Practical tools for survival analysis with heterogeneity-induced competing risks

ACC Coolen
Institute for Mathematical and Molecular Biomedicine
King’s College London

- Survival analysis and competing risks
- Individual versus cohort level risk
- Modelling heterogeneity-induced competing risks
- Applications:
  - synthetic data
  - prostate and colorectal cancer data
Practical tools for survival analysis with heterogeneity-induced competing risks

ACC Coolen
Institute for Mathematical and Molecular Biomedicine
King’s College London

- Survival analysis and competing risks
- Individual versus cohort level risk
- Modelling heterogeneity-induced competing risks
- Applications:
  - synthetic data
  - prostate and colorectal cancer data
Practical tools for survival analysis with heterogeneity-induced competing risks

ACC Coolen
Institute for Mathematical and Molecular Biomedicine
King’s College London

- Survival analysis and competing risks
- Individual versus cohort level risk
  - Modelling heterogeneity-induced competing risks
- Applications:
  - synthetic data
  - prostate and colorectal cancer data
Practical tools for survival analysis with heterogeneity-induced competing risks

ACC Coolen
Institute for Mathematical and Molecular Biomedicine
King’s College London

- Survival analysis and competing risks
- Individual versus cohort level risk
- Modelling heterogeneity-induced competing risks

Applications:
- synthetic data
- prostate and colorectal cancer data
Practical tools for survival analysis with heterogeneity-induced competing risks

ACC Coolen
Institute for Mathematical and Molecular Biomedicine
King’s College London

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

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

$$z_i = (z_i^1, \ldots, z_i^p) : \text{values of } p \text{ covariates}$$

$$t_i \geq 0 : \text{time of first event}$$

$$r_i \in \{1, \ldots, R\} : \text{type of first event}$$

Question:

- Extract regularities that relate covariates to risks
  - associations between risks and modifiable covariates
  - covariate-conditioned survival prediction
  - targeted treatment for those most at risk
Survival analysis and competing risks

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

  $\mathbf{z}_i = (z^i_1, \ldots, z^i_p)$: values of $p$ covariates
  $t_i \geq 0$: time of first event
  $r_i \in \{1, \ldots, R\}$: type of first event

**Question:**

- Extract regularities that relate covariates to risks
  - associations between risks and modifiable covariates
  - covariate-conditioned survival prediction
  - targeted treatment for those most at risk
competing risk problem

- Traditional methods (e.g. Kaplan-Meier, Cox) assume risks have uncorrelated event times (censoring is noninformative)

\[ P(t_1, \ldots, t_R|z) = P(t_1|z)P(t_2, \ldots, t_R|z) \]

- If correlated event times: informative censoring
  - primary hazard rate contaminated by non-primary risks (‘false protectivity’, ‘false aetiology’)

predicted survival probabilities can be misleading ...

Graphs showing survival probabilities over time for primary risk only and primary+secondary scenarios.
competing risk problem

- Traditional methods (e.g. Kaplan-Meier, Cox) assume risks have uncorrelated event times (censoring is noninformative)

\[ P(t_1, \ldots, t_R | \mathbf{z}) = P(t_1 | \mathbf{z}) P(t_2, \ldots, t_R | \mathbf{z}) \]

- If correlated event times: informative censoring
  primary hazard rate contaminated by non-primary risks (‘false protectivity’, ‘false aetiology’)

predicted survival probabilities can be misleading ...
competing risk problem

- Traditional methods (e.g. Kaplan-Meier, Cox) assume risks have uncorrelated event times (censoring is noninformative)

\[ P(t_1, \ldots, t_R | z) = P(t_1 | z)P(t_2, \ldots, t_R | z) \]

- If correlated event times: informative censoring primary hazard rate contaminated by non-primary risks (‘false protectivity’, ‘false aetiology’)

predicted survival probabilities can be misleading ...
What would be ‘decontaminated’ primary risk survival function, if other risks were disabled?

follows from \( P(t_1|z) = \int_0^\infty \cdots \int_0^\infty dt_2 \cdots dt_R P(t_1, \ldots, t_R|z) \)

Tsiatis (1975):
without further assumptions one cannot infer \( P(t_1, \ldots, t_R|z) \) or \( P(t_1|z) \) from survival data

Possible routes
– assume risk independence
  (Cox, KM, frailty & random effects models)
– don’t try to decontaminate ... (Fine & Gray)
– constrain math form of \( P(t_1, \ldots, t_R|z) \)
  (Heckmannn & Honoré, Abbrinng & vd Berg)
– Bayesian methods
  (multiple possible explanantions, but not equally probable ...)

ACC Coolen (IMMB@KCL)
What would be ‘decontaminated’ primary risk survival function, *if other risks were disabled*?

follows from \( P(t_1 | z) = \int_0^\infty \cdots \int_0^\infty dt_2 \cdots dt_R P(t_1, \ldots, t_R | z) \)

Tsiatis (1975):

without further assumptions one cannot infer \( P(t_1, \ldots, t_R | z) \) or \( P(t_1 | z) \) from survival data

Possible routes

– assume risk independence
  
  \textit{(Cox, KM, frailty & random effects models)}

– don’t try to decontaminate ... \textit{(Fine & Gray)}

– constrain math form of \( P(t_1, \ldots, t_R | z) \)

  \textit{(Heckmannn & Honoré, Abbring & vd Berg)}

– Bayesian methods

  \textit{(multiple possible explanations, but not equally probable ...)}
What would be ‘decontaminated’ primary risk survival function, if other risks were disabled?

follows from  
\[ P(t_1 | z) = \int_0^\infty \cdots \int_0^\infty dt_2 \cdots dt_R P(t_1, \ldots, t_R | z) \]

Tsiatis (1975):
without further assumptions one cannot infer  
\( P(t_1, \ldots, t_R | z) \) or \( P(t_1 | z) \) from survival data

Possible routes
– assume risk independence  
  \((Cox, KM, frailty & random effects models)\)
– don’t try to decontaminate ... \((Fine & Gray)\)
– constrain math form of \( P(t_1, \ldots, t_R | z) \)  
  \((Heckmannn & Honoré, Abbring & vd Berg)\)
– Bayesian methods  
  (multiple possible explanations, but not equally probable ...)
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:

  - $(h_A^\uparrow, h_B^\uparrow)$
    - 480
  - $(h_A^\downarrow, h_B^\uparrow)$
    - 20
  - $(h_A^\uparrow, h_B^\downarrow)$
    - 20
  - $(h_A^\downarrow, h_B^\downarrow)$
    - 480
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:
  
  - $(h_A \uparrow, h_B \uparrow)$: 480
  - $(h_A \downarrow, h_B \uparrow)$: 20
  - $(h_A \uparrow, h_B \downarrow)$: 20
  - $(h_A \downarrow, h_B \downarrow)$: 480
to make progress:

model all risks and their relations at individual and cohort level

event time statistics:
\[ P(t_1, \ldots, t_R) \]  
\[ P_i(t_1, \ldots, t_R) \]

cause-specific hazard rates:
\[ h_r(t) \]  
\[ h^i_r(t) \]

cause-specific survival functions:
\[ S_r(t) \]  
\[ S^i_r(t) \]

links:
\[ P(t_1, \ldots, t_R) = \frac{1}{N} \sum_{i=1}^{N} P_i(t_1, \ldots, t_R) \]
\[ S_r(t) = \frac{1}{N} \sum_{i=1}^{N} S^i_r(t) \]

\[ h_r(t) = \frac{\sum_{i=1}^{N} h^i_r(t)e^{-\sum_{r'=1}^{R} \int_0^t ds h^i_{r'}(s)}}{\sum_{i=1}^{N} e^{-\sum_{r'=1}^{R} \int_0^t ds h^i_{r'}(s)}} \]
to make progress:

model all risks and their relations at individual and cohort level

event time statistics:

cohort: $\mathcal{P}(t_1, \ldots, t_R)$

individual $i$: $\mathcal{P}_i(t_1, \ldots, t_R)$

cause-specific hazard rates:

cohort: $h_r(t)$

individual $i$: $h^i_r(t)$

cause-specific survival functions:

cohort: $S_r(t)$

individual $i$: $S^i_r(t)$

links:

\[
\mathcal{P}(t_1, \ldots, t_R) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{P}_i(t_1, \ldots, t_R)
\]

\[
S_r(t) = \frac{1}{N} \sum_{i=1}^{N} S^i_r(t)
\]

\[
h_r(t) = \frac{\sum_{i=1}^{N} h^i_r(t) e^{-\sum_{r'=1}^{R} \int_0^t ds \ h^i_{r'}(s)}}{\sum_{i=1}^{N} e^{-\sum_{r'=1}^{R} \int_0^t ds \ h^i_{r'}(s)}}
\]
to make progress:

model all risks and their relations
at individual and cohort level

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:
\[
\mathcal{P}(t_1, \ldots, t_R) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{P}_i(t_1, \ldots, t_R)
\]
\[
S_r(t) = \frac{1}{N} \sum_{i=1}^{N} S_r^i(t)
\]
\[
h_r(t) = \frac{\sum_{i=1}^{N} h_r^i(t) e^{-\sum_{r'=1}^{R} \int_0^t ds h_r^{i'}(s)}}{\sum_{i=1}^{N} e^{-\sum_{r'=1}^{R} \int_0^t ds h_r^{i'}(s)}}
\]
Cohort complexity

**level 1**: homogeneous cohort, no competing risks
\[
P_i(t_1, \ldots, t_R) = \prod_r P(t_r | z_i)
\]
\[
P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z)
\]

**level 2**: heterogeneous cohort, no competing risks
\[
P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r)
\]
\[
P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z)
\]

**level 3**: heterogeneity-induced competing risks
\[
P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r)
\]
\[
P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z)
\]

**level 4**: individual and cohort level competing risks
\[
P_i(t_1, \ldots, t_R) \neq \prod_{r=1} P_i(t_r)
\]
\[
P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z)
\]
Cohort complexity

**level 1** : homogeneous cohort, no competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P(t_r|z_i) \]
\[ P(t_1, \ldots, t_R|z) = \prod_r P(t_r|z) \]

**level 2** : heterogeneous cohort, no competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R|z) = \prod_r P(t_r|z) \]

**level 3** : heterogeneity-induced competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R|z) \neq \prod_r P(t_r|z) \]

**level 4** : individual *and* cohort level competing risks
\[ P_i(t_1, \ldots, t_R) \neq \prod_{r=1}^{R} P_i(t_r) \]
\[ P(t_1, \ldots, t_R|z) \neq \prod_r P(t_r|z) \]
Cohort complexity

level 1: homogeneous cohort, no competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P(t_r | z_i) \]
\[ P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z) \]

level 2: heterogeneous cohort, no competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z) \]

level 3: heterogeneity-induced competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z) \]

level 4: individual and cohort level competing risks
\[ P_i(t_1, \ldots, t_R) \neq \prod_{r=1} P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z) \]
Cohort complexity

**level 1**: homogeneous cohort, no competing risks

\[ P_i(t_1, \ldots, t_R) = \prod_r P(t_r | z_i) \]
\[ P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z) \]

**level 2**: heterogeneous cohort, no competing risks

\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z) \]

**level 3**: heterogeneity-induced competing risks

\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z) \]

**level 4**: individual and cohort level competing risks

\[ P_i(t_1, \ldots, t_R) \neq \prod_{r=1} P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z) \]
Cohort complexity

level 1 : homogeneous cohort, no competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P(t_r | z_i) \]
\[ P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z) \]

level 2 : heterogeneous cohort, no competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z) \]

level 3 : heterogeneity-induced competing risks
\[ P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z) \]

level 4 : individual and cohort level competing risks
\[ P_i(t_1, \ldots, t_R) \neq \prod_{r=1}^R P_i(t_r) \]
\[ P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z) \]
Cohort complexity

**level 1**: homogeneous cohort, no competing risks
\[
\mathcal{P}_i(t_1, \ldots, t_R) = \prod_r \mathcal{P}(t_r | z_i) \\
\mathcal{P}(t_1, \ldots, t_R | z) = \prod_r \mathcal{P}(t_r | z)
\]

**level 2**: heterogeneous cohort, no competing risks
\[
\mathcal{P}_i(t_1, \ldots, t_R) = \prod_r \mathcal{P}_i(t_r) \\
\mathcal{P}(t_1, \ldots, t_R | z) = \prod_r \mathcal{P}(t_r | z)
\]

**level 3**: heterogeneity-induced competing risks
\[
\mathcal{P}_i(t_1, \ldots, t_R) = \prod_r \mathcal{P}_i(t_r) \\
\mathcal{P}(t_1, \ldots, t_R | z) \neq \prod_r \mathcal{P}(t_r | z)
\]

**level 4**: individual and cohort level competing risks
\[
\mathcal{P}_i(t_1, \ldots, t_R) \neq \prod_{r=1} \mathcal{P}_i(t_r) \\
\mathcal{P}(t_1, \ldots, t_R | z) \neq \prod_r \mathcal{P}(t_r | z)
\]
Cohort complexity

level 1 : homogeneous cohort, no competing risks
\[ \mathcal{P}_i(t_1, \ldots, t_R) = \prod_r \mathcal{P}(t_r | z_i) \]
\[ \mathcal{P}(t_1, \ldots, t_R | z) = \prod_r \mathcal{P}(t_r | z) \]

level 2 : heterogeneous cohort, no competing risks
\[ \mathcal{P}_i(t_1, \ldots, t_R) = \prod_r \mathcal{P}_i(t_r) \]
\[ \mathcal{P}(t_1, \ldots, t_R | z) = \prod_r \mathcal{P}(t_r | z) \]

level 3 : heterogeneity-induced competing risks
\[ \mathcal{P}_i(t_1, \ldots, t_R) = \prod_r \mathcal{P}_i(t_r) \]
\[ \mathcal{P}(t_1, \ldots, t_R | z) \neq \prod_r \mathcal{P}(t_r | z) \]

level 4 : individual and cohort level competing risks
\[ \mathcal{P}_i(t_1, \ldots, t_R) \neq \prod_{r=1} \mathcal{P}_i(t_r) \]
\[ \mathcal{P}(t_1, \ldots, t_R | z) \neq \prod_r \mathcal{P}(t_r | z) \]
Cohort complexity

level 1: homogeneous cohort, no competing risks
\[
P_i(t_1, \ldots, t_R) = \prod_r P(t_r | z_i)
\]
\[
P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z)
\]

level 2: heterogeneous cohort, no competing risks
\[
P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r)
\]
\[
P(t_1, \ldots, t_R | z) = \prod_r P(t_r | z)
\]

level 3: heterogeneity-induced competing risks
\[
P_i(t_1, \ldots, t_R) = \prod_r P_i(t_r)
\]
\[
P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z)
\]

level 4: individual and cohort level competing risks
\[
P_i(t_1, \ldots, t_R) \neq \prod_{r=1} P_i(t_r)
\]
\[
P(t_1, \ldots, t_R | z) \neq \prod_r P(t_r | z)
\]
Heterogeneity-induced competing risks

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

\[
\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}] = \frac{\sum_{i,z_i=\mathbf{z}} \prod_r \delta_F[h_r-h^i_r]}{\sum_{i,z_i=\mathbf{z}} 1}
\]

risk \( r \) hazard rate of individual \( i \)

\[
h^i_r = \{h^i_r(t)\}
\]

disabling non-primary risks:

\[
h^i_r \to 0 \quad \text{for all } r > 1
\]

\[
\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}] \to \mathcal{W}[h_1|\mathbf{z}] \prod_{r>1} \delta_F[h_r] \quad \mathcal{W}[h_1|\mathbf{z}] = \frac{\sum_{i,z_i=\mathbf{z}} \delta_F[h_1-h^i_1]}{\sum_{i,z_i=\mathbf{z}} 1}
\]

data log-likelihood:

\[
\mathcal{L}(D|\mathcal{W}) = \sum_{i=1}^N \log \int \{dh_1 \ldots dh_R\} \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}_i] h^i_r(t_i) e^{-\sum_{r=1}^R \int_0^{t_i} ds h_r(s)}
\]
Heterogeneity-induced competing risks

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

\[ \mathcal{W}[h_1, \ldots, h_R | \mathbf{z}] = \frac{\sum_{i,z_i=\mathbf{z}} \prod_r \delta_F[h_r - h^i_r]}{\sum_{i,z_i=\mathbf{z}} 1} \]

risk \( r \) hazard rate of individual \( i \)

disabling non-primary risks:

\[ h^i_r \to 0 \quad \text{for all } r > 1 \]

\[ \mathcal{W}[h_1, \ldots, h_R | \mathbf{z}] \to \mathcal{W}[h_1 | \mathbf{z}] \prod_{r>1} \delta_F[h_r] \]

\[ \mathcal{W}[h_1 | \mathbf{z}] = \frac{\sum_{i,z_i=\mathbf{z}} \delta_F[h_1 - h^i_1]}{\sum_{i,z_i=\mathbf{z}} 1} \]

data log-likelihood:

\[ \mathcal{L}(D | \mathcal{W}) = \sum_{i=1}^{N} \log \int \{dh_1 \ldots dh_R\} \mathcal{W}[h_1, \ldots, h_R | \mathbf{z}_i] h^i_r(t_i) e^{-\sum_{r=1}^{R} \int_{0}^{t_i} ds h_r(s)} \]
Heterogeneity-induced competing risks

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

$$\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}] = \frac{\sum_{i,z_i=\mathbf{z}} \prod_r \delta_F[h_r - h^i_r]}{\sum_{i,z_i=\mathbf{z}} 1}$$

hazard rate of individual $i$

risk $r$ hazard rate

disabling non-primary risks:

$h^i_r \to 0 \quad \text{for all } r > 1$

$$\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}] \to \mathcal{W}[h_1|\mathbf{z}] \prod_{r>1} \delta_F[h_r] \quad \mathcal{W}[h_1|\mathbf{z}] = \frac{\sum_{i,z_i=\mathbf{z}} \delta_F[h_1 - h^i_1]}{\sum_{i,z_i=\mathbf{z}} 1}$$

data log-likelihood:

$$\mathcal{L}(D|\mathcal{W}) = \sum_{i=1}^N \log \int \{dh_1 \ldots dh_R\} \mathcal{W}[h_1, \ldots, h_R|\mathbf{z}_i] h^i_r(t_i) e^{-\sum_{r=1}^R \int_0^{t_i} ds h_r(s)}$$
Decontamination formulae

‘crude’ cause-specific quantities:

\[
S_r(t|z) = e^{- \int_0^t ds \ h_r(s|z)}
\]

\[
h_r(t|z) = \frac{\int \{dh_1 \ldots dh_R\} \ W[h_1, \ldots, h_R|z] \ h_r(t) e^{- \sum_{r'} \int_0^t ds \ h_{r'}(s)}}{\int \{dh_1 \ldots dh_R\} \ W[h_1, \ldots, h_R|z] \ e^{- \sum_{r'} \int_0^t ds \ h_{r'}(s)}}
\]

decontaminated:

\[
\tilde{S}_r(t|z) = \int \{dh_1 \ldots dh_R\} \ W[h_1, \ldots, h_R|z] \ e^{- \int_0^t ds \ h_r(s)}
\]

\[
\tilde{h}_r(t|z) = \frac{\int \{dh_1 \ldots dh_R\} \ W[h_1, \ldots, h_R|z] \ h_r(t) e^{- \int_0^t ds \ h_r(s)}}{\int \{dh_1 \ldots dh_R\} \ W[h_1, \ldots, h_R|z] \ e^{- \int_0^t ds \ h_r(s)}}
\]
Decontamination formulae

‘crude’ cause-specific quantities:

\[
S_r(t|z) = e^{-\int_0^t ds \ h_r(s|z)}
\]

\[
h_r(t|z) = \frac{\int\{dh_1 \ldots dh_R\} \ \mathcal{W}[h_1, \ldots, h_R|z] \ h_r(t) e^{-\sum_{r'} \int_0^t ds \ h_{r'}(s)}}{\int\{dh_1 \ldots dh_R\} \ \mathcal{W}[h_1, \ldots, h_R|z] e^{-\sum_{r'} \int_0^t ds \ h_{r'}(s)}}
\]

decontaminated:

\[
\tilde{S}_r(t|z) = \int\{dh_1 \ldots dh_R\} \ \mathcal{W}[h_1, \ldots, h_R|z] \ e^{-\int_0^t ds \ h_r(s)}
\]

\[
\tilde{h}_r(t|z) = \frac{\int\{dh_1 \ldots dh_R\} \ \mathcal{W}[h_1, \ldots, h_R|z] \ h_r(t) e^{-\int_0^t ds \ h_r(s)}}{\int\{dh_1 \ldots dh_R\} \ \mathcal{W}[h_1, \ldots, h_R|z] \ e^{-\int_0^t ds \ h_r(s)}}
\]
Parametrisations of $\mathcal{W}[h_1, \ldots, h_R | z]$

proportional hazards at level of individuals

$$\mathcal{W}[h_1, \ldots, h_R | z] = \int d\beta_1 \ldots d\beta_R \int \{d\lambda_1 \ldots d\lambda_R\} \mathcal{M}(\beta_1, \ldots, \beta_R; \lambda_1, \ldots, \lambda_R)$$

$$\times \prod r \delta_F \left[ h_r - \lambda_r e^{\beta_0^r + \sum_{\mu=1}^p \beta^r_{\mu} z_\mu} \right]$$

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

- e.g. latent class heterogeneity:

$$\mathcal{M}(\beta_1, \ldots, \beta_R; \lambda_1, \ldots, \lambda_R) = \mathcal{M}(\beta_1, \ldots, \beta_R) \prod_{r=1}^R \delta_F[\lambda_r - \hat{\lambda}_r]$$

$$\mathcal{M}(\beta_1, \ldots, \beta_R) = \sum_{\ell=1}^L w_\ell \prod_{r=1}^R \delta(\beta_r - \hat{\beta}_r^{\ell})$$

$$\hat{\beta}_r^{\ell} = (\hat{\beta}_r^{\ell 0}, \ldots, \hat{\beta}_r^{\ell p})$$
parametrizations of \( \mathcal{W}[h_1, \ldots, h_R|z] \)

proportional hazards at level of individuals

\[
\mathcal{W}[h_1, \ldots, h_R|z] = \int d\beta_1 \ldots d\beta_R \int \{d\lambda_1 \ldots d\lambda_R\} \mathcal{M}(\beta_1, \ldots, \beta_R; \lambda_1, \ldots, \lambda_R) \\
\times \prod_r \delta_F\left[h_r - \lambda_r e^{\beta_0^r + \sum_{\mu=1}^p \beta_\mu^r z_\mu}\right]
\]

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

- e.g. latent class heterogeneity:

\[
\mathcal{M}(\beta_1, \ldots, \beta_R; \lambda_1, \ldots, \lambda_R) = \mathcal{M}(\beta_1, \ldots, \beta_R) \prod_{r=1}^R \delta_F[\lambda_r - \hat{\lambda}_r] \\
\mathcal{M}(\beta_1, \ldots, \beta_R) = \sum_{\ell=1}^L \mathcal{W}_\ell \prod_{r=1}^R \delta(\beta_r - \hat{\beta}_\ell^r) \\
\hat{\beta}_r^\ell = (\hat{\beta}_r^{\ell0}, \ldots, \hat{\beta}_r^{\ell p})
\]
Applications – synthetic data

$S^R_{KM} :$ Kaplan-Meier

$S_1 :$ crude survival curve

red dashed: true survival curves
Applications – synthetic data

\( S_{1}^{\text{KM}} \): Kaplan-Meier
\( S_{1} \): crude survival curve
\( \tilde{S}_{1} \): decontaminated curves

*red dashed*: true survival curves
retrospective class identification

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

Data:

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

\[\beta_1^1 = (0.5, 0.5, 0.5) + (2, 0, 2)\]
\[\beta_2^1 = (0.5, 0.5, 0.5) + (-2, -2, 0)\]
\[\beta_3^1 = (0.5, 0.5, 0.5) + (0, 2, -2)\]

each individual \(i\):

point \((p_{1i}^i, p_{2i}^i, p_{3i}^i)\) in \(\mathbb{R}^3\)

\[p_{\ell i}^i = P(\ell | t_i, r_i, z_i)\]
retrospective class identification

\[ P(\ell|t, r, z) = \frac{W_\ell \, e^{\hat{\beta}_r \cdot z - \sum_{r'}^R \exp(\hat{\beta}_{r'} \cdot z) \int_0^t ds \, \hat{\lambda}_{r'}(s)}}{\sum_{\ell'}^L W_{\ell'} \, e^{\hat{\beta}_{r'} \cdot z - \sum_{r'}^R \exp(\hat{\beta}_{r'} \cdot z) \int_0^t ds \, \hat{\lambda}_{r'}(s)}} \]

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 \( (p_{1i}, p_{2i}, p_{3i}) \) in \( \mathbb{R}^3 \)

\[ p_\ell^i = P(\ell|t_i, r_i, z_i) \]
retrospective class identification

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

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 \( (p_{1i}, p_{2i}, p_{3i}) \) in \( \mathbb{R}^3 \)

\[ p_{\ell i} = P(\ell | t_i, r_i, z_i) \]
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:

\[
\begin{array}{cc}
\beta_1 &= 0.14 \\
\beta_2 &= -0.15 \\
\beta_3 &= 0.20 \\
\beta_4 &= -0.09 \\
\beta_5 &= -0.08 \\
\end{array}
\]

\[
HR_\mu = \exp(2\beta_\mu)
\]
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:

<table>
<thead>
<tr>
<th>BMI</th>
<th>selenium</th>
<th>phys1</th>
<th>phys2</th>
<th>smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = 0.14$</td>
<td>$\beta_2 = -0.15$</td>
<td>$\beta_3 = 0.20$</td>
<td>$\beta_4 = -0.09$</td>
<td>$\beta_5 = -0.08$</td>
</tr>
</tbody>
</table>

$HR_\mu = \exp(2\beta_\mu)$
<table>
<thead>
<tr>
<th>CLASSES</th>
<th>PRIMARY RISK</th>
<th>SECONDARY RISK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>208 events</td>
<td>910 events</td>
</tr>
<tr>
<td></td>
<td>BMI selen phys1 phys2 smok</td>
<td>BMI selen phys1 phys2 smok</td>
</tr>
</tbody>
</table>

| 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 | $w_2 = 0.49$ -0.07 -0.16 0.19 -0.10 -0.27 |
| frailties| $\beta_{10}^1 - \beta_{10}^2 = -4.61$ (HR 0.010) | $\beta_{20}^1 - \beta_{20}^2 = -4.06$ (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 effect?)
\( S_1^{KM} \): 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, \n**\( S_1 \): crude survival curves, \n**\( \tilde{S}_1 \): 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)
colorectal cancer trial, nr of patients $N = 155$

139 events, times $t = 260 \pm 160$
16 censoring, times $t = 750 \pm 220$

covariates:
FRET efficiency for Her2-Her3 dimer
Her3 concentration
Her2-Her3 dimer concentration
Her2 concentration
Cetuximab treatment, 1=no, 2=yes
KRAS mutation, 0=no, 1=yes

Cox regression hazard ratios:

<table>
<thead>
<tr>
<th>FRET</th>
<th>Her3</th>
<th>Her2-Her3</th>
<th>Her2</th>
<th>Cetuximab</th>
<th>KRAS mut</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>1.8</td>
<td>1.1</td>
<td>0.7</td>
<td>1.7</td>
</tr>
</tbody>
</table>
Applications – COIN colorectal cancer data set

colorectal cancer trial, nr of patients $N = 155$

139 events, times $t = 260 \pm 160$
16 censoring, times $t = 750 \pm 220$

covariates:
FRET efficiency for Her2-Her3 dimer
Her3 concentration
Her2-Her3 dimer concentration
Her2 concentration
Cetuximab treatment, 1=no, 2=yes
KRAS mutation, 0=no, 1=yes

Cox regression hazard ratios:

<table>
<thead>
<tr>
<th>FRET</th>
<th>Her3</th>
<th>Her2-Her3</th>
<th>Her2</th>
<th>Cetuximab</th>
<th>KRAS mut</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>1.8</td>
<td>1.1</td>
<td>0.7</td>
<td>1.7</td>
</tr>
</tbody>
</table>
Bayesian model selection

**model 1**

**model 2**

**model 3**

model 1: classes differ in overall frailties only
model 2: classes differ in overall frailties and association pars
model 3: classes differ in full base hazard rates and association pars

most probable explanation of COIN data:
model 2, with $L = 2$ and $K = 3$

$L$: nr of classes

$K$: complexity of base hazard rates
Two sub-cohorts, with similar base hazard rates, but distinct overall frailties and associations.

Hazard ratios:
> 1: elevated risk, < 1: reduced risk

<table>
<thead>
<tr>
<th></th>
<th>FRET</th>
<th>Her3</th>
<th>Her2-Her3</th>
<th>Her2</th>
<th>Cetuximab</th>
<th>KRAS mut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox</td>
<td>0.5</td>
<td>1.0</td>
<td>1.8</td>
<td>1.1</td>
<td>0.7</td>
<td>1.7</td>
</tr>
<tr>
<td>new model:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>class I, 40%</td>
<td>0.7</td>
<td>1.5</td>
<td>3.7</td>
<td>1.1</td>
<td>0.3</td>
<td>2.5</td>
</tr>
<tr>
<td>class II, 60%</td>
<td>0.6</td>
<td>1.2</td>
<td>0.7</td>
<td>0.9</td>
<td>1.1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

significantly higher overall risk in class II

new quantitative tools to identify a priori the responders to Cetuximab?
Competing risk problem solvable if we assume risk correlations are caused by *residual heterogeneity* (heterogeneity not captured by covariates).

Natural parametrisation of $\mathcal{W}[h_1, \ldots, h_R|z]$, includes standard methods as special cases (Cox, frailty models, random effects models, ...)

Practical tools:
- Formulae for decontaminated survival curves,
- Formulae for retrospective Bayesian class assignment

Synthetic data:
Method detects structure, parameters, and survival curves correctly

ULSAM and COIN cancer data:
New intuitive explanations for previously unexplained results, useful aid in discovery of relevant new biomarkers.
competing risk problem solvable if we assume risk correlations are caused by residual heterogeneity (heterogeneity not captured by covariates)

natural parametrisation of $\mathcal{W}[h_1, \ldots, h_R|z]$, includes standard methods as special cases (Cox, frailty models, random effects models, ...)

practical tools:
– formulae for decontaminated survival curves,
– formulae for retrospective Bayesian class assignment

synthetic data:
method detects structure, parameters, and survival curves correctly

ULSAM and COIN cancer data:
new intuitive explanations for previously unexplained results, useful aid in discovery of relevant new biomarkers
competing risk problem solvable if we assume risk correlations are caused by *residual heterogeneity* (heterogeneity not captured by covariates)

natural parametrisation of $\mathcal{W}[h_1, \ldots, h_R | \mathbf{z}]$, includes standard methods as special cases (Cox, frailty models, random effects models, ...)

practical tools:
– formulae for decontaminated survival curves,
– formulae for retrospective Bayesian class assignment

synthetic data:
method detects structure, parameters, and survival curves correctly

ULSAM and COIN cancer data:
new intuitive explanations for previously unexplained results, useful aid in discovery of relevant new biomarkers


- competing risk problem solvable if we assume risk correlations are caused by *residual heterogeneity* (heterogeneity not captured by covariates)

- natural parametrisation of $\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]$, includes standard methods as special cases (Cox, frailty models, random effects models, ...)

- practical tools:
  - formulae for decontaminated survival curves,
  - formulae for retrospective Bayesian class assignment

- synthetic data:
  method detects structure, parameters, and survival curves correctly

- ULSAM and COIN cancer data:
  new intuitive explanations for previously unexplained results, useful aid in discovery of relevant new biomarkers
competing risk problem solvable if we assume risk correlations are caused by *residual heterogeneity* (heterogeneity not captured by covariates)

natural parametrisation of $\mathcal{W}[h_1, \ldots, h_R|\mathbf{z}]$, includes standard methods as special cases (Cox, frailty models, random effects models, ...)

practical tools:
- formulae for decontaminated survival curves,
- formulae for retrospective Bayesian class assignment

synthetic data:
method detects structure, parameters, and survival curves correctly

ULSAM and COIN cancer data:
new intuitive explanations for previously unexplained results, useful aid in discovery of relevant new biomarkers
Thanks to

Collaborators
Hans van Baardewijk, Hans Garmö, Birgitta Grundmark, Mieke van Hemelrijck, Lars Holmberg, Mark Rowley, Björn Zethelius

Discussions
Shola Agbaje, Salma Ayis, James Barrett, Maria D’Iorio, Niels Keiding, Katherine Lawler

Funding