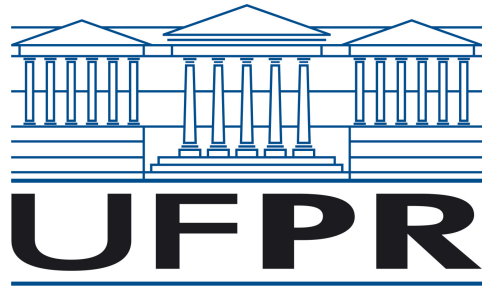
Modeling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing

paper presentation



Henrique Laureano
http://leg.ufpr.br/~henrique

Last modification on 2020-03-31 21:34:40

"Modeling the **cumulative incidence function** of **multivariate competing risks data** allowing for **within-cluster dependence** of risk and timing"

what?

"Modeling the **cumulative incidence function** of **multivariate competing risks data** allowing for **within-cluster dependence** of risk and timing"

what?

» cause-specific **cumulative incidence function** (CIF).

$$\text{for a type } j \text{ failure,} \quad F_j(t \mid X) = \mathbb{P}[T \leq t, J = j \mid X]$$
$$= \int_0^t f_j(u \mid X) du, \quad t > 0,$$

where $f_j(t \mid X) = \lambda_j(t \mid X) \times S(t \mid X)$ is the (sub)density for the time to a type j failure.

» ...

"Modeling the **cumulative incidence function** of **multivariate competing risks data** allowing for **within-cluster dependence** of risk and timing"

what?

- ...
- **multivariate competing risks**,

  *we have more than one, that is why it is **multivariate**, cause of interest **competing** to be responsible by the failure (if not censor).*

"Modeling the **cumulative incidence function** of **multivariate competing risks data** allowing for **within-cluster dependence** of risk and timing"

what?

- ...
- **multivariate competing risks**,

  *we have more than one, that is why it is **multivariate**, cause of interest **competing** to be responsible by the failure (if not censor).*

- multivariate competing risks **data**, i.e.,

  we'll not do a multivariate competing risks model,
  we'll do a model for multivariate competing risks data!

"Modeling the **cumulative incidence function** of **multivariate competing risks data** allowing for **within-cluster dependence** of risk and timing"

what?

- ...
- **within-cluster dependence**, i.e., a random/latent effect structure for
  - risk: how a failure occurrence relates to other;
  - timing: some failures aren't likely to happen equally all time and the failure time distribution may vary between clusters.

paper structure

- **intro**: ideas, motivation and 'selling the fish';
- **model**: model specification, likelihood, estimation and extras;
- simulation results;
- application: **Danish register-based family data on breast cancer**;
- final remarks.

ideas, motivation and 'selling the fish'

ideas, motivation and 'selling the fish'

---

focus: **family studies** and why a random effects approach

- the within-cluster dependence, which is here a **within-family dependence**, is often the key point of interest or at least as important as determining the role of different risk factors;

ideas, motivation and 'selling the fish'

---

focus: **family studies** and why a random effects approach

- the within-cluster dependence, which is here a **within-family dependence**, is often the key point of interest or at least as important as determining the role of different risk factors;

- the within-family dependence can be viewed as an expression of familial aggregation and may reflect both disease **heritability** and the impact of shared **environmental effects**.

About the **model approach**: *what we could do?*

- a frailty-based two-stage approach, where the marginal CIFs are estimated in the 1st stage and a dependence parameter is estimated in the 2nd stage using an Archimedean **copula**.

And why we don't do that?

- The necessary to **adjust for right-censoring**. This is done through modeling of the censoring distribution and employment of inverse probability of censoring weights (IPCWs);
- If the **censoring distribution is misspecified**, the weighting may introduce bias.

The idea is to model the cluster-specific CIF as a product

The idea is to model the cluster-specific CIF as a product

For two competing causes of failure we write

$$F_i(t \mid X, \eta_i, u_1, u_2) = \underbrace{\pi_i(X, u_1, u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[\alpha_i\{g(t)\} - X^\top \gamma_i - \eta_i]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad i = 1, 2,$$

where

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ u_1 \\ u_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\eta_1} & \varrho_{\eta_1,\eta_2} & \varrho_{\eta_1,u_1} & \varrho_{\eta_1,u_2} \\ & \sigma^2_{\eta_2} & \varrho_{\eta_2,u_1} & \varrho_{\eta_2,u_2} \\ & & \sigma^2_{u_1} & \varrho_{u_1,u_2} \\ & & & \sigma^2_{u_2} \end{pmatrix} \right].$$

- The cluster-specific survivor function is given as $S(t \mid X, \boldsymbol{\eta}, \boldsymbol{u}) = 1 - F_i(t \mid X, \eta_i, \boldsymbol{u})$, $i = 1, 2$.

The idea is to model the cluster-specific CIF as a product

For two competing causes of failure we write

$$F_i(t \mid X, \eta_i, u_1, u_2) = \underbrace{\pi_i(X, u_1, u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[\alpha_i\{g(t)\} - X^\top \gamma_i - \eta_i]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad i = 1, 2,$$

where

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ u_1 \\ u_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\eta_1} & \varrho_{\eta_1,\eta_2} & \varrho_{\eta_1,u_1} & \varrho_{\eta_1,u_2} \\ & \sigma^2_{\eta_2} & \varrho_{\eta_2,u_1} & \varrho_{\eta_2,u_2} \\ & & \sigma^2_{u_1} & \varrho_{u_1,u_2} \\ & & & \sigma^2_{u_2} \end{pmatrix} \right].$$

- The cluster-specific survivor function is given as $S(t \mid X, \boldsymbol{\eta}, \boldsymbol{u}) = 1 - F_i(t \mid X, \eta_i, \boldsymbol{u})$, $i = 1, 2$.

Why modeling the CIF?

Proposing a model for the CIF provides a framework for exploring and making inference about the distribution of age at disease onset.

cluster-specific CIF

$$F_i(t \mid X, \eta_i, u_1, u_2) = \underbrace{\pi_i(X, u_1, u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[\alpha_i\{g(t)\} - X^\top \gamma_i - \eta_i]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad i = 1, 2.$$

This separation of the CIF is possible via the transformation of the time variable t, given as

$$g(t) = \text{arctanh}\left(\frac{t - \delta/2}{\delta/2}\right) \quad \text{for} \quad t \in]0, \delta[.$$

cluster-specific CIF

$$F_i(t \mid X, \eta_i, u_1, u_2) = \underbrace{\pi_i(X, u_1, u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[\alpha_i\{g(t)\} - X^\top \gamma_i - \eta_i]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad i = 1, 2.$$

This separation of the CIF is possible via the transformation of the time variable t, given as

$$g(t) = \text{arctanh}\left(\frac{t - \delta/2}{\delta/2}\right) \quad \text{for} \quad t \in]0, \delta[.$$

- With this transformation the value of the cluster-specific **failure time trajectory** will equal 1 at time δ, a fixed time point at which all individuals still at risk are censored.

cluster-specific CIF

$$F_i(t \mid X, \eta_i, u_1, u_2) = \underbrace{\pi_i(X, u_1, u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[\alpha_i\{g(t)\} - X^\top \gamma_i - \eta_i]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad i = 1, 2.$$

This separation of the CIF is possible via the transformation of the time variable t, given as

$$g(t) = \text{arctanh}\left(\frac{t - \delta/2}{\delta/2}\right) \quad \text{for} \quad t \in]0, \delta[.$$

- With this transformation the value of the cluster-specific **failure time trajectory** will equal 1 at time δ, a fixed time point at which all individuals still at risk are censored.
- $\alpha_i(x)$, $i = 1, 2$, are monotonically increasing functions of x and known up to a parameter vector, w_i, $i = 1, 2$. e.g., monotonically increasing B-spline or piecewise linear functions.

cluster-specific CIF

$$F_i(t \mid X, \eta_i, u_1, u_2) = \underbrace{\pi_i(X, u_1, u_2)}_{\substack{\text{cluster-specific} \\ \text{risk level}}} \times \underbrace{\Phi[\alpha_i\{g(t)\} - X^\top \gamma_i - \eta_i]}_{\substack{\text{cluster-specific} \\ \text{failure time trajectory}}}, \quad i = 1, 2.$$

The cluster-specific **risk levels** are modeled using a **multinomial logistic regression model with random effects**, i.e.

$$\pi_i(X, \boldsymbol{u}) = \frac{\exp\{X^\top \beta_i + u_i\}}{1 + \sum_{j=1}^{2} \exp\{X^\top \beta_j + u_j\}}, \quad i = 1, 2.$$

- We employ that **multinomial model** to ensure that the sum of the predicted CIFs do not exceed 1.

nice aspects or consequences

» at time δ, where $F_i(\delta \mid X, \eta_i, \boldsymbol{u}) = \pi_i(X, \boldsymbol{u})$, the cluster-specific **survival function** is given by

$$S(\delta \mid X, \boldsymbol{\eta}, \boldsymbol{u}) = \frac{1}{1 + \sum_{j=1}^{2} \exp\{X^\top \beta_j + u_j\}};$$

- at time δ, where $F_i(\delta \mid X, \eta_i, \boldsymbol{u}) = \pi_i(X, \boldsymbol{u})$, the cluster-specific **survival function** is given by

$$S(\delta \mid X, \boldsymbol{\eta}, \boldsymbol{u}) = \frac{1}{1 + \sum_{j=1}^{2} \exp\{X^\top \beta_j + u_j\}};$$

- the interpretation of the regression coefficients $\boldsymbol{\beta}$s is given by the traditional **odds-ratio**, but now in a **multinomial version**;

- at time δ, where $F_i(\delta \mid X, \eta_i, \boldsymbol{u}) = \pi_i(X, \boldsymbol{u})$, the cluster-specific **survival function** is given by

$$S(\delta \mid X, \boldsymbol{\eta}, \boldsymbol{u}) = \frac{1}{1 + \sum_{j=1}^{2} \exp\{X^\top \beta_j + u_j\}};$$

- the interpretation of the regression coefficients $\boldsymbol{\beta}$s is given by the traditional **odds-ratio**, but now in a **multinomial version**;

- the regression coefficients $\boldsymbol{\gamma}$s reflect how covariates affect the **failure time trajectories**, i.e., the shape of the CIFs;

- ...

Likelihood

I'm already telling you, this paper has a bad and not very explained, notation

---

To accommodate the censorship and the cluster structures, the chosen approach is the **pairwise composite likelihood** given as

$$L(\theta; \boldsymbol{T}, \boldsymbol{\epsilon}, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{u}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i - 1} \prod_{k=j+1}^{n_i} L(\theta; T_{ij}, \epsilon_{ij}, X_{ij}, T_{ik}, \epsilon_{ik}, X_{ik}, \boldsymbol{\eta}_i, \boldsymbol{u}_i),$$

where $\theta = \{\beta_1, \beta_2, \gamma_1, \gamma_2, w_1, w_2, \Sigma_{\boldsymbol{\eta u}}\}^\top$. i.e., still considering just two competing causes.

Likelihood

I'm already telling you, this paper has a bad and not very explained, notation

---

To accommodate the censorship and the cluster structures, the chosen approach is the **pairwise composite likelihood** given as

$$L(\theta; \boldsymbol{T}, \boldsymbol{\epsilon}, \boldsymbol{X}, \boldsymbol{\eta}, \boldsymbol{u}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i - 1} \prod_{k=j+1}^{n_i} L(\theta; T_{ij}, \epsilon_{ij}, X_{ij}, T_{ik}, \epsilon_{ik}, X_{ik}, \boldsymbol{\eta}_i, \boldsymbol{u}_i),$$

where $\theta = \{\beta_1, \beta_2, \gamma_1, \gamma_2, w_1, w_2, \Sigma_{\boldsymbol{\eta u}}\}^\top$. i.e., still considering just two competing causes.

1st step? The same as always.
Integrating out in the random effects **to get the marginal** and using the Bayes'rule to reduce the dimensionality of the integral, we have

$$L_M(\theta; \boldsymbol{T}, \boldsymbol{\epsilon}, \boldsymbol{X}) = \int \pi(\boldsymbol{T}, \boldsymbol{\epsilon} \mid \boldsymbol{X}, \boldsymbol{u}) \pi(\boldsymbol{u}) d\boldsymbol{u}.$$

pairwise composite likelihood

Ignoring the cluster subscript i, the likelihood contribution of the pair j, k to the *pairwise composite likelihood* is given as

$$
\begin{aligned}
L_{jk}(\theta; T_j, \epsilon_j, X_j, T_k, \epsilon_k, X_k, \boldsymbol{\eta}, \boldsymbol{u}) = & \left\{ \prod_{h=1}^{2} \prod_{l=1}^{2} f_h(T_j \mid X_j, \eta_h, \boldsymbol{u}) f_l(T_k \mid X_k, \eta_l, \boldsymbol{u}) \right\} \\
& \times \left\{ \prod_{h=1}^{2} f_h(T_j \mid X_j, \eta_h, \boldsymbol{u}) S(T_k \mid X_k, \boldsymbol{\eta}, \boldsymbol{u}) \right\} \\
& \times \left\{ \prod_{l=1}^{2} S(T_j \mid X_j, \boldsymbol{\eta}, \boldsymbol{u}) f_l(T_k \mid X_k, \eta_l, \boldsymbol{u}) \right\} \\
& \times \left\{ S(T_j \mid X_j, \boldsymbol{\eta}, \boldsymbol{u}) S(T_k \mid X_k, \boldsymbol{\eta}, \boldsymbol{u}) \right\}.
\end{aligned}
$$

The indicator functions were omitted,
but the equation is still 'clear' and readable.

pairwise composite likelihood

In that likelihood, we have four contributions:

- the one when **both** individuals experience failure (either cause);
- two for the case when **only one** individual experiences failure;
- and one for the case when **both** individuals don't experience failure.

We have basically two quantities: $f(\cdot)$ and $S(\cdot)$, i.e, the CIFs derivative wrt t and the survival function.

pairwise composite likelihood

In that likelihood, we have four contributions:

- the one when **both** individuals experience failure (either cause);
- two for the case when **only one** individual experiences failure;
- and one for the case when **both** individuals don't experience failure.

We have basically two quantities: $f(\cdot)$ and $S(\cdot)$, i.e, the CIFs derivative wrt t and the survival function.

Writing down each of the four components (the 2nd and 3rd are symmetric) and, again, integrating out in the random effects based in a Bayes'rule, we obtain the contributions to the conditional densities necessary in the marginal.

- The resulting contributions are basically products of the **multinomial logistic regression model** with **univariate** or **bivariate normals**.

pairwise composite likelihood
↳ ESTIMATION

As expected, the marginal likelihood doesn't have a closed-form. The numerical approach chosen for parameter estimation is the **adaptive Gaussian quadrature** (AGQ) with Gauss-Hermite rules.

pairwise composite likelihood
↳ ESTIMATION

As expected, the marginal likelihood doesn't have a closed-form. The numerical approach chosen for parameter estimation is the **adaptive Gaussian quadrature** (AGQ) with Gauss-Hermite rules.

- the likelihood contributions to the **composite likelihood** from pairs within the same cluster are not independent, as consequence:
  - the Fisher information needs to be substituted by the so-called **sandwich estimator** when estimating the variance of the parameter estimates.

Danish register-based family data on breast cancer

The cohort study is too big (1 292 051 families) and the model is also too complicated (plus the extra computational cost of the **AGQ** approximation).

Danish register-based family data on breast cancer

The cohort study is too big (1 292 051 families) and the model is also too complicated (plus the extra computational cost of the **AGQ** approximation).

Solution? Sampling.

Danish register-based family data on breast cancer

The cohort study is too big (1 292 051 families) and the model is also too complicated (plus the extra computational cost of the **AGQ** approximation).

Solution? Sampling.

Characteristics? Divide the dataset into strata of similar characteristics, and sample from each with desired probabilities with the goal of building a representative sample. i.e., with a good portion of each event type and censoring.

Danish register-based family data on breast cancer

The cohort study is too big (1 292 051 families) and the model is also too complicated (plus the extra computational cost of the **AGQ** approximation).

Solution? Sampling.

Characteristics? Divide the dataset into strata of similar characteristics, and sample from each with desired probabilities with the goal of building a representative sample. i.e., with a good portion of each event type and censoring.

Consequences? We need an estimator of the pairwise composite log-likelihood - a **weighted log-likelihood**; the score function changes and we need a new sandwich estimator for the variance of the parameter estimates.

1000 populations of 50 000 clusters of size three and two competing causes

1000 populations of 50 000 clusters of size three and two competing causes

- parameters fixed, no covariates (just intercept), simple functions and failure probabilities given by the **multinomial logistic model**;

1000 populations of 50 000 clusters of size three and two competing causes

- parameters fixed, no covariates (just intercept), simple functions and failure probabilities given by the **multinomial logistic model**;
- a random number ς from the standard uniform distribution was sampled and the **failure time** found by isolating t in the expression $\varsigma = \Phi[\alpha_i\{g(t)\} - X\gamma_i - \eta_i]$, $i = 1, 2$;

1000 populations of 50 000 clusters of size three and two competing causes

- parameters fixed, no covariates (just intercept), simple functions and failure probabilities given by the **multinomial logistic model**;
- a random number ς from the standard uniform distribution was sampled and the **failure time** found by isolating t in the expression $\varsigma = \Phi[\alpha_i\{g(t)\} - X\gamma_i - \eta_i]$, $i = 1, 2$;
- approximately 50% censoring;

1000 populations of 50 000 clusters of size three and two competing causes

- parameters fixed, no covariates (just intercept), simple functions and failure probabilities given by the **multinomial logistic model**;
- a random number ς from the standard uniform distribution was sampled and the **failure time** found by isolating t in the expression $\varsigma = \Phi[\alpha_i\{g(t)\} - X\gamma_i - \eta_i]$, $i = 1, 2$;
- approximately 50% censoring;
- AGQ approximation with five quadrature points and Gauss-Hermite rules.

1000 populations of 50 000 clusters of size three and two competing causes

- parameters fixed, no covariates (just intercept), simple functions and failure probabilities given by the **multinomial logistic model**;
- a random number ς from the standard uniform distribution was sampled and the **failure time** found by isolating t in the expression $\varsigma = \Phi[\alpha_i\{g(t)\} - X\gamma_i - \eta_i]$, $i = 1, 2$;
- approximately 50% censoring;
- AGQ approximation with five quadrature points and Gauss-Hermite rules.

Overall, the model performed well, the parameter estimates were unbiased and the coverage rates were good.

irregularities: caused by the numerical derivative of the AGQ approximation *or* the number of quadrature points.

Danish register-based family data on breast cancer
↳ more details

*family data on **breast cancer** among women - the most common malignancy in women.*

The cohort consisted of 1 292 051 families and 3 029 653 individuals:

- 908 002 (70.3%) families with a mother and a single daughter;
- 322 547 (25.0%) families with a mother and two daughters;
- 61 502 (4.8%) families with a mother and three daughters.

Danish register-based family data on breast cancer
↳ more details

*family data on **breast cancer** among women - the most common malignancy in women.*

The cohort consisted of 1 292 051 families and 3 029 653 individuals:

- 908 002 (70.3%) families with a mother and a single daughter;
- 322 547 (25.0%) families with a mother and two daughters;
- 61 502 (4.8%) families with a mother and three daughters.

---

aim: investigate the cumulative incidence of breast cancer and the **dependence between mothers and daughters**. Hence, the dependence between sisters was not looked at.

Danish register-based family data on breast cancer
↳ more details

*family data on **breast cancer** among women - the most common malignancy in women.*

The cohort consisted of 1 292 051 families and 3 029 653 individuals:

- 908 002 (70.3%) families with a mother and a single daughter;
- 322 547 (25.0%) families with a mother and two daughters;
- 61 502 (4.8%) families with a mother and three daughters.

---

aim: investigate the cumulative incidence of breast cancer and the **dependence between mothers and daughters**. Hence, the dependence between sisters was not looked at.

failure causes: 1) breast cancer; 2) death; and 3) other cancers, which were grouped together.

Danish register-based family data on breast cancer
↳ general results

- AGQ approximation with five quadrature points and Gauss-Hermite rules;
- *found*: within-cluster dependence of breast cancer with regard to both risk and timing;
- The model provides a framework for exploring and making inference about the distribution of age at disease onset and to investigate how absolute risk of disease is related to age at onset.

```
@article{PAPER_PRESENTED,
   title = "Modeling the cumulative incidence function of
            multivariate competing risks data allowing for
            within-cluster dependence of risk and timing",
   author = "Luise Cederkvist and Holst, {Klaus K.} and
             Andersen, {Klaus K.} and Scheike, {Thomas H.}",
   year = "2019",
   volume = "20",
   pages = "199--217",
   journal = "Biostatistics",
   publisher = "Oxford University Press",
   number = "2",
}
```