How to include time-varying exposures prone to measurement error in survival analyses?

TG4 subgroup:
Cécile Proust-Lima, Viviane Philipps, Veronika Deffner, Hendrieke Boshuizen, Laurence Freedman, Anne Thiébaut

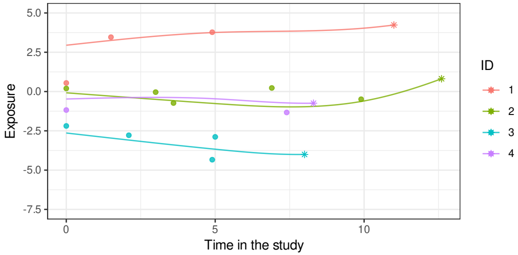
INSERM U1219, Bordeaux Population Health Research Center, Bordeaux, France
Univ. Bordeaux, ISPED, Bordeaux, France
cecile.proust-lima@inserm.fr

ISCB - STRATOS Mini-Symposium - August 31, 2023

Context

- Association between a time-varying exposure and a time to event:
  - BMI and incidence of breast cancer
  - Physical activity and incidence of Parkinson disease
  - Blood Pressure and cardiovascular event
  - ...

- Exposure data are **measures of an underlying continuous-time process**:
  - measured with error
  - measured at sparse and irregular times
  - stopped by the event occurrence



Central statistical issue

- Cox model with time-varying covariate dedicated to
  - continuously observed time-varying covariate
    (value known at each observed survival time (event/censored))
  - observed without error
  - covariate (and observation process) not impacted
    by the event occurrence:"external" exposure

Central statistical issue

- Cox model with time-varying covariate dedicated to
  - continuously observed time-varying covariate (value known at each observed survival time (event/censored)) ✗ sparse
  - observed without error ✗ error-prone
  - covariate (and observation process) not impacted by the event occurrence:"external" exposure ✗ internal/ truncation

✗ these assumptions rarely apply in health studies (Prentice 1982; Andersen 2002)

Statistical model envisaged

- Within the Cox modeling framework, the target model for time to event T_i is:

$$\lambda_i(t) = \lambda_0(t) \exp(X_i(t)\gamma) \qquad t > 0$$

  - $X_i(t)$ is the "true" exposure process

- Available data: exposure measurements $\tilde{X}_{ij}$ at sparse times t_{ij}
  - with generally truncation at the event time: $\max(t_{ij}) < T_i$
  - with random measurement error:

$$\tilde{X}_{ij} = X_i(t_{ij}) + \varepsilon_{ij} \quad \text{with} \quad \epsilon_{ij} \underset{iid}{\sim} \mathscr{D}$$

How to leverage sparse and error-prone observations of $X_i(t)$ to correctly estimate γ?

Solutions identified in the literature

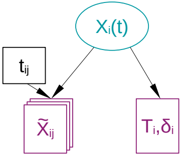
- Towards satisfying Cox model properties?
  - sparse: extrapolation/interpolation of values at all time points:
    - Last Value Carried Forward (LOCF)
    - predictions from a regression model
  - error-prone: regression model to separate observations from the underlying process
  - internal / truncation: account for the truncation induced by the event

Solutions identified in the literature

- Towards satisfying Cox model properties?
  - sparse: extrapolation/interpolation of values at all time points:
    - ★ Last Value Carried Forward (LOCF)
    - ★ predictions from a regression model
  - error-prone: regression model to separate observations from the underlying process
  - internal / truncation: account for the truncation induced by the event
- Methods identified in the literature

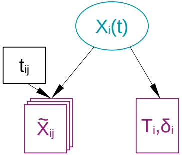
	LOCF	Regression	Multiple	Joint
Reference		Ye Biometrics 2008	Moreno-Betancur Biostat 2018	Wulfsohn Biometrics 1997
sparse	✓	✓	✓	✓
error-prone	✗	✓	✓	✓
internal / truncation	✗	✗	✓	✓

Simulation study: Comparison of methods



- Samples of 500 subjects ; 500 replications
- Generation process "true" model for subject i
  - True exposure process: $X_i(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \quad \forall t \in \mathbb{R}^+$ with $\boldsymbol{u_i} \sim \mathcal{N}(0, \boldsymbol{B})$

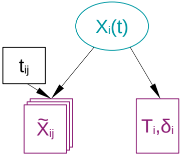
Simulation study: Comparison of methods



- Samples of 500 subjects ; 500 replications
- Generation process "true" model for subject i
  - True exposure process: $X_i(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \ \ \forall t \in \mathbb{R}^+$ with $\boldsymbol{u_i} \sim \mathcal{N}(0, \boldsymbol{B})$
  - Visit process j every y years (y=1,2) until administrative censoring at 10 years:

$$t_{ij} = j + \tau_{ij} \quad \text{with} \quad \tau_{ij} \sim \mathcal{U}(-1, 1)$$

Simulation study: Comparison of methods



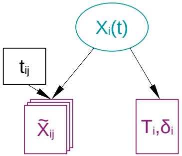
- Samples of 500 subjects ; 500 replications
- Generation process "true" model for subject i
  - True exposure process: $X_i(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \ \ \forall t \in \mathbb{R}^+$ with $\boldsymbol{u_i} \sim \mathcal{N}(0, \boldsymbol{B})$
  - Visit process j every y years (y=1,2) until administrative censoring at 10 years:

$$t_{ij} = j + \tau_{ij} \quad \text{with} \quad \tau_{ij} \sim \mathcal{U}(-1, 1)$$

  - Repeated exposure observations at visit times:

$$\tilde{X}_{ij} = X(t_{ij}) + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Simulation study: Comparison of methods



- Samples of 500 subjects ; 500 replications
- Generation process "true" model for subject i
  - True exposure process: $X_i(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \;\; \forall t \in \mathbb{R}^+$ with $\boldsymbol{u_i} \sim \mathcal{N}(0, \boldsymbol{B})$
  - Visit process j every y years (y=1,2) until administrative censoring at 10 years:

$$t_{ij} = j + \tau_{ij} \quad \text{with} \quad \tau_{ij} \sim \mathcal{U}(-1,1)$$

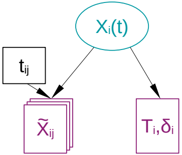
  - Repeated exposure observations at visit times:

$$\tilde{X}_{ij} = X(t_{ij}) + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

  - Survival outcome (T_i, δ_i) with hazard

$$\lambda_i(t) = \lambda_0(t) \exp(X_i(t)\gamma) \quad \text{with a Weibull } \lambda_0(t)$$

Simulation study: Comparison of methods



- Samples of 500 subjects ; 500 replications
- Generation process "true" model for subject i
  - True exposure process: $X_i(t) = \boldsymbol{F}(t)(\boldsymbol{\beta} + \boldsymbol{u_i}) \;\; \forall t \in \mathbb{R}^+$ with $\boldsymbol{u_i} \sim \mathcal{N}(0, \boldsymbol{B})$
  - Visit process j every y years (y=1,2) until administrative censoring at 10 years:

$$t_{ij} = j + \tau_{ij} \quad \text{with} \quad \tau_{ij} \sim \mathcal{U}(-1,1)$$

  - Repeated exposure observations at visit times:

$$\tilde{X}_{ij} = X(t_{ij}) + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

  - Survival outcome (T_i, δ_i) with hazard

$$\lambda_i(t) = \lambda_0(t) \exp(X_i(t)\gamma) \quad \text{with a Weibull } \lambda_0(t)$$

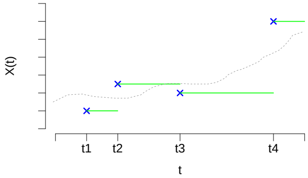
+ Eventually, truncation of $\tilde{X}$ at the event time
  (are indicated in red the parameters that changed according to the scenarios)

Simulations: Estimation models/techniques

- Naive LOCF (Last Observation Carried Forward) Cox model:

$$\lambda_i(t) = \lambda_0(t)\exp(\tilde{X}_i(t)\gamma)$$

with $\tilde{X}_i(t) = \tilde{X}_i(t_{ij})$ with $j = \max(k; t_{ik} \leq t)$

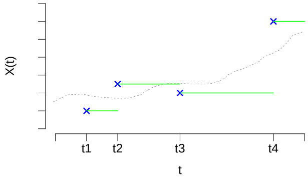


Simulations: Estimation models/techniques

- Naive LOCF (Last Observation Carried Forward) Cox model:

$$\lambda_i(t) = \lambda_0(t)\exp(\tilde{X}_i(t)\gamma)$$

with $\tilde{X}_i(t) = \tilde{X}_i(t_{ij})$ with $j = \max(k; t_{ik} \leq t)$

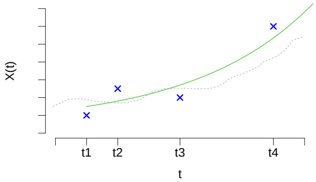


- Regression Calibration:

$$\lambda_i(t) = \lambda_0(t)\exp(\hat{X}_i(t)\gamma)$$

with $\hat{X}_i(t)$ predicted from Linear Mixed Model: $\tilde{X}_{ij} = \underbrace{\mathbf{F}(\mathbf{t})(\boldsymbol{\beta} + \boldsymbol{u_i})}_{X_i(t_{ij})} + \varepsilon_{ij}$

and $\hat{X}_i(t) = \mathbb{E}(X_i(t)|(X_{ij})_{j=1,...,n_i})$



  - For the simulations, two settings:
    - ★ classical RC: estimation of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ based on $\tilde{X}_{ij} < T_i$
    - ★ external RC: estimation of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ based on $\tilde{X}_{ij}$ even after T_i

Simulations: Estimation models/techniques (cont'd)

- Multiple Imputation:

$$\lambda_i(t) = \lambda_0(t)\exp(\hat{X}_i^m(t)\gamma)$$

with a modified Linear Mixed Model: $\tilde{X}_{ij} = \underbrace{\mathbf{F}(\mathbf{t})(\boldsymbol{\beta} + \boldsymbol{u_i}) + \beta_D D_{ij} + \beta_\Lambda \Lambda(T_i)}_{X_i(t_{ij})} + \varepsilon_{ij}$

and $\hat{X}_i^m(t)$ draws $(m = 1, ..., M)$ from the posterior distribution of $\mathbb{E}(X_i(t)|(X_{ij})_{j=1,...,n_i})$

Simulations: Estimation models/techniques (cont'd)

- Multiple Imputation:

$$\lambda_i(t) = \lambda_0(t)\exp(\hat{X}_i^m(t)\gamma)$$

with a modified Linear Mixed Model: $\tilde{X}_{ij} = \underbrace{\mathbf{F}(\mathbf{t})(\boldsymbol{\beta}+\boldsymbol{u_i}) + \beta_D D_{ij} + \beta_\Lambda \Lambda(T_i)}_{X_i(t_{ij})} + \varepsilon_{ij}$

and $\hat{X}_i^m(t)$ draws $(m = 1,...,M)$ from the posterior distribution of $\mathbb{E}(X_i(t)|(X_{ij})_{j=1,...,n_i})$

- Joint model of both processes:

$$\lambda_i(t) = \lambda_0(t)\exp(X_i(t)\gamma) \qquad \& \qquad \tilde{X}_{ij} = \underbrace{\mathbf{F}(\mathbf{t})(\boldsymbol{\beta}+\boldsymbol{u_i})}_{X_i(t_{ij})} + \varepsilon_{ij}$$

Variance estimation in the two-stage approaches

⚠ For RC and MI methods:
Parametric bootstrap with the Rubin's rule to account for first-stage variability

Variance estimation in the two-stage approaches

⚠ For RC and MI methods:
Parametric bootstrap with the Rubin's rule to account for first-stage variability

- parameters in the LMM noted $\theta = (\beta, \text{vec}(B))$

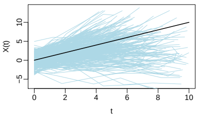
- Internal, external regression calibration :
  - for b=1,...,500 draws: $\theta^b \sim \mathcal{N}(\hat{\theta}, \hat{V}(\hat{\theta}))$
  - BLUP $\hat{u}_i{}^b$ computed in θ^b
  - $\hat{X}^b(t)$ computed from θ^b and $\hat{u}_i{}^b$
  - Cox model estimated using $\hat{X}^b(t)$
  - Rubin's rule on $\hat{\gamma}^b$

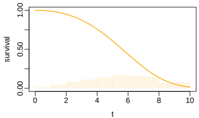
- Multiple Imputation:
  - for b=1,..., 500 draws $\theta^b \sim \mathcal{N}(\hat{\theta}, \hat{V}(\hat{\theta}))$
  - draw of $\hat{u}_i{}^b \sim \mathcal{N}(\hat{u}_i(\theta^b), \hat{V}(\hat{u}_i(\theta^b)))$
  - $\hat{X}^b(t)$ computed from θ^b and $\hat{u}_i{}^b$
  - Cox model estimated using $\hat{X}^b(t)$
  - Rubin's rule on $\hat{\gamma}^b$

⚠ in Moreno-Betancur (2018): draws for fixed effects only, not for variance parameters

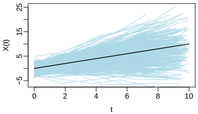
Linear, weak asso, small measurement error

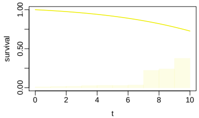
Medium Survival (417 events, 3.4 measures /subject)



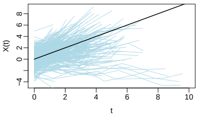


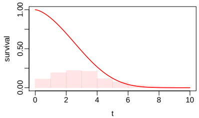
Higher survival (132 events, 4.8 measures /subject)

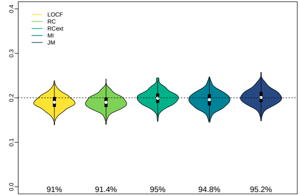


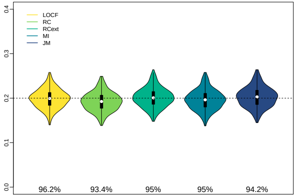


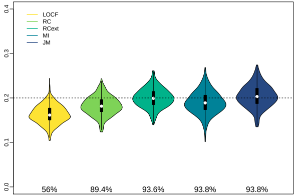
Lower survival (489 events, 2.1 measures /subject)





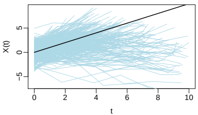


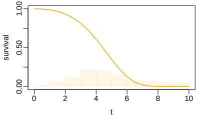




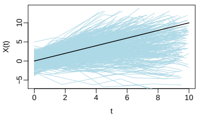
Linear, Strong asso, small measurement error

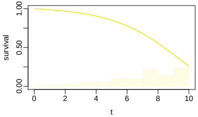
medium survival: (446 events, 2.9 measures /subject)



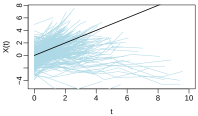


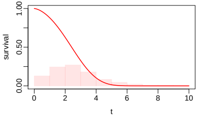
Higher survival: (277 events, 4.2 measures /subject)

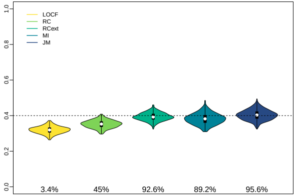


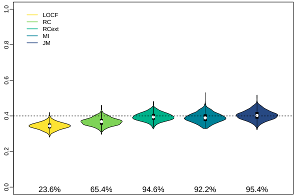


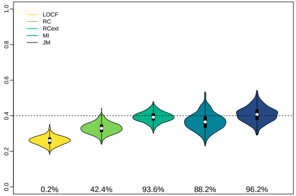
lower survival: (487 events, 1.9 measures /subject)





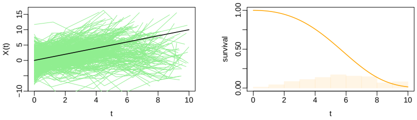




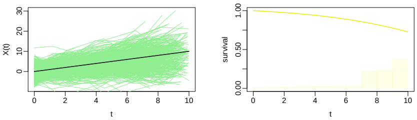


Linear, weak asso, large measurement error

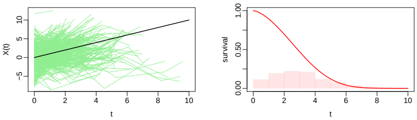
medium survival: 389 events, 3.5 measures /subject:

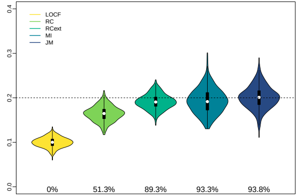


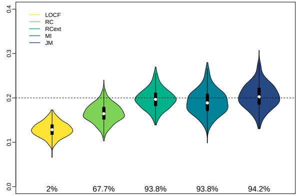
Higher Survival: 133 events, 4.9 measures/subject:

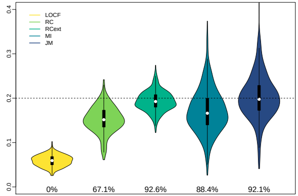


Lower Survival: 489 events, 2.1 measures /subject:



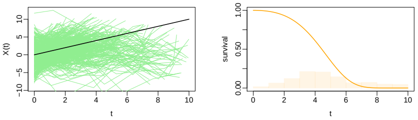




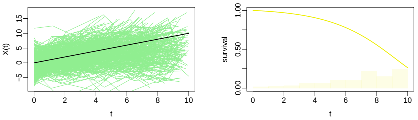


Linear, Strong asso, larger measurement error

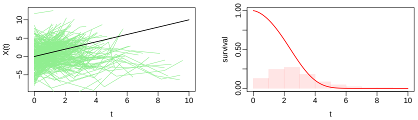
medium survival:: 407 events, 3.1 measures /subject:

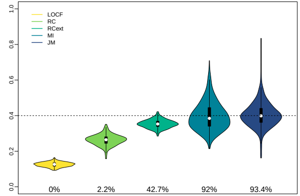


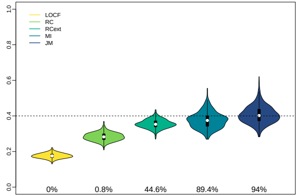
Higher Survival: 277 events, 4.2 measures / subject:



Lower Survival: 487 events, 1.9 measures /subject:







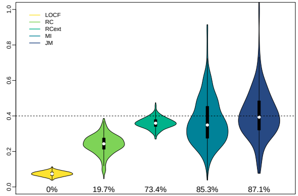
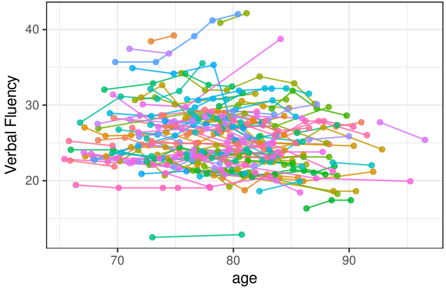


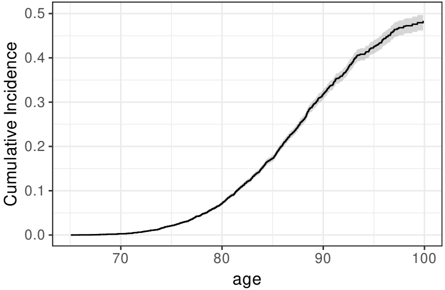
Illustration in dementia Research

- Association of time-dependent covariates with the instantaneous risk of dementia
  - Population-based 3C study with 17 years of follow-up, visits every 2-3 years, N=8193
  - Adjustment for city and gender
  - Trajectory over age approximated with natural cubic splines

Adiposity - Body Mass Index



Diagnosis of dementia



Log Hazard Ratios for Adiposity

Adopisity

BMI (in kg / m^2)

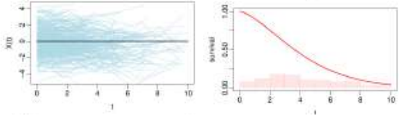
Method	log HR*	SE	p
LOCF	-0.0160	0.0077	0.0372
RC	-0.0138	0.0080	0.0830
MI	-0.0159	0.0081	0.0502
JM	-0.0142	0.0081	0.0774

* adjusted for gender, center

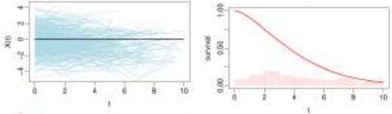
Back to simulations:

Constant trajectory, lower survival

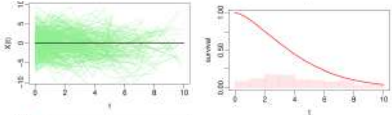
Smaller association, small error (455 events, 2.6 meas/subj)

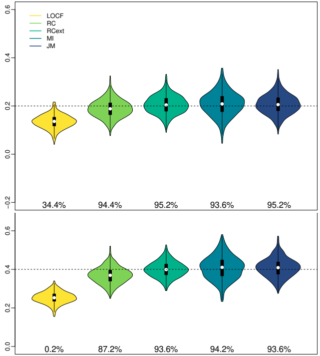


Larger association, small error (441 events, 2.6 meas/subj)



Smaller association, large error (455 events, 2.6 meas/subj)





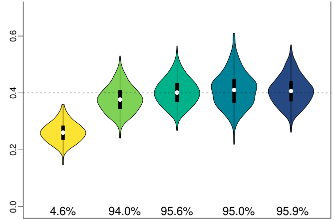
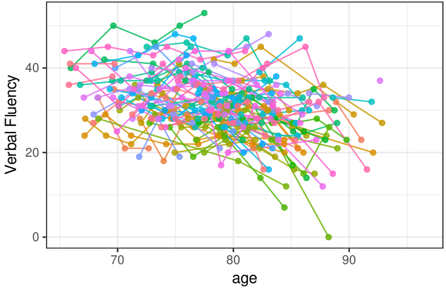


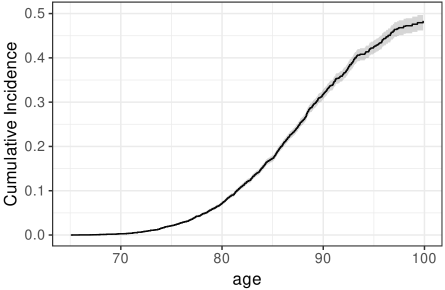
Illustration in dementia Research

- Association of time-dependent covariates with the instantaneous risk of dementia
  - Population-based 3C study with 17 years of follow-up, visits every 2-3 years, N=8193
  - Adjustment for city, gender and education
  - Trajectory over age approximated with natural cubic splines

Verbal Fluency - Isaacs Set Test (IST) Score



Diagnosis of dementia



Log Hazard Ratios for Adiposity and Verbal Fluency

Adopisity

BMI (in kg / m^2)

Method	log HR*	SE	p
LOCF	-0.0160	0.0077	0.0372
RC	-0.0138	0.0080	0.0830
MI	-0.0159	0.0081	0.0502
JM	-0.0142	0.0081	0.0774

* adjusted for gender, center

Verbal Fluency

IST sumscore in points (score from 0 to 40)

Method	log HR*	SE	p
LOCF	-0.125	0.005	<0.0001
RC	-0.222	0.007	<0.0001
MI	-0.199	0.009	<0.0001
JM	-0.255	0.008	<0.0001

* adjusted for gender, education, center

Conclusions

- Take home message:
  - LOCF strongly biased
  - Approximation with Two-stage methods valid if they account for early truncation by the event:
    - ★ using data available after the event if external (Regression Calibration)
    - ★ incorporating information on the event (Multiple Imputation)
  - JM works very well (expected as the generation model)

⚠ Results obtained under correct specification!

  - be careful with the functional form (nonlinear effect, lag, other features, ...)
  - be careful with the modelled trajectory

- Technical remarks:
  - Variance estimation with RC and MI using Rubin's rule
  - Same results with 10% MCAR data, different measure frequencies, nonlinear trajectory
  - Same results expected with other functional forms in the survival model

Acknowledgements and references

Topic Group 4 "Measurement error and Classification"

STRATOS INITIATIVE

Investigators of the 3C study

References:

Issue of time-varying covariates in Cox models:

Andersen PK, Liestøl K. Attenuation caused by infrequently updated covariates in survival analysis. Biostatistics. 1 oct 2003;4(4):633-49.

Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika. 1 août 1982;69(2):331-42.

Regression Calibration:

Ye W, Lin X, Taylor JMG. Semiparametric Modeling of Longitudinal Measurements and Time-to-Event Data-A Two-Stage Regression Calibration Approach. Biometrics. 2008;64(4):1238-46.

+ Albert PS, Shih JH. On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. Biometrics. sept 2010;983-91.

Multiple Imputation:

Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas SK, Peeters A, Wolfe R. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). Biostatistics. 1 oct 2018;19(4):479-96.

Joint Models:

Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman & Hall/CRC Biostatistics Series. 2012.