

# Modeling the cumulative incidence function of clustered competing risks data: a multinomial GLMM approach

master thesis defense



Henrique Laureano ([.github.io](https://github.io))  
LEG @ UFPR

April 19, 2021

- 1 Data
- 2 Model
- 3 TMB: Template Model Builder
- 4 Simulation study
- 5 Conclusion
- 6 References

# Clustered competing risk data



Key terms:

- 1 **Clustered**: groups with a dependence structure (e.g. families);
- 2 Causes **competing** by *something*.

Something?

- **Failure** of an industrial or electronic component;
- **Occurrence** or **cure** of a disease or some biological process;
- **Progress** of a patient clinic state.

Independent of the application, always the same framework

Cluster	ID	Cause 1	Cause 2	Censorship	Time	Feature
1	1	Yes	No	No	10	A
1	2	No	No	Yes	8	A
2	1	No	No	Yes	7	B
2	2	No	Yes	No	5	A

# Big picture: Failure time data

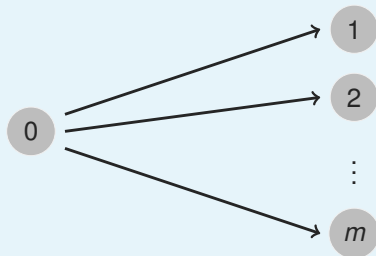


Failure time process



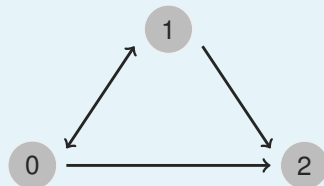
*Same methodologies,  
different names.*

Competing risks process



**Survival analysis** Biomedical studies;  
**Reliability analysis** Industrial life testing.

Multistate process



A comprehensive reference is Kalbfleisch and Prentice (2002)'s book.

- 1 Data
- 2 Model
- 3 TMB: Template Model Builder
- 4 Simulation study
- 5 Conclusion
- 6 References

# Modeling clustered competing risks data



What?



Why?



How?

# Failure time data → Survival models



First of all, we have to choose which **scale** we model the **survival experience**.

① Usually, is in the

$$\text{hazard (failure rate) scale : } \lambda(t | \text{features}) = \lambda_0(t) \times c(\text{features}). \quad (1)$$

We have a Equation 1 for each competing cause.

The cluster dependence is something actually not measured...

Not measured dependence → **random/latent effects** → Frailty models.

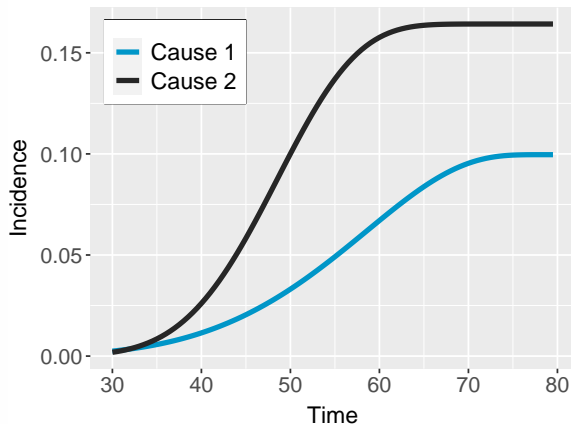
Full likelihood analysis with frailty models for competing risks data is generally complicated, when not impracticable.

② *Not* usually, the **probability scale**.

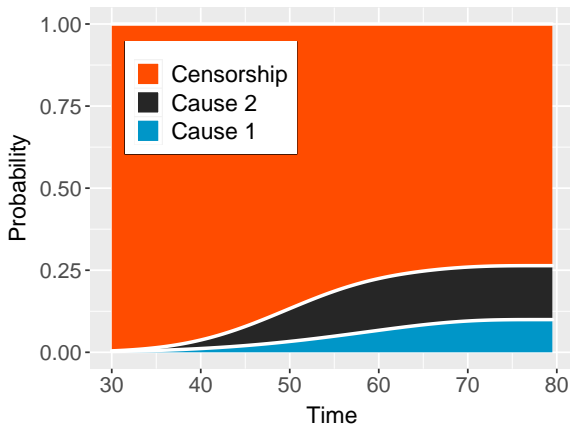
# Probability scale → Cause-specific CIF



## Cumulative Incidence Function (CIF)



## All CIFs sum up to 1



i.e.,  $\text{CIF} = \mathbb{P}[\text{failure time} \leq t, \text{ a given cause} \mid \text{features \& latent effects}]$ .



for a cause-specific of failure  $k$ ,  
the cumulative incidence function (CIF) is defined as

$$\begin{aligned}
 F_k(t | \mathbf{x}) &= \mathbb{P}[T \leq t, K = k | \mathbf{x}] \\
 &= \int_0^t f_k(z | \mathbf{x}) \, dz \quad (f_k(t | \mathbf{x}) \text{ is the (sub)density for the time to a type } k \text{ failure}) \\
 &= \int_0^t \underbrace{\lambda_k(z | \mathbf{x})}_{\text{cause-specific hazard function}} \underbrace{S(z | \mathbf{x})}_{\text{overall survival function}} \, dz, \quad t > 0, \quad k = 1, \dots, K.
 \end{aligned}$$



Again, a comprehensive reference is Kalbfleisch and Prentice (2002)'s book.



Here, we use the same CIF specification of Cederkvist et al. (2019).

# Cederkvist et al. (2019)'s CIF specification



For two competing causes of failure,  
the cause-specific CIFs are specified in the following manner

$$F_k(t \mid \mathbf{x}, u_1, u_2, \eta_k) = \underbrace{\pi_k(\mathbf{x}, u_1, u_2)}_{\text{cluster-specific risk level}} \times \underbrace{\Phi[w_k g(t) - \mathbf{x}\gamma_k - \eta_k]}_{\text{cluster-specific failure time trajectory}}, \quad t > 0, \quad k = 1, 2, \quad (2)$$

with

- ❶  $\pi_k(\mathbf{x}, \mathbf{u}) = \exp\{\mathbf{x}\beta_k + u_k\} / \left(1 + \sum_{m=1}^{K-1} \exp\{\mathbf{x}\beta_m + u_m\}\right), \quad k = 1, 2, \quad K = 3;$
- ❷  $\Phi(\cdot)$  is the cumulative distribution function of a standard Gaussian distribution;
- ❸  $g(t) = \text{arctanh}(2t/\delta - 1), \quad t \in (0, \delta), \quad g(t) \in (-\infty, \infty).$



In Cederkvist et al. (2019), this CIF specification is modeled under a *complicated* pairwise composite likelihood approach (Lindsay 1988; Varin, Reid, and Firth 2011).

# Our contribution: a full likelihood analysis



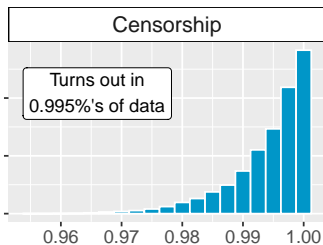
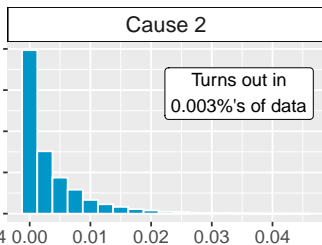
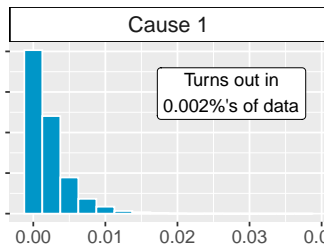
For two competing causes of failure, a subject  $i$ , in the cluster  $j$ , in time  $t$ , we have

$$y_{ijt} \mid \underbrace{\{u_{1j}, u_{2j}, \eta_{1j}, \eta_{2j}\}}_{\text{latent effects}} \sim \text{Multinomial}(p_{1ijt}, p_{2ijt}, p_{3ijt})$$

$$\begin{bmatrix} u_1 \\ u_2 \\ \eta_1 \\ \eta_2 \end{bmatrix} \sim \text{Multivariate Normal} \left( \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_1}^2 & \text{cov}(u_1, u_2) & \text{cov}(u_1, \eta_1) & \text{cov}(u_1, \eta_2) \\ & \sigma_{u_2}^2 & \text{cov}(u_2, \eta_1) & \text{cov}(u_2, \eta_2) \\ & & \sigma_{\eta_1}^2 & \text{cov}(\eta_1, \eta_2) \\ & & & \sigma_{\eta_2}^2 \end{bmatrix} \right)$$

$$\begin{aligned} p_{kijt} &= \frac{\partial}{\partial t} F_k(t \mid \mathbf{x}, \mathbf{u}, \eta_k) \\ &= \frac{\exp\{\mathbf{x}_{kij}\beta_k + u_{kj}\}}{1 + \sum_{m=1}^{K-1} \exp\{\mathbf{x}_{mij}\beta_m + u_{mj}\}} \\ &\quad \times w_k \frac{\delta}{2\delta t - 2t^2} \phi \left( w_k \text{arctanh} \left( \frac{t - \delta/2}{\delta/2} \right) - \mathbf{x}_{kij}\gamma_k - \eta_{kj} \right), \quad k = 1, 2. \end{aligned} \quad (3)$$

# Simulating from the model



bandwidth=0.0025

# Marginal likelihood function for two competing causes



$$\begin{aligned}
 L(\theta; \mathbf{y}) &= \prod_{j=1}^J \int_{\Re^4} \pi(\mathbf{y}_j | \mathbf{r}_j) \times \pi(\mathbf{r}_j) \, d\mathbf{r}_j \\
 &= \prod_{j=1}^J \int_{\Re^4} \underbrace{\left\{ \prod_{i=1}^{n_j} \prod_{t=1}^{n_{ij}} \left( \frac{(\sum_{k=1}^K y_{kijt})!}{y_{1ijt}! y_{2ijt}! y_{3ijt}!} \prod_{k=1}^K p_{kijt}^{y_{kijt}} \right) \right\}}_{\text{fixed effect component}} \times \\
 &\quad \underbrace{(2\pi)^{-2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{r}_j^\top \Sigma^{-1} \mathbf{r}_j \right\}}_{\text{latent effect component}} d\mathbf{r}_j \\
 &= \prod_{j=1}^J \int_{\Re^4} \underbrace{\left\{ \prod_{i=1}^{n_j} \prod_{t=1}^{n_{ij}} \prod_{k=1}^K p_{kijt}^{y_{kijt}} \right\}}_{\text{fixed effect}} \underbrace{(2\pi)^{-2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{r}_j^\top \Sigma^{-1} \mathbf{r}_j \right\}}_{\text{latent effect component}} d\mathbf{r}_j, \quad (4)
 \end{aligned}$$

with  $p_{kijt}$  from Equation 3 and where  $\theta = [\beta \ \gamma \ \mathbf{w} \ \sigma^2 \ \rho]^\top$  is the parameters vector.

- 1 Data
- 2 Model
- 3 TMB: Template Model Builder**
- 4 Simulation study
- 5 Conclusion
- 6 References



Kristensen et al. (2016).

*An R (R Core Team 2021) package for the quickly implementation of complex random effect models through simple C++ templates.*

## Workflow

- 1 Write your objective function in a .cpp through a `#include <TMB.hpp>`;
- 2 Compile and load it in R via `TMB::compile()` and `base::dyn.load(TMB::dynlib())`;
- 3 Compute your objective function derivatives with `obj <- TMB::MakeADFun()`;
- 4 Perform the model fitting, `opt <- base::nlminb(obj$par, obj$fn, obj$gr)`;
- 5 Compute the parameters standard deviations, `TMB::sdreport(obj)`.



For details about TMB, AD, and Laplace approximation: Laureano (2021).

- 1 Data
- 2 Model
- 3 TMB: Template Model Builder
- 4 Simulation study**
- 5 Conclusion
- 6 References



## Risk model

Latent effects only on the risk level  
i.e.,

$$\Sigma = \begin{bmatrix} \sigma_{u_1}^2 & \text{COV}_{u_1, u_2} \\ & \sigma_{u_2}^2 \end{bmatrix}.$$

## Time model

Latent effects only on the failure  
time trajectory level i.e.,

$$\Sigma = \begin{bmatrix} \sigma_{\eta_1}^2 & \text{COV}_{\eta_1, \eta_2} \\ & \sigma_{\eta_2}^2 \end{bmatrix}.$$

## Block-diag model

Latent effects on the risk and time levels  
without cross-correlations i.e.,

$$\Sigma = \begin{bmatrix} \sigma_{u_1}^2 & \text{COV}_{u_1, u_2} & 0 & 0 \\ & \sigma_{u_2}^2 & 0 & 0 \\ & & \sigma_{\eta_1}^2 & \text{COV}_{\eta_1, \eta_2} \\ & & & \sigma_{\eta_2}^2 \end{bmatrix}.$$

## Complete model

A *complete* latent effects structure  
i.e.,

$$\Sigma = \begin{bmatrix} \sigma_{u_1}^2 & \text{COV}_{u_1, u_2} & \text{COV}_{u_1, \eta_1} & \text{COV}_{u_1, \eta_2} \\ & \sigma_{u_2}^2 & \text{COV}_{u_2, \eta_1} & \text{COV}_{u_2, \eta_2} \\ & & \sigma_{\eta_1}^2 & \text{COV}_{\eta_1, \eta_2} \\ & & & \sigma_{\eta_2}^2 \end{bmatrix}.$$

# Simulation study setup



**Four** latent effects structures:

- 1** Risk model;
- 2** Time model;
- 3** Block-diag model;
- 4** Complete model.

**Two** CIF configurations:

**Low** max incidence  $\approx 0.15$ ;

**High** max incidence  $\approx 0.60$ .

For each of those  $4 \times 2 = 8$  scenarios, we vary the sample and cluster sizes:

## *5000 data points*

- 2500 clusters of **size 2**;
- 1000 clusters of **size 5**;
- 500 clusters of **size 10**.

## *30000 data points*

- 15000 clusters of **size 2**;
- 6000 clusters of **size 5**;
- 3000 clusters of **size 10**.

## *60000 data points*

- 30000 clusters of **size 2**;
- 12000 clusters of **size 5**;
- 6000 clusters of **size 10**.

Totalizing,  $8 \times 3 \times 3 = 72$  scenarios.

For each scenario, we simulate **400** samples, totalizing  $72 \times 400 = 28800$  model fittings.

First of all, the **time**.

- The *non-complete* models (2D Laplace aprox.) are kind of fast, taking always **less than 5 min**.
- In the most expensive scenarios (30K 4D Laplaces), **the complete model takes 30 min**.  
In a **full R** implementation with 10K 4D Laplaces, it **took 30hrs**. **TMB is fast**.
- We also did a Bayesian analysis via Stan/NUTS-HMC (Stan Development Team [2020](#)).
  - **1 week of parallelized processing** for a 2500 size 2 clusters scenario with tuned NUTS.  
This just reinforces the MCMC impracticability for some complex models.

Model **identifiability**.

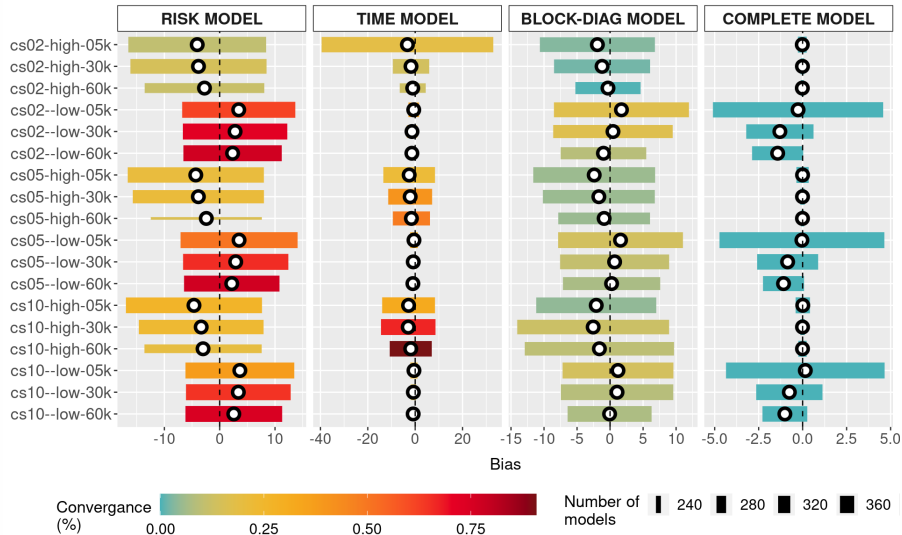
- The *non-complete* models fail to learn the data.  
They appear to be *not structured enough* to capture the data characteristics.

# Some simulation study results



Parameter:  $\beta_1$

with  $\pm 1.96$  standard deviations

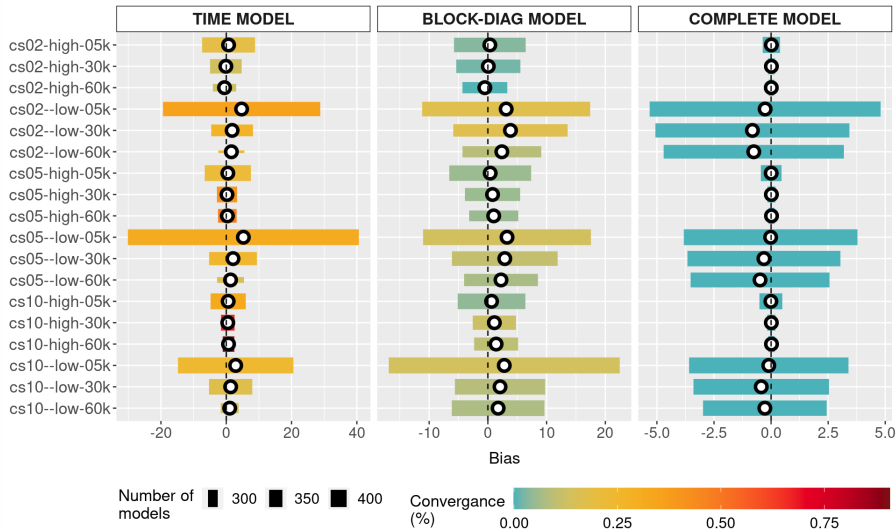


# Some simulation study results



Parameter:  $\log(\sigma_4^2)$

with  $\pm 1.96$  standard deviations

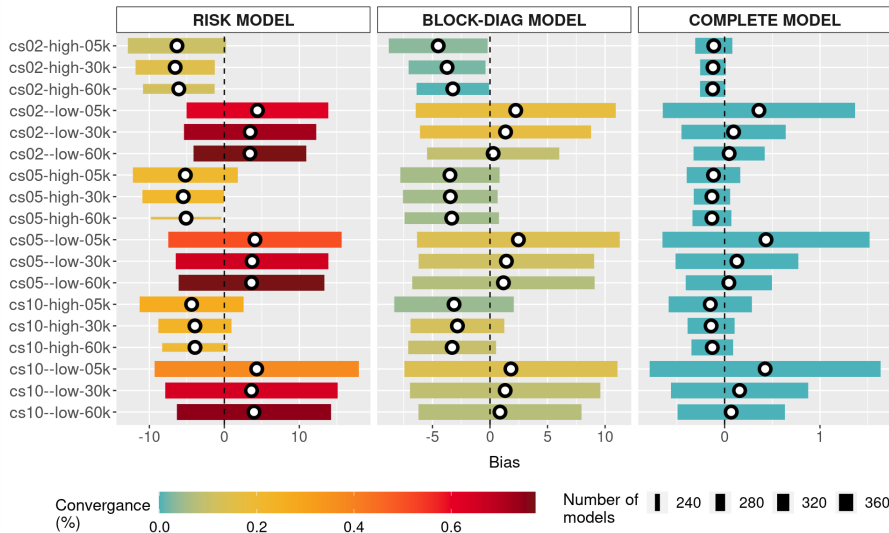


# Some simulation study results



Parameter:  $z(\rho_{12})$

with  $\pm 1.96$  standard deviations

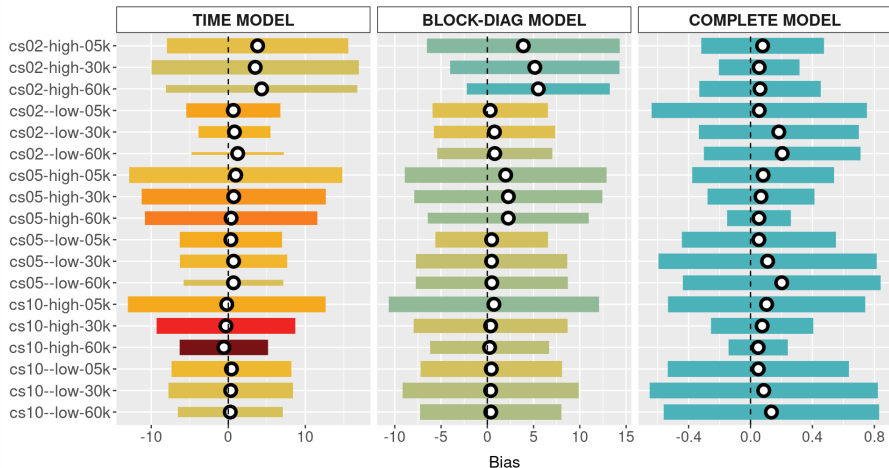


# Some simulation study results



Parameter:  $z(\rho_{34})$

with  $\pm 1.96$  standard deviations



Number of models: 300 350 400

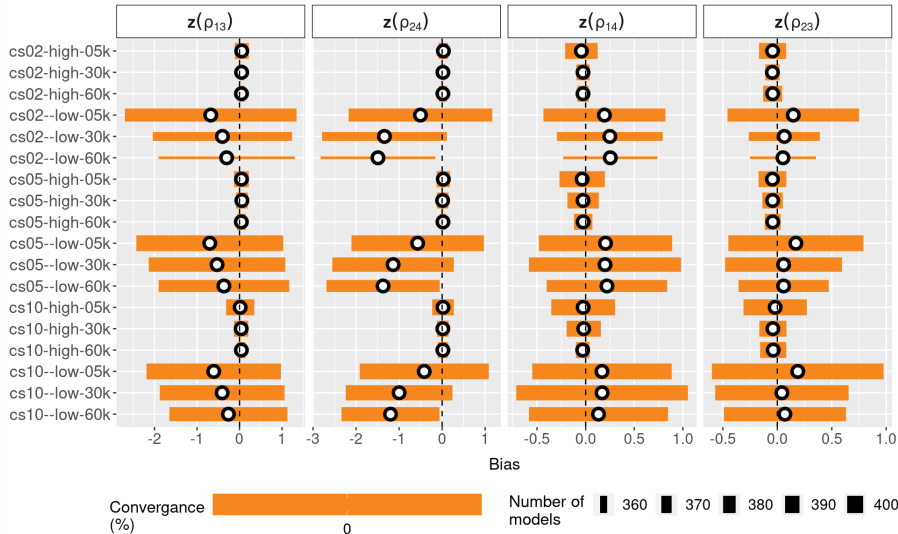
Convergence (%) 0.00 0.25 0.50 0.75

# Some simulation study results



## Complete model's cross-correlations

with  $\pm 1.96$  standard deviations



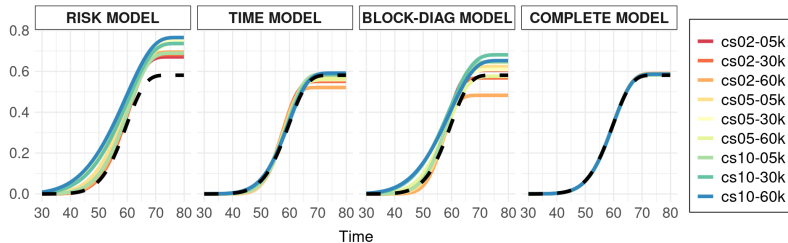


# Simulation study results: High CIF scenario



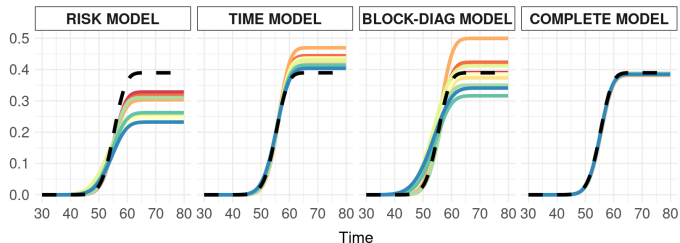
## CIF of failure cause 1

True curve in dashed black



## CIF of failure cause 2

True curve in dashed black

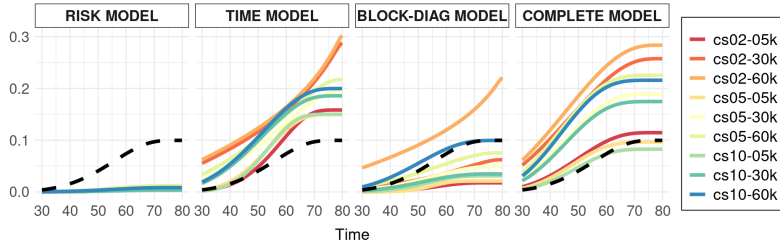


# Simulation study results: Low CIF scenario



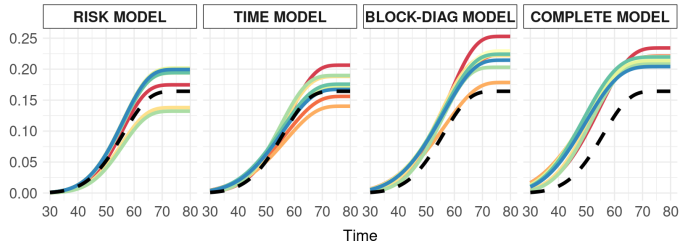
## CIF of failure cause 1

True curve in dashed black



## CIF of failure cause 2

True curve in dashed black



- 1 Data
- 2 Model
- 3 TMB: Template Model Builder
- 4 Simulation study
- 5 Conclusion**
- 6 References

**The complete model works.** It's not magnificent, but it works.

- 1 It works better in the high CIF scenarios;
- 2 As expected, as the sample size increases the results get better;
- 3 Given the low data representativity, the model needs a considerable amount of data to perform well;
- 4 In standard multinomial GLMMs, as bigger the clusters better the results. In our CIF-based formulation, this characteristic is not so clear.

What else can we do?

- 1 Instead of a conditional approach (latent effects model), we can try a marginal approach e.g., an McGLM (Bonat and Jørgensen 2016);
- 2 We can also try a copula (Embrechts 2009), on maybe two fronts:  
1) for a full specification; 2) to accommodate the within-cluster dependence.



For more read Laureano (2021) master thesis.

# Thanks for watching and have a great day



Special thanks to



**PPGMNE**

Programa de Pós-Graduação em  
Métodos Numéricos em Engenharia



Joint work with

Wagner H. Bonat

<http://leg.ufpr.br/~wagner>

Paulo Justiniano Ribeiro Jr.

<http://leg.ufpr.br/~paulojus>



[henriquelaureano.github.io](https://henriquelaureano.github.io)

- 1 Data
- 2 Model
- 3 TMB: Template Model Builder
- 4 Simulation study
- 5 Conclusion
- 6 References**

- Bonat, W. H., and B. Jørgensen. 2016. "Multivariate Covariance Generalized Linear Models." *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 65 (5): 649–75.
- Cederkvist, L., K. K. Holst, K. K. Andersen, and T. H. Scheike. 2019. "Modeling the Cumulative Incidence Function of Multivariate Competing Risks Data Allowing for Within-Cluster Dependence of Risk and Timing." *Biostatistics* 20 (2): 199–217.
- Embrechts, P. 2009. "Copulas: A Personal View." *The Journal of Risk and Insurance* 76 (3): 639–50.
- Kalbfleisch, J. D., and R. L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. Second Edition. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Kristensen, K., A. Nielsen, C. W. Berg, H. J. Skaug, and B. M. Bell. 2016. "TMB: Automatic Differentiation and Laplace Approximation." *Journal of Statistical Software* 70 (5): 1–21.
- Laureano, H. A. 2021. "Modeling the Cumulative Incidence Function of Clustered Competing Risks Data: A Multinomial Glmm Approach." Master's thesis, Federal University of Paraná (UFPR).
- Lindsay, B. G. 1988. "Composite Likelihood Methods." *Contemporary Mathematics* 80 (1): 221–39.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Stan Development Team. 2020. "RStan: The R Interface to Stan." <https://mc-stan.org/>.
- Varin, C., N. Reid, and D. Firth. 2011. "An Overview of Composite Likelihood Methods." *Statistica Sinica* 21 (1): 5–42.