Modélisation conjointe de données longitudinales et de temps d'événements sous R

Joint modeling of longitudinal and time-to-event data in R

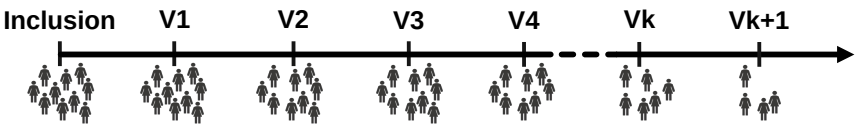
Cécile Proust-Lima
joint works with Viviane Philipps, Tiphaine Saulnier, Anthony Devaux, Robin Genuer

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

10ème rencontres R - Vannes - June 12, 2024

Epidemiological studies

- Overal cohort design:

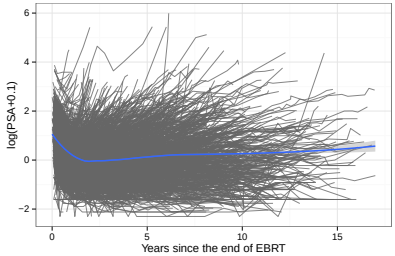


- Target population:
  - Whole population in a certain window of age
    - ★ 3-City Study: elderly - ELFE: young children - CONSTANCES: adults
  - Population with a certain diagnosis (e.g., cancer, Multiple System Atrophy)
    - ★ Clinical prospective cohort: monitoring the population for prognosis
    - ★ Clinical trial: testing an intervention in randomized groups

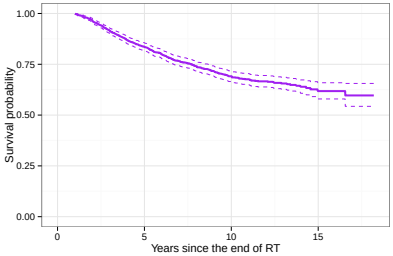
- Available data
  - at baseline (exposures, confounders, participant characteristics)
  - over follow-up (exposures, health indicators, events)

Progression of health phenomena studied through

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



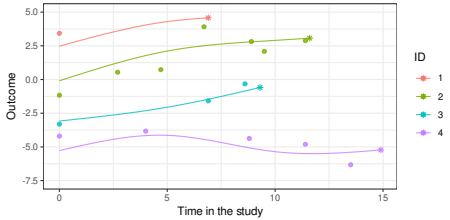
- time to health outcome (e.g., death, diagnosis, progression, dropout)



provide inter-related information that need to be analyzed together (jointly)

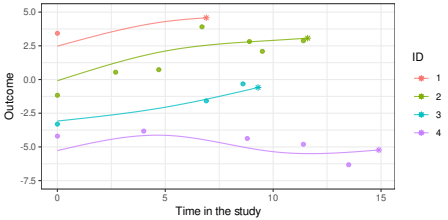
The endogenous nature of time-varying variables

- Marker/Exposure data are **measures of an underlying process**:
  - measured with error
  - measured at sparse and irregular times
  - influenced by the event occurence: endogenous / internal



The endogenous nature of time-varying variables

- Marker/Exposure data are **measures of an underlying process**:
  - measured with error
  - measured at sparse and irregular times
  - influenced by the event occurence: endogenous / internal



- Dedicated biostatistical model = **mixed models / random-effect models**

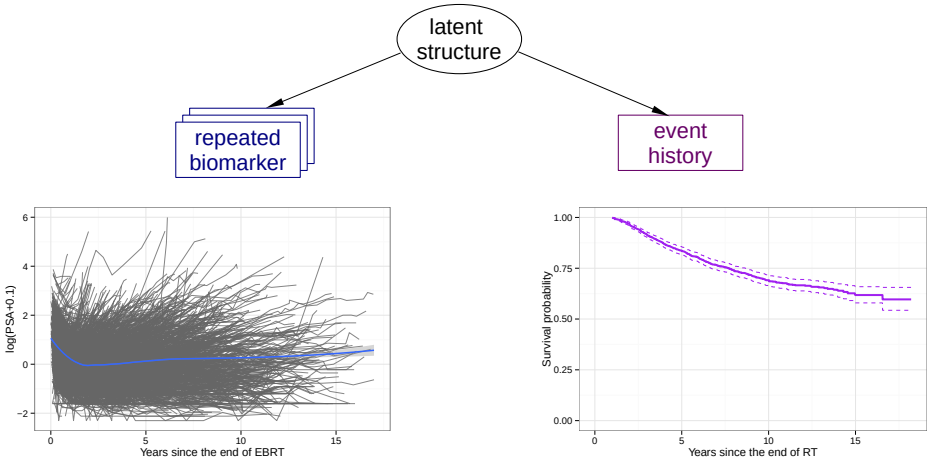
- Underlying process of interest $Y^*(t)$ defined at any time $t \in \mathbb{R}$

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

- Observations Y_{ij} at sparse times t_{ij}
  - ★ with generally truncation at the event time: $\max(t_{ij}) < T_i$
  - ★ with random measurement error: $Y_{ij} = Y_i^*(t_{ij}) + \varepsilon_{ij}$ with $\varepsilon_{ij} \underset{iid}{\sim} \mathcal{D}$

- Estimation in R:
  lme (nlme),
  lmer (lme4),
  hlme (lcmm),
  ...

Joint modelling principle



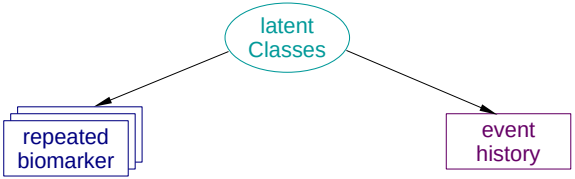
Simultaneous modelling of correlated longitudinal and survival data

(Classical) Research Questions addressed by joint models

- quantify the association of a endogenous marker with the risk of event
- predict the risk of clinical endpoint using the biomarker information
  - individual dynamic prediction and screening optimization
- describe the trajectory of the biomarker stopped by the clinical progression
  - and evaluate its determinants
- explore/understand the association between the two processes
  - variability / heterogeneity in the disease progression

Joint latent class models (JLCM) (Proust-Lima, 2014)

- Latent class c_i: $P(c_i = g) = \pi_{ig} = \frac{e^{\xi_{0g} + X_{Ci}{}^\top \xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l} + X_{Ci}{}^\top \xi_{1l}}}$ (with $\xi_{0G} = 0$ & $\xi_{1G} = \mathbf{0}$)



- Class-specific linear mixed model for the biomarker trajectory:

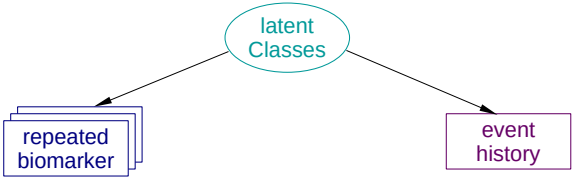
$$\begin{aligned} Y_{ij}|_{c_i=g} &= Y_{ig}^*(t_{ij}) + \epsilon_{ij} \\ &= Z_i(t_{ij})^T b_i|_{c_i=g} + X_{Li}(t_{ij})^\top \beta_g + \epsilon_{ij} \\ b_i|_{c_i=g} &\sim \mathcal{N}(\mu_g, B_g),\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2) \end{aligned}$$

- proportional hazard model for the event:

$$\lambda_i(t \mid c_i = g) = \lambda_{0g}(t) \exp(X_{Ti}(t)\delta_g)$$

Joint latent class models (JLCM) (Proust-Lima, 2014)

- Latent class c_i: $P(c_i = g) = \pi_{ig} = \dfrac{e^{\xi_{0g} + X_{Ci}{}^\top \xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l} + X_{Ci}{}^\top \xi_{1l}}}$ (with $\xi_{0G} = 0$ & $\xi_{1G} = \mathbf{0}$)



- Class-specific linear mixed model for the biomarker trajectory:

$$\begin{aligned} Y_{ij}\,|_{c_i=g} &= Y^*_{ig}(t_{ij}) + \epsilon_{ij} \\ &= Z_i(t_{ij})^T b_i|_{c_i=g} + X_{Li}(t_{ij})^\top \beta_g + \epsilon_{ij} \\ b_i|_{c_i=g} &\sim \mathcal{N}\left(\mu_g, B_g\right),\ \epsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2_\epsilon\right) \end{aligned}$$

- proportional hazard model for the event:

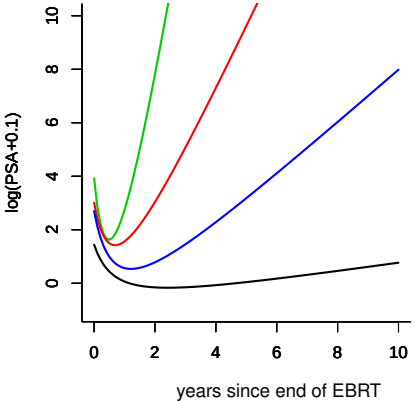
$$\lambda_i(t \mid c_i = g) = \lambda_{0g}(t) \exp(X_{Ti}(t)\delta_g)$$

→ describes the processes as made of homogenous subgroups

→ descriptive approach appropriate for *a priori* heterogenous populations

JLCM Illustration in Prostate Cancer (Proust-Lima, SMMR 2014)

- Four patterns of PSA trajectory and risk of any clinical recurrence
  - N=459 men from the University of Michigan Hospital Cohort
  - after a radiation therapy (EBRT)



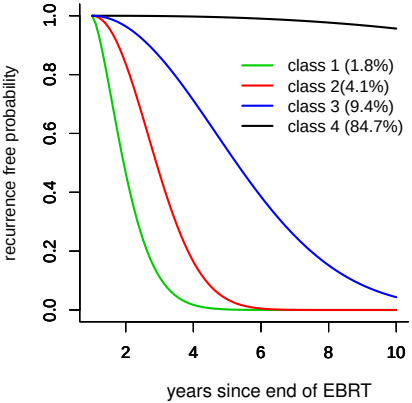
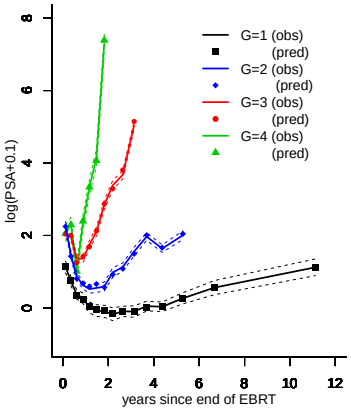
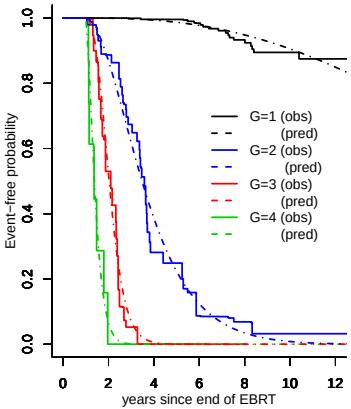


Illustration in Prostate Cancer (Proust-Lima, SMMR 2014)

- Very close to the observations:
  - high discrimination (mean probability of latent class membership > 92%)
  - excellent fit to the data compared to other joint models





Estimation in lcmm R package (Proust-Lima, JSS 2017)

- Maximum Likelihood Estimates

$$\mathscr{L}_i(\boldsymbol{\theta}) = \sum_{g=1}^{G} f(Y_i \mid c_i = g; \boldsymbol{\theta}) \;\; f(T_i \mid c_i = g; \boldsymbol{\theta}) \;\; P(c_i = g; \boldsymbol{\theta})$$

- Optimization algorithm: Marquardt-Levenberg Algorithm with marqLevAlg R package (Philipps R Journal 2022)
  - Newton-like optimization
  - Strict convergence criteria (parameter stability, likelihood stability, first and second derivatives)
  - Parallel numerical computations of the derivatives

- Management of local maxima
  - Grid search = B estimations from random initial values

- Variance-covariance matrix given by the inverse of the Hessian matrix

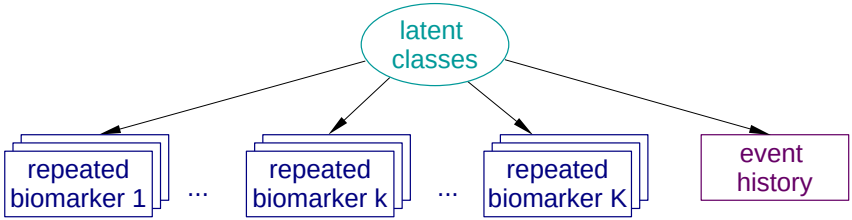
lcmm in practice: jlcmm or Jointlcmm function

```
 # G=1
m1 <- jlcmm( fixed = logPSA ~ I((time + 1)^(-1.5)) + time,
             random =~ I((time + 1) ^(-1.5)) + time, subject = "ID",
             survival = Surv(tsurv,event) ~ tstage2 + tstage34,
             hazard = "splines",
             data = cohort)


 # G=4
m4 <-
     jlcmm( fixed = logPSA ~ I((time + 1)^(-1.5)) + time,
             random =~ I((time + 1) ^(-1.5)) + time, subject="ID",
             mixture =~ I((time + 1)^(-1.5)) + time,
             survival = Surv(tsurv,event) ~ tstage2 + tstage34,
             hazard = "splines", hazardtype = "PH",
             ng = 4, data = cohort, B = random(m1))
```

lcmm in practice: jlcmm or Jointlcmm function

```
 # G=1
m1 <- jlcmm( fixed = logPSA ~ I((time + 1)^(-1.5)) + time,
             random =~ I((time + 1) ^(-1.5)) + time, subject = "ID",
             survival = Surv(tsurv,event) ~ tstage2 + tstage34,
             hazard = "splines",
             data = cohort)


 # G=4
m4 <- gridsearch(
      jlcmm( fixed = logPSA ~ I((time + 1)^(-1.5)) + time,
             random =~ I((time + 1) ^(-1.5)) + time, subject="ID",
             mixture =~ I((time + 1)^(-1.5)) + time,
             survival = Surv(tsurv,event) ~ tstage2 + tstage34,
             hazard = "splines", hazardtype = "PH",
             ng = 4, data = cohort,
             rep = 100, maxiter = 30, minit = m1, cl = 10)
```

Extensions to more complex data structure: longitudinal / competing causes

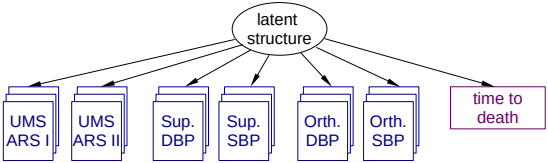


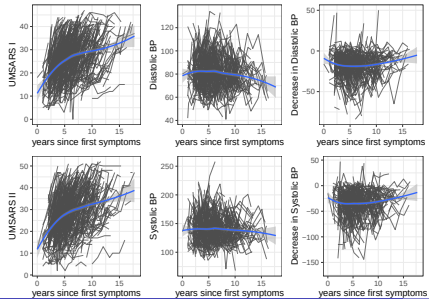
Class-and-marker-specific mixed model
(Proust-Lima, Stat Med 2023)

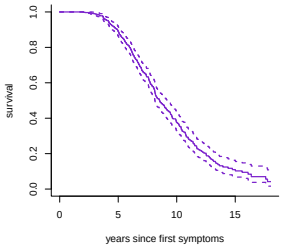
cause-and-class proportional hazard model
(Proust-Lima, Stat Med 2016)

- estimation in R: mpjlcmm function in package lcmm

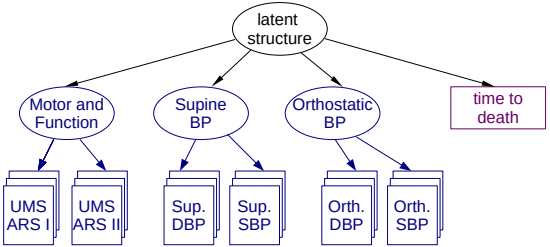
Example in Multiple System Atrophy (MSA)

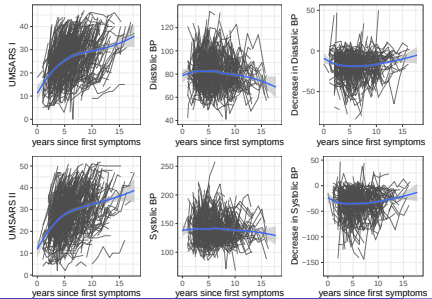


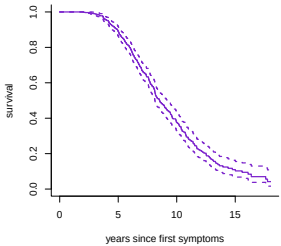




Example in Multiple System Atrophy (MSA)





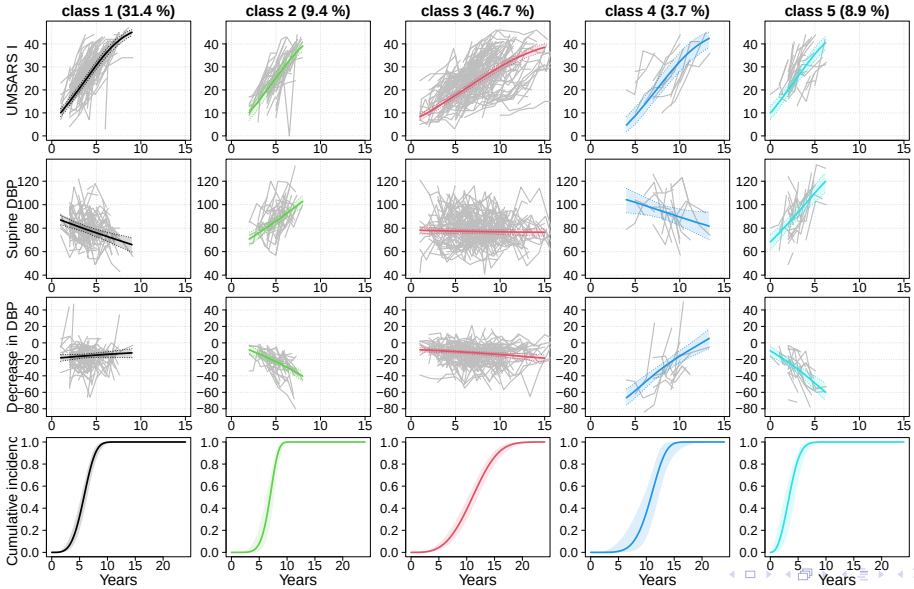


Actual structure of mpjlcmm R function

(Proust-Lima, Stat Med 2023)

- Latent process mixed model with latent classes
- Marker-specific measurement model

5 latent classes identified



(Classical) Research Questions addressed by joint models

- quantify the association of a endogenous marker with the risk of event

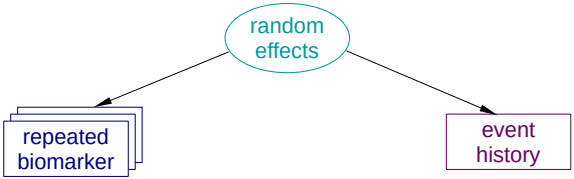
$$\lambda_i(t) = \lambda_0(t) \exp(Y^*(t)\boldsymbol{\eta})$$

- predict the risk of clinical endpoint using the biomarker information
  - individual dynamic prediction and screening optimization
- describe the trajectory of the biomarker stopped by the clinical progression
  - and evaluate its determinants
- explore/understand the association between the two processes
  - variability / heterogeneity in the disease progression

Shared Random-Effect Models (SREM) (Rizopoulos, 2012)

- Shared random-effects b_i distribution:

$$\boldsymbol{b}_i \sim \mathcal{N}(0, \boldsymbol{B})$$



- linear mixed model for the biomarker trajectory:

$$\begin{aligned} Y_{ij} \,|\boldsymbol{b}_i &= Y_i^*(t_{ij}) + \epsilon_{ij} \\ &= \boldsymbol{X}_{Li}(t_{ij})^\top \boldsymbol{\beta} \;+\; \boldsymbol{Z}_i(t_{ij})^\top \boldsymbol{b}_i \;+\; \epsilon_{ij} \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2) \end{aligned}$$

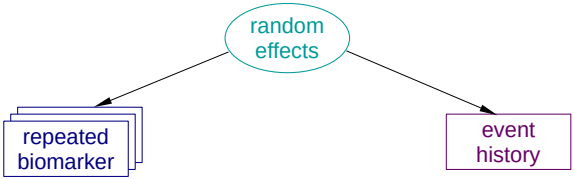
- proportional hazard model for the event:

$$\lambda_i(t; \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{X}_{Ti}(t)^\top \boldsymbol{\delta} + \; Y_i^*(t)\; \eta\right)$$

Shared Random-Effect Models (SREM) (Rizopoulos, 2012)

- Shared random-effects b_i distribution:

$$\boldsymbol{b}_i \sim \mathcal{N}(0, \boldsymbol{B})$$



- linear mixed model for the biomarker trajectory:

$$\begin{aligned} Y_{ij} \,|_{\boldsymbol{b}_i} &= Y_i^*(t_{ij}) + \epsilon_{ij} \\ &= \boldsymbol{X}_{Li}(t_{ij})^\top \boldsymbol{\beta} \;+\; \boldsymbol{Z}_i(t_{ij})^\top \boldsymbol{b}_i \;+\; \epsilon_{ij} \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2) \end{aligned}$$

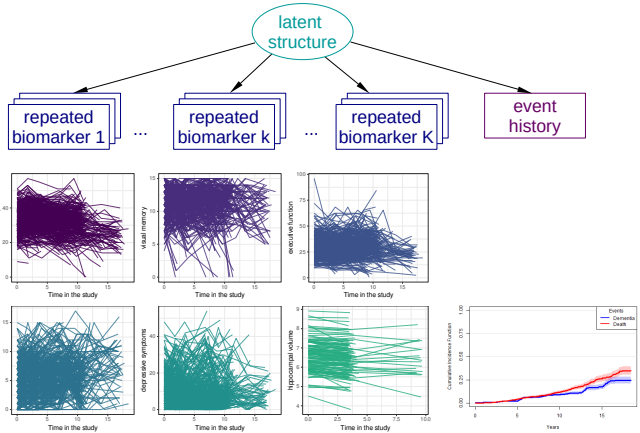
- proportional hazard model for the event:

$$\lambda_i(t; \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{X}_{Ti}(t)^\top \boldsymbol{\delta} + \boldsymbol{f}\ (t, \boldsymbol{b}_i, ...)\ \boldsymbol{\eta}\right)$$

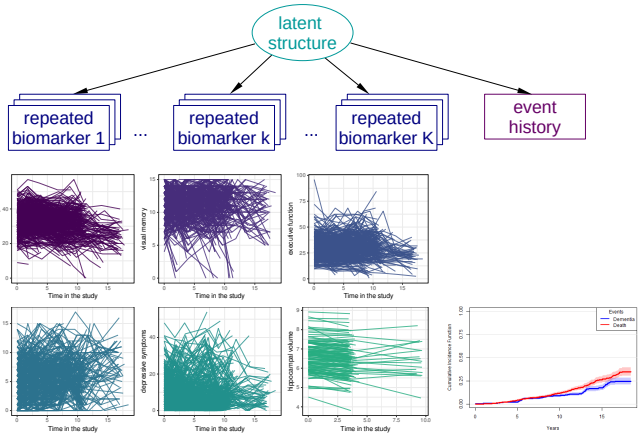
→ $\boldsymbol{\eta}$ quantifies the effect of the biomarker on the risk of event

→ biomarker trajectory corrected for the informative truncation by the event

Joint models with multivariate longitudinal / survival data



Joint models with multivariate longitudinal / survival data



- K different linear mixed models
  - a big vector of random effects: $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}, ..., \boldsymbol{b}_{ik}, ..., \boldsymbol{b}_{iK})$
- P cause-specific survival models

$$\lambda_{ip}(t; \boldsymbol{b}_i) = \lambda_{0p}(t) \times \exp\left(\boldsymbol{X}_{Ti}(t)\delta_p + \sum_{k=1}^{K} \boldsymbol{f}_k(t, \boldsymbol{b}_{ik}, ...)\, \boldsymbol{\eta}_{kp}\right)$$

- Examples:
  - Effect of a marker/exposure adjusted for other time-varying variables
  - Prediction of the event based on all the information available

Estimation in R (not exhaustive list!)

	inference	algorithm	integration	distributions	multiple markers
JM	Freq	EM/optim/MLA	paGH	Gaussian	✗
JMbayes	Bayes	MCMC		Exp. family	✓
JMbayes2	Bayes	MCMC		Exp. family	✓
joineR	Freq	EM	GH	Gaussian	✗
joineRML	Freq	MCEM	qMC	Gaussian	✓
rstanArm	Bayes	STAN		Exp. Family	✓
INLAJoint	Bayes	INLA		Gauss. process	✓
JLPM	Freq	MLA	qMC	Bin/ord/curvi	✗ /✓
frailtyPack	Freq	MLA	aGH	Gaussian	✗
JMBordo	Freq	MLA	qMC	Exp. Family	✓
... and saemix, BeQut and others ...					

qMC = quasi Monte Carlo; (p)aGH = (pseudo) adaptive Gauss Hermite

→ **Various specifications:** mixed models, survival models, dependence structure, data nature, ...

Example with JMbayes2

Structure common to packages of Rizopoulos's group : JM, JMbayes, JMbayes2

```
library("JMbayes2")
# specification of the longitudinal model(s)
LongModel <- lme(log(serBilir) ~ year * sex, data = pbc2, random = ~ year | id)

# specification of the survival model
CoxModel <- coxph(Surv(years, status2) ~ sex, data = pbc2.id)

# estimation of the joint model (by default, current value)
jointFit1 <- jm(CoxModel, LongModel, time_var = "year")
```

Numerical issues in SREM with multiple longitudinal biomarkers

1. Log-likelihood computation becomes rapidly untractable (huge numerical integration)

$$\mathscr{L}_i(\boldsymbol{\theta}) = \int_{\boldsymbol{b_i}} f(Y_i \mid \boldsymbol{b_i}; \boldsymbol{\theta}) \; f(T_i \mid \boldsymbol{b_i}; \boldsymbol{\theta}) \; f(\boldsymbol{b_i}; \boldsymbol{\theta}) \; d\boldsymbol{b_i}$$

Solutions:

- (quasi) Monte Carlo integration
- Bayesian inference (e.g., MCMC, INLA)

Numerical issues in SREM with multiple longitudinal biomarkers

1. Log-likelihood computation becomes rapidly untractable (huge numerical integration)

$$\mathcal{L}_i(\boldsymbol{\theta}) = \int_{\boldsymbol{b}_i} f(Y_i \mid \boldsymbol{b}_i;\boldsymbol{\theta})\ f(T_i \mid \boldsymbol{b}_i;\boldsymbol{\theta})\ f(\boldsymbol{b}_i;\boldsymbol{\theta})\ d\boldsymbol{b}_i$$

   Solutions:
   - (quasi) Monte Carlo integration
   - Bayesian inference (e.g., MCMC, INLA)

2. Large number of predictors in the survival model with $\sum_{k=1}^{K} \boldsymbol{f}_k\ (\ t, \boldsymbol{b}_{ik}, ...)\ \boldsymbol{\eta}_{kp}$

   Solutions:
   - Regularization in the survival model - Lasso - (e.g., Andrinopoulou and Rizopoulos SiM 2012, Chen and Wang SiM 2017)

Numerical issues in SREM with multiple longitudinal biomarkers

1. Log-likelihood computation becomes rapidly untractable (huge numerical integration)

$$\mathscr{L}_i(\boldsymbol{\theta}) = \int_{\boldsymbol{b_i}} f(Y_i \mid \boldsymbol{b_i};\boldsymbol{\theta})\ f(T_i \mid \boldsymbol{b_i};\boldsymbol{\theta})\ f(\boldsymbol{b_i};\boldsymbol{\theta})\ d\boldsymbol{b_i}$$

   Solutions:
   - (quasi) Monte Carlo integration
   - Bayesian inference (e.g., MCMC, INLA)

2. Large number of predictors in the survival model with $\sum_{k=1}^{K} \boldsymbol{f_k}\ (\ t, \boldsymbol{b}_{ik}, ...)\ \boldsymbol{\eta}_{kp}$

   Solutions:
   - Regularization in the survival model - Lasso - (e.g., Andrinopoulou and Rizopoulos SiM 2012, Chen and Wang SiM 2017)

3. Too high number of parameters for simultaneous estimation (for K long. + P surv. regressions)

   Solutions:
   - 2-step methods / regression calibration (Ye et al., 2008, Signorelli et al., 2021, Mauff et al. 2020)

Numerical issues in SREM with multiple longitudinal biomarkers

1. Log-likelihood computation becomes rapidly untractable (huge numerical integration)

$$\mathscr{L}_i(\boldsymbol{\theta}) = \int_{\boldsymbol{b}_i} f(Y_i \mid \boldsymbol{b}_i;\boldsymbol{\theta})\ f(T_i \mid \boldsymbol{b}_i;\boldsymbol{\theta})\ f(\boldsymbol{b}_i;\boldsymbol{\theta})\ d\boldsymbol{b}_i$$

   Solutions:
   - (quasi) Monte Carlo integration
   - Bayesian inference (e.g., MCMC, INLA)

2. Large number of predictors in the survival model with $\sum_{k=1}^{K} \boldsymbol{f}_k\ (\ t, \boldsymbol{b}_{ik}, ...)\ \boldsymbol{\eta}_{kp}$

   Solutions:
   - Regularization in the survival model - Lasso - (e.g., Andrinopoulou and Rizopoulos SiM 2012, Chen and Wang SiM 2017)

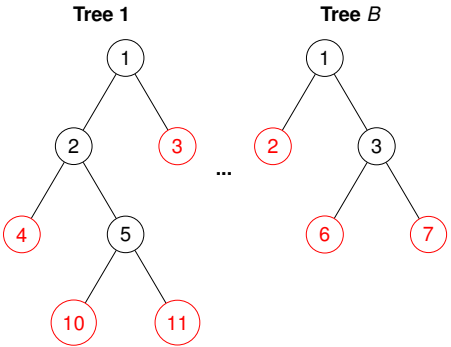
3. Too high number of parameters for simultaneous estimation (for K long. + P surv. regressions)

   Solutions:
   - 2-step methods / regression calibration (Ye et al., 2008, Signorelli et al., 2021, Mauff et al. 2020)

What about totally changing the framework and use random forests?

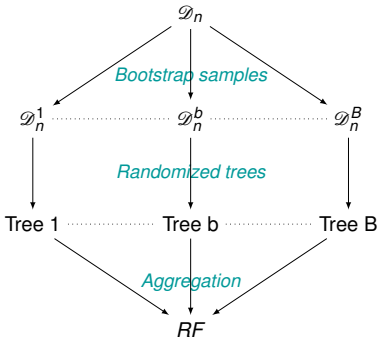
Random survival forests for competing causes of events

- Random survival forest principle (Ishwaran et al. 2008, 2014)
  - Ensemble of decision trees that partition the subjects into homogeneous leaves regarding survival



Random survival forests for competing causes of events

- Random survival forest principle (Ishwaran et al. 2008, 2014)
  - Ensemble of decision trees that partition the subjects into homogeneous leaves regarding survival

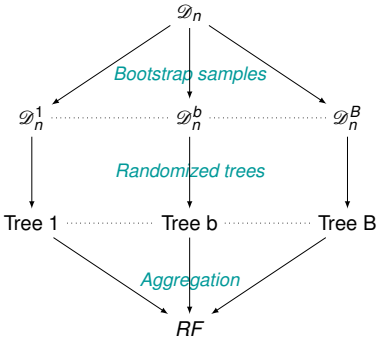


Random survival forests for competing causes of events

- Random survival forest principle (Ishwaran et al. 2008, 2014)
  - Ensemble of decision trees that partition the subjects into homogeneous leaves regarding survival

- Pros:
  - Useful for individual prediction
  - Designed for high-dimensional data (i.e., large number of predictors)
  - Handle complex relationship between predictors and event

- Cons:
  - ⚠ Limited to time-independent predictors

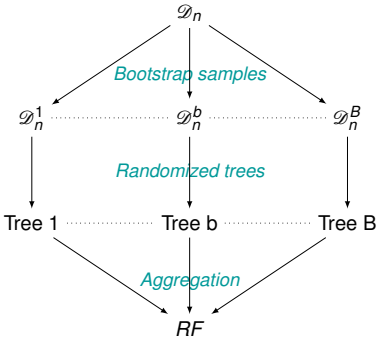


Random survival forests for competing causes of events

- Random survival forest principle (Ishwaran et al. 2008, 2014)
  - Ensemble of decision trees that partition the subjects into homogeneous leaves regarding survival

- Pros:
  - Useful for individual prediction
  - Designed for high-dimensional data (i.e., large number of predictors)
  - Handle complex relationship between predictors and event

- Cons:
  - ⚠ Limited to time-independent predictors

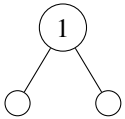


- Our solution: **DynForest**
  - Incorporate time-dependent predictors in the tree building process

Splitting rule in random survival forests

Find two groups of subjects which maximize the difference in event probability:

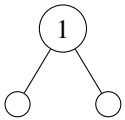
1. Randomly draw `mtry` predictors
2. Build two groups from each predictor's values
3. Compute the statistic to quantify the distance (e.g. Fine & Gray for the event probability)

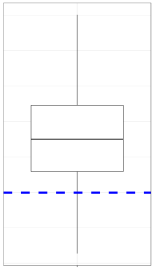


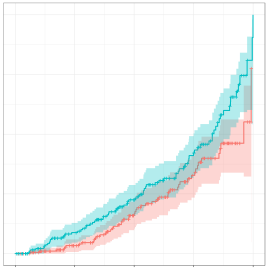
Splitting rule in random survival forests

Find two groups of subjects which maximize the difference in event probability:

1. Randomly draw `mtry` predictors
2. Build two groups from each predictor's values
3. Compute the statistic to quantify the distance (e.g. Fine & Gray for the event probability)



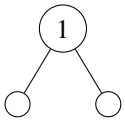


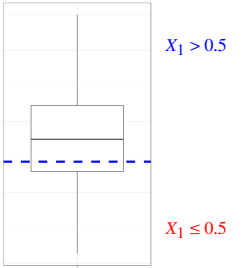


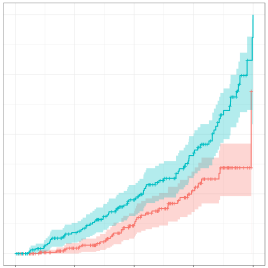
Splitting rule in random survival forests

Find two groups of subjects which maximize the difference in event probability:

1. Randomly draw `mtry` predictors
2. Build two groups from each predictor's values
3. Compute the statistic to quantify the distance (e.g. Fine & Gray for the event probability)



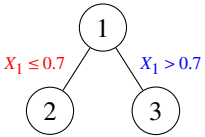


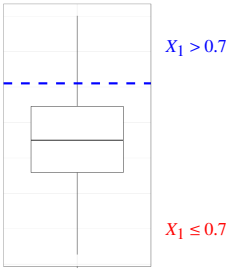


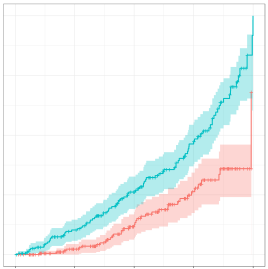
Splitting rule in random survival forests

Find two groups of subjects which maximize the difference in event probability:

1. Randomly draw `mtry` predictors
2. Build two groups from each predictor's values
3. Compute the statistic to quantify the distance (e.g. Fine & Gray for the event probability)





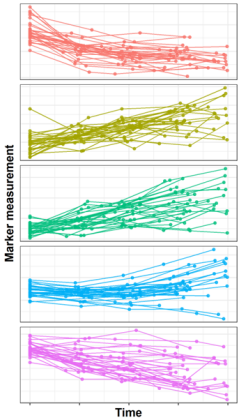


How to incorporate time-dependent predictors?

At each node, transform time-dependent predictors Y_k into time-fixed features :

1. Model $\boldsymbol{Y}_k$ trajectory using mixed models:

$$Y_{ik}(t_{ijk}) = \boldsymbol{Z}_{ik}(t_{ijk})^\top(\boldsymbol{\beta}_k + \boldsymbol{b}_{ik}) + \epsilon_{ijk}$$



How to incorporate time-dependent predictors?

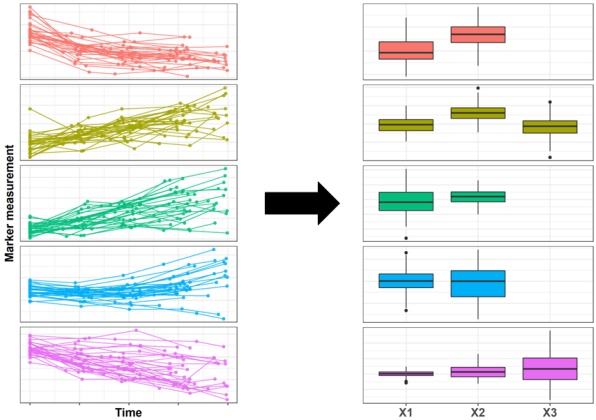
At each node, transform time-dependent predictors Y_k into time-fixed features :

1. Model $\boldsymbol{Y}_k$ trajectory using mixed models:

$$Y_{ik}(t_{ijk}) = \boldsymbol{Z}_{ik}(t_{ijk})^\top(\boldsymbol{\beta}_k + \boldsymbol{b}_{ik}) + \epsilon_{ijk}$$

2. Compute individual random-effects:

$$\hat{\boldsymbol{b}}_{ik} = \mathbb{E}(\boldsymbol{b}_{ik}|\boldsymbol{Y}_i)$$



How to incorporate time-dependent predictors?

At each node, transform time-dependent predictors Y_k into time-fixed features :

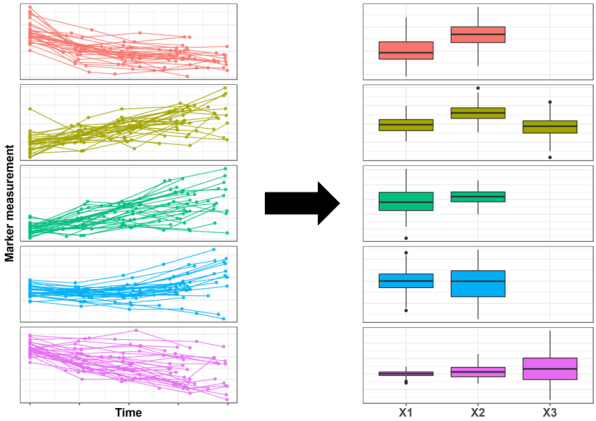
1. Model $\boldsymbol{Y}_k$ trajectory using mixed models:

$$Y_{ik}(t_{ijk}) = \boldsymbol{Z}_{ik}(t_{ijk})^\top(\boldsymbol{\beta}_k + \boldsymbol{b}_{ik}) + \epsilon_{ijk}$$

2. Compute individual random-effects:

$$\hat{\boldsymbol{b}}_{ik} = \mathbb{E}(\boldsymbol{b}_{ik}|\boldsymbol{Y}_i)$$

3. Consider them as splitting variable candidates

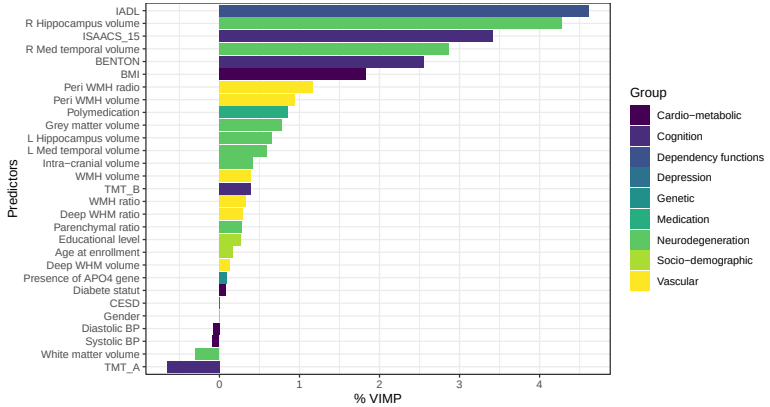


DynForest in practice: example of call

```r
library(DynForest)

# definitions of the objects:
Y <- list(type = "surv", Y = unique(pbc2_train[,c("id","years","event")]))
fixedData_train <- unique(pbc2_train[,c("id","age","drug","sex")])
timeData_train <- pbc2_train[,c("id","time","serBilir","SGOT","albumin","alkaline")]

# definitions of mixed models:
timeVarModel <- list(serBilir = list(fixed = serBilir ~ time, random = ~ time),
                     SGOT = list(fixed = SGOT ~ time + I(time^2), random = ~ time + I(time^2)),
                     albumin = list(fixed = albumin ~ time, random = ~ time),
                     alkaline = list(fixed = alkaline ~ time,random = ~ time))

# training of the forest:
res_dyn <- DynForest(timeData = timeData_train,
                     fixedData = fixedData_train,
                     timeVar = "time", idVar = "id",
                     timeVarModel = timeVarModel, Y = Y,
                     ntree = 200, mtry = 3, nodesize = 2, minsplit = 3,
                     cause = 2, ncores = 7, seed = 1234)
```

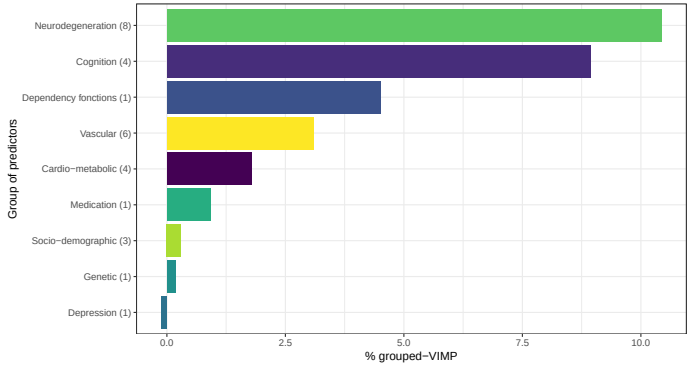
Application to dementia from multi-modal repeated data in 3C

Quantification of the variable importances for the prediction: 24 time-dependent predictors, 5 time-fixed predictors



Application to dementia from multi-modal repeated data (by groups

Quantification of the variable importances for the prediction **by group**



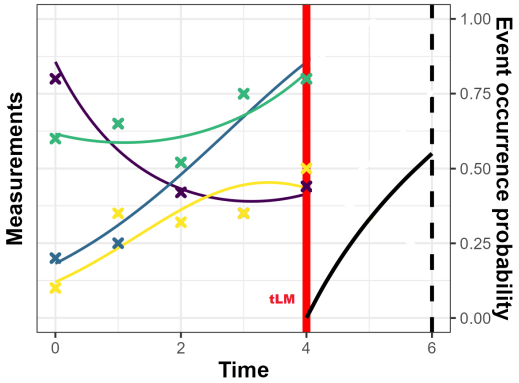
(Classical) Research Questions addressed by joint models

- quantify the association of a endogenous marker with the risk of event
- predict the risk of clinical endpoint using the biomarker information
  - individual dynamic prediction and screening optimization
- describe the trajectory of the biomarker stopped by the clinical progression
  - and evaluate its determinants
- explore/understand the association between the two processes
  - variability / heterogeneity in the disease progression

Dynamic prediction for a new subject ⋆

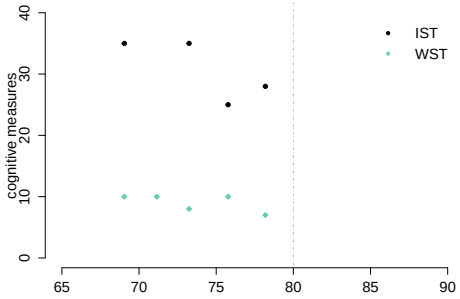
- Predicted probability from landmark s at horizon t:

$$\pi^{\star}(s,t) = \mathbb{P}\left(T_{\star} < s+t, \delta_{\star} = p | T_{\star} > s, \mathscr{Y}_{\star}(s), \mathscr{X}_{\star}\right)$$



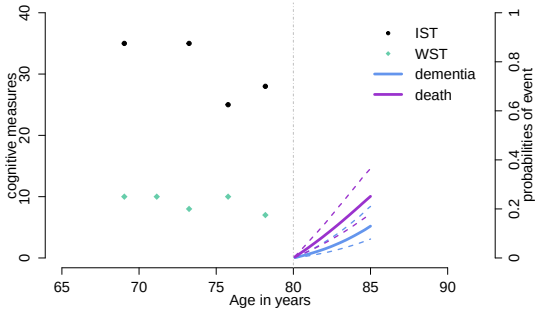
Dynamic prediction for a new subject $\star$: joint models

- Direct posterior computation (Bayes):
  - Monte Carlo approximation of the posterior distribution
  - e.g. with **`dynpred`** function in lcmm, with **`predict`** function in JMbayes2
- Performances evaluation:
  - **`riskRegression`** R package for AUC, Brier Score (Gerds & Kattan, 2021; Blanche, 2015)
  - in lcmm, **`epoce`** function for UACV (Commenges, Bcs 2011)



Dynamic prediction for a new subject ⋆: joint models

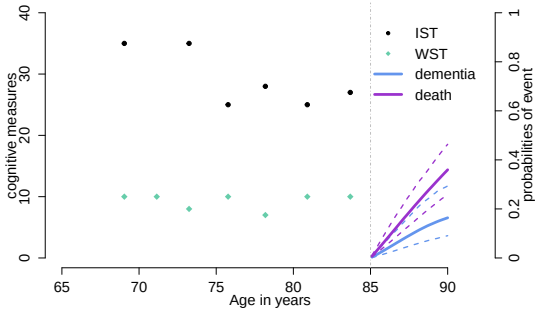
- Direct posterior computation (Bayes):
  - Monte Carlo approximation of the posterior distribution
  - e.g. with **`dynpred`** function in lcmm, with **`predict`** function in JMbayes2
- Performances evaluation:
  - **`riskRegression`** R package for AUC, Brier Score (Gerds & Kattan, 2021; Blanche, 2015)
  - in lcmm, **`epoce`** function for UACV (Commenges, Bcs 2011)



	at 80 years old
5-year probability of dementia (%):	13.0 [7.7,21.0]
5-year probability of death (%) :	25.1 [18.1,36.5]

Dynamic prediction for a new subject ⋆: joint models

- Direct posterior computation (Bayes):
  - Monte Carlo approximation of the posterior distribution
  - e.g. with **`dynpred`** function in lcmm, with **`predict`** function in JMbayes2
- Performances evaluation:
  - **`riskRegression`** R package for AUC, Brier Score (Gerds & Kattan, 2021; Blanche, 2015)
  - in lcmm, **`epoce`** function for UACV (Commenges, Bcs 2011)



	at 80 years old	at 85 years old
5-year probability of dementia (%):	13.0 [7.7,21.0]	16.4 [9.1,29.4]
5-year probability of death (%) :	25.1 [18.1,36.5]	36.0 [25.9,46.3]

Dynamic prediction for a new subject $\star$: random forests

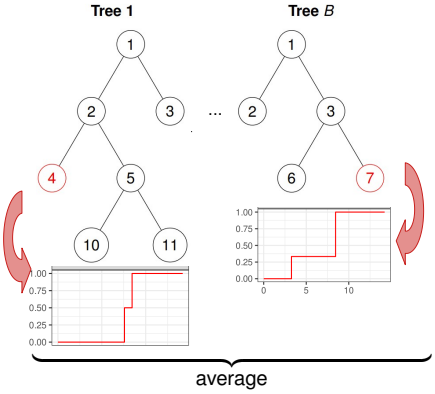
- Predicted probability from landmark s at horizon t:

$$\pi^{\star}(s,t) = \mathbb{P}\left(T_{\star} < s+t, \delta_{\star} = p | T_{\star} > s, \mathscr{Y}_{\star}(s), \mathscr{X}_{\star}\right)$$

- Drop down the new subject $\star$ into the trees using:
  - the history of time-dependent predictors $\mathscr{Y}_{\star}(s)$ up to landmark time s
  - time-fixed covariates $\mathscr{X}_{\star}$

- Average the leaf-and-tree-specific cumulative incidence functions $\hat{\pi}^{(tree,leaf)}(s,t)$ across trees:

$$\hat{\pi}_{\star}(s,t) = \frac{1}{B}\left(\hat{\pi}_{\star}^{(1,4)}(s,t) + ... + \hat{\pi}_{\star}^{(B,7)}(s,t)\right)$$



Concluding remarks

- Joint models = central technique in health studies (and probably beyond)
  - understanding of etiology, natural history and progression
  - individual dynamic prediction
  - correct for informative dropout

- Different solutions/implementations available in R
  - latent classes /shared random effects
  - different parametric assumptions (baseline risk, distribution of outcomes)
  - numerical limitations with many longitudinal markers

- `DynForest`: example of promising alternative from statistical learning
  - Not a two-step approach!
    - ★ separate mixed models for the longitudinal markers at each node
    - ★ naturally handles informative censoring of biomarker data as estimated on homogeneous nodes
  - accounts for nonlinear associations, interactions, etc.
  - current extensions with FPCA tools (Segalas 2024), distances, and other splitting rules

- Fundings:

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

- References

- **JLCM:** Proust-Lima et al (2014). JLCM for longitudinal and time-to-event data: A review. SMMR 23, 74-90
  https://cecileproust-lima.github.io/lcmm/
- **SREM:** Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman & Hall/CRC 2012
  https://drizopoulos.github.io/JMbayes2
- **DynForest**: Devaux et al (2023). Random survival forests with multivariate longitudinal endogenous covariates, SMMR 32, 2331-2346
  https://github.com/anthonydevaux/DynForest
- **marqLevAlg:** Philipps et al (2021). Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg. The R Journal 13(2), 365-379.
  https://github.com/VivianePhilipps/marqLevAlgParallel

Other references

Random Survival Forests:

Ishwaran et al. (2008) *Annals Applied Stat*, 2(3), 841-60
Ishwaran et al. (2014) *Biostatistics*, 15(4), 757-73.
Segalas et al. (2024) *arXiv*
`https://arxiv.org/abs/2402.10624`

Regression Calibration / 2-Stage:

Signorelli et al. (2021) *Statistics in medicine*, 40(27), 6178-96
Ye et al. (2008) *Biometrics*, 64(4), 1238-46
Devaux et al. (2022) *BMC Med Res Methodol*, 22(1), 188

Error of Prediction:

Blanche et al. (2015). *Biometrics*, 71, 102-13.
Gerds & Kattan (2021) *R. Chapman & Hall/CRC*

SREM:

Andrinopoulou, Rizopoulos (2016) *Stat Med*, 35(26), 4813-23.
Chen, Wang(2017)*Stat Med*, 36(24), 3820-9
Ferrer et al. (2016) *Stat Med*, 35(22), 3933-48
Mauff et al. (2017)*Stat Med*, 36(23), 3746-59
Rizopoulos (2012) *CRC Press*
Rouanet et al. (2016) *Biometrics*, 72(4), 1123-35
Rustand et al. (2023) *Biostatistics*, 2024, 25(2), 429-448

JLCM:

Rouanet et al. (2016) *Biometrics*, 72(4), 1123-35
Proust-Lima et al. (2016). *Statistics in Medicine*, 35(3), 382-398
Proust-Lima et al. (2017). *Journal of Statistical Software*, 78(2), 1-56
Proust-Lima et al. (2023). *Statistics in Medicine*, 42(22), 3996-4014