Joint modeling of longitudinal and competing-risk data using cumulative incidence functions accounting for failure cause misclassification

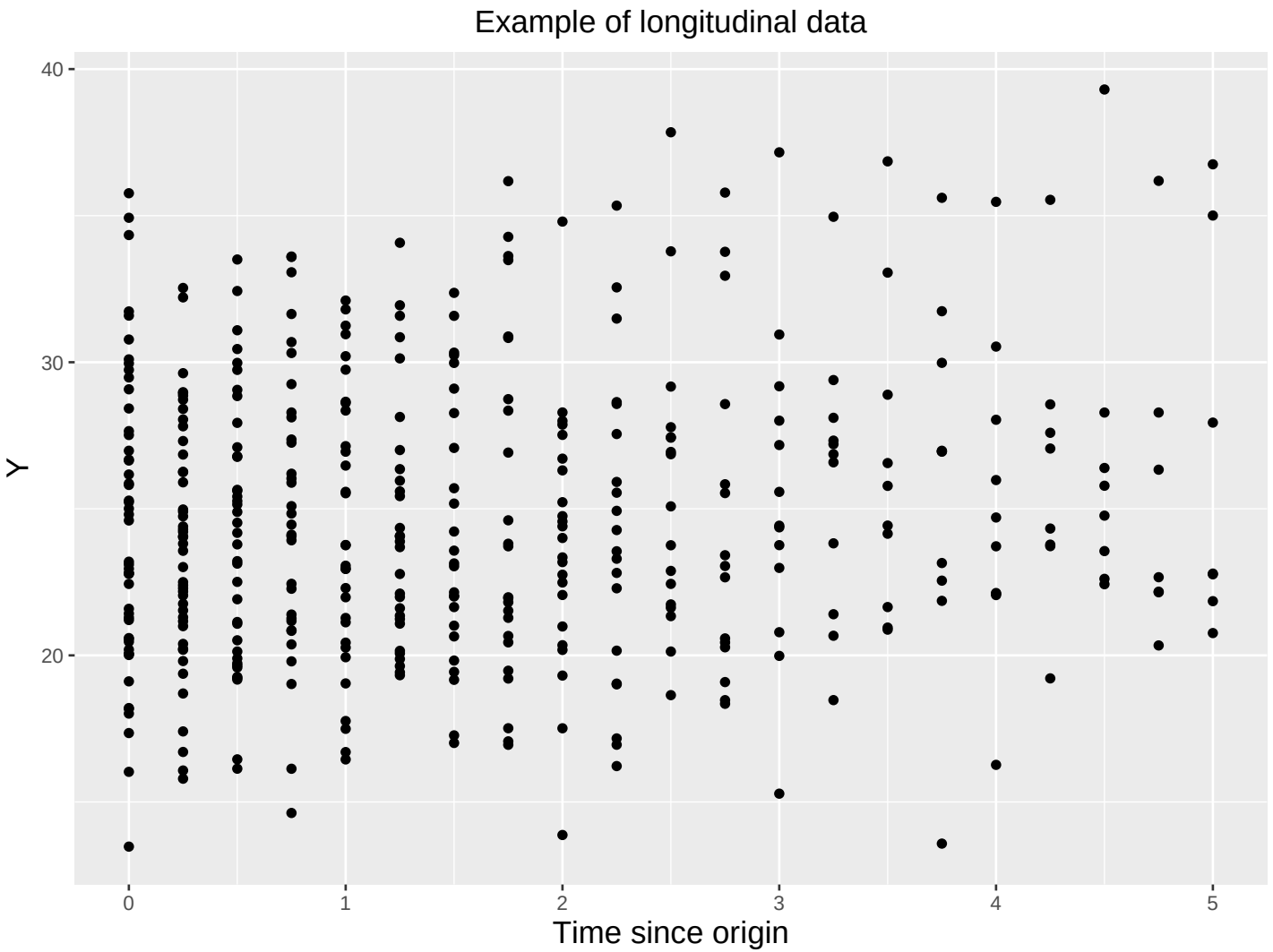
Christos Thomadakis[1], Loukia Meligkotsidou[2], Constantin T. Yiannoutsos[3], and Giota Touloumi[1]

[1]Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece
[2]Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece
[3]Department of Biostatistics and Health Data Science, Indiana University, 410 West 10th Street, Suite 3000, Indianapolis, IN 46202, USA

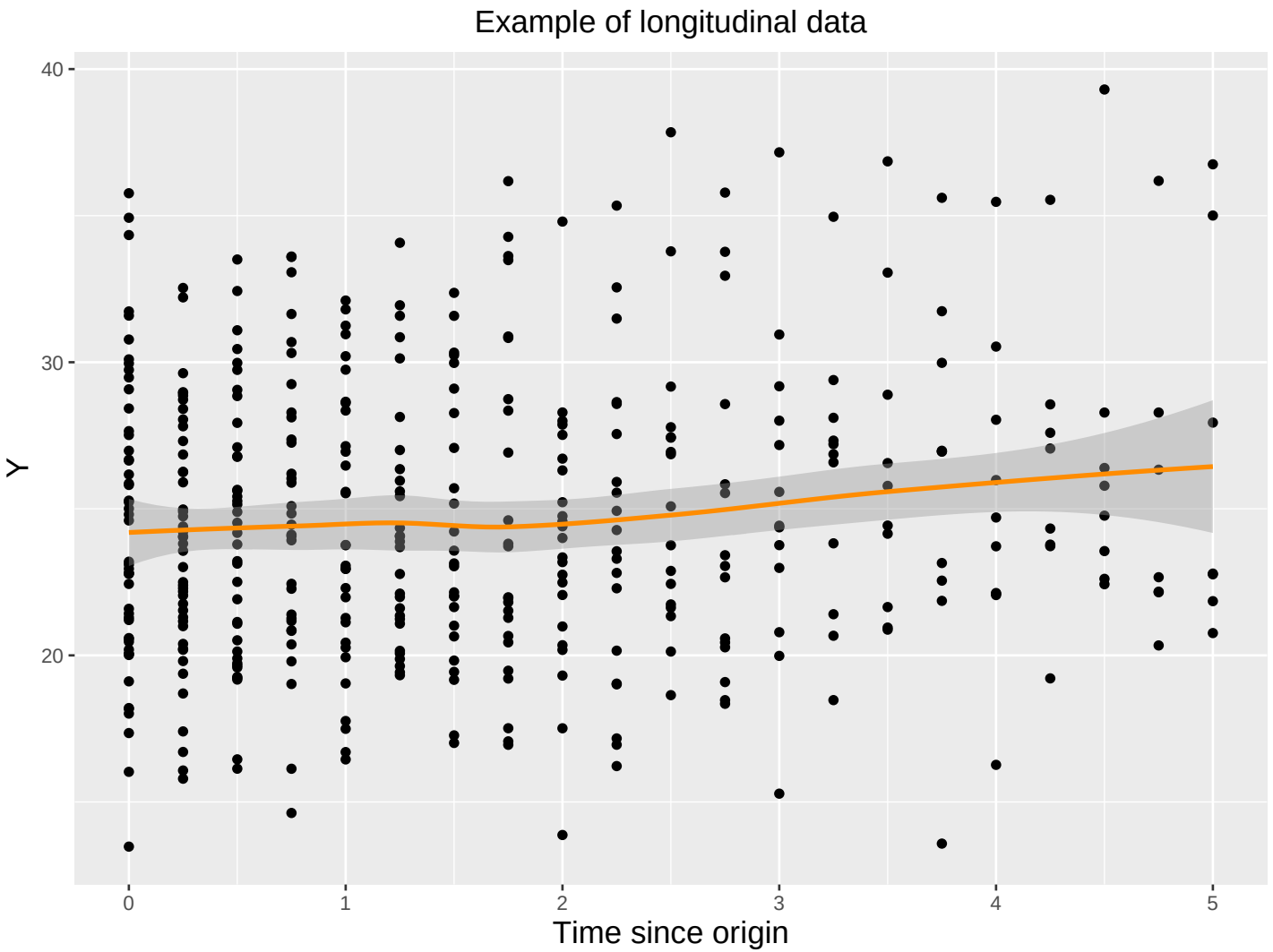
Longitudinal studies

- Longitudinal data consist of repeated measurements over time on the same individuals.
- In medical research, the values of biochemical markers related to some disease are typically recorded at each clinic visit to keep track of disease progression.
- This is in contrast to cross-sectional studies where, for each individual data, on a single time point are collected.
- There is often great variability across subjects, e.g. due to unmeasured characteristics such as genetic or environmental factors.
- In longitudinal studies, the effects of such factors are cancelled out, helping us identify causal relationships under certain assumptions.

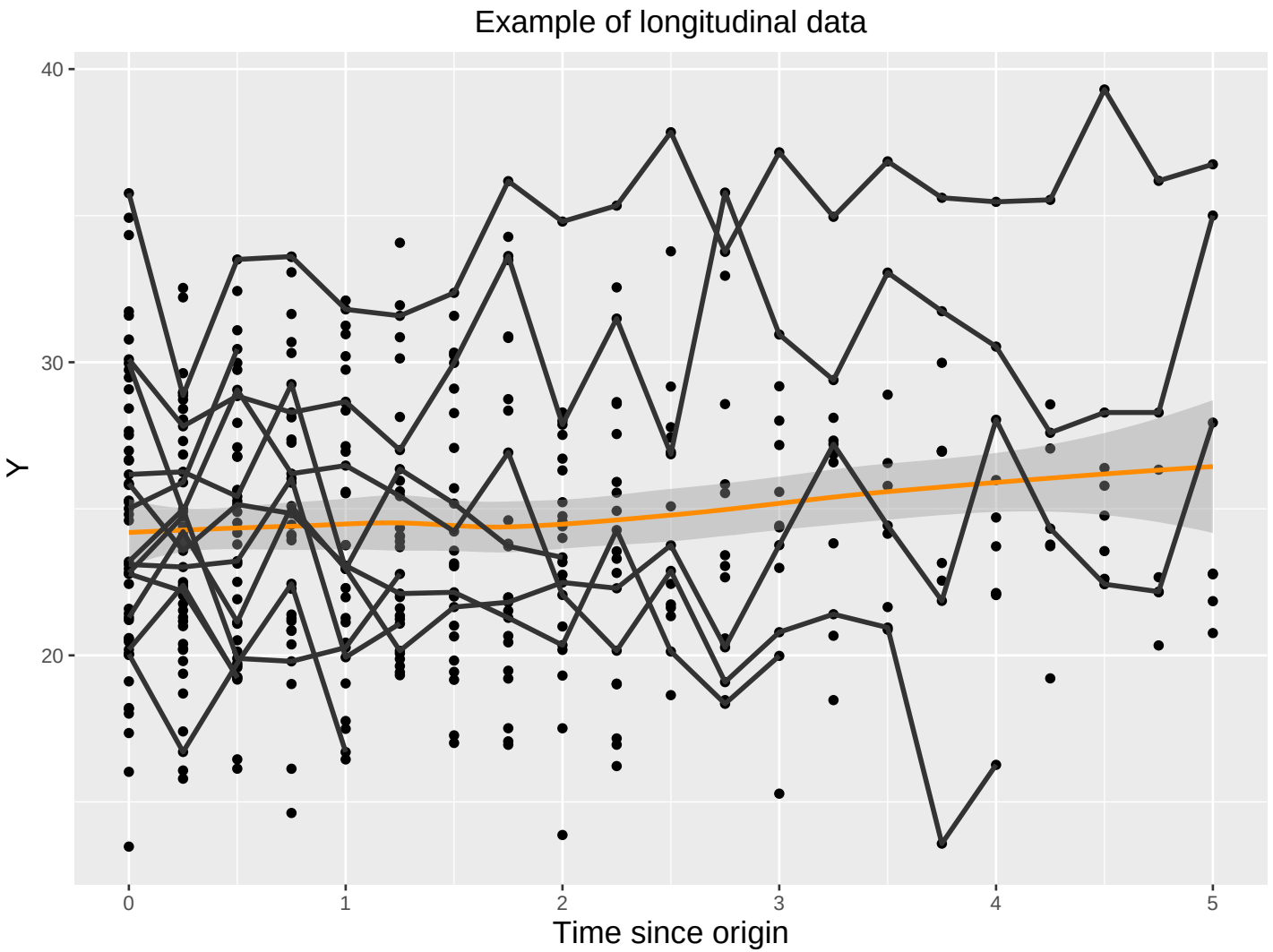
Example of longitudinal data



Example of longitudinal data



Example of longitudinal data



Linear mixed models (LMMs)

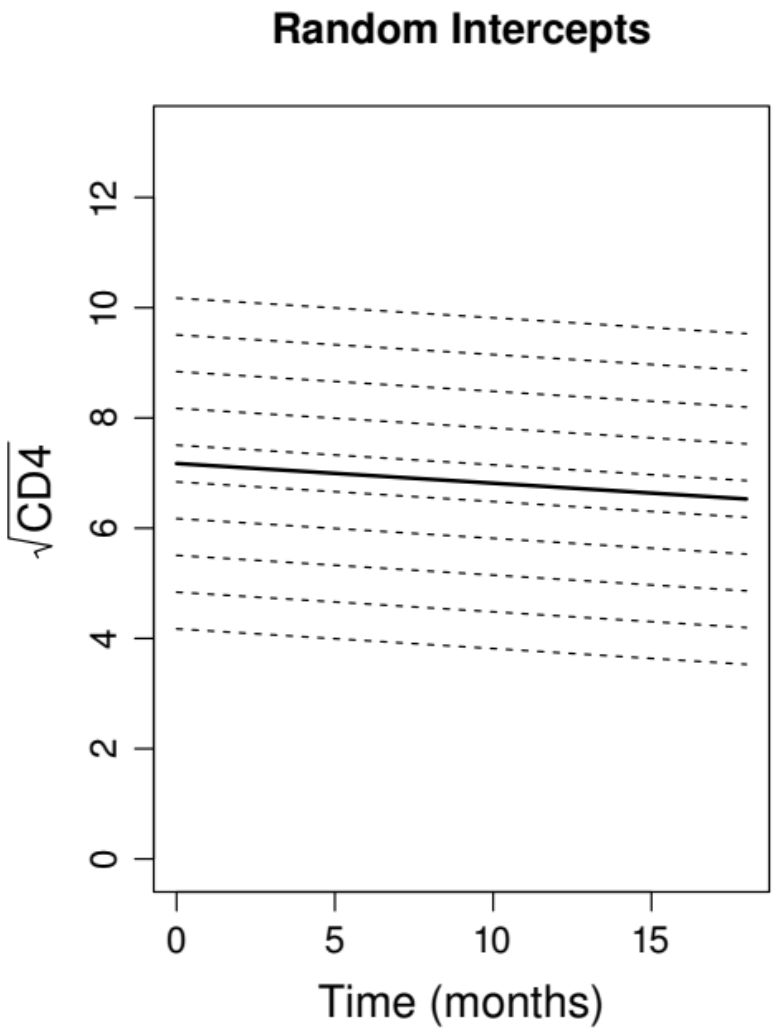
- Linear mixed models (LMMs) are frequently applied in longitudinal data analysis.
- The random intercept and slope model is particularly popular

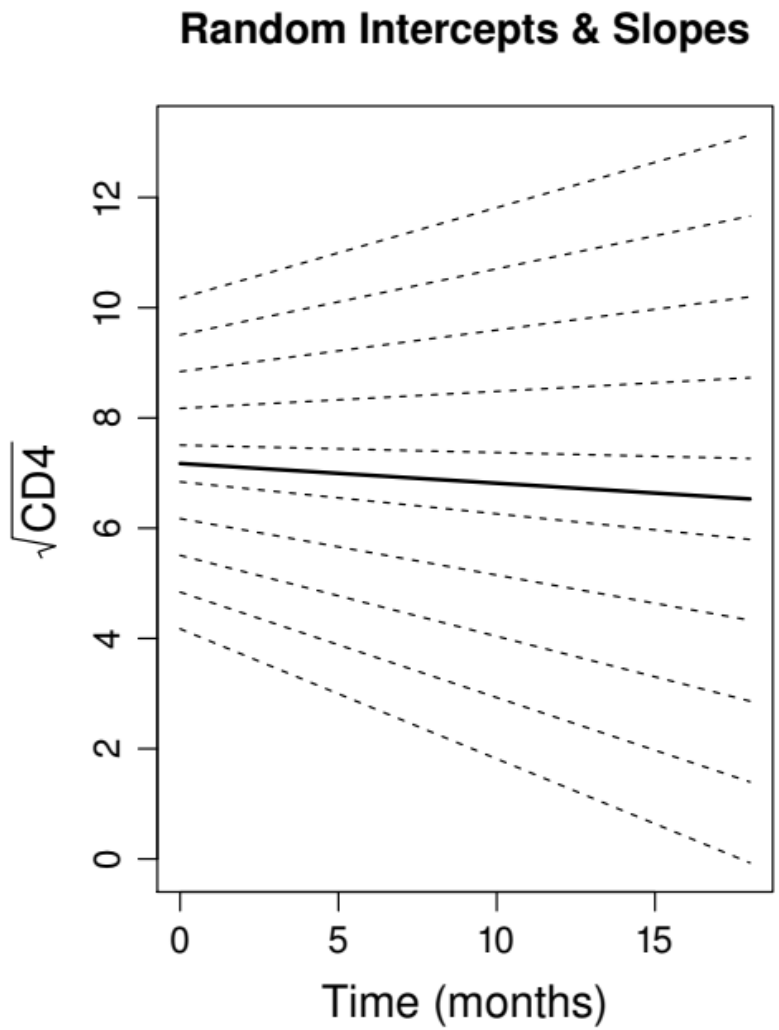
$$Y_j = (\beta_0 + b_0) + (\beta_1 + b_1)t_j + \epsilon_j, \quad j = 1, 2, \ldots, Q$$

- $(b_0, b_1)^\top \sim N(\mathbf{0}, \boldsymbol{D}) \rightarrow$ random intercept and slope
- $(\beta_0, \beta_1)^\top \rightarrow$ fixed effects
- $\epsilon_j \sim N(0, \omega^{-1}) \rightarrow$ the within-individual error

- The idea is that individuals have their own intercepts and rates of change over time (slopes).

Random intercept and slope





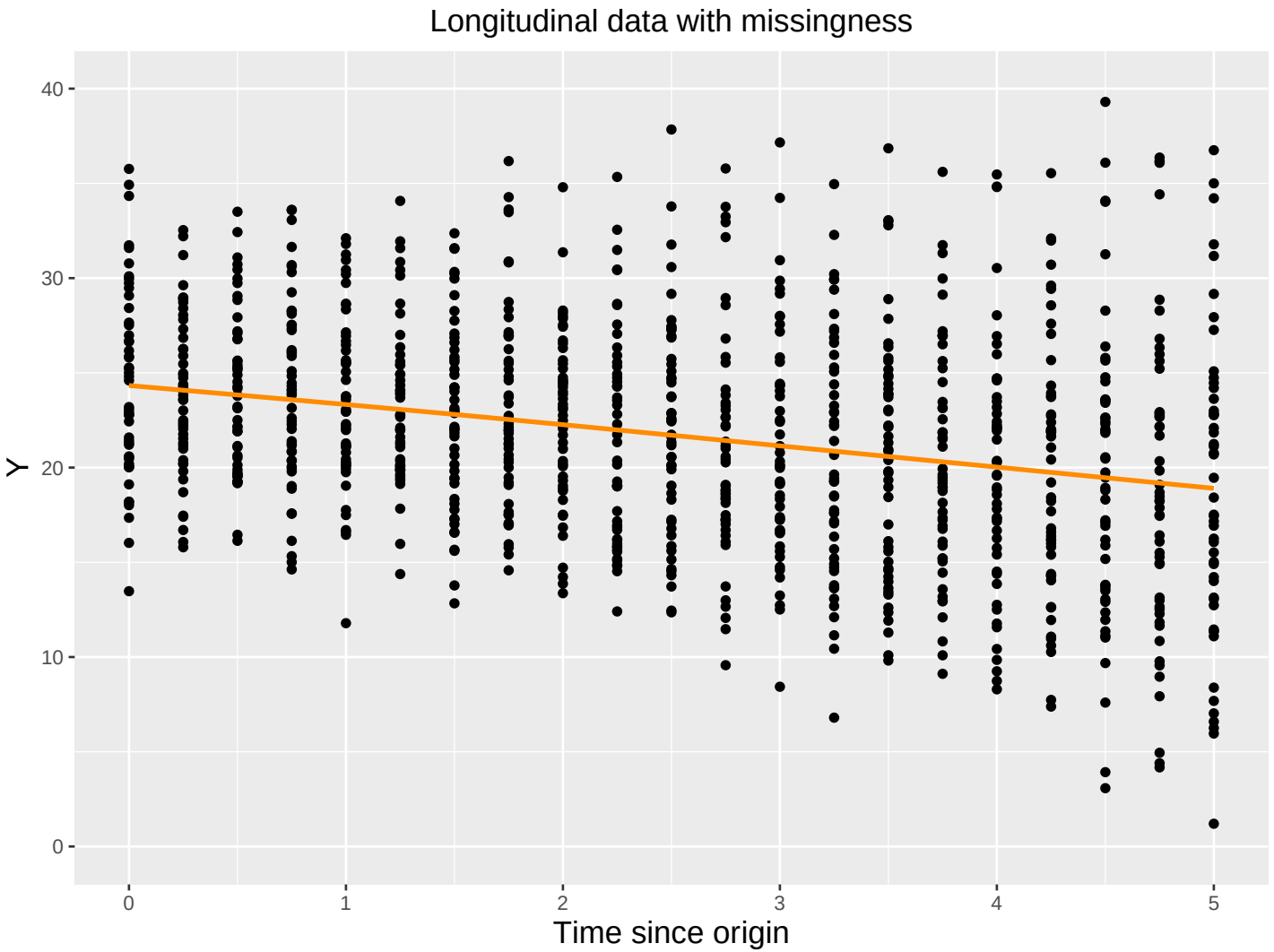
Linear mixed models - II

- In the general case,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon},$$

  where $\boldsymbol{\beta}$ and $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{D})$ are the fixed and random effects, with associated design matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively. $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \omega^{-1}\boldsymbol{b}_i)$.
- $\boldsymbol{X}$ and $\boldsymbol{Z}$ are typically functions of time.
- $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $Var(\boldsymbol{Y}) = \sigma^2(\boldsymbol{b}_i + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\top)$.
- Hence, integrating out the random effects, a specific model for the covariance structure is implied.
- Both Bayesian and Frequentist inferences are straightforward.

Example of longitudinal data with missingness



Types of missing data

Subject	Visits				
	1	2	3	4	5
1	x	x	x	x	x
2	x	x	x	?	?
3	?	x	x	x	x
4	?	x	?	x	?

Figure: Examples of missing data patterns in a hypothetical longitudinal study with five planned measurements; "x" denotes an observed measurement and "?" a missing longitudinal response

Missing data in longitudinal studies

- In practice, some values in $\boldsymbol{Y} = (Y_1, \ldots, Y_Q)^\top$ can be missing.
- Dropout: measurements after a certain time point are missing.
- $M = \sum_{j=1}^{Q} R_j$, the number of the observed measurements, where R_j denotes an indicator of Y_j being observed/recorded.
- The quality of the inference depends crucially on the association between $\boldsymbol{Y}$ and M.
- Given $M = j$, $\boldsymbol{Y}^\top = (\boldsymbol{Y}_{(j)}^\top, \boldsymbol{Y}_{(\bar{j})}^\top)$, with

$$\boldsymbol{Y}_{(j)} = (Y_1, \ldots, Y_j)^\top \quad \text{(observed data)}$$
$$\boldsymbol{Y}_{(\bar{j})} = (Y_{j+1}, \ldots, Y_Q)^\top \quad \text{(missing data)}.$$

- Bayesian/Frequentist inferences are based on the joint distribution of the observed data and the dropout process, $(\boldsymbol{Y}_{(M)}, M)$.

Missing data mechanisms Little & Rubin (1987)

- Missing completely at random (**MCAR**): all valid methods unbiased

$$\Pr(M = j|\boldsymbol{Y}; \boldsymbol{\theta}_t) = \Pr(M = j; \boldsymbol{\theta}_t).$$

- Missing at random (**MAR**): likelihood-based methods (e.g. linear mixed models) modelling all the observed data unbiased

$$\Pr(M = j|\boldsymbol{Y}; \boldsymbol{\theta}_t) = \Pr(M = j|\boldsymbol{Y}_{(j)}; \boldsymbol{\theta}_t).$$

- Missing not at random (**MNAR**): needs joint modelling of the marker and the dropout process

$$\Pr(M = j|\boldsymbol{Y}; \boldsymbol{\theta}_t) = \Pr(M = j|\boldsymbol{Y}_{(j)}, \boldsymbol{Y}_{(\bar{j})}; \boldsymbol{\theta}_t).$$

Definite discrimination between MAR and MNAR is difficult as it relies on modelling assumptions (Molenberghs et al. 2008).

Shared random effects models (SREMs)

- Shared random effects models (SREMs) are a specific subclass of joint models.
- **Key assumption:** The marker model and the time-to-dropout model are linked through the random effects $\boldsymbol{b}$. They specify models for both
  - the observed responses $\boldsymbol{Y}|\boldsymbol{b}$, e.g. through LMMs: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$;
  - the time to dropout $T|\boldsymbol{b}$
- Given $\boldsymbol{b}$, the two processes are assumed to be independent.
- They are joint models corresponding to a MNAR mechanism.
  - Dependence on the missing observations is introduced by integrating out the random effects.
- Numerical/stochastic integration methods are usually required for inference.

Examples of SREMs

In many SREMs, it is considered that

$$m(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{Z}(t)\boldsymbol{b}$$

is the "true" marker value at t.

Examples of SREMs

In many SREMs, it is considered that

$$m(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{Z}(t)\boldsymbol{b}$$

is the "true" marker value at t.

- **Current value parameterization:** The hazard of dropout depends on the "true" marker value

$$h\{t|m(t); \boldsymbol{\theta}_t\} = h_0(t; \boldsymbol{\psi}) \exp\{\alpha m(t)\}.$$

Examples of SREMs

In many SREMs, it is considered that

$$m(t) = \boldsymbol{X}(t)\boldsymbol{\beta} + \boldsymbol{Z}(t)\boldsymbol{b}$$

is the "true" marker value at t.

- **Current value parameterization:** The hazard of dropout depends on the "true" marker value

$$h\{t|m(t);\boldsymbol{\theta}_t\} = h_0(t;\boldsymbol{\psi})\exp\{\alpha m(t)\}.$$

- **Random-effects parameterization:** The hazard of dropout depends directly on the random effects

$$h(t|\boldsymbol{b};\boldsymbol{\theta}_t) = h_0(t;\boldsymbol{\psi})\exp\left(\boldsymbol{\alpha}^\top \boldsymbol{b}\right).$$

Competing risk SREMs

- Most of the research in joint modeling assumes that there is a single cause of failure (event).
- Multiple failure causes exist in many applications.
- When occurrence of one event precludes the occurrence of other events (or substantially alters the probability of observing the other events) $\rightarrow$ Competing risks.
- Joint modelling of longitudinal data and competing-risk survival data has also gained attention in the last decade.
- In principle, competing-risk data can be analyzed through either cause-specific hazards or cumulative incidence functions (CIFs).

Cause-specific hazards

Let $T_i^\star$ be the survival time and $K_i \in \{1, \ldots, K\}$ the failure cause

- **Cause-specific hazards**: $\rightarrow$ the rate of failure from a particular cause at a specific time point given that the individual has survived up to that point.

$$\alpha_{ik}(t) = \lim_{h \to 0} \frac{P(t < T_i^\star \leq t + h, K_i = k | T_i > t)}{h}$$

- Proportional cause-specific hazards are usually applied in practice, $\alpha_{ik}(t) = \alpha_{0k}(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$.

Cause-specific hazards

Let $T_i^\star$ be the survival time and $K_i \in \{1, \ldots, K\}$ the failure cause

- **Cause-specific hazards**: $\rightarrow$ the rate of failure from a particular cause at a specific time point given that the individual has survived up to that point.

$$\alpha_{ik}(t) = \lim_{h \to 0} \frac{P(t < T_i^\star \leq t + h, K_i = k | T_i > t)}{h}$$

- Proportional cause-specific hazards are usually applied in practice, $\alpha_{ik}(t) = \alpha_{0k}(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$.
- Likelihood function for a sample $\{(t_i, K_i)\}_{i=1}^N$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \alpha_{ik}(t_i)^{\delta_{ik}} \exp\left\{ - \int_0^{t_i} \alpha_{ik}(u) du \right\}$$

Cause-specific hazards

Let $T_i^\star$ be the survival time and $K_i \in \{1, \ldots, K\}$ the failure cause

- **Cause-specific hazards**: $\rightarrow$ the rate of failure from a particular cause at a specific time point given that the individual has survived up to that point.

$$\alpha_{ik}(t) = \lim_{h \to 0} \frac{P(t < T_i^\star \leq t + h, K_i = k | T_i > t)}{h}$$

- Proportional cause-specific hazards are usually applied in practice, $\alpha_{ik}(t) = \alpha_{0k}(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$.
- Likelihood function for a sample $\{(t_i, K_i)\}_{i=1}^N$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \alpha_{ik}(t_i)^{\delta_{ik}} \exp\left\{ - \int_0^{t_i} \alpha_{ik}(u) du \right\}$$

$\Rightarrow$ models for each cause can be fitted separately by treating the other failure causes as non-informative right censoring.

Cumulative incidence function (CIF)

Definition: The probability of occurrence of a specific cause over time

$$F_{ik}(t) = P(T_i^\star \leq t, K_i = k) = \int_0^t \alpha_{ik}(u) \exp\left\{ - \int_0^u \sum_{k=1}^{K} \alpha_{ik}(w) dw \right\} du$$

- Complex function of cause-specific hazards
- Semiparametric modeling of subdistribution hazards, $\lambda_{ik}(t)$, proposed by Fine & Gray (1999) is typically performed for the event of interest as

$$F_{ik}(t) = 1 - \exp\left\{ - \int_0^t \lambda_{ik}(u) du \right\}$$

i.e. there is an **1-1 relationship** between $F_{ik}(t)$ and $\lambda_{ik}(t)$.

Cause-specific hazards or Cumulative incidence in SREMs?

- Aetiological-type research questions $\rightarrow$ cause-specific hazards
- Prognosis of a disease and prediction purposes $\rightarrow$ CIF
- Cause-specific hazards more frequent in joint modeling, probably due to the much easier implementation (recall the likelihood factorizes into K independent components).
- The CIFs can be obtained from the CSHs, requiring though **complex integration**.
  - becomes even more difficult due to the presence of random effects
- Therefore, SREMs in terms of the CIFs would be more natural and could substantially reduce the computational burden of formally deriving CIF estimates based on cause-specific hazard estimates.

Boundedness constraint in CIF-based modeling

Issue with CIF-based modeling

The all-cause CIF should be bounded by 1.

Approaches to deal with it in standard Survival Analysis

1. Ignore the constraint (Fine & Gray 1999, Jeong & Fine 2006, Mozumder et al. 2018)
2. Model the the baseline asymptote for one cause-specific CIF (Shi et al. 2013)
3. Add a small positive number to force the survival function to be positive (Mao & Lin 2017)
4. Incorporating a formal (nonlinear) boundedness constraint in the maximization process (Bakoyannis et al. 2017), e.g. through the Augmented Lagrangian Adaptive Barrier Minimization Algorithm (`alabama` library in `R`).

Boundedness constraint in SREMs

- However, how to impose such a constraint in SREMs is not so clear as SPMs are defined conditionally on the random effects, and integration over the prior distribution of the random effects is required to obtain the observed data likelihood.
- Under the Bayesian paradigm, Gelfand et al. (1992) suggested that when the **constraints involve the data** (as it is the case in CIF modelling), it is more natural to build the constraints into the likelihood function rather than into the prior distribution.

Motivating example: CD4 cell counts

- CD4 cell count, an immunological biormarker, has been widely used to keep track of HIV progression.
- CD4 counts increase rapidly after ART initiation, reaching in most cases normal levels within a few years.
- Robust CD4 recovery is important both at the individual and population level as lower CD4 counts are associated with higher mortality.

Motivation: CD4 modelling after ART initiation

- CD4 data are censored due to death in care and disengagement from care (competing risks).
- Significant under-reporting of deaths, more often in resource-constrained countries.
- Deceased patients can be incorrectly classified as disengaged from care $\Rightarrow$ biased estimates.
- Thus, disengagement from care is different from non-informative censoring (e.g. administrative censoring).
- Competing risk SREMs have been proposed in the literature, with most approaches based on cause-specific hazards.
  - However, the cumulative probability of an event over time, i.e. the cumulative incidence function (CIF), could be more relevant from a clinical perspective.

Failure cause misclassification

- One solution: **Double sampling**
  - The true failure cause is ascertained in a small random sample of individuals initially classified as disengaged from care.
- Various approaches to deal with outcome misclassification:
  - Bakoyannis et al. (2019): missing absorbing states in a multi-state model through pseudo-likelihood.
  - Daniel Paulino et al. (2003): misclassification in Binomial regression using MCMC.

AIMs

1. To propose a unified and flexible approach to jointly model a continuous disease marker over time and competing risks using CIFs for the survival submodels, accounting also for misspecified failure cause.

Proposed model structure

- **Longitudinal submodel**: a standard LMM

$$y_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i + \epsilon_i(t),$$

with $m_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i$ the **"true" marker value** for the ith individual at time t and $M_i(t) = \{m_i(s) : 0 \leq s \leq t\}$.

Proposed model structure

- **Longitudinal submodel**: a standard LMM

$$y_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i + \epsilon_i(t),$$

with $m_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i$ the **"true" marker value** for the ith individual at time t and $M_i(t) = \{m_i(s) : 0 \leq s \leq t\}$.

- **Competing risks submodel**: We simultaneously model the CIFs for all causes conditionally on the history of true marker values, $M_i(t)$:

$$F_{ik}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = \Pr\{T_i^\star \leq t, K_i = k|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\},$$

where $\boldsymbol{w}_{ik}$ denotes baseline covariates and $\boldsymbol{\theta}_{tk}$ the parameters of the kth CIF.

Models for CIFs

$$F_{ik}^{M}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = 1 - \exp\left\{-\int_0^t e^{\boldsymbol{B}_k^\top(s)\boldsymbol{\psi}_k + \boldsymbol{\gamma}_k^\top \boldsymbol{w}_{ik} + \alpha_k m_i(s)} ds\right\}, \textbf{SREM-CIF-1}$$

$$F_{ik}^{M}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = 1 - \left\{1 + c_k \int_0^t e^{\boldsymbol{B}_k^\top(s)\boldsymbol{\psi}_k + \boldsymbol{\gamma}_k^\top \boldsymbol{w}_{ik} + \alpha_k m_i(s)} ds\right\}^{-1/c_k} \textbf{SREM-CIF-2}$$

where $\boldsymbol{B}_k(t)$ is a B-splines basis matrix for cause k at time t and α_k is the parameter linking the "true" marker values to the CIF for cause k. Also, $\boldsymbol{\theta}_{tk}^\top = (\boldsymbol{\psi}_k^\top, \boldsymbol{\gamma}_k^\top, \alpha_k)$.

- **SREM-CIF-1** $\rightarrow$ proportional subdistribution hazards joint model (Deslandes & Chevret 2010), whereas **SREM-CIF-2** is an extension of **SREM-CIF-1** based on the generalized odds rate transformation (Jeong & Fine 2007, Bakoyannis et al. 2017).
- **SREM-CIF-2** reduces to **SREM-CIF-1** as $c_k \searrow 0$, thus the model proposed by Deslandes & Chevret (2010) is a special case.

Addressing boundness constraints

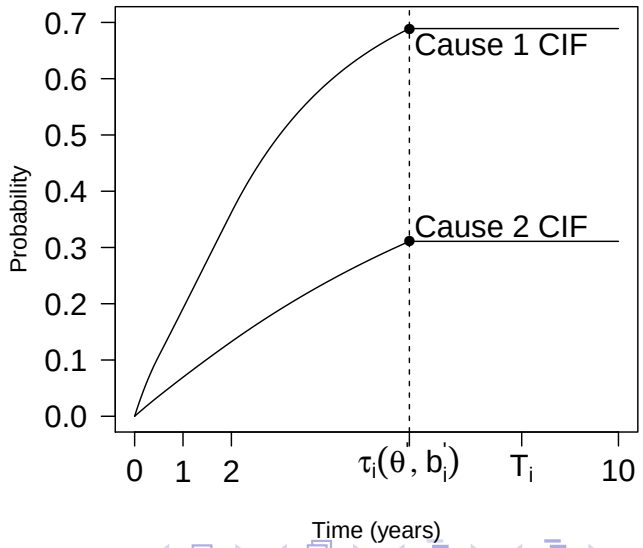
The sum of all cause-specific CIFs should be bounded by 1 at each failure time. To account for that, we assumed that

$$F_{ik}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = \begin{cases} F_{ik}^{M}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}, 0 \leq t \leq \tau_i \\ F_{ik}^{M}\{\tau_i|M_i(\tau_i), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}, t > \tau_i \end{cases},$$

i.e. we allowed the CIFs to increase up to a certain time point

$$\tau_i = \sup\left[t : \sum_{k=1}^{K} F_{ik}^{M}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} \leq 1\right]$$

- $\tau_i \equiv \tau_i(\boldsymbol{\theta}, \boldsymbol{b}_i)$ is the upper limit for the survival time $T_i^{\star}$; $\tau_i = \infty$ if the constraint is met $\forall t > 0$.
- If some specific parameter values $(\boldsymbol{\theta}, \boldsymbol{b}_i)$ do not meet this constraint $\Rightarrow$ **zero** likelihood $\Rightarrow$ **zero** posterior.



Bayesian estimation

- Bayesian estimation procedure using MCMC.
- Letting $\boldsymbol{\theta} = (\boldsymbol{\theta}_L^\top, \boldsymbol{\theta}_t^\top)^\top$ be the whole parameter vector and $\mathcal{D}_{obs}$ be the observed data, the posterior of all unknown parameters, $f(\boldsymbol{\theta}, \boldsymbol{b}|\mathcal{D}_{obs})$, is proportional to

$$f(\boldsymbol{\theta}) \prod_{i=1}^{N} \left(f(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}_L) f(\boldsymbol{b}_i;\boldsymbol{\theta}_L) \prod_{k=1}^{K} f_{ik}\{T_i|M_i(T_i), \boldsymbol{w}_{ik};\boldsymbol{\theta}_{tk}\}^{\delta_{ik}} \right.$$
$$\left. \left[1 - \sum_{k=1}^{K} F_{ik}\{T_i|M_i(T_i), \boldsymbol{w}_{ik};\boldsymbol{\theta}_{tk}\} \right]^{1-\delta_i} \right),$$

where $\delta_{ik} = I(K_i = k)$, $\delta_i = \sum_{k=1}^{K} \delta_{ik}$, $T_i = \min(T_i^\star, C_i)$ and $f_{ik}\{t|M_i(t), \boldsymbol{w}_{ik};\boldsymbol{\theta}_{tk}\}$ denotes the derivative of $F_{ik}\{t|M_i(t), \boldsymbol{w}_{ik};\boldsymbol{\theta}_{tk}\}$ over t.

Calculation of the upper limit τ_i?

- Since $F_{ik}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = \begin{cases} F_{ik}^M\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}, 0 \leq t \leq \tau_i \\ F_{ik}^M\{\tau_i|M_i(\tau_i), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}, t > \tau_i \end{cases}$ the density function is equal to

$$f_{ik}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = I(0 < t < \tau_i)\frac{\partial F_{ik}^M\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}}{\partial t}$$

- Therefore, the posterior distribution is equivalent to including the model-based CIF, $F_{ik}^M\{T_i|M_i(T_i), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}$, and its derivative along with the indicator function

$$I\left[\sum_{k=1}^{K} F_{ik}^M\{T_i|M_i(T_i), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}\right]$$

- Thus, calculation of $\tau_i(\boldsymbol{\beta}, \boldsymbol{\theta}_t, \boldsymbol{b}_i)$ is not required within the MCMC algorithm (more on the usefulness of τ_i later).

MCMC details

Some conditional distributions can be conjugate:

- Assuming $\omega \sim \text{Gamma}(\lambda_1, \lambda_2) \Rightarrow$ the corresponding conditional posterior distribution is

$$\text{Gamma}\left\{\frac{n}{2} + \lambda_1, \lambda_2 + \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)^\top(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)\right\}$$

- Assuming $\boldsymbol{D} \sim IW(\boldsymbol{A}, df)$ (Inverse-Wishart), the corresponding conditional posterior distribution is equal to

$$IW\left(\boldsymbol{A} + \sum_{i=1}^{N}\boldsymbol{b}_i\boldsymbol{b}_i^\top, df + N\right)$$

Conditional posterior of $\boldsymbol{\beta}$

Proposal distribution: $q(\boldsymbol{\beta}^{can}|\mathcal{D}, \boldsymbol{b}; \omega) \sim N(\boldsymbol{\mu}_1, \boldsymbol{C}_1)$ (the posterior distribution given the longitudinal model only)

$$\begin{aligned} \boldsymbol{C}_1 &= \left(\boldsymbol{C}_0^{-1} + \omega \sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{X}_i \right)^{-1} \\ \boldsymbol{\mu}_1 &= \boldsymbol{C}_1 \left\{ \boldsymbol{C}_0^{-1} \boldsymbol{\mu}_0 + \omega \sum_{i=1}^{N} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{Z}_i \boldsymbol{b}_i) \right\} \end{aligned}$$

Then the acceptance probability is equal to

$$p = \min \left\{ 1, \frac{\prod_{i=1}^{N} f\{T_i, K_i | M_i^{can}(T_i), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}}{\prod_{i=1}^{N} f\{T_i, K_i | M_i(T_i), \boldsymbol{w}_i; \boldsymbol{\theta}_{tk}\}} \right\},$$

where $M_i^{can}(T_i)$ and $M_i(T_i)$ denote the "true" marker values up to T_i evaluated at the candidate, $\boldsymbol{\beta}^{can}$, and current MCMC value, $\boldsymbol{\beta}$, respectively.

Conditional posterior of $\boldsymbol{b}_i$

$$
\begin{aligned}
f(\boldsymbol{b}_i|\mathcal{D};\boldsymbol{\theta}) \quad &\propto \quad \exp\left\{-\frac{1}{2}\boldsymbol{b}_i^\top(\boldsymbol{D}^{-1}+\omega\boldsymbol{Z}_i^\top\boldsymbol{Z}_i)\boldsymbol{b}_i+\omega\boldsymbol{b}_i^\top\boldsymbol{Z}_i^\top(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})\right\} \\
&\times \quad \prod_{k=1}^{K} f_{ik}^{M}\{T_i|M_i(T_i),\boldsymbol{w}_{ik};\boldsymbol{\theta}_{tk}\}^{\delta_{ik}}\left[S_i^{M}\{T_i|M_i(T_i),\boldsymbol{w}_i;\boldsymbol{\theta}_t\}\right]^{1-\delta_i} \\
&\times \quad I\left[\sum_{k=1}^{K} F_{ik}^{M}\{T_i|M_i(T_i),\boldsymbol{w}_{ik};\boldsymbol{\theta}_{tk}\}<1\right]
\end{aligned}
$$

- Starting from the posterior mode using only the marker model, $\boldsymbol{\mu}_{\boldsymbol{b}_i}$, we carry out a single Newton Raphson step

$$\boldsymbol{b}_i^\star = \boldsymbol{\mu}_{\boldsymbol{b}_i} + \mathcal{I}(\boldsymbol{\mu}_{\boldsymbol{b}_i})^{-1}\mathcal{U}(\boldsymbol{\mu}_{\boldsymbol{b}_i})$$

where $\mathcal{U}(\boldsymbol{b}_i) = \frac{\partial \log f(\boldsymbol{b}_i|\mathcal{D};\boldsymbol{\theta})}{\partial \boldsymbol{b}_i}$ and $\mathcal{I}(\boldsymbol{b}_i) = -\frac{\partial^2 \log f(\boldsymbol{b}_i|\mathcal{D};\boldsymbol{\theta})}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^\top}$.

- Boundness constraint was ignored in $\mathcal{U}(\boldsymbol{b}_i)$ and $\mathcal{I}(\boldsymbol{b}_i)$

Conditional posterior of $\boldsymbol{b}_i$

- **Proposal density:** $q(\boldsymbol{b}_i^{can}|\mathcal{D};\boldsymbol{\theta}) \sim N\left\{\boldsymbol{b}_i^{\star}, (\boldsymbol{D}^{-1} + \omega \boldsymbol{Z}_i^{\top}\boldsymbol{Z}_i)^{-1}\right\}$
- Does not depend on the current value of $\boldsymbol{b}_i$, though it does depend on the current values of the remaining parameters, $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_L$.
- Metropolis-Hastings acceptance probability:

$$p = \min\left\{1, \frac{f(\boldsymbol{b}_i^{can}|\mathcal{D};\boldsymbol{\theta})}{f(\boldsymbol{b}_i|\mathcal{D};\boldsymbol{\theta})} \times \frac{q(\boldsymbol{b}_i|\mathcal{D};\boldsymbol{\theta})}{q(\boldsymbol{b}_i^{can}|\mathcal{D};\boldsymbol{\theta})}\right\}.$$

- If the all-cause CIF is not bounded by 1 at $\boldsymbol{b}_i^{can}$, the acceptance probability is equal to zero as $f(\boldsymbol{b}_i^{can}|\mathcal{D};\boldsymbol{\theta})$ equals zero.
- A similar approach was adopted to update the values of $\boldsymbol{\theta}_t$, but we performed a low number of BFGS steps instead of Newton-Raphson to avoid calculation of the Hessian.

Integral in the definition of CIFs

SREM-CIF-1

$$F_{ik}^{M}\{t|M_i(t),\boldsymbol{w}_{ik};\boldsymbol{\theta}_{tk}\} = 1-\exp\left\{-\int_0^t e^{\boldsymbol{B}_k^{\top}(s)\boldsymbol{\psi}_k+\boldsymbol{\gamma}_k^{\top}\boldsymbol{w}_{ik}+\alpha_k m_i(s)}ds\right\}$$

- To calculate this one-dimensional integral, we used Gauss-Legendre rules with 30 nodes.
- That is, we first transformed the integration limits to $(-1,1)$

$$\int_a^b g(x)dx = \frac{b-a}{2}\int_{-1}^{1} g\left\{\frac{(b-a)u}{2}+\frac{a+b}{2}\right\}du$$

which can approximated by $\sum_{j=1}^{30} w_j g\left\{\frac{(b-a)x_j}{2}+\frac{a+b}{2}\right\}$.

- The pairs $\{(x_j,w_j)\}_{j=1}^{30}$ are predetermined to yield an exact solution to the integral if the integrand can be expressed in the form of any polynomial of degree $(2\times 30)-1$ or less that interpolates the abscissas.

Inference under misclassified causes of failure

- Let K_i be the true failure cause and $\tilde{K}_i$ be the observed one.
- **Misclassification probabilities**:

$$\pi_{jk}(\mathcal{D}_{misc,i}) = \Pr(\tilde{K}_i = j | K_i = k, \mathcal{D}_{misc,i}; \boldsymbol{\theta}_{misc}),$$

$\sum_{j=1}^{K} \pi_{jk}(\mathcal{D}_{misc,i}) = 1$ for all $k = 1, 2, \ldots, K$.

- $\mathcal{D}_{misc,i}$ observed data up to the event time, T_i.

Assumptions

1. Non-informative right censoring (e.g. administrative censoring) is correctly classified, i.e. $K_i = 0 \Leftrightarrow \tilde{K}_i = 0$
2. $\tilde{K}_i$ always observed, but K_i available only in a random sample (double sampling) ($R_i = 1$).

Missing failure status

In this context, the observed data are

$$\mathcal{D}_{obs} = \begin{cases} (\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, T_i, K_i, \tilde{K}_i, \boldsymbol{w}_i, \mathcal{D}_{misc,i}, R_i) \text{ if } R_i = 1, \ i = 1, \ldots, N, \\ (\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, T_i, K_i, \tilde{K}_i, \boldsymbol{w}_i, \mathcal{D}_{misc,i}, R_i) \text{ if } R_i = 0, \ i = 1, \ldots, N. \end{cases}$$

where R_i is an indicator function of the ith individual being doubly sampled.

- We assume MAR for the probability of being in the double sampling:

$$\Pr\{K_i = k | \tilde{K}_i = j, T_i = t, M_i(t), \boldsymbol{w}_i, \mathcal{D}_{misc,i}; \boldsymbol{\theta}, \boldsymbol{\theta}_{misc}\} =$$
$$\Pr\{K_i = k | \tilde{K}_i = j, T_i = t, M_i(t), \boldsymbol{w}_i, R_i, \mathcal{D}_{misc,i}; \boldsymbol{\theta}, \boldsymbol{\theta}_{misc}\}.$$

- The true failure cause should not depend on whether $R_i = 1$ or $R_i = 0$.

Missing failure status

- Due to MAR we are able to predict the missing true failure cause based on the observed data only.
- In fact, this is an alternative definition of MAR (using simplified notation)

$$
\begin{aligned}
f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, R_i) &= \frac{f(\boldsymbol{y}_i^m, \boldsymbol{y}_i^o, R_i)}{f(\boldsymbol{y}_i^o, R_i)} = \frac{f(R_i|\boldsymbol{y}_i^m, \boldsymbol{y}_i^o) f(\boldsymbol{y}_i^m, \boldsymbol{y}_i^o)}{f(R_i|\boldsymbol{y}_i^o) f(\boldsymbol{y}_i^o)} \\
&= \frac{f(R_i|\boldsymbol{y}_i^o)}{f(R_i|\boldsymbol{y}_i^o)} f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o) = f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o).
\end{aligned}
$$

Imputation step - I

Due to MAR, as shown above, the missing failure causes can be predicted by the model through

$$\Pr\{K_i = k|\tilde{K}_i = j, T_i^\star = t, M_i(t), \boldsymbol{w}_i, \mathcal{D}_{misc,i}; \boldsymbol{\theta}, \boldsymbol{\theta}_{misc}\}$$

By the assumptions that

1. the failure cause probabilities do not depend on $\mathcal{D}_{misc,i}$ and $\boldsymbol{\theta}_{misc}$
2. and the misclassification probabilities $\pi_{jk}(\mathcal{D}_{misc,i})$ are independent of the random effects and the parameters of interest, $\boldsymbol{\theta}$, thus independent of $M_i(t)$ and $\boldsymbol{w}_i$

and using the law of total probability, it can be shown that $\Pr\{K_i = k|\tilde{K}_i = j, T_i^\star = t, M_i(t), \boldsymbol{w}_i, \mathcal{D}_{misc,i}; \boldsymbol{\theta}, \boldsymbol{\theta}_{misc}\}$ is equal to

$$\frac{\Pr\{K_i = k|T_i^\star = t, M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}\pi_{jk}(\mathcal{D}_{misc,i})}{\sum_{k=1}^{K} \Pr\{K_i = k|T_i^\star = t, M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}\pi_{jk}(\mathcal{D}_{misc,i})}$$

Imputation step - II

It also follows that the failure cause probabilities conditionally on the survival time $T_i^\star = t$ are equal to

$$\Pr\{K_i = k | T_i^\star = t, M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\} = \frac{\alpha_{ik}\{t | M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}}{\sum_{k=1}^{K} \alpha_{ik}\{t | M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}},$$

where $\alpha_{ik}\{t | M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}$ denotes the kth cause-specific hazard function for individual i.

Imputation step - II

It also follows that the failure cause probabilities conditionally on the survival time $T_i^\star = t$ are equal to

$$\Pr\{K_i = k | T_i^\star = t, M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\} = \frac{\alpha_{ik}\{t | M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}}{\sum_{k=1}^{K} \alpha_{ik}\{t | M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}},$$

where $\alpha_{ik}\{t | M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}$ denotes the kth cause-specific hazard function for individual i. By definition,

$$F_{ik}\{t | M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = \int_0^t \alpha_{ik}\{u | M_i(u), \boldsymbol{w}_i; \boldsymbol{\theta}_t\} S_i\{u | M_i(u), \boldsymbol{w}_i; \boldsymbol{\theta}_t\} du,$$

thus it follows that the missing failure causes can be predicted by

$$\frac{f_{ik}\{t | M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} \pi_{jk}(\mathcal{D}_{misc,i})}{\sum_{k=1}^{K} f_{ik}\{t | M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} \pi_{jk}(\mathcal{D}_{misc,i})}.$$

Imputation step - Remarks

- Since a failure cause is always reported, **both the joint model and the misclassification model** are required to predict the missing causes of failure.
- K_i can be repeatedly imputed within the MCMC algorithm based on the current values of $\boldsymbol{\theta}$ and $\pi_{jk}(\mathcal{D}_{misc,i})$ (**data augmentation**), (e.g. Daniel Paulino et al. 2003).
- In general, to model $\pi_{jk}(\mathcal{D}_{misc,i})$ conditional on the observed information (i.e. $\mathcal{D}_{misc,i}$), multinomial logistic regression could be used for example.
- $\mathcal{D}_{misc,i}$ can include the observed marker values, the event time, or other auxiliary information that is not included in the scientific model of interest (joint model).

MCMC for misclassified failure causes

1. Choose adequate initial values $\boldsymbol{\theta}^{(0)}, \boldsymbol{b}^{(0)}, \{K_i^{(0)} : i \in \mathcal{I}_{mis}\}, \boldsymbol{\theta}_{misc}^{(0)}$, meeting the likelihood constraints for all individuals. For $l = 1, 2, \ldots, L$

MCMC for misclassified failure causes

1. Choose adequate initial values $\boldsymbol{\theta}^{(0)}, \boldsymbol{b}^{(0)}, \{K_i^{(0)} : i \in \mathcal{I}_{mis}\}, \boldsymbol{\theta}_{misc}^{(0)}$, meeting the likelihood constraints for all individuals. For $l = 1, 2, \ldots, L$
2. **Posterior step**
   - Update $(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{b}^{(l-1)})$ to $(\boldsymbol{\theta}^{(l)}, \boldsymbol{b}^{(l)})$ according to the posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{b} | \{K_i^{(l-1)} : i \in \mathcal{I}_{mis}\}, \mathcal{D}_{obs})$, i.e. the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{b})$ given all the observed data, with the missing failure causes being equal to their current values.

MCMC for misclassified failure causes

1. Choose adequate initial values $\boldsymbol{\theta}^{(0)}, \boldsymbol{b}^{(0)}, \{K_i^{(0)} : i \in \mathcal{I}_{mis}\}, \boldsymbol{\theta}_{misc}^{(0)}$, meeting the likelihood constraints for all individuals. For $l = 1, 2, \ldots, L$
2. **Posterior step**
   - Update $(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{b}^{(l-1)})$ to $(\boldsymbol{\theta}^{(l)}, \boldsymbol{b}^{(l)})$ according to the posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{b} | \{K_i^{(l-1)} : i \in \mathcal{I}_{mis}\}, \mathcal{D}_{obs})$, i.e. the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{b})$ given all the observed data, with the missing failure causes being equal to their current values.
   - Update $\boldsymbol{\theta}_{misc}^{(l-1)}$ to $\boldsymbol{\theta}_{misc}^{(l)}$ according to $f(\boldsymbol{\theta}_{misc} | \{K_i^{(l-1)} : i \in \mathcal{I}_{mis}\}, \{K_i : i \notin \mathcal{I}_{mis}\}, \{\mathcal{D}_{misc,i}, i = 1, \ldots, N\})$. If $\mathcal{D}_{misc,i}$ is an empty set, the Dirichlet distribution (Beta for $K = 2$) leads to conditional conjugacy.

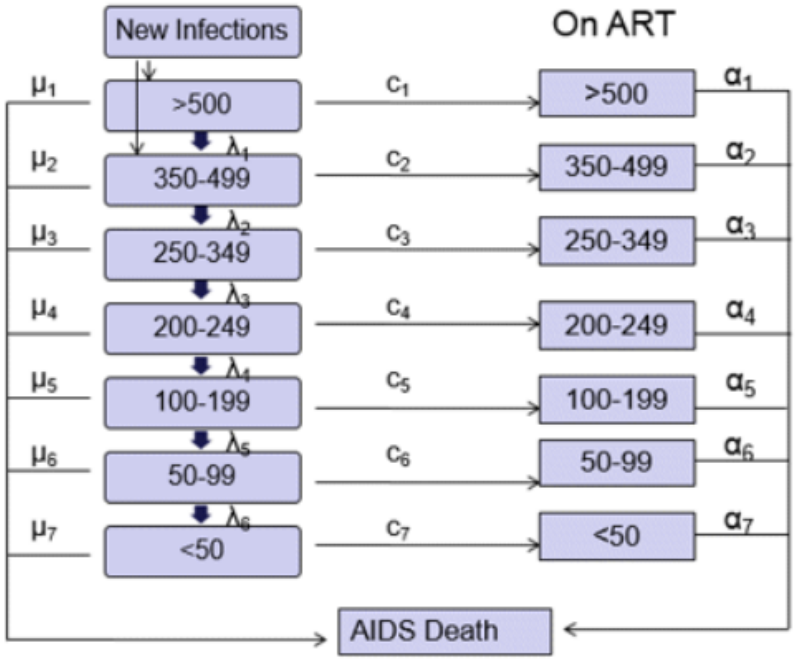
MCMC for misclassified failure causes

1. Choose adequate initial values $\boldsymbol{\theta}^{(0)}, \boldsymbol{b}^{(0)}, \{K_i^{(0)} : i \in \mathcal{I}_{mis}\}, \boldsymbol{\theta}_{misc}^{(0)}$, meeting the likelihood constraints for all individuals. For $l = 1, 2, \ldots, L$
2. **Posterior step**
   - Update $(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{b}^{(l-1)})$ to $(\boldsymbol{\theta}^{(l)}, \boldsymbol{b}^{(l)})$ according to the posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{b} | \{K_i^{(l-1)} : i \in \mathcal{I}_{mis}\}, \mathcal{D}_{obs})$, i.e. the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{b})$ given all the observed data, with the missing failure causes being equal to their current values.
   - Update $\boldsymbol{\theta}_{misc}^{(l-1)}$ to $\boldsymbol{\theta}_{misc}^{(l)}$ according to $f(\boldsymbol{\theta}_{misc} | \{K_i^{(l-1)} : i \in \mathcal{I}_{mis}\}, \{K_i : i \notin \mathcal{I}_{mis}\}, \{\mathcal{D}_{misc,i}, i = 1, \ldots, N\})$. If $\mathcal{D}_{misc,i}$ is an empty set, the Dirichlet distribution (Beta for $K = 2$) leads to conditional conjugacy.
3. **Imputation step**
   - Sample $\{K_i^{(l)} : i \in \mathcal{I}_{mis}\}$ directly from its posterior distribution.

Motivation - UNAIDS mortality estimates

- The United Nations (UN) Joint Programme in HIV/AIDS (UNAIDS) produces various estimates of parameters relevant to the worldwide HIV epidemic.
- E.g. Progression to next CD4 category, mortality by CD4 category, among many others.
- A CIF-based joint modeling approach could directly inform some parameters of the Spectrum software.
- The relevant statistical literature is sparse (Hu et al. 2012).

Figure: A portion of the Spectrum software.



Definition of longitudinal and survival states

- In medical research, e.g. in the *Spectrum* software of UNAIDS, it is common to discretize the marker values into non-overlapping intervals

$$\{[s_0, s_1), \ldots, [s_{J-1}, s_J)\}$$

and define mutually-exclusive states based on survival and (discretized) marker data.

For any $t > 0$,

$$\{m_i(t) \in S_h, T_i^\star > t\}, h = 1, \ldots, J \quad \text{(Marker states)}$$
$$\{T_i^\star \leq t, K_i = k\}, k = 1, \ldots, K \quad \text{(Survival states)}$$

where $S_h = [s_{h-1}, s_h)$.

- As the focus is often on describing the "true" biological process, states have been defined in terms of the "true" marker values.

Monitoring the cohort evolution through states

Progression of the whole cohort can be easily monitored by a series of estimated multistate probabilities

- $\Pr\{m_i(t) \in S_h, T_i^\star > t | \boldsymbol{w}_i; \boldsymbol{\theta}\}$, $h = 1, \ldots, J$
  - Latent marker state probability, which expresses the probability of being event free and having "true" marker values in S_h.
- $\Pr(T_i^\star \leq t, K_i = k | \boldsymbol{w}_{ik}; \boldsymbol{\theta})$, $k = 1, \ldots, K$
  - The population-averaged CIF for a particular cause

The above estimates can be visualized through a multistate probability plot.

Transition probabilities by baseline marker state

To get better insight into the dynamics of the processes, one may be also interested in transitions by baseline marker stats. It can be easily shown that

$$\Pr\{T_i^\star \leq t, K_i = k | m_i(0) \in S_g, \boldsymbol{w}_{ik}; \boldsymbol{\theta}\}$$
$$= \int_{m_i(0)\in S_g} F_{ik}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} \frac{f(\boldsymbol{b}_i; \boldsymbol{\theta}_L)}{\Pr\{m_i(0) \in S_g; \boldsymbol{\theta}_L\}} d\boldsymbol{b}_i$$
$$\Pr\{m_i(t) \in S_h, T_i^\star > t | m_i(0) \in S_g, \boldsymbol{w}_i; \boldsymbol{\theta}\}$$
$$= \int_{m_i(0)\in S_g, m_i(t)\in S_h} S_i\{t|M_i(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\} \frac{f(\boldsymbol{b}_i; \boldsymbol{\theta}_L)}{\Pr\{m_i(0) \in S_g; \boldsymbol{\theta}_L\}} d\boldsymbol{b}_i$$

Inference involves two distinct problems (i) approximation of the integral over the random effects and (ii) accounting for the variability in $\boldsymbol{\theta}$.

Estimation of the transition probabilities (given $\boldsymbol{\theta}$)

$\Pr\{T_i^\star \leq t, K_i = k | m_i(0) \in S_g, \boldsymbol{w}_{ik}; \boldsymbol{\theta}\}$ can be approximated using $\{\boldsymbol{b}_{ig}^{(j)}\}_{j=1}^{N_{mc}} \sim N(\boldsymbol{0}, \boldsymbol{D})$ given that $m_i(0) \in S_g$, by

$$N_{mc}^{-1} \sum_{j=1}^{N_{mc}} F_{ik}\{t | M_{ig}^{(j)}(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} \tag{1}$$

where $m_{ig}^{(j)}(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_{ig}^{(j)}$ and $M_{ig}^{(j)}(t) = \{m_{ig}^{(j)}(s) : 0 \leq s \leq t\}$.

Estimation of the transition probabilities (given $\boldsymbol{\theta}$)

Similarly, after multiplying and dividing the integrand by $\Pr\{m_i(t) \in S_h, m_i(0) \in S_g; \boldsymbol{\theta}\}$ (**importance sampling**), $\Pr\{m_i(t) \in S_h, T_i^\star > t | m_i(0) \in S_g, \boldsymbol{w}_i; \boldsymbol{\theta}\}$ can be approximated using $\{\boldsymbol{b}_{igh}^{(j)}\}_{j=1}^{N_{mc}} \sim N(\boldsymbol{0}, \boldsymbol{D})$ given that $m_i(0) \in S_g$ and $m_i(t) \in S_h$, i.e.

$$\frac{\Pr\{m_i(t) \in S_h, m_i(0) \in S_g; \boldsymbol{\theta}\}}{\Pr\{m_i(0) \in S_g; \boldsymbol{\theta}\} N_{mc}} \sum_{j=1}^{N_{mc}} S_i\{t | M_{igh}^{(j)}(t), \boldsymbol{w}_i; \boldsymbol{\theta}_t\}, \qquad (2)$$

where $m_{igh}^{(j)}(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_{igh}^{(j)}$ and $M_{igh}^{(j)}(t) = \{m_{igh}^{(j)}(s) : 0 \leq s \leq t\}$.

Monte Carlo integration for estimating the transition probabilities

Samples from multivariate normal under linear inequality constraints

Samples $\{\boldsymbol{b}_{ig}^{(j)}\}_{j=1}^{N_{mc}}$ for $\boldsymbol{b}_i$ from the $N(\boldsymbol{0}, \boldsymbol{D})$ distribution under the linear constraint $m_i(0) \in S_g$ can be simulated, among many other options (e.g. Gibbs sampling), very efficiently through Hamiltonian Monte Carlo (Pakman 2015).

Monte Carlo integration for estimating the transition probabilities

Samples from multivariate normal under linear inequality constraints

Samples $\{\boldsymbol{b}_{ig}^{(j)}\}_{j=1}^{N_{mc}}$ for $\boldsymbol{b}_i$ from the $N(\boldsymbol{0}, \boldsymbol{D})$ distribution under the linear constraint $m_i(0) \in S_g$ can be simulated, among many other options (e.g. Gibbs sampling), very efficiently through Hamiltonian Monte Carlo (Pakman 2015).

- Note that if $\sum_{k=1}^{K} F_{ik}^{M}\{t|M_{ig}^{(j)}(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} > 1$,

$$F_{ik}\{t|M_{ig}^{(j)}(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} = F_{ik}\{t'|M_{ig}^{(j)}(t'), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\},$$

  where $t' = \tau_i(\boldsymbol{\beta}, \boldsymbol{\theta}_t, \boldsymbol{b}_{ig}^{(j)})$.
- Thus, calculation of the upper bound is required only for the random draws that do not fulfil the boundedness constraint.

Posterior samples for multistate/transition probabilities

A posterior sample for the transition probabilities can be obtained by

1. drawing $\boldsymbol{\theta}^{(l)} \sim f(\boldsymbol{\theta}|\mathcal{D}_{obs})$, $l = 1, 2, \ldots, L$ and
2. approximating $\Pr\{m_i(t) \in S_h, T_i^\star > t | m_i(0) \in S_g, \boldsymbol{w}_i; \boldsymbol{\theta}^{(l)}\}$ and $\Pr\{m_i(t) \in S_h, T_i^\star > t | m_i(0) \in S_g, \boldsymbol{w}_i; \boldsymbol{\theta}^{(l)}\}$, for each $l = 1, 2, \ldots, L$, using the formulas previously described.

Posterior samples for population-averaged CIFs and latent marker state probabilities through

$$\Pr(T_i^\star \leq t, K_i = k | \boldsymbol{w}_{ik}; \boldsymbol{\theta})$$

$$= \sum_{g=1}^{J} \Pr\{T_i^\star \leq t, K_i = k | m_i(0) \in S_g, \boldsymbol{w}_{ik}; \boldsymbol{\theta}\} \Pr\{m_i(0) \in S_g; \boldsymbol{\theta}\}$$

$$\Pr\{m_i(t) \in S_h, T_i^\star > t | \boldsymbol{w}_i; \boldsymbol{\theta}\}$$

$$= \sum_{g=1}^{J} \Pr\{m_i(t) \in S_h, T_i^\star > t | m_i(0) \in S_g, \boldsymbol{w}_i; \boldsymbol{\theta}\} \Pr\{m_i(0) \in S_g; \boldsymbol{\theta}\}$$

CIF estimates conditional on observed marker states

- In a clinical application, estimating the population-averaged CIF conditional on the observed marker state could be valuable for making projections about the future cohort evolution.
- Thus, CIFs given observed baseline state, $\Pr\{T_i^\star \leq t, K_i = k | y_i(0) \in S_g, \boldsymbol{w}_{ik}; \boldsymbol{\theta}\}$, $g = 1, \ldots, J$, could be of interest.
- By similar probabilistic arguments, $\Pr\{T_i^\star \leq t, K_i = k | y_i(0) \in S_g, \boldsymbol{w}_{ik}; \boldsymbol{\theta}\}$ can be shown to be equal to

$$\int_{y_i(0) \in S_g} \int F_{ik}\{t | M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} \frac{f\{y_i(0), \boldsymbol{b}_i; \boldsymbol{\theta}\}}{\Pr\{y_i(0) \in S_g; \boldsymbol{\theta}\}} \, d\boldsymbol{b}_i \, dy_i(0).$$

CIF estimates conditional on observed marker states

$$\int_{y_i(0)\in S_g} \int F_{ik}\{t|M_i(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\} \frac{f\{y_i(0), \boldsymbol{b}_i; \boldsymbol{\theta}\}}{\Pr\{y_i(0) \in S_g; \boldsymbol{\theta}\}} \, d\boldsymbol{b}_i \, dy_i(0).$$

which can be estimated by drawing samples $\{y_{ig}^{(j)}(0), \boldsymbol{b}_{ig}^{(j)}\}_{j=1}^{N_{mc}}$ for $\{y_i(0), \boldsymbol{b}_i\}$ from the

$$N\left\{\begin{pmatrix} \boldsymbol{x}_i^\top(0)\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2 + \boldsymbol{z}_i^\top(0)\boldsymbol{D}\boldsymbol{z}_i(0) & \boldsymbol{z}_i^\top(0)\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{z}_i(0) & \boldsymbol{D} \end{pmatrix}\right\},$$

distribution, constrained such that $y_i(0) \in S_g$, i.e.

$$\Pr\{T_i^\star \leq t, K_i = k | y_i(0) \in S_g, \boldsymbol{w}_{ik}; \boldsymbol{\theta}\}$$

can be approximated by $N_{mc}^{-1} \sum_{j=1}^{N_{mc}} F_{ik}\{t|M_{ig}^{(j)}(t), \boldsymbol{w}_{ik}; \boldsymbol{\theta}_{tk}\}$, where $m_{ig}^{(j)}(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_{ig}^{(j)}$ and $M_{ig}^{(j)}(t) = \{m_{ig}^{(j)}(s) : 0 \leq s \leq t\}$.

CIF estimates conditional on history of observed marker states

- Similarly, one may be also interested in CIFs conditional on being in certain observed states at specific time points.
- In this case, it would be reasonable to also condition on survival up to the last time point and the baseline state, i.e. $\Pr\{T_i^\star \leq t, K_i = k | T_i^\star > s, y_i(0) \in S_g, y_i(s) \in S_h, \boldsymbol{w}_i; \boldsymbol{\theta}\}$, for $0 \leq s < t$ and $g, h \in \{1, 2, \ldots, J\}$.
- Estimation becomes more involved requiring evaluation of two integrals...
- Such estimates could be useful for identifying certain subsets of the population who are event free and at high risk for developing any of the events.

Simulation study design

- Marker data generated by an LMM assuming piece-wise linear evolution over time
  - 10 year study duration with 2 obs/year.
- Two competing risks: $K = 1$ (death in care) and $K = 2$ (disengagement from care), with the CIFs based on

$$F_{ik}\{t|M_i(t), W_i; \boldsymbol{\theta}_{tk}\} = 1 - \exp\left\{-\int_0^t e^{u_k(s)+\gamma_1 W_i+\alpha_k m_i(s)} ds\right\}, \textbf{SREM-CIF-1}$$

$$F_{ik}\{t|M_i(t), W_i; \boldsymbol{\theta}_{tk}\} = 1 - \left\{1 + c_k \int_0^t e^{u_k(s)+\gamma_1 W_i+\alpha_k m_i(s)} ds\right\}^{-1/c_k} \textbf{SREM-CIF-2}$$

- $u_k(t)$ is a complex polynomial, and $c_k = 1$ in **SREM-CIF-2**.
- A binary covariate (W_i) effect on both CIFs was assumed.
- For each scenario, both models were fitted.

Simulation study design

- **Misclassification:** $\pi_{11} = 0.75$ and $\pi_{22} = 0.90$, i.e. the first event (death) more likely to be misclassified.
- 20% of individuals who failed from any event were included in the double sampling.
- Population CIFs and latent marker state estimates were also recorded for each replication.
- Based on our motivating example, we considered 7 latent marker states: [0,50), [50,100), [100,200), [200,250), [250,350), [350,500) and [500,∞) cells/μL.
- Just as an example we present estimates for the population CIFs and the [350,500) cells/μL latent marker state at 10 years.

Scenario-1: Results under SREM-CIF-1

Parameter	True[1]	Median	Bias	ASD	MCSD	Cov.	Median	Bias	ASD	MCSD	Cov.
Longitudinal	**Results from SREM-CIF-1**						**Results from SREM-CIF-2**				
Intercept	12.850	12.856	0.006	0.126	0.122	94.200	12.856	0.006	0.126	0.122	94.000
Slope1 (β_1)	6.030	6.027	-0.003	0.109	0.104	95.600	6.020	-0.010	0.109	0.104	95.200
Slope2 (β_2)	0.770	0.769	-0.001	0.031	0.030	94.800	0.767	-0.003	0.031	0.030	95.000
Slope3 (β_3)	0.000	-0.001	-0.001	0.017	0.017	94.200	-0.001	-0.001	0.017	0.017	94.200
Cause 1 (e.g. death)											
"True" marker value (α_1)	-0.160	-0.161	-0.001	0.016	0.017	94.600	-0.182		0.019	0.020	
Binary covariate (γ_1)	0.150	0.147	-0.003	0.147	0.149	94.200	0.159		0.167	0.168	
CIF1 $t = 10$, $w = 1$ (%)	15.604	15.246	-0.357	1.667	1.693	92.600	15.210	-0.394	1.664	1.677	93.000
CIF1 $t = 10$, $w = 0$ (%)	13.673	13.410	-0.263	1.613	1.642	94.600	13.433	-0.240	1.600	1.634	93.400
Cause 2 (e.g. disengagement)											
"True" marker value (α_2)	-0.020	-0.021	-0.001	0.010	0.010	94.800	-0.026		0.012	0.012	
Binary covariate (γ_2)	-0.150	-0.152	-0.002	0.088	0.087	94.600	-0.184		0.108	0.106	
CIF2 $t = 10$, $w = 1$ (%)	37.431	37.514	0.083	2.076	2.145	92.400	37.775	0.343	2.041	2.089	92.200
CIF2 $t = 10$, $w = 0$ (%)	41.997	42.143	0.146	2.083	2.168	92.800	42.138	0.141	2.029	2.100	93.000
Misclassification par.											
π_{11} (%)	75.000	73.873	-1.127	4.829	4.884	95.000	73.968	-1.032	4.815	4.881	94.800
π_{22} (%)	90.000	89.160	-0.840	1.935	2.051	91.600	89.079	-0.921	1.934	2.046	91.600
Marker states											
State 6,[2] $w = 1$ (%)	12.389	12.373	-0.015	0.501	0.508	94.200	12.289	-0.100	0.499	0.505	93.400
State 6, $w = 0$ (%)	11.687	11.635	-0.052	0.491	0.519	92.400	11.620	-0.067	0.487	0.519	91.200
State 6 to 7,[3] $w = 1$ (%)	41.984	42.127	0.143	1.830	1.818	95.800	42.054	0.070	1.825	1.822	94.800
State 6 to 7, $w = 0$ (%)	39.495	39.592	0.097	1.849	1.751	96.200	39.582	0.086	1.820	1.751	96.200

[1] "True" denotes the true parameter values; "Median" the mean of posterior medians over the 500 replications; "Bias" the mean bias for posterior median estimates; "ASD" the average posterior standard deviation, "MCSD" the empirical Monte carlo deviation of estimates and "Cov." the empirical coverage probability of posterior credible intervals.

[2] $\{\sqrt{350} \leq m_i(10) < \sqrt{500}\} \cap \{T_i^\star > 10\}$.

[3] $\{\sqrt{350} \leq m_i(0) < \sqrt{500}\} \rightarrow \{m_i(10) > \sqrt{500}\} \cap \{T_i^\star > 10\}$.

Scenario-2: Results under SREM-CIF-2

Parameter	True[1]	Median	Bias	ASD	MCSD	Cov.	Median	Bias	ASD	MCSD	Cov.
Longitudinal	**Results from SREM-CIF-1**						**Results from SREM-CIF-2**				
Intercept	12.850	12.846	-0.004	0.126	0.126	95.600	12.846	-0.004	0.126	0.126	96.000
Slope1 (β_1)	6.030	6.034	0.004	0.110	0.108	95.400	6.028	-0.002	0.110	0.108	96.000
Slope2 (β_2)	0.770	0.772	0.002	0.031	0.032	93.400	0.770	-0.000	0.031	0.032	93.400
Slope3 (β_3)	0.000	0.001	0.001	0.017	0.017	95.400	0.001	0.001	0.017	0.017	95.600
Cause 1 (e.g. death)											
"True" marker value (α_1)	-0.160	-0.143		0.016	0.016		-0.163	-0.003	0.019	0.019	95.600
Binary covariate (γ_1)	0.150	0.141		0.150	0.164		0.159	0.009	0.168	0.179	93.600
CIF1 $t = 10$, $w = 1$ (%)	15.521	15.264	-0.256	1.707	1.881	90.800	15.315	-0.206	1.713	1.839	92.200
CIF1 $t = 10$, $w = 0$ (%)	13.765	13.461	-0.304	1.643	1.677	94.000	13.492	-0.273	1.638	1.641	94.600
Cause 2 (e.g. disengagement)											
"True" marker value (α_2)	-0.020	-0.016		0.010	0.010		-0.019	0.001	0.012	0.012	95.400
Binary covariate (γ_2)	-0.150	-0.121		0.088	0.091		-0.152	-0.002	0.108	0.109	93.600
CIF2 $t = 10$, $w = 1$ (%)	37.837	37.932	0.095	2.090	2.258	91.400	38.046	0.209	2.064	2.201	91.600
CIF2 $t = 10$, $w = 0$ (%)	41.417	41.613	0.196	2.088	2.192	93.400	41.649	0.232	2.043	2.124	93.400
Misclassification par.											
π_{11} (%)	75.000	73.856	-1.144	4.890	5.123	91.800	73.819	-1.181	4.881	5.104	91.800
π_{22} (%)	90.000	89.096	-0.904	1.982	1.969	92.600	89.032	-0.968	1.977	1.963	92.600
Marker states											
State 6,[2] $w = 1$ (%)	12.257	12.238	-0.019	0.499	0.526	93.400	12.170	-0.087	0.498	0.524	93.000
State 6, $w = 0$ (%)	11.794	11.742	-0.052	0.493	0.504	93.600	11.722	-0.072	0.489	0.496	94.400
State 6 to 7,[3] $w = 1$ (%)	40.956	40.926	-0.029	1.827	1.833	94.400	40.919	-0.037	1.816	1.821	95.000
State 6 to 7, $w = 0$ (%)	39.052	39.007	-0.045	1.839	1.929	93.400	39.020	-0.032	1.809	1.902	93.400

[1] "True" denotes the true parameter values; "Median" the mean of posterior medians over the 500 replications; "Bias" the mean bias for posterior median estimates; "ASD" the average posterior standard deviation, "MCSD" the empirical Monte carlo deviation of estimates and "Cov." the empirical coverage probability of posterior credible intervals.

[2] $\{\sqrt{350} \leq m_i(10) < \sqrt{500}\} \cap \{T_i^\star > 10\}$.

[3] $\{\sqrt{350} \leq m_i(0) < \sqrt{500}\} \rightarrow \{m_i(10) > \sqrt{500}\} \cap \{T_i^\star > 10\}$.

Application to East Africa IeDEA data

- Data derived from the East Africa IeDEA cohort study.
- A 60% random sample from [35-45) years old women was selected leading to 8005 individuals.
- CD4 evolution since ART initiation.
- Two competing risks: (i) death in care ($K = 1$) and (ii) disengagement from care ($K = 2$).
- **Unidirectional misclassification**: a true disengagement cannot be an observed death.
- 3275 (40.9%) and 273 (3.4%) observed disengagements from care and deaths, respectively.
- 443 (13.5%) disengaged patients included in double sampling,
  - of whom, 80 (18.1%) were actually deceased.

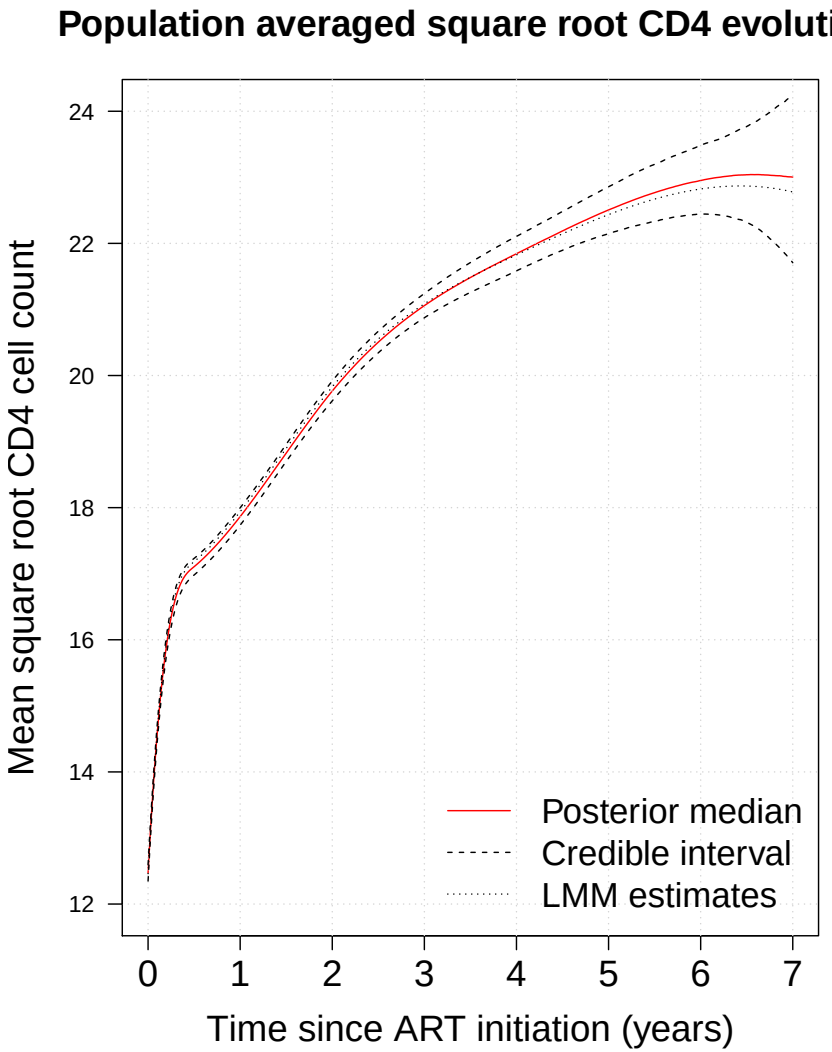
Application to East Africa IeDEA data

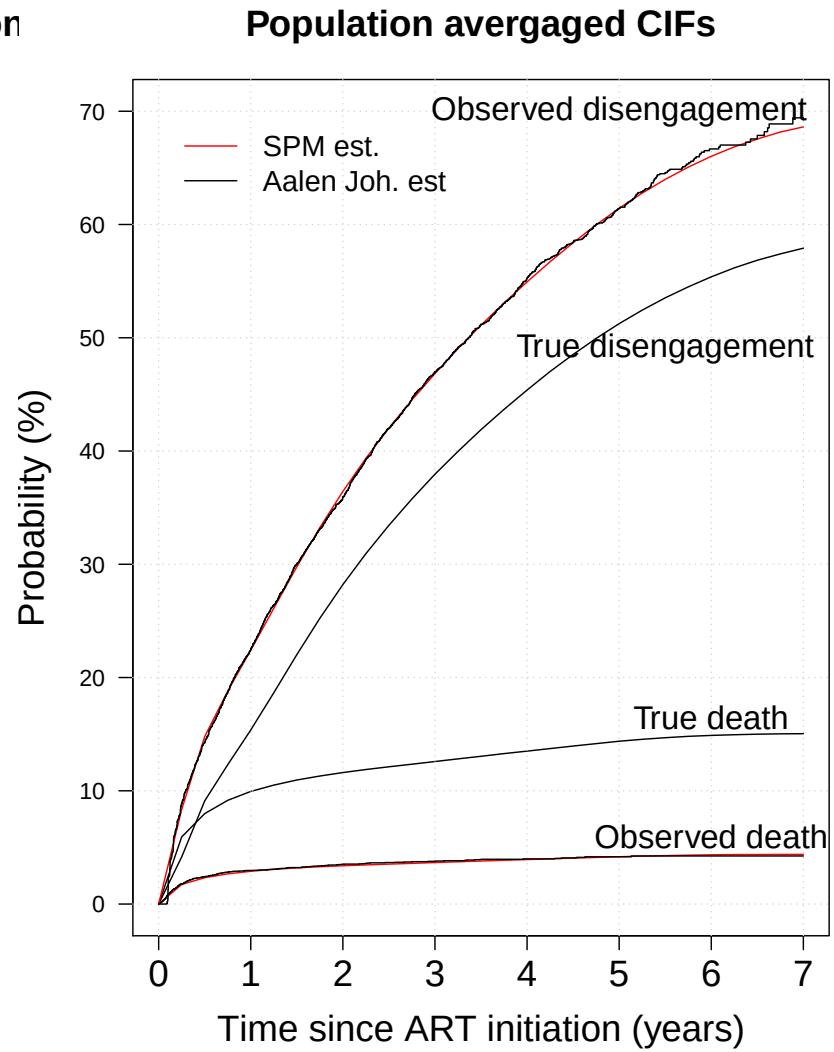
- B-splines (3 internal knots) for the square root CD4 evolution.
- Optimal fit based on the marginalized DIC when $c_1 = 1.5$ and $c_2 = 1e-05$ (effectively a subdistribution hazards model).

	Misclassification				No Misclassification			
Parameter	**Median**[1]	**SD**	**LB**	**UB**	**Median**	**SD**	**LB**	**UB**
Longitudinal								
Intercept	12.48	0.06	12.35	12.60	12.47	0.06	12.35	12.59
β_1	4.32	0.10	4.13	4.51	4.36	0.10	4.17	4.55
β_2	4.81	0.11	4.59	5.03	4.84	0.11	4.63	5.06
β_3	8.07	0.15	7.78	8.36	8.07	0.15	7.78	8.36
β_4	9.62	0.27	9.10	10.17	9.52	0.28	8.96	10.05
β_5	10.76	0.44	9.88	11.61	10.51	0.44	9.64	11.38
β_6	10.52	0.65	9.25	11.77	10.24	0.66	8.94	11.51
Cause 1 (Death)								
"True" marker value, α_1	-0.20	0.01	-0.23	-0.17	-0.18	0.02	-0.22	-0.15
Cause 2 (Disengagement)								
"True" marker value sHR, $\exp(\alpha_2)$	1.04	0.01	1.03	1.06	1.00	<0.01	0.99	1.01
π_{11} (%)	29.21	1.99	25.56	33.32				

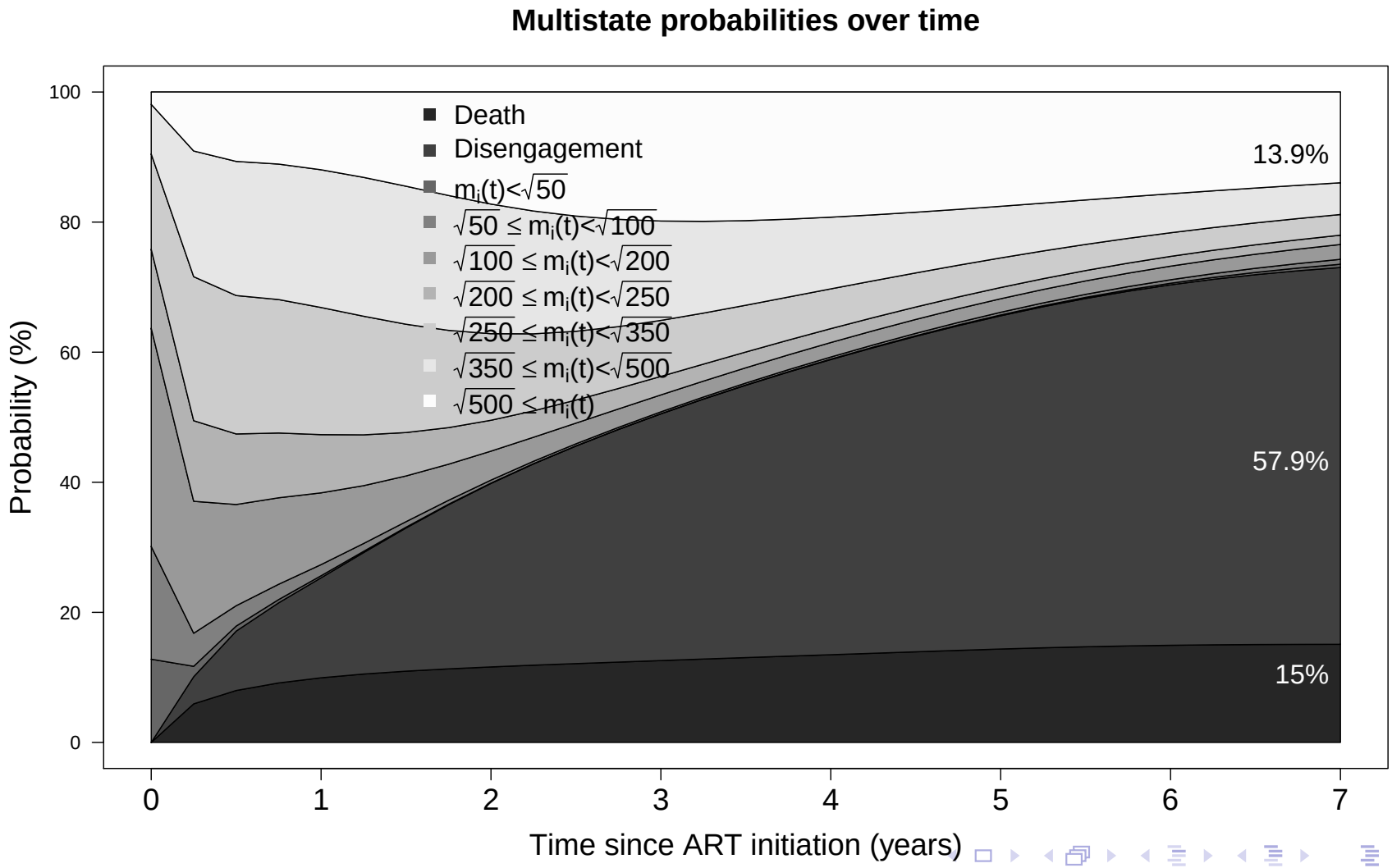
[1] "Median", "SD", "LB", and "UB" denote the posterior median, standard deviation, 2.5% and 97.5% quantiles, respectively. "sHR" denotes the subdistribution hazard ratio.

Results from the fitted SREM





Results from the fitted SREM



Conclusions - I

- We propose a flexible CIF-based joint modeling approach which could be a useful alternative to a cause-specific-hazard-based one.
- Our proposed model has been extended to accommodate failure cause misclassification through a double sampling approach.
- Based solely on the joint model, we also derive posterior samples for multistate probabilities defined jointly by marker and competing risk data.
  - Some transition probabilities between states are also derived.
  - Observed or latent marker states can be defined.

Conclusions - II

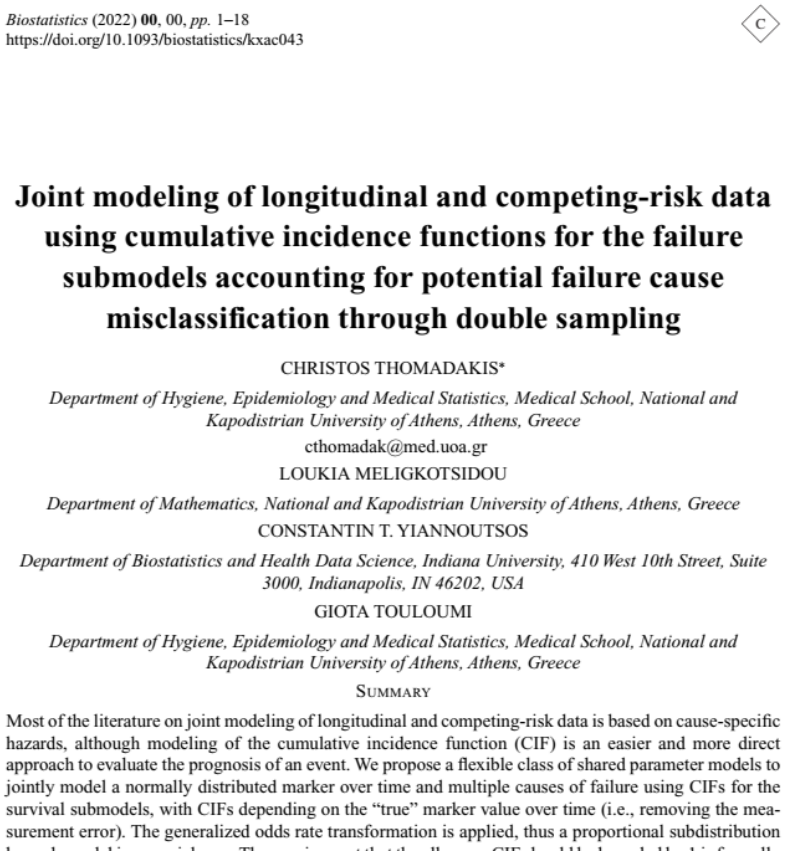
- The requirement that the all-cause CIF should be bounded by 1 is formally considered.
  - No random effects $\rightarrow$ it can be dealt with in the maximization process.
  - Not trivial in the presence of random effects.
- Our model assumes an upper bound of the survival time $\rightarrow$ zero likelihood when the constraint is violated $\Leftrightarrow$ introducing an indicator function in the likelihood.
- However, to estimate multistate probabilities, CIFs should be evaluable at any random effect value drawn from its prior.
- Thus, having an explicitly defined model for the CIFs accounting for the constraints, population-averaged quantities can be estimated directly.

Conclusions - III

- Simulations have shown that it works well with a 20% doubly sampled individuals.
- Population CIFs and latent marker states over time estimates were also derived and evaluated in the simulation study.
- **Extensions**
  - Dynamic failure probabilities (Rizopoulos 2012).
  - More flexible functional forms for the dependence on $m_i(t)$.
  - Include the marker slope $\frac{\partial m_i(t)}{\partial t}$ as well?

Publications

- The whole work has been published in *Biostatistics*.



Biostatistics (2022) **00**, 00, *pp.* 1–18
https://doi.org/10.1093/biostatistics/kxac043

Joint modeling of longitudinal and competing-risk data using cumulative incidence functions for the failure submodels accounting for potential failure cause misclassification through double sampling

CHRISTOS THOMADAKIS*
Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece
cthomadak@med.uoa.gr
LOUKIA MELIGKOTSIDOU
Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece
CONSTANTIN T. YIANNOUTSOS
Department of Biostatistics and Health Data Science, Indiana University, 410 West 10th Street, Suite 3000, Indianapolis, IN 46202, USA
GIOTA TOULOUMI
Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece

SUMMARY

Most of the literature on joint modeling of longitudinal and competing-risk data is based on cause-specific hazards, although modeling of the cumulative incidence function (CIF) is an easier and more direct approach to evaluate the prognosis of an event. We propose a flexible class of shared parameter models to jointly model a normally distributed marker over time and multiple causes of failure using CIFs for the survival submodels, with CIFs depending on the "true" marker value over time (i.e., removing the measurement error). The generalized odds rate transformation is applied, thus a proportional subdistribution

- Parts of this work were presented as oral presentations in the ISCB conference, Leuven (2019) and Lyon (2021) and in the EMR-IBS (2021) conference.

Acknowledgements

- The publication of the article in OA mode was financially supported by HEAL-Link.
- National Institutes of Health (NIH) grants: U01AI069911 and R21AI145662.

Bibliography

Bakoyannis, G., Yu, M. & Yiannoutsos, C. T. (2017), 'Semiparametric regression on cumulative incidence function with interval-censored competing risks data', *Statistics in Medicine* **36**(23), 3683–3707.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7350*

Bakoyannis, G., Zhang, Y. & Yiannoutsos, C. T. (2019), 'Nonparametric inference for markov processes with missing absorbing state', *Statistica Sinica* **29**(4), 2083–2104.

Daniel Paulino, C., Soares, P. & Neuhaus, J. (2003), 'Binomial regression with misclassification', *Biometrics* **59**(3), 670–675.

Deslandes, E. & Chevret, S. (2010), 'Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: application to icu data', *BMC Medical Research Methodology* **10**(1), 69.

Fine, J. & Gray, R. (1999), 'A proportional hazards model for the subdistribution of a competing risk', *Journal of the American Statistical Association* **94**(446), 496–509.

Gelfand, A. E., Smith, A. F. M. & Lee, T.-M. (1992), 'Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling', *Journal of the American Statistical Association* **87**(418), 523–532.
URL: *http://www.jstor.org/stable/2290286*

Hu, B., Li, L., Wang, X. & Greene, T. (2012), 'Nonparametric multistate representations of survival and longitudinal data with measurement error', *Statistics in Medicine* **31**(21), 2303–2317.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5369*

Jeong, J.-H. & Fine, J. P. (2006), 'Parametric regression on cumulative incidence function', *Biostatistics* **8**(2), 184–196.
URL: *https://doi.org/10.1093/biostatistics/kxj040*

Jeong, J.-H. & Fine, J. P. (2007), 'Parametric regression on cumulative incidence function', *Biostatistics* **8**(2), 184–196.

Mao, L. & Lin, D. Y. (2017), 'Efficient estimation of semiparametric transformation models for the cumulative incidence of competing risks', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(2), 573–587.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12177*

Molenberghs, G., Beunckens, C., Sotto, C. & Kenward, M. G. (2008), 'Every missingness not at random model has a missingness at random counterpart with equal fit', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **70**(2), 371–388.

Mozumder, S. I., Rutherford, M. & Lambert, P. (2018), 'Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach', *Statistics in Medicine* **37**(1), 82–97.