Dealing with Missing Values in Multivariate Joint Models for Longitudinal and Survival Data

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

N_Erler   www.nerler.com   NErler

ISCB 2020

Erasmus MC
University Medical Center Rotterdam

Chronic Hepatitis C

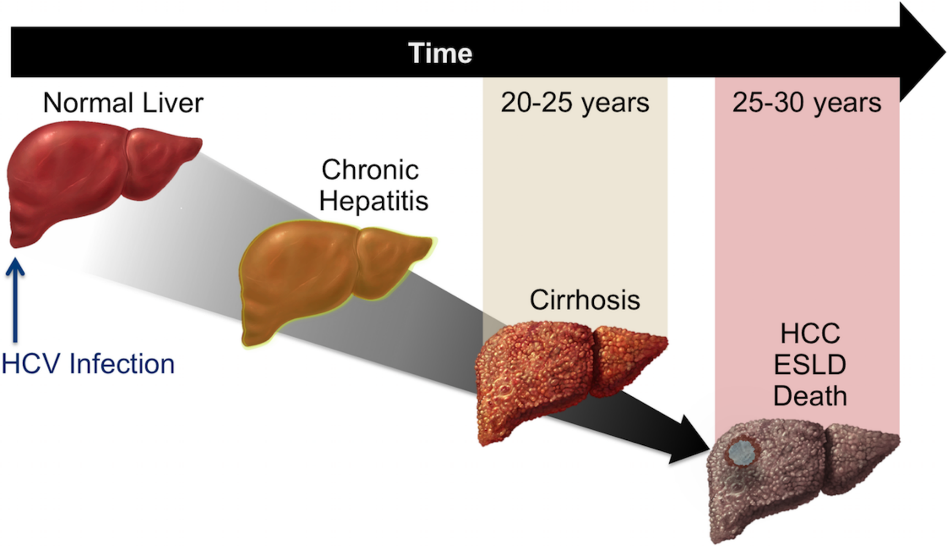
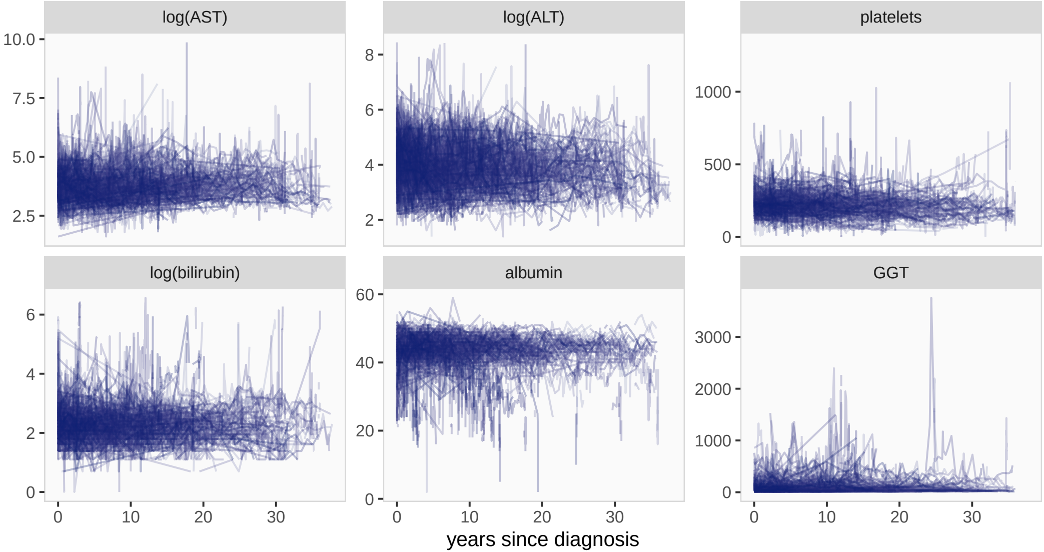
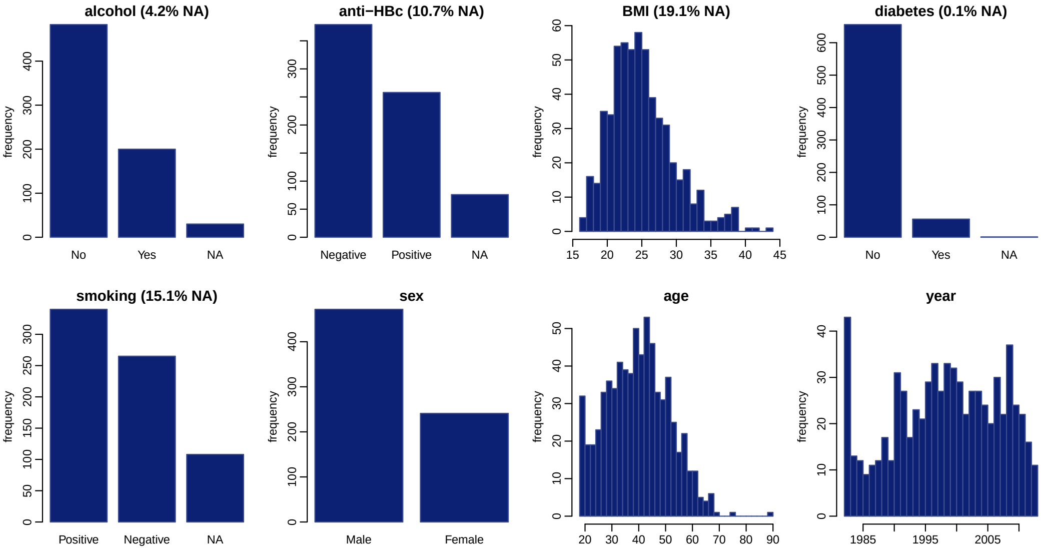


Image: https://www.hepatitisc.uw.edu/go/evaluation-staging-monitoring/natural-history/core-concept/all

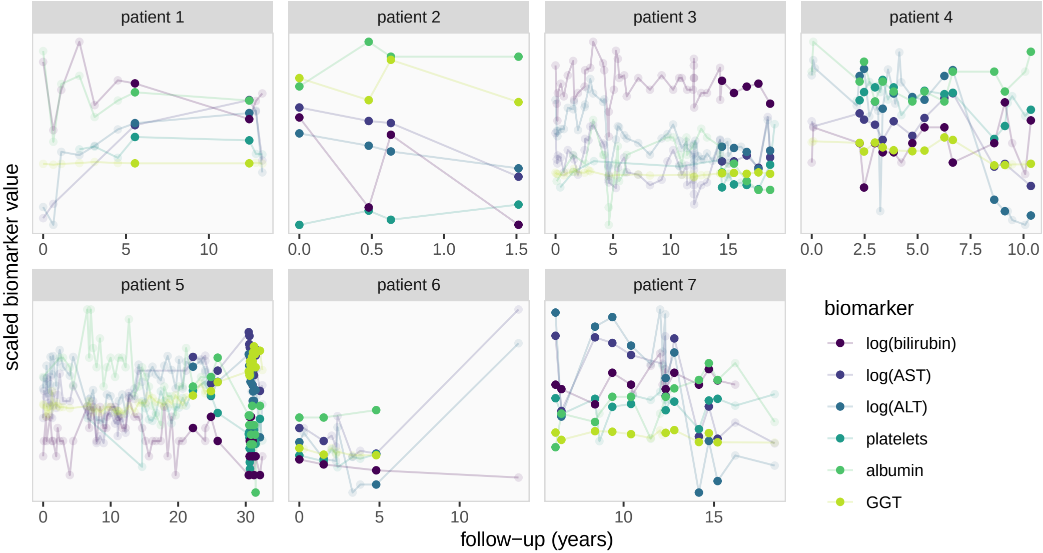
Longitudinal Covariates



Baseline Covariates



Missing Values in Longitudinal Covariates



Multivariate Joint Model

Proportional hazards model for time until event:

$$h_i(t) = h_0(t) \exp\left(\underbrace{\mathbf{x}_i^\top \beta^{(tc)}}_{\text{time constant}} + \sum_{k=1}^{K} \underbrace{\eta_{ki}(t)^\top \beta_k^{(tv)}}_{\text{time varying}} \right)$$

Multivariate Joint Model

Proportional hazards model for time until event:

$$h_i(t) = h_0(t) \exp\left(\underbrace{\mathbf{x}_i^\top \beta^{(tc)}}_{\text{time constant}} + \boxed{\sum_{k=1}^{K} \underbrace{\eta_{ki}(t)^\top \beta_k^{(tv)}}_{\text{time varying}}}\right)$$

Longitudinal (mixed) model for each biomarker $k = 1, \ldots K$:

$$\begin{aligned} \mathbb{E}(y_{ki}(t) \mid \mathbf{b}_{ki}) &= \eta_{ki}(t) \\ &= \underbrace{\mathbf{x}_{ki}(t)^\top \beta^{(k)}}_{\text{fixed effects}} + \underbrace{\mathbf{z}_{ki}(t)^\top \mathbf{b}_{ki}}_{\text{random effects}} \end{aligned}$$

Multivariate Joint Model

Proportional hazards model for time until event:

$$h_i(t) = h_0(t) \exp\left(\underbrace{\mathbf{x}_i^\top \beta^{(tc)}}_{\text{time constant}} + \sum_{k=1}^{K} \underbrace{\eta_{ki}(t)^\top \beta_k^{(tv)}}_{\text{time varying}}\right)$$

Longitudinal (mixed) model for each biomarker $k = 1, \ldots K$:

$$\begin{aligned}\mathbb{E}(y_{ki}(t) \mid \mathbf{b}_{ki}) &= \eta_{ki}(t)\\ &= \underbrace{\mathbf{x}_{ki}(t)^\top \beta^{(k)}}_{\text{fixed effects}} + \underbrace{\mathbf{z}_{ki}(t)^\top \mathbf{b}_{ki}}_{\text{random effects}}\end{aligned}$$

Missing values in (baseline) covariates.

Imputation of Missing Covariates

Imputation of a (baseline) variable x_i:

➡ sample from the **predictive distribution** of the missing values given the observed values

Imputation of Missing Covariates

Imputation of a (baseline) variable x_i:

➡ sample from the **predictive distribution** of the missing values given the observed values

$$p(x_i \mid \underbrace{\textbf{everything else}})$$

- ☺ other baseline variables
- ☹ repeatedly measured variables (incl. outcomes)
- ☹ survival outcome

Imputation of Missing Covariates

Imputation of a (baseline) variable x_i:

➡ sample from the **predictive distribution** of the missing values given the observed values

$$p(x_i \mid \underbrace{\textbf{everything else}})$$

- ☺ other baseline variables
- ☹ repeatedly measured variables (incl. outcomes)
- ☹ survival outcome

➡ **We cannot directly specify the (correct) imputation model!**

Imputation of Missing Covariates

Idea:

- specify the joint distribution $p(\textbf{everything})$
- derive $p(x_i \mid \textbf{everything else})$ from $p(\textbf{everything})$

Imputation of Missing Covariates

Idea:

- specify the joint distribution $p(\textbf{everything})$
- derive $p(x_i \mid \textbf{everything else})$ from $p(\textbf{everything})$

But:

$$\begin{aligned} p(\textbf{everything}) &= p(\text{survival outcome,}\\ &\qquad \text{longitudinal outcomes,}\\ &\qquad \text{longitudinal covariates,}\\ &\qquad \text{baseline covariates,}\\ &\qquad \text{random effects,}\\ &\qquad \text{parameters)}\\ &= p(\mathbf{T}, \mathbf{D}, \mathbf{y}, \mathbf{X}, \mathbf{b}, \boldsymbol{\theta}) \end{aligned}$$

Does this really solve anything?

Imputation of Missing Covariates

Idea:

- specify the joint distribution $p(\textbf{everything})$
- derive $p(x_i \mid \textbf{everything else})$ from $p(\textbf{everything})$

But:

$$\begin{aligned}
p(\textbf{everything}) = p(&\text{survival outcome},\\
&\text{longitudinal outcomes},\\
&\text{longitudinal covariates},\\
&\text{baseline covariates},\\
&\text{random effects},\\
&\text{parameters})\\
= p(&\mathbf{T}, \mathbf{D}, \mathbf{y}, \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})
\end{aligned}$$

Does this really solve anything? 👍**Yes, it does!**

Fully Bayesian Analysis & Imputation

From probability theory:

$$p(\mathbf{A}, \mathbf{B}) = p(\mathbf{A} \mid \mathbf{B})\, p(\mathbf{B})$$

Fully Bayesian Analysis & Imputation

From probability theory:

$$p(\mathbf{A}, \mathbf{B}) = p(\mathbf{A} \mid \mathbf{B})\ p(\mathbf{B})$$

Joint distribution

$$p(\mathbf{T}, \mathbf{D}, \mathbf{y}, \mathbf{X}, \mathbf{b}, \theta) = p(\mathbf{T}, \mathbf{D} \mid \mathbf{X}, \mathbf{b}, \theta) \quad p(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \theta) \quad p(\mathbf{X} \mid \theta) \quad p(\mathbf{b} \mid \theta) \quad p(\theta)$$

Fully Bayesian Analysis & Imputation

From probability theory:

$$p(\mathbf{A}, \mathbf{B}) = p(\mathbf{A} \mid \mathbf{B}) \; p(\mathbf{B})$$

Joint distribution

$$p(\mathbf{T}, \mathbf{D}, \mathbf{y}, \mathbf{X}, \mathbf{b}, \boldsymbol{\theta}) = \underbrace{p(\mathbf{T}, \mathbf{D} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})}_{\substack{\text{survival} \\ \text{model}}} \quad p(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta}) \quad p(\mathbf{X} \mid \boldsymbol{\theta}) \quad p(\mathbf{b} \mid \boldsymbol{\theta}) \quad p(\boldsymbol{\theta})$$

Fully Bayesian Analysis & Imputation

From probability theory:

$$p(\mathbf{A}, \mathbf{B}) = p(\mathbf{A} \mid \mathbf{B})\; p(\mathbf{B})$$

Joint distribution

$$p(\mathbf{T}, \mathbf{D}, \mathbf{y}, \mathbf{X}, \mathbf{b}, \boldsymbol{\theta}) = \underbrace{p(\mathbf{T}, \mathbf{D} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})}_{\text{survival model}} \quad \underbrace{p(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})}_{\text{multivariate longitudinal model}} \quad p(\mathbf{X} \mid \boldsymbol{\theta}) \quad p(\mathbf{b} \mid \boldsymbol{\theta}) \quad p(\boldsymbol{\theta})$$

Fully Bayesian Analysis & Imputation

From probability theory:

$$p(\mathbf{A}, \mathbf{B}) = p(\mathbf{A} \mid \mathbf{B})\; p(\mathbf{B})$$

Joint distribution

$$p(\mathbf{T}, \mathbf{D}, \mathbf{y}, \mathbf{X}, \mathbf{b}, \boldsymbol{\theta}) = \underbrace{\underbrace{p(\mathbf{T}, \mathbf{D} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})}_{\text{survival model}} \quad \underbrace{p(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})}_{\text{multivariate longitudinal model}}}_{\text{analysis model}} \quad p(\mathbf{X} \mid \boldsymbol{\theta}) \quad \underbrace{p(\mathbf{b} \mid \boldsymbol{\theta})}_{\text{random effects}} \quad \underbrace{p(\boldsymbol{\theta})}_{\text{priors}}$$

Fully Bayesian Analysis & Imputation

From probability theory:

$$p(\mathbf{A}, \mathbf{B}) = p(\mathbf{A} \mid \mathbf{B})\; p(\mathbf{B})$$

Joint distribution

$$p(\mathbf{T}, \mathbf{D}, \mathbf{y}, \mathbf{X}, \mathbf{b}, \theta) = \underbrace{\underbrace{p(\mathbf{T}, \mathbf{D} \mid \mathbf{X}, \mathbf{b}, \theta)}_{\substack{\text{survival}\\ \text{model}}} \;\; \underbrace{p(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \theta)}_{\substack{\text{multivariate}\\ \text{longitudinal}\\ \text{model}}}}_{\text{analysis model}} \;\; \underbrace{p(\mathbf{X} \mid \theta)}_{\substack{\text{imputation}\\ \text{part}}} \;\; \underbrace{p(\mathbf{b} \mid \theta)}_{\substack{\text{random}\\ \text{effects}}} \;\; \underbrace{p(\theta)}_{\text{priors}}$$

Fully Bayesian Analysis & Imputation

Imputation part

$$
\begin{aligned}
p(\mathbf{X} \mid \theta) = p(\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{X}_{compl.} \mid \boldsymbol{\theta}) \quad = \quad & p(\mathbf{x}_1 \mid \mathbf{X}_{compl.}, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_p, \boldsymbol{\theta}) \\
& p(\mathbf{x}_2 \mid \mathbf{X}_{compl.}, \mathbf{x}_3, \ldots, \mathbf{x}_p, \boldsymbol{\theta}) \\
& \vdots \\
& p(\mathbf{x}_p \mid \mathbf{X}_{compl.}, \boldsymbol{\theta})
\end{aligned}
$$

Fully Bayesian Analysis & Imputation

Imputation part

$$
\begin{aligned}
p(\mathbf{X} \mid \theta) = p(\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{X}_{compl.} \mid \boldsymbol{\theta}) \quad = \quad & p(\mathbf{x}_1 \mid \mathbf{X}_{compl.}, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_p, \boldsymbol{\theta}) \\
& p(\mathbf{x}_2 \mid \mathbf{X}_{compl.}, \mathbf{x}_3, \ldots, \mathbf{x}_p, \boldsymbol{\theta}) \\
& \vdots \\
& p(\mathbf{x}_p \mid \mathbf{X}_{compl.}, \boldsymbol{\theta})
\end{aligned}
$$

Estimation:

via MCMC ➡ **Gibbs sampling** (using Metropolis-Hastings, …)

Fully Bayesian Analysis & Imputation

Imputation part

$$\begin{aligned} p(\mathbf{X} \mid \theta) = p(\mathbf{x}_1, \ldots, \mathbf{x}_p, \mathbf{X}_{compl.} \mid \boldsymbol{\theta}) \quad = \quad & p(\mathbf{x}_1 \mid \mathbf{X}_{compl.}, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_p, \boldsymbol{\theta}) \\ & p(\mathbf{x}_2 \mid \mathbf{X}_{compl.}, \mathbf{x}_3, \ldots, \mathbf{x}_p, \boldsymbol{\theta}) \\ & \vdots \\ & p(\mathbf{x}_p \mid \mathbf{X}_{compl.}, \boldsymbol{\theta}) \end{aligned}$$

Estimation:

via MCMC ➡ **Gibbs sampling** (using Metropolis-Hastings, ...)

Software:

Implemented in the **R** package **JointAI** (using JAGS)

In Practice: Analysis of the HCV Data

```r
library("JointAI")
library("splines")

fmla <- list(
  # formula for survival model
  Surv(etime, event) ~ age + sex + alc + smoke + BMI + DM + year +
    logBili + logALT + logAST + Plt,

  # formulas for the longitudinal outcomes
  logBili ~ age + sex + time + (time | id),
  logAST  ~ age + sex + ns(time, df = 5) + (ns(time, df = 5) | id),
  logALT  ~ age + sex + ns(time, df = 3) + (ns(time, df = 3) | id),
  Plt     ~ age + sex + ns(time, df = 3) + (ns(time, df = 3) | id)
)
```

In Practice: Analysis of the HCV Data

```r
library("JointAI")
library("splines")

fmla <- list(
  # formula for survival model
  Surv(etime, event) ~ age + sex + alc + smoke + BMI + DM + year +
    logBili + logALT + logAST + Plt,

  # formulas for the longitudinal outcomes
  logBili ~ age + sex + time + (time | id),
  logAST  ~ age + sex + ns(time, df = 5) + (ns(time, df = 5) | id),
  logALT  ~ age + sex + ns(time, df = 3) + (ns(time, df = 3) | id),
  Plt     ~ age + sex + ns(time, df = 3) + (ns(time, df = 3) | id)
)
```

```r
mod <- JM_imp(fmla,
              data = HCVdata,
              timevar = "time",
              n.iter = 2000)
```

In Practice: Analysis of the HCV Data

Additional options:

- covariate **model types**
- **hyper-parameters**
- number of **chains & thinning** interval
- ...

Additional features:

- use of **auxiliary** variables
- use of ridge **shrinkage** priors
- **multi-level** settings (e.g., multi-center)
- ...

For more info, see **https://nerler.github.io/JointAI**

Connecting Models

Longitudinal ➡ Survival

Longitudinal ➡ Longitudinal

Connecting Models

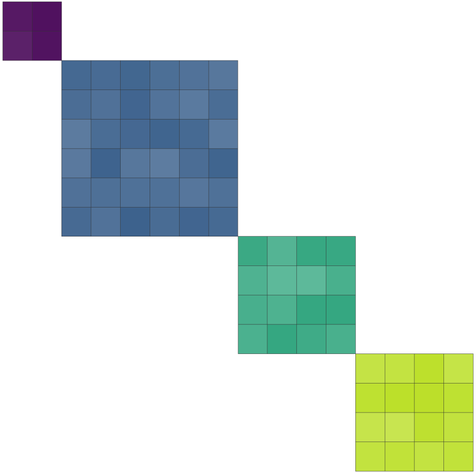
Longitudinal ➡ Survival

type of association

- underlying value $\eta_{ki}(t)$
- slope
- cumulative effect
- time-lag
- ...
- combination of the above

Longitudinal ➡ Longitudinal

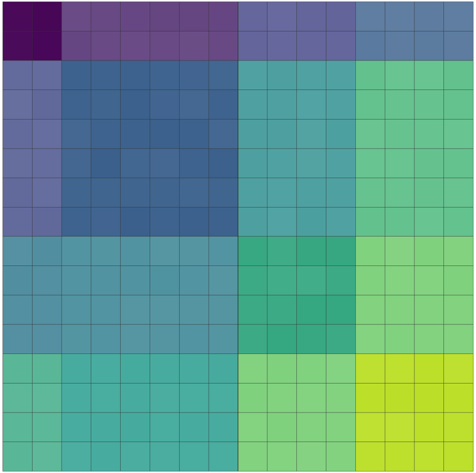
Connecting Models



Longitudinal ➡ Longitudinal

- ► **independent**

Connecting Models



Longitudinal ➡ Longitudinal

- **independent**
- **correlated** random effects
  - ☺ endogenous
  - ☹ dimensionality (136 elements!)
  - ☹ linear association

Connecting Models

Longitudinal ➡ Survival

type of association

- underlying value $\eta_{ki}(t)$
- slope
- cumulative effect
- time-lag
- ...
- combination of the above

Longitudinal ➡ Longitudinal

- **independent**
- **correlated** random effects
  - ☺ endogenous
  - ☹ dimensionality (136 elements!)
  - ☹ linear association
- **fixed effects**
  - ☺ potentially non-linear
  - ☹ exogenous

```
logBili ~ logAST + logALT + Plt + ...
logAST  ~ logALT + Plt + ...
logALT  ~ Plt + ...
Plt     ~ ...
```

(Interim) Conclusion: Does it work?

- **Theoretically: 👍**

  (if no assumptions are violated)

(Interim) Conclusion: Does it work?

- **Theoretically: 👍**

  (if no assumptions are violated)

- **Practically: 👍**

  R available in software

  ✓ computationally feasible:

  - 👥 713
  - ☰ ~20,000 rows
  - ⚙ complex model (4 dependent outcomes)
  - 🔄 10,000 iterations
  - 🕘 6.5h

(Interim) Conclusion: Does it work?

- **Theoretically:** 👍

  (if no assumptions are violated)

- **Practically:** 👍
  - R available in software
  - ✓ computationally feasible:
    - 👥 713
    - ☰ ∼20,000 rows
    - ⚙ complex model (4 dependent outcomes)
    - 🔄 10,000 iterations
    - 🕘 6.5h

- **Empirically:** to be continued... 📐🛠🔨

Thank you for your attention.

✉ n.erler@erasmusmc.nl
N_Erler
NErler
www.nerler.com

Erasmus MC
University Medical Center Rotterdam