Promoting good practice in handling measurement error and misclassification using simulations: two case examples

Cécile Proust-Lima,

Maris Dussartre, Viviane Philipps, Paul Gustafson, Pamela Shaw,
Laurence Freedman, Veronika Deffner, Hendriek Boshuizen, Anne Thiébaut
for TG4 of the STRATOS initiative

INSERM U1219, Bordeaux Population Health Research Center, Bordeaux, France
Univ. Bordeaux, ISPED, Bordeaux, France
cecile.proust-lima@inserm.fr

19th International Biometric Conference
Atlanta, US - December, 2024

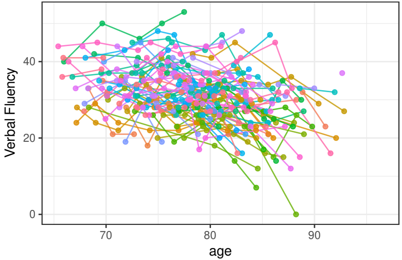
I have no current or past relationship with commercial entities, except for having taught shortcourses

Context

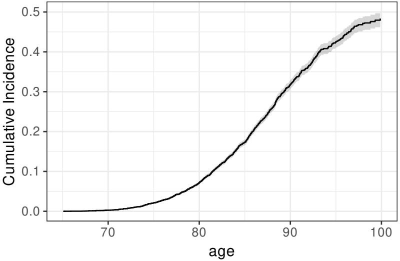
- Constant biostatistical developments:
  - to handle all the data imperfections
  - and correctly approach statistical inference
  - with associated software
- Why not largely used in the epidemiological community?
  - statistical solutions remain complicated
  - usually require advanced statistical skills
  - not enough communication:
    - ★ **what is the problem?**
    - ★ **what are the solutions?**
    - ★ **why should we care?**

Example 1: Time-varying covariates in survival analyses

- repeated measures of marker (e.g., blood biomarker, MRI features, PRO / QoL scales) or exposure (e.g., blood pressure, BMI)



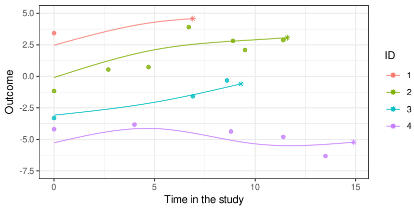
- time to health outcome (e.g., death, diagnosis, progression, dropout)



Target "Cox" model: $\lambda_i(t) = \lambda_0(t) \exp(X_i^*(t)\eta)$

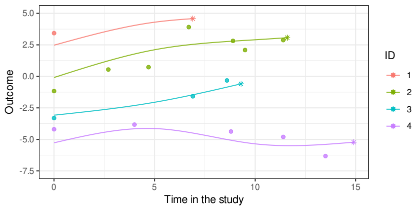
Example 1: Gap between observations and true exposure

- Exposure data = **measures of an underlying process**:
  - measured with error
  - measured at sparse and irregular times
  - observation stopped by the event occurence



Example 1: Gap between observations and true exposure

- Exposure data = **measures of an underlying process**:
  - measured with error
  - measured at sparse and irregular times
  - observation stopped by the event occurence



- Dedicated biostatistical model = **joint models**

★ Mixed model:

- underlying process of interest $X^*(t)$ at any time t

$$X_i^*(t) = \boldsymbol{W_i(t)}^\top \boldsymbol{\beta} + \boldsymbol{Z}_i(t)^\top \boldsymbol{b}_i \quad \text{with} \quad \boldsymbol{b}_i \sim \mathcal{N}(0, \boldsymbol{B})$$

- noisy observations X_{ij} at sparse times t_{ij} ($< T_i$)

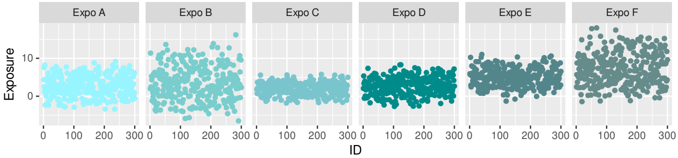
$$X_{ij} = X_i^*(t_{ij}) + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \underset{iid}{\sim} \mathcal{D}$$

★ Cox model:

$$\lambda_i(t) = \lambda_0(t) \exp(X_i^*(t)\eta)$$

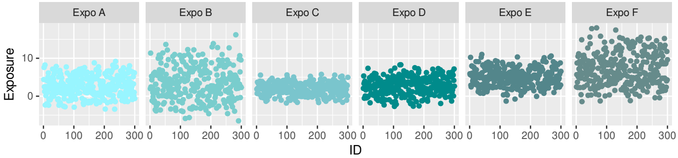
Example 2: Use of latent classes to summarize complex information

1. multi-dimensional exposures at baseline: e.g., cardiometabolic health (obesity, activity, glycemia, blood pressure, cholesterol)

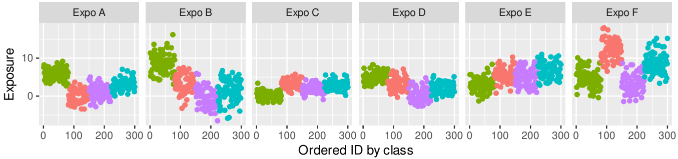


Example 2: Use of latent classes to summarize complex information

1. multi-dimensional exposures at baseline: e.g., cardiometabolic health (obesity, activity, glycemia, blood pressure, cholesterol)

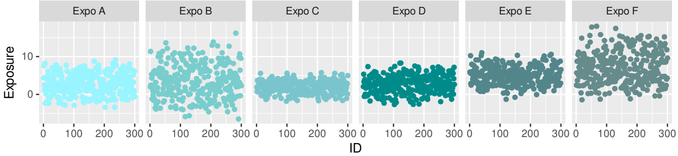


2. Estimate a latent class model and create a classification by assigning each subject to a fitted class:

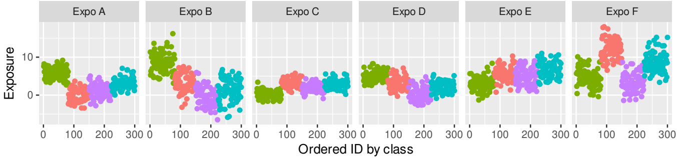


Example 2: Use of latent classes to summarize complex information

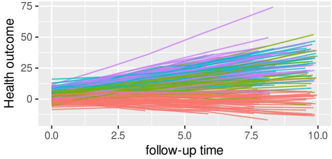
1. multi-dimensional exposures at baseline: e.g., cardiometabolic health (obesity, activity, glycemia, blood pressure, cholesterol)

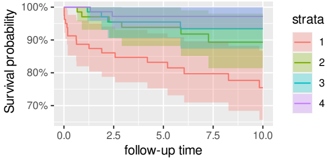


2. Estimate a latent class model and create a classification by assigning each subject to a fitted class:

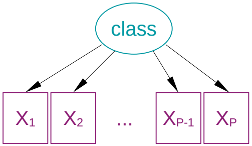


3. Assess the association of the latent classes with outcomes: e.g., cognitive trajectory, stroke in subsequent analyses





Example 2: Inherent uncertainty of estimated latent class structures

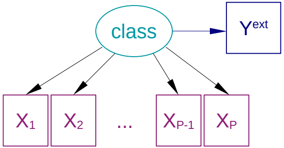


- Latent class model:
  - ⋆ Latent class: $\boldsymbol{c_i = g}$ with $\boldsymbol{\pi_{ig} = P(c_i = g)}$
  - ⋆ Distribution of the exposures $\boldsymbol{X_i} = (X_{i1}, ..., X_{iP})^\top$ in each latent class g:

$$\boldsymbol{X}_i | c_i = g \sim \mathscr{D}(\mu_{ig}, V_g)$$

⚠ with $\mathscr{D}$ adapted to the nature of the data

Example 2: Inherent uncertainty of estimated latent class structures



- Latent class model:
  - ⋆ Latent class: $c_i = g$ with $\pi_{ig} = P(c_i = g)$
  - ⋆ Distribution of the exposures $X_i = (X_{i1}, ..., X_{iP})^\top$ in each latent class g:

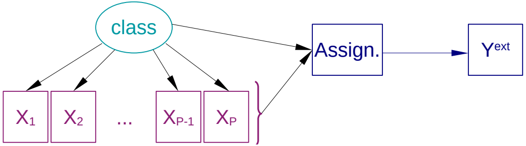
$$X_i | c_i = g \sim \mathcal{D}(\mu_{ig}, V_g)$$

  - ⋆ Distribution of the distal outcome Y_i^{ext} in each latent class g:

$$Y_i^{\text{ext}} | c_i = g \sim \mathcal{D}(\nu_{ig}, B_g)$$

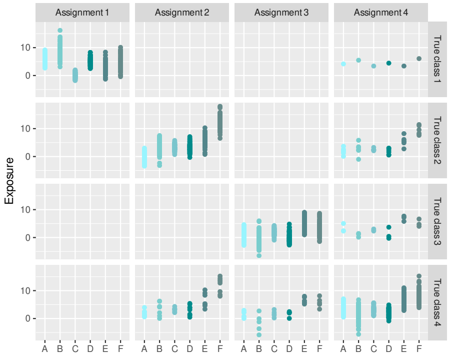
⚠ with $\mathcal{D}$ adapted to the nature of the data

Example 2: Inherent uncertainty of estimated latent class structures



- Latent class model:
  - ⋆ Latent class: $c_i = g$ with $\pi_{ig} = P(c_i = g)$
  - ⋆ Distribution of the exposures $X_i = (X_{i1}, \ldots, X_{iP})^\top$ in each latent class g:

$$X_i | c_i = g \sim \mathcal{D}(\mu_{ig}, V_g)$$



⚠ **Assignment ≠ Truth**

For $k \neq g$,

$$P(\text{assignment} = k \mid \text{true class} = g) \neq 0$$

Objective

As part of the STRATOS (STRengthening Analytical Thinking for Observational Studies) Topic Group "measurement error and misclassification":

Use simulation studies to illustrate

data challenges and benchmark pragmatic solutions

when some targetted truth is sought

⚠ In both cases:

- the whole joint model is the target, not the solution of interest
- alert about the problem
- focus on easily feasible approximation techniques from the literature

Simulation Strategy

Morris, White, Crowther (2019). Using simulation studies to evaluate statistical methods. Stat Med `https://doi.org/10.1002/sim.8086`

Aim = assess the correct inference of proxy methods in a target model

Data Generation = the whole joint model with varying scenarios

Estimands = regression parameter of interest

Methods = based on literature review

Performances = Bias, Coverage Rate of 95% CI, MSE

Example 1: time-varying covariate in survival model

Data Generation: joint model = linear mixed model for the exposure + Cox model for the event

Mixed model:

$$X_i^*(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \quad \forall t \in \mathbb{R}^+$$

with $\boldsymbol{F}(t)$ linear, quadratic

Noisy observations:

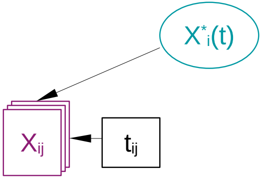
$$X_{ij} = X_i^*(t_{ij}) + \sigma \varepsilon_{ij}$$

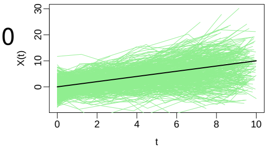
with $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ and $\sigma =$ 1,3

Visits j every y=1,2 up to 10

$$\begin{cases} t_{ij} = j + \tau_{ij} \\ \max(t_{ij}) < T_i \end{cases}$$

N=500 subjects





Example 1: time-varying covariate in survival model

Data Generation: joint model = linear mixed model for the exposure + Cox model for the event

Mixed model:

$$X_i^*(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \quad \forall t \in \mathbb{R}^+$$

with $\boldsymbol{F}(t)$ linear, quadratic

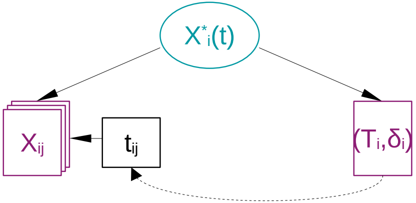
Noisy observations:

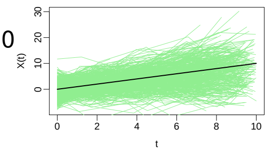
$$X_{ij} = X_i^*(t_{ij}) + \sigma\varepsilon_{ij}$$

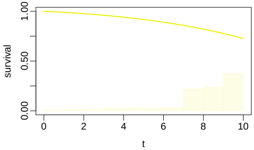
with $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ and σ= 1,3

Visits j every y=1,2 up to 10

$$\begin{cases} t_{ij} = j + \tau_{ij} \\ \max(t_{ij}) < T_i \end{cases}$$







N=500 subjects

Proportional Hazard Model

$$\lambda_i(t) = \lambda_0(t)\exp(X_i^*(t)\eta)$$

with 2 asso. η = 0.2, 0.4

with 3 Weibull $\lambda_0(t)$

Example 1: time-varying covariate in survival model

Data Generation: joint model = linear mixed model for the exposure + Cox model for the event

Mixed model:

$$X_i^*(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \quad \forall t \in \mathbb{R}^+$$

with $\boldsymbol{F}(t)$ linear, quadratic

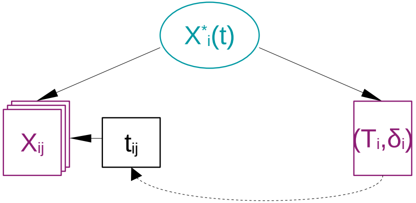
Noisy observations:

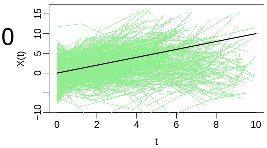
$$X_{ij} = X_i^*(t_{ij}) + \sigma\varepsilon_{ij}$$

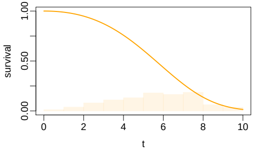
with $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ and σ= 1,3

Visits j every y=1,2 up to 10

$$\begin{cases} t_{ij} = j + \tau_{ij} \\ \max(t_{ij}) < T_i \end{cases}$$







N=500 subjects

Proportional Hazard Model

$$\lambda_i(t) = \lambda_0(t)\exp(X_i^*(t)\eta)$$

with 2 asso. η = 0.2, 0.4

with 3 Weibull $\lambda_0(t)$

Example 1: time-varying covariate in survival model

Data Generation: joint model = linear mixed model for the exposure + Cox model for the event

Mixed model:

$$X_i^*(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \quad \forall t \in \mathbb{R}^+$$

with $\boldsymbol{F}(t)$ linear, quadratic

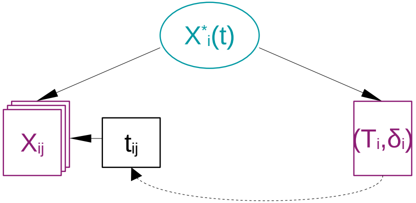
Noisy observations:

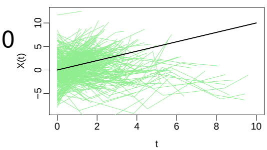
$$X_{ij} = X_i^*(t_{ij}) + \sigma\varepsilon_{ij}$$

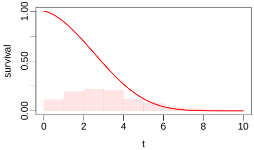
with $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ and σ= 1,3

Visits j every y=1,2 up to 10

$$\begin{cases} t_{ij} = j + \tau_{ij} \\ \max(t_{ij}) < T_i \end{cases}$$







N=500 subjects

Proportional Hazard Model

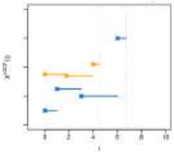
$$\lambda_i(t) = \lambda_0(t)\exp(X_i^*(t)\eta)$$

with 2 asso. η = 0.2, 0.4

with 3 Weibull $\lambda_0(t)$

Estimands = $\boldsymbol{\eta}$

Example 1: time-varying covariate in survival model



Methods= Approximation methods from the literature for:

$$\lambda_i(t) = \lambda_0(t)\exp(X_i^*(t)\eta)$$

- Last Observation Carried Forward (LOCF)
  - ★ use the last observation X_{ij} until a new one is available

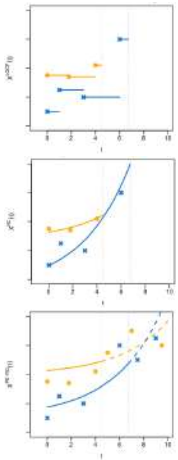
Example 1: time-varying covariate in survival model

Methods= Approximation methods from the literature for:

$$\lambda_i(t) = \lambda_0(t) \exp(X_i^*(t)\eta)$$

- Last Observation Carried Forward (LOCF)
  - ★ use the last observation X_{ij} until a new one is available
- Regression Calibration
  - ★ estimate a mixed model on available X
  - ★ compute the expected value $\hat{X}_i^*(t)$
  - ★ include $\hat{X}_i^*(t)$ in the survival model

⚠ 2 cases: truncation of Y at the event (RC) or access to posterior data (Post-Event RC / PE-RC)



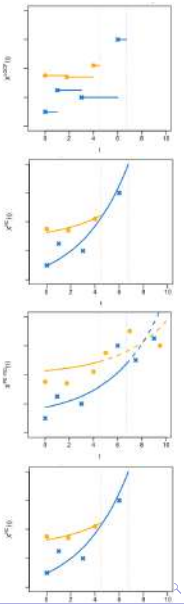
Example 1: time-varying covariate in survival model

Methods= Approximation methods from the literature for:

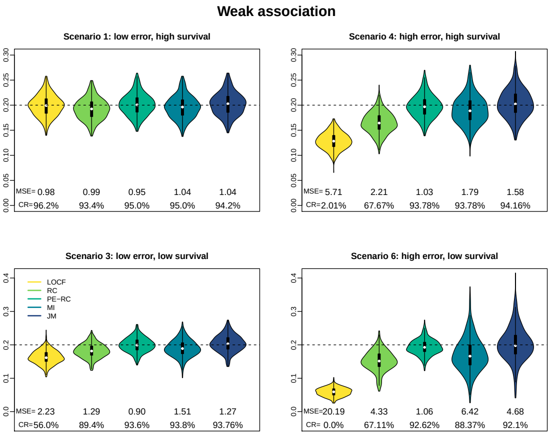
$$\lambda_i(t) = \lambda_0(t)\exp(X_i^*(t)\eta)$$

- Last Observation Carried Forward (LOCF)
  - ★ use the last observation X_{ij} until a new one is available
- Regression Calibration
  - ★ estimate a mixed model on available X
  - ★ compute the expected value $\hat{X}_i^*(t)$
  - ★ include $\hat{X}_i^*(t)$ in the survival model

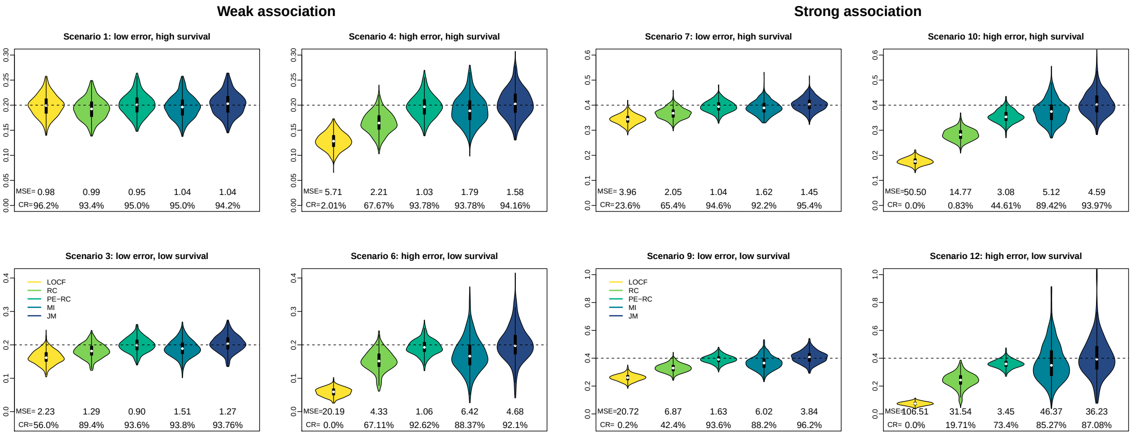
  ⚠ 2 cases: truncation of Y at the event (RC) or access to posterior data (Post-Event RC / PE-RC)
- Multiple Imputation (MI) (Moreno-Betancur, 2018)
  - ★ estimate a mixed model on available X using information on T
  - ★ draw values $X_i^{*(b)}(t)$
  - ★ include $X_i^{*(b)}(t)$ in the survival model



Example 1: some results (on 500 replicates)



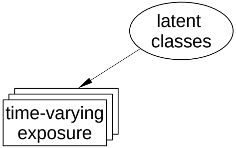
Example 1: some results (on 500 replicates)



Example 2: estimated latent class structure in subsequent analysis

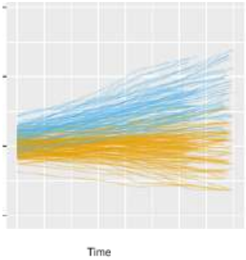
Data Generation = Simultaneous generation of the total information

2 classes
(50% / 50%)



2 sample sizes
(N=200, 1000)

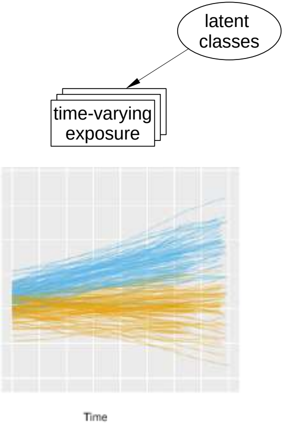
3 separation levels in the trajectories
(entropy=65%, 75%, 85%)



Example 2: estimated latent class structure in subsequent analysis

Data Generation = Simultaneous generation of the total information

2 classes
(50% / 50%)



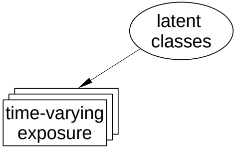
2 sample sizes
(N=200, 1000)

3 separation levels in the trajectories
(entropy=65%, 75%, 85%)

Example 2: estimated latent class structure in subsequent analysis

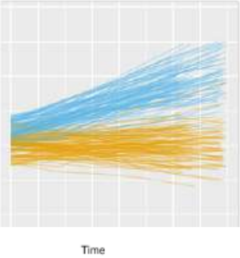
Data Generation = Simultaneous generation of the total information

2 classes
(50% / 50%)



2 sample sizes
(N=200, 1000)

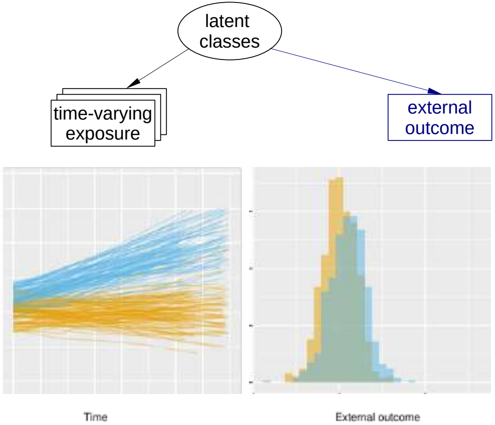
3 separation levels in the trajectories (entropy=65%, 75%, 85%)



Example 2: estimated latent class structure in subsequent analysis

Data Generation = Simultaneous generation of the total information

2 classes
(50% / 50%)



2 sample sizes
(N=200, 1000)

3 separation levels in the trajectories
(entropy=65%, 75%, 85%)

3 distances between classes:

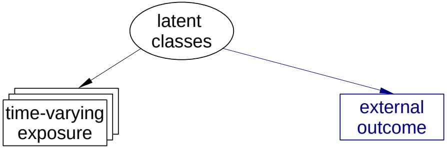
$$Y_i^{\text{ext}} = \beta_1 \mathbb{1}_{c_i=1} + \beta_2 \mathbb{1}_{c_i=2} + \sigma\epsilon_i$$

$\beta_2 - \beta_1$ = 0.5, 2 or 5

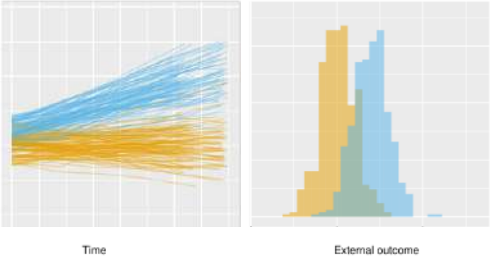
Example 2: estimated latent class structure in subsequent analysis

Data Generation = Simultaneous generation of the total information

2 classes
(50% / 50%)



2 sample sizes
(N=200, 1000)



3 separation levels in the trajectories
(entropy=65%, 75%, 85%)

3 distances between classes:

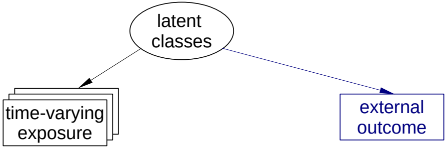
$$Y_i^{\text{ext}} = \beta_1 \mathbb{1}_{c_i=1} + \beta_2 \mathbb{1}_{c_i=2} + \sigma\epsilon_i$$

$\beta_2 - \beta_1$ = 0.5, 2 or 5

Example 2: estimated latent class structure in subsequent analysis

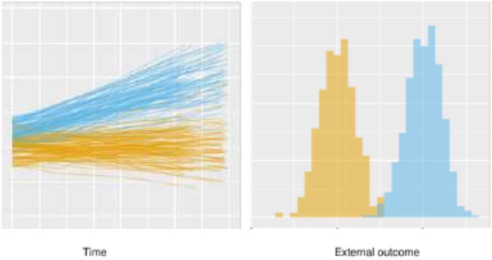
Data Generation = Simultaneous generation of the total information

2 classes
(50% / 50%)



2 sample sizes
(N=200, 1000)

3 separation levels in the trajectories
(entropy=65%, 75%, 85%)



3 distances between classes:

$$Y_i^{\text{ext}} = \beta_1 \mathbb{1}_{c_i=1} + \beta_2 \mathbb{1}_{c_i=2} + \sigma\epsilon_i$$

$\beta_2 - \beta_1$ = 0.5, 2 or 5

Estimands = β_2, β_1, σ

Example 2: estimated latent class structure in subsequent analysis

Methods = Approximation methods from the literature for:

$$Y_i^{\text{ext}}|_{c_i} = \beta_1 \mathbb{1}_{c_i=1} + \beta_2 \mathbb{1}_{c_i=2} + \sigma\epsilon_i$$

- The Naive modal method:
  Assignment $\hat{c}$ as covariate as if there was a perfect classification

Example 2: estimated latent class structure in subsequent analysis

Methods = Approximation methods from the literature for:

$$Y_i^{\text{ext}}|_{c_i} = \beta_1 \mathbb{1}_{c_i=1} + \beta_2 \mathbb{1}_{c_i=2} + \sigma\epsilon_i$$

- The Naive modal method:
  Assignment $\hat{c}$ as covariate as if there was a perfect classification

- The Naive proportional method:
  Assignment $\hat{c}$ as covariate weighted by the class-membership posterior probability $\mathbb{P}(c = g|X_i)$

- The Weighting correction method (Bolck 2004, Bakk 2013):
  Assignment $\hat{c}$ as covariate weighted by the probability of misclassification $\mathbb{P}(\hat{c} = k|c = g)$

Example 2: estimated latent class structure in subsequent analysis

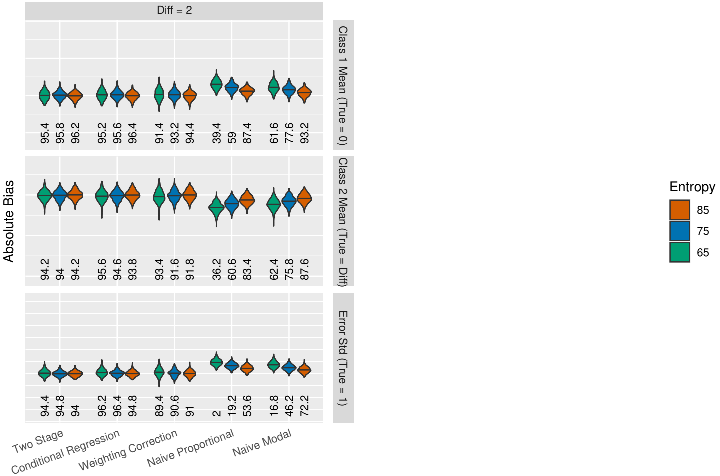
Methods = Approximation methods from the literature for:

$$Y_i^{\text{ext}}|_{c_i} = \beta_1 \mathbb{1}_{c_i=1} + \beta_2 \mathbb{1}_{c_i=2} + \sigma\epsilon_i$$

- The Naive modal method:
  Assignment $\hat{c}$ as covariate as if there was a perfect classification
- The Naive proportional method:
  Assignment $\hat{c}$ as covariate weighted by the class-membership posterior probability $\mathbb{P}(c = g|X_i)$
- The Weighting correction method (Bolck 2004, Bakk 2013):
  Assignment $\hat{c}$ as covariate weighted by the probability of misclassification $\mathbb{P}(\hat{c} = k|c = g)$
- The conditional regression on the true classes (Vermunt 2010, Bakk 2013):
  Regression rewritten as a latent class model according to our target classes
- The two-stage method (Xue et Bandeen-Roche 2002, Bakk et Kuha 2018, Proust-Lima 2023):
  Estimate the parameters of the regression for Y^{ext} using the whole joint likelihood $\mathscr{L}(X, Y^{\text{ext}})$ with parameters from X model fixed

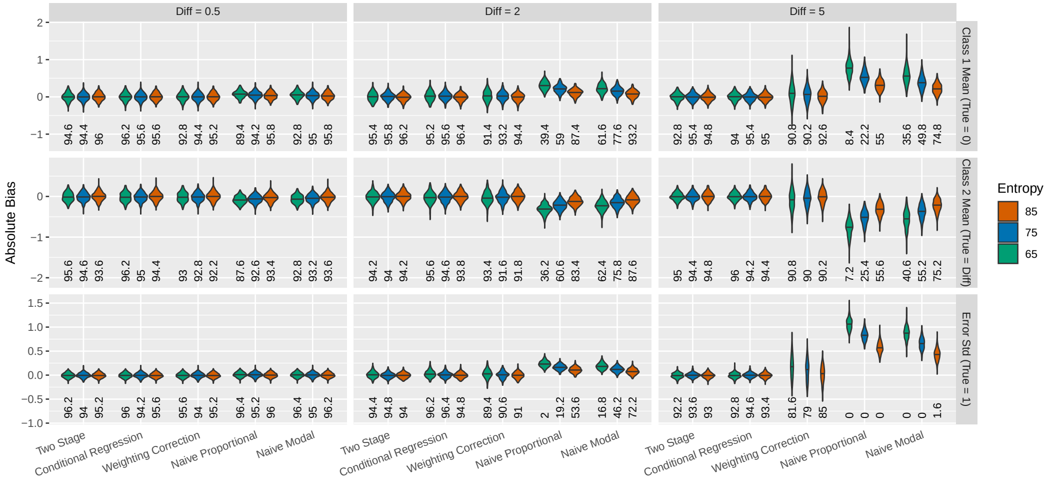
Example 2: some results of performances - N=200 (on 500 replicates)

- 3 parameters to examine: mean in each class + variance of the error



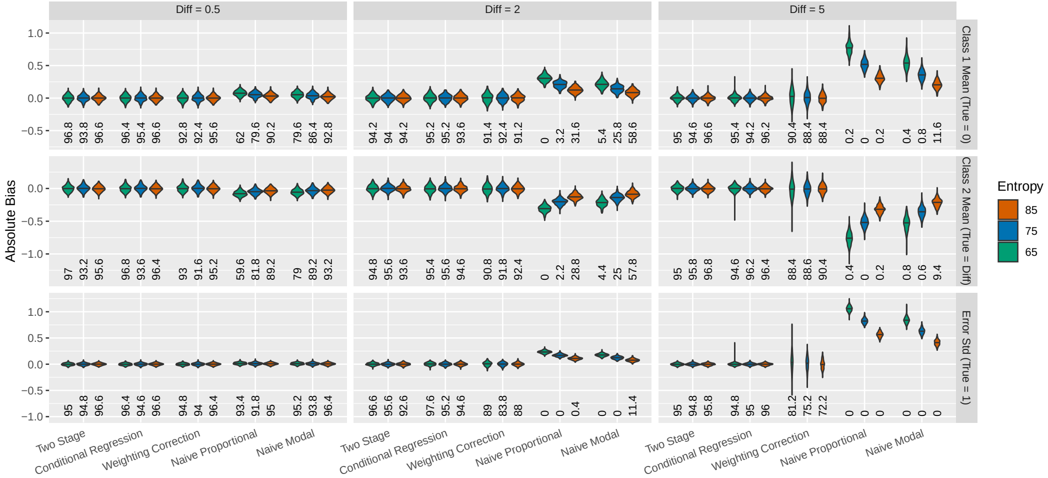
Example 2: some results of performances - N=200 (on 500 replicates)

- 3 parameters to examine: mean in each class + variance of the error



Example 2: some results of performances - N=1000 (on 500 replicates)

- 3 parameters to examine: mean in each class + variance of the error



Concluding remarks

Essential step for strengthening the statistical analysis in epidemiological studies

Great potential but such a difficult exercise!

- Most critical challenges
  - convince without being too technical
  - impossible to reach exhaustivity
  - lack of (user-friendly) implementations
  - sometimes/often, only sophisticated techniques work
- Simulations or Applications?
  - Complementary roles
  - Simulations are the perfect setting to raise awareness
  - Applications remain much more tangible - necessary to convince the reluctants ("OK, but should I really care?")
  - But too many other things happen - e.g., misspecification

Acknowledgements and references

Topic Group 4 "Measurement error and Classification"

STRATOS INITIATIVE

Project ID3M, Fondation vaincre Alzheimer

FONDATION VAINCRE ALZHEIMER

Univ. Bordeaux IdEx "Investments for the Future" (Research Network Public Health Data Science)

INVESTIR L'AVENIR

Réseau de Recherche Impulsion PHDS | Public Health Data Science Bordeaux Network

université de BORDEAUX

References:

Slides on my webpage:
`https://www.bordeaux-population-health.center/author/cecile.proust-lima/`

Example 1:

- Andersen, Liestøl (2003). Attenuation caused by infrequently updated covariates in survival analysis. Biostatistics 4(4):633-49
- Moreno-Betancur, Carlin et al (2018). Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). Biostatistics 19(4):479-96

Example 2:

- Bakk & Kuha (2021). Relating latent class membership to external variables: An overview. Br J Math Stat Psychol 74:340-62
- Proust-Lima, Saulnier et al (2023). Describing complex disease progression using joint latent class models for multivariate longitudinal markers and clinical endpoints. Stat Med 42:3996-4014