 = (0.5, 0.5, 0.5) + (0, 2, -2)$

each individual i:
point (p_1^i, p_2^i, p_3^i) in $\mathbb{R}^3$
$p_\ell^i = P(\ell|t_i, r_i, \mathbf{z}_i)$



retrospective class identification

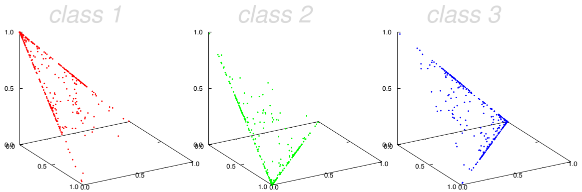
$$P(\ell|t,r,\mathbf{z}) \;=\; \frac{w_\ell\; e^{\hat{\boldsymbol{\beta}}_r^\ell\cdot\mathbf{z}-\sum_{r'=1}^R \exp(\hat{\boldsymbol{\beta}}_{r'}^\ell\cdot\mathbf{z})\int_0^t ds\;\hat{\lambda}_{r'}(s)}}{\sum_{\ell'=1}^L w_{\ell'}\; e^{\hat{\boldsymbol{\beta}}_r^{\ell'}\cdot\mathbf{z}-\sum_{r'=1}^R \exp(\hat{\boldsymbol{\beta}}_{r'}^{\ell'}\cdot\mathbf{z})\int_0^t ds\;\hat{\lambda}_{r'}(s)}}$$

Data:

3 classes,
$w_1 = w_2 = w_3 = \frac{1}{3}$
2 competing risks

$$\beta_1^1 = (0.5, 0.5, 0.5) + (2, 0, 2)$$
$$\beta_1^2 = (0.5, 0.5, 0.5) + (-2, -2, 0)$$
$$\beta_1^3 = (0.5, 0.5, 0.5) + (0, 2, -2)$$

each individual i:
point (p_1^i, p_2^i, p_3^i) in $\mathbb{R}^3$
$p_\ell^i = P(\ell|t_i, r_i, \mathbf{z}_i)$



retrospective class identification

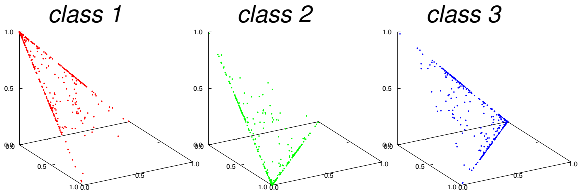
$$P(\ell|t,r,\mathbf{z}) \;=\; \frac{w_\ell\; e^{\hat{\boldsymbol{\beta}}_r^\ell\cdot\mathbf{z}-\sum_{r'=1}^R \exp(\hat{\boldsymbol{\beta}}_{r'}^\ell\cdot\mathbf{z})\int_0^t \mathrm{d}s\;\hat{\lambda}_{r'}(s)}}{\sum_{\ell'=1}^L w_{\ell'}\; e^{\hat{\boldsymbol{\beta}}_r^{\ell'}\cdot\mathbf{z}-\sum_{r'=1}^R \exp(\hat{\boldsymbol{\beta}}_{r'}^{\ell'}\cdot\mathbf{z})\int_0^t \mathrm{d}s\;\hat{\lambda}_{r'}(s)}}$$

Data:

3 classes,
$w_1=w_2=w_3=\frac{1}{3}$
2 competing risks

$$\beta_1^1=(0.5,0.5,0.5)+(2,0,2)$$
$$\beta_1^2=(0.5,0.5,0.5)+(-2,-2,0)$$
$$\beta_1^3=(0.5,0.5,0.5)+(0,2,-2)$$

each individual i:
point (p_1^i,p_2^i,p_3^i) in $\mathbb{R}^3$
$p_\ell^i=P(\ell|t_i,r_i,\mathbf{z}_i)$



Applications – ULSAM prostate cancer data set

$N = 2047$,
primary events: 208
death (non-PC): 910
end of trial: 929

covariates:		
	body mass index	(real-valued)
	serum selenium level	(integer)
	physical activity, leisure time	(0/1/2)
	physical activity, work	(0/1/2)
	smoking	(0/1/2)

Cox regression:

BMI	*selenium*	*phys1*	*phys2*	*smoking*
$\beta_1 = 0.14$	$\beta_2 = -0.15$	$\beta_3 = 0.20$	$\beta_4 = -0.09$	$\beta_5 = -0.08$

$$\mathrm{HR}_\mu = \exp(2\beta_\mu)$$

$N = 2047$,
primary events: 208
death (non-PC): 910
end of trial: 929

covariates:		
	body mass index	(real-valued)
	serum selenium level	(integer)
	physical activity, leisure time	(0/1/2)
	physical activity, work	(0/1/2)
	smoking	(0/1/2)

Cox regression:

BMI	*selenium*	*phys1*	*phys2*	*smoking*
$\beta_1 = 0.14$	$\beta_2 = -0.15$	$\beta_3 = 0.20$	$\beta_4 = -0.09$	$\beta_5 = -0.08$

$$\mathrm{HR}_\mu = \exp(2\beta_\mu)$$

	CLASSES	PRIMARY RISK					SECONDARY RISK				
		208 events					910 events				
		BMI	*selen*	*phys1*	*phys2*	*smok*	*BMI*	*selen*	*phys1*	*phys2*	*smok*
Cox		0.14	-0.15	0.20	-0.09	-0.08					
new	$w_1 = 0.51$	1.22	-0.41	0.73	-0.01	1.43	0.82	-0.42	-0.31	-0.14	1.35
	$w_2 = 0.49$	-0.07	-0.16	0.19	-0.10	-0.27	0.10	-0.07	-0.07	0.04	0.18
	frailties:	$\beta^1_{10} - \beta^2_{10} = -4.61$ (HR 0.010)					$\beta^1_{20} - \beta^2_{20} = -4.06$ (HR 0.017)				

healthy class: strong effects of covariates,
BMI and smoking important risk factors

frail class: weak effects of covariates,
BMI and smoking weakly protective
(reverse causal effects?)

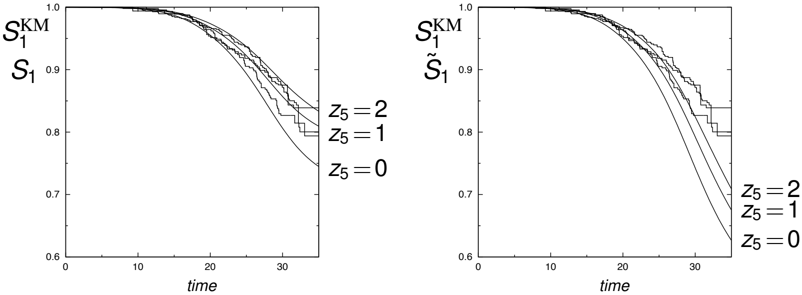
phys1? (leisure time physical activity)

	CLASSES	PRIMARY RISK					SECONDARY RISK				
		208 events					910 events				
		BMI	*selen*	*phys1*	*phys2*	*smok*	*BMI*	*selen*	*phys1*	*phys2*	*smok*
Cox		0.14	-0.15	0.20	-0.09	-0.08					
new	$w_1 = 0.51$	1.22	-0.41	0.73	-0.01	1.43	0.82	-0.42	-0.31	-0.14	1.35
	$w_2 = 0.49$	-0.07	-0.16	0.19	-0.10	-0.27	0.10	-0.07	-0.07	0.04	0.18
	frailties:	$\beta^1_{10} - \beta^2_{10} = -4.61$ (HR 0.010)					$\beta^1_{20} - \beta^2_{20} = -4.06$ (HR 0.017)				

healthy class: strong effects of covariates,
BMI and smoking important risk factors

frail class: weak effects of covariates,
BMI and smoking weakly protective
(reverse causal effects?)

phys1? (leisure time physical activity)



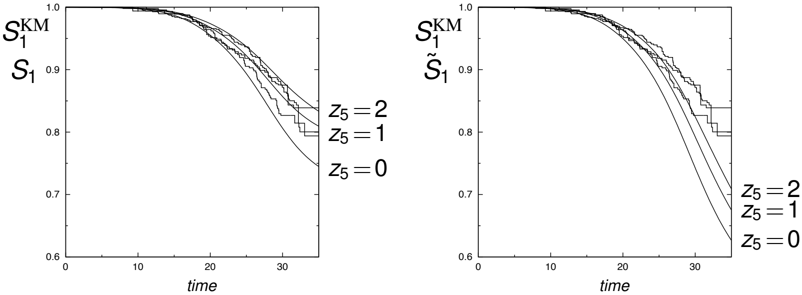
S_1^{KM} : Kaplan-Meier ,
S_1 : crude survival curves,
$\tilde{S}_1$: decontaminated curves

$z_5 = 0$: non-smokers
$z_5 = 1$: ex-smokers
$z_5 = 2$: smokers

false protectivity due to competing risks

Cox/KM underestimate PC risk

BMI & smoking important risk factors in *healthy class*, *frail class* dominate Cox regression and survival curves (due to larger nr of events)



S_1^{KM}: Kaplan-Meier ,
S_1: crude survival curves,
$\tilde{S}_1$: decontaminated curves

$z_5=0$: non-smokers
$z_5=1$: ex-smokers
$z_5=2$: smokers

false protectivity due to competing risks

Cox/KM underestimate PC risk

BMI & smoking important risk factors in *healthy class*,
frail class dominate Cox regression and survival curves
(due to larger nr of events)

Summary

- competing risk problem can be solved if we assume risk correlations are caused by *residual heterogeneity* (heterogeneity not captured by covariates)
- formulae for decontaminated survival curves, expressed in terms of $\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]$ *covariate-conditioned joint distribution of hazard rates for all risks*
- Natural parametrisation of $\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]$, includes standard methods as special cases (Cox, frailty models, random effects models, ...)
- Application to synthetic data with competing risks: method detects structure, parameters, and survival curves correctly
- Application to ULSAM cancer data: new intuitive explanations for previously unexplained results

Thanks to

Collaborators

Hans van Baardewijk, Hans Garmö
Mieke van Hemelrijck, Lars Holmberg

Discussions

Shola Agbaje, Salma Ayis,
Maria D'Iorio, Niels Keiding,
Katherine Lawler

Funding

SEVENTH FRAMEWORK PROGRAMME

Prostate Action