WEB APPENDIX

5-YEAR DYNAMIC PREDICTION OF DEMENTIA USING REPEATED MEASURES OF COGNITIVE TESTS AND A DEPENDENCY SCALE

Céline Ben-Hassen, Catherine Helmer, Claudine Berr and Hélène Jacqmin-Gadda

Contents:

Web appendix 1 : Detailed dementia diagnosis procedure

In the two cohorts, the procedures to classify dementia required an evaluation by a neurologist or geriatrician followed by a classification committee with experts who reviewed all available information for a participant in order to obtain a consensus on the diagnosis. There were however some differences between the two cohorts and within the 3C cohort. 1/ In PAQUID the DSM-IIIR criteria were used whereas the DSM-IV criteria were used in 3C; 2/ In PAQUID the expert committee was local, whereas it was national in 3C, the same for all the 3C centers (Bordeaux and Montpellier); 3/ In PAQUID and in 3C-Bordeaux, there was a first selection step, and only participants suspected of having dementia, based on either their neuropsychological performance or decline relative to a previous examination, were then examined by a senior neurologist to establish a clinical diagnosis. In 3C Montpellier all the participants were examined by the neurologist. However for both 3C Bordeaux and Montpellier the same committee reviewed all the suspected cases for classification.

Web appendix 2 : Details of the statistical methods

The joint model. The equation of the joint model transcript as follows:

$$Y_{ijk} = \tilde{Y}_{ik}(t_{ijk}) + \varepsilon_{ijk} = X_{1ik}^{T}(t_{ijk})\beta_k + Z_{ik}^{T}(t_{ijk})b_{ik} + \varepsilon_{ijk} \quad [1a]$$

$$\log E\,(Y_{ij4}) = \tilde{Y}_{i4}(t_{ij4}) = X_{1i4}^{T}(t_{ij4})\beta_4 + Z_{i4}^{T}(t_{ij4})b_{i4} \quad [1b]$$

$$\alpha_i(t) = \alpha_0(t)\exp\left(X_{2_i}^{T}\gamma + \sum_{k=1}^{4}\tilde{Y}_{ik}(t)\eta_k + \sum_{k=1}^{4}\tilde{Y}_{ik}(t)'\,\xi_k\right) \quad [2]$$

Where $Y_{ijk}$is the value of marker k for subject i at timepoint j (k = 4 for IADL4); X_1, X_2, and Z are vectors of covariates (including functions of time); b_{ik} is a subject and marker-specific vector of random effects; ε_{ijk} the measurement error; $\alpha_i(t)$ the hazard function; $\alpha_0(t)$ the baseline hazard function; $\tilde{Y}_{ik}(t)$ the expected value of Y_{ik} at time t and $\tilde{Y}_{ik}(t)'$ its first derivative (*i.e*, the slope of the trajectory of marker k at time t); β, γ, η and ξ the regression coefficients.

Model estimation procedure : The strategy of analysis had two steps. First, "unimarker" joint models were estimated on repeated measures of each marker and time-to-dementia using the PAQUID sample (Formulas *[1a]*, *[1b]* and *[2]* above). The mixed-effect sub-models and the time-to-dementia sub-model were both adjusted on the age at inclusion (standardized as such: $(age - 65)/10$) for the mixed-effect sub-models), sex, and education status (1 if the subject had a primary school diploma, 0 if not). We compared different modeling of the time-trend for the markers: linear, quadratic, splines with 2 or 4 degrees of freedom and we test for an interaction between age at inclusion and time. The best unimarker models were selected based on goodness-of-fit and predictive abilities (AUC and Brier Score) on the training sample. The selected unimarker models included a linear time trend for IADL and splines functions of time for the cognitive tests (with 4 degrees of freedom for the IST and BVRT and 2 degrees of

freedom for the MMSE) and an interaction term between time and age at inclusion was added for the MMSE and the IADL4 scale.

All joint models were estimated using the Bayesian algorithm implemented in the R package JMbayes (1). Convergence was evaluated through MCMC diagnostic plots to check the mixing of several chains without any trend.

Second, the multimarkers model was estimated. As estimation of joint models with several longitudinal markers is challenging due to the number of parameters and random effects. As convergence was not reached with the most complex 4-markers model, we removed the BVRT test which was by far the least predictive among the four unimarker models and slightly simplified the modeling of marker trajectories : the number of degrees of freedom of the splines of the IST longitudinal process was reduced from 4 to 2 and the interaction term was removed of the estimation of the MMSE and IADL longitudinal process. In the final model, the time trend of the cognitive tests was modeled through cubic splines with 1 interior node, while that of IADL4 was assumed to be linear (on the log scale). The three mixed models included correlated random intercept and coefficients for time with an unconstraint correlation structure between the 8 random effects.

Predictions. For each subject *i* still at risk of dementia at a given time *s*, we estimated the probability to develop dementia before $s + t$, given all the markers measures ($\mathcal{Y}(s)$) for the subject *i* up to time *s*, the values of the covariates ($w = (X_1, X_2, Z)$), and the data from the training dataset ($\mathcal{D}_n$), using the following formula (1) :

$$\pi_i(s,t) = 1 - P(T_i \geq s + t \mid T_i \geq s, \mathcal{Y}_i(s), w_i, \mathcal{D}_n) \quad [3]$$

$$= 1 - \int P(T_i \geq s + t \mid T_i \geq s, \mathcal{Y}_i(s), w_i, \theta)\, P(\theta \mid \mathcal{D}_n)\, d\theta,$$

T_i being the time of event, θ the vector of parameters of the model k, and $P(\theta \mid \mathcal{D}_n)$ the posterior distribution of θ given the training data.

Evaluation of predictive abilities

AUC and Brier Scores. Predictive abilities of the models are evaluated by computing AUC and BS. The AUC is the probability that a randomly selected case has a larger predicted probability than a randomly selected control. In the competing risks framework, two definitions of the controls are available leading to two definitions of AUC:

$$AUC1(s,t) = P\left(\pi_i\ (s,t) > \pi_j\ (s,t) \middle|\ \delta_i(s,t) = 1, \delta_j(s,t) = 0, T_i > s, T_j > s\right) \quad [4a]$$

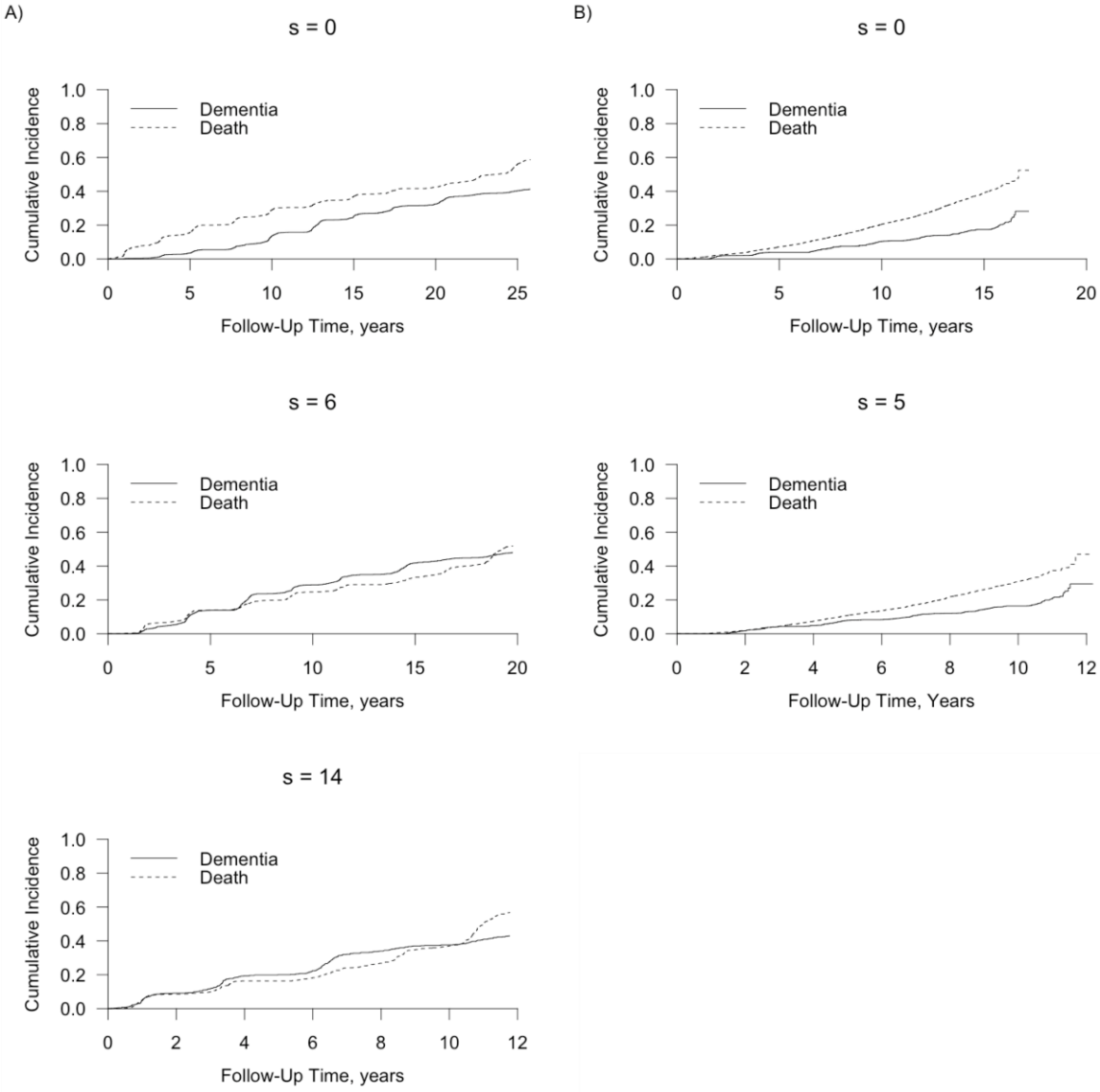
$$AUC2(s,t) = P\left(\pi_i(s,t) > \pi_j(s,t) \middle|\ \delta_i(s,t) = 1, \delta_j(s,t) \neq 1, T_i > s, T_j > s\right) \quad [4b]$$

$\delta(s,t)$ being the event indicator which equals to 0 if the subject is free of the main event at time $s+t$, 1 in case of dementia, and 2 in case of death.

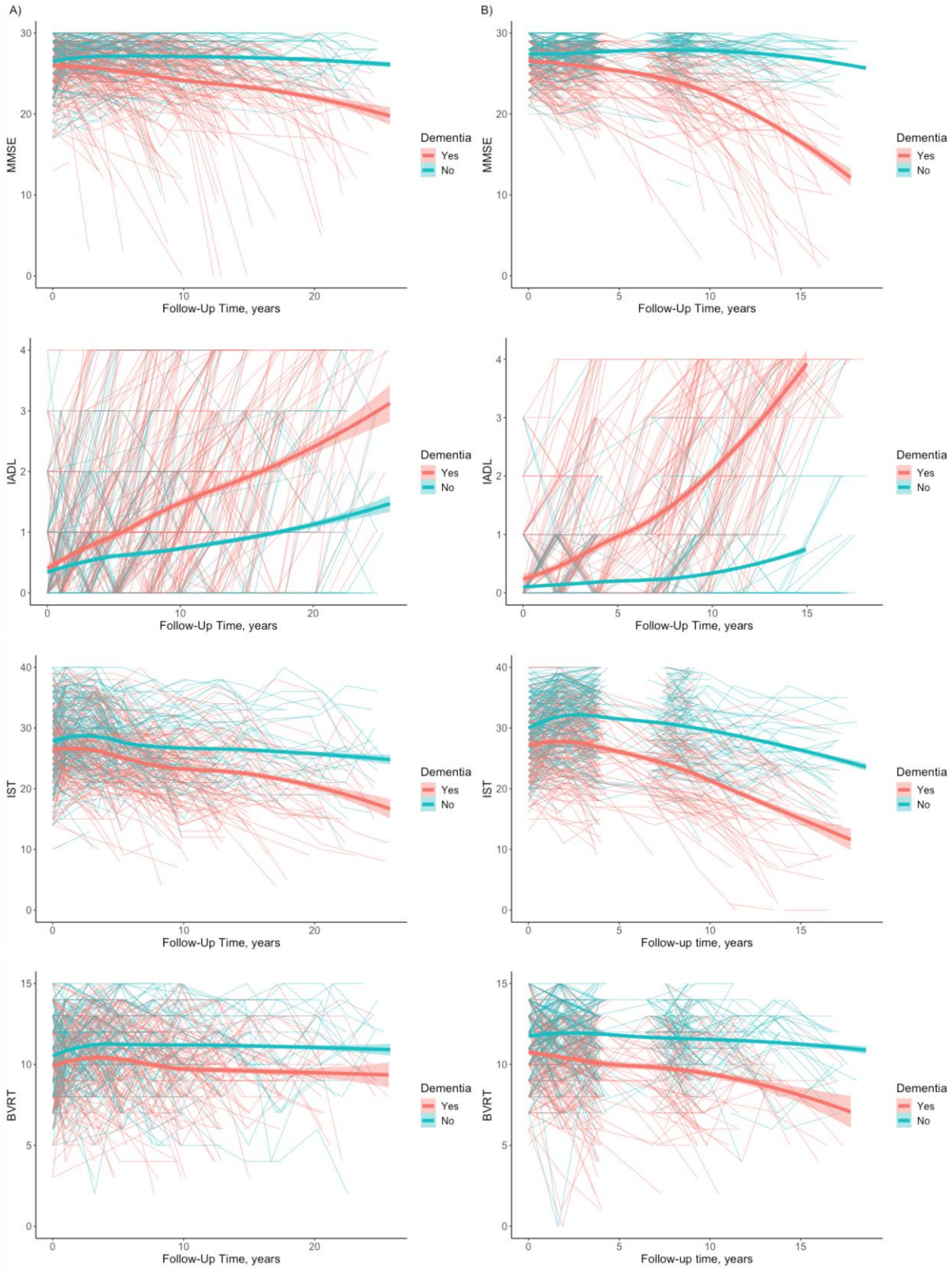
The general formula for the Brier Score is:

$$BS(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i(s,t) - \pi_i\ (s,t)\right)^2 \quad [5]$$

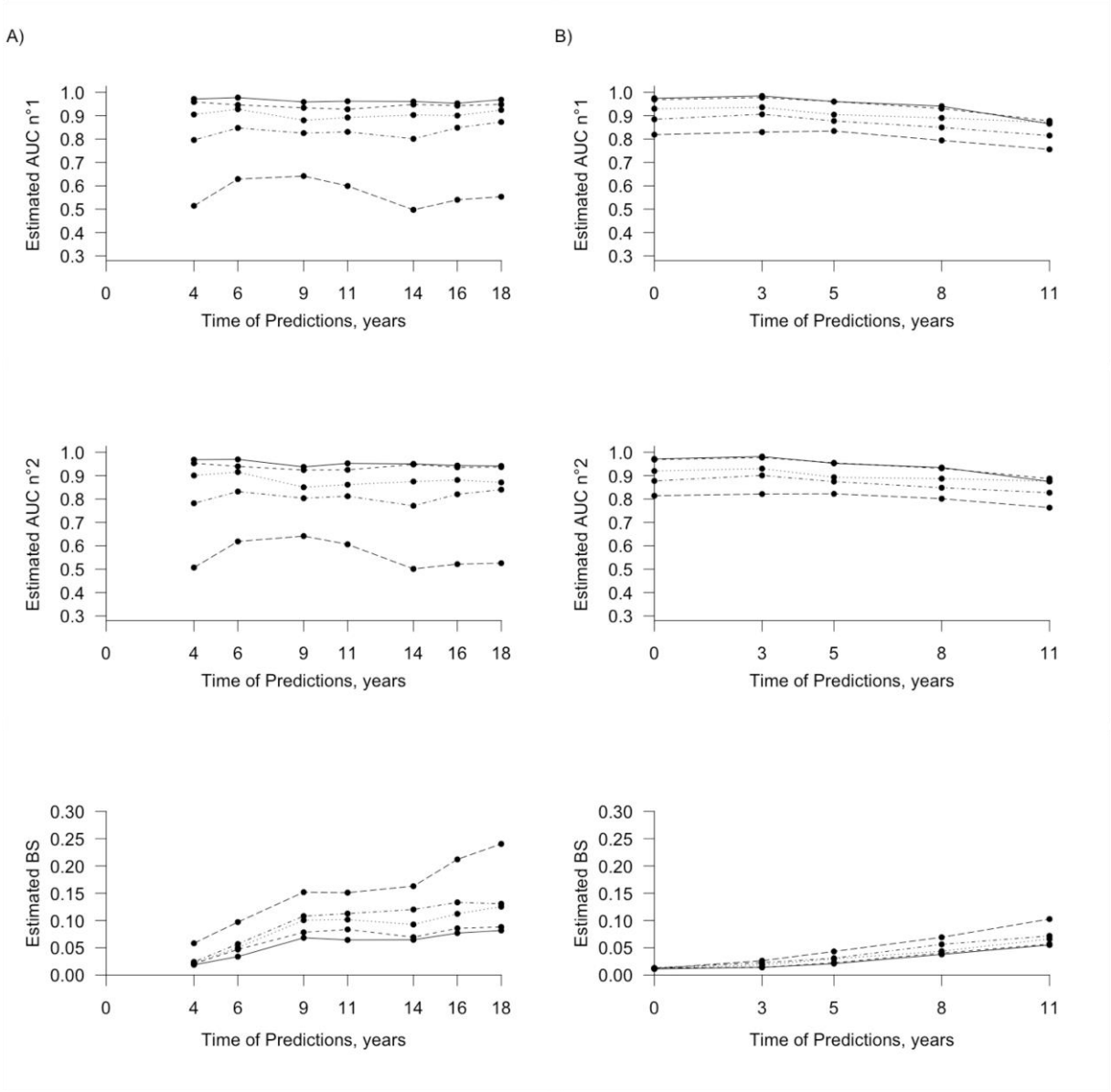
To account for right censoring, the Inverse probability of Censoring Weighting (IPCW) estimators of AUC and BS were used (2).



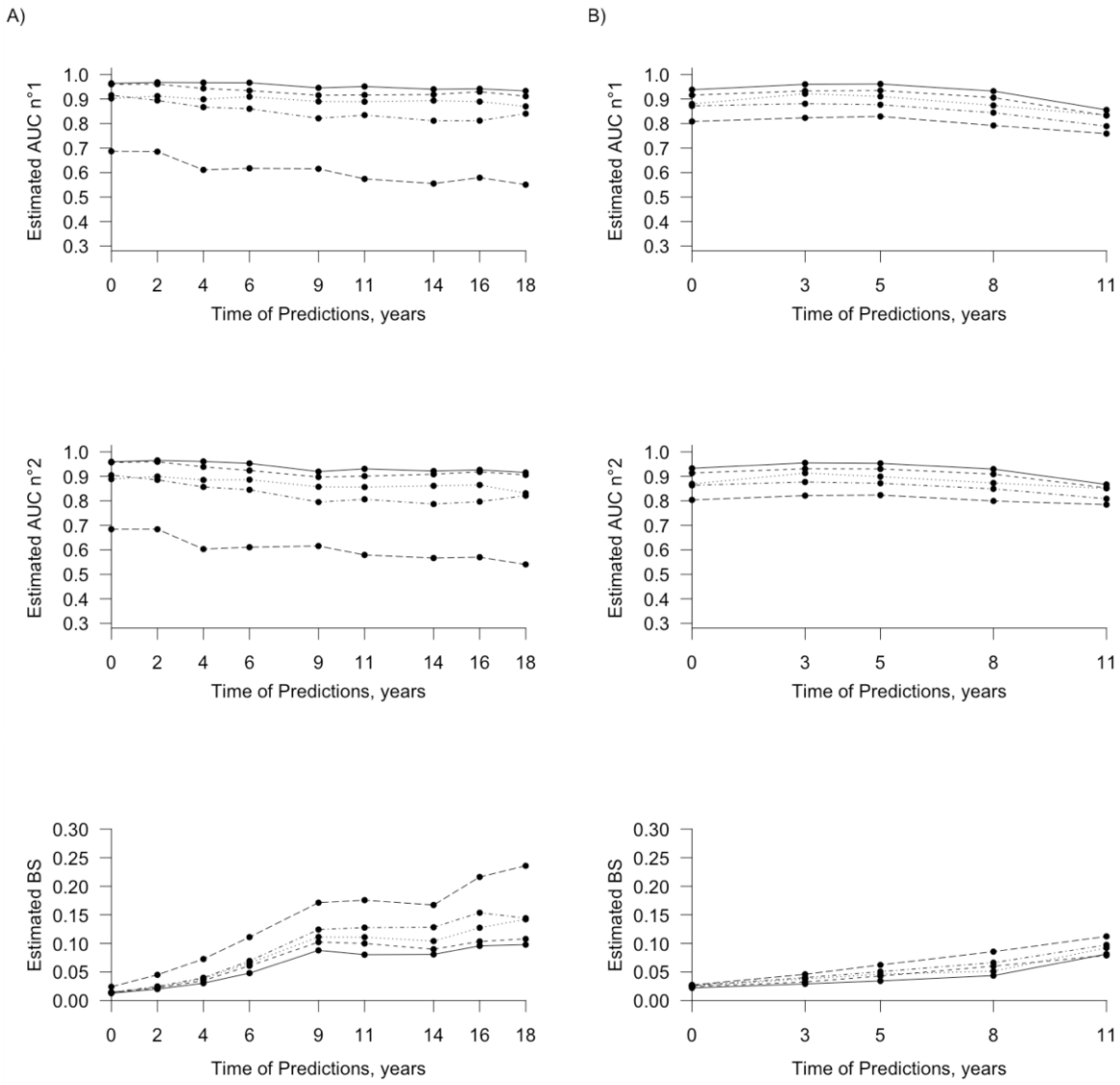
Web Figure 1: Cumulative incidence of dementia and death without dementia estimated by the non-parametric Aalen Johansen estimator from time s= 0, 6 and 14 for PAQUID (A) et s= 0 and 5 for 3C (B).



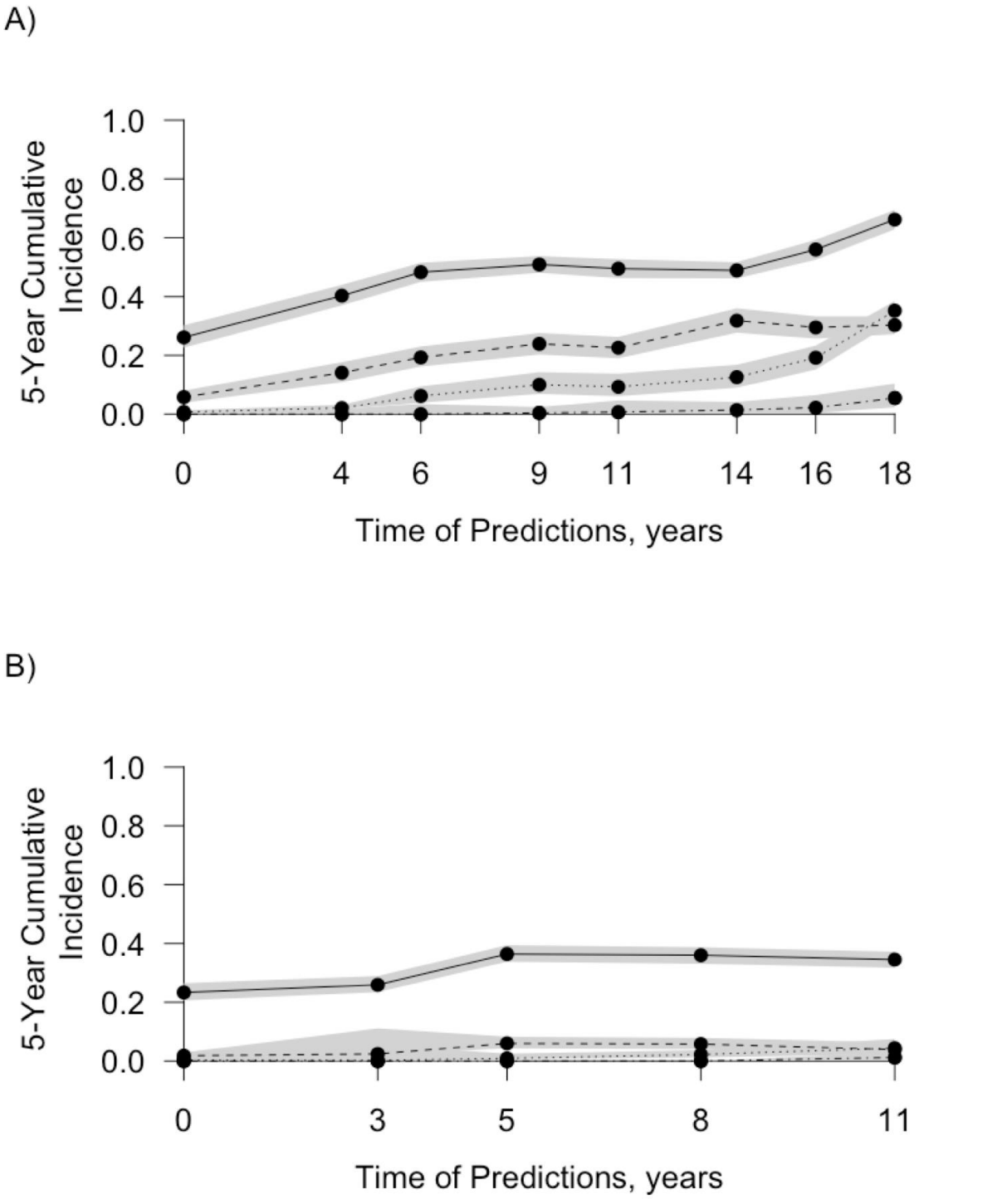
Web Figure 2: Spaghetti plots of observed markers values of 200 randomly chosen subjects with a dementia diagnosis and 200 randomly chosen subjects without a dementia diagnosis for both PAQUID (A) and 3C (B) with smoothed mean curves of the evolution estimated on the whole samples.



Web Figure 3: Estimation of the AUC and the Brier Scores for 5-year dementia prediction from time s on PAQUID (5-fold cross-validation) (A) and 3C (external validation) (B) for individuals with a high cognitive level (MMSE>28, 3rd quartile in PAQUID at inclusion) and no dependency at study entry (IADL4=0): N= 598 subjects for PAQUID (293 men, 305 women, with a median follow-up time equal to 12.4 years, and 153 incident dementia cases), and N=1,125 subjects for 3C (463 men, 662 women, with a median follow-up time equal to 12 years, and 107 incident dementia cases). Definitions of AUC n°1 and AUC n°2 are identical to Figure 3 in the main manuscript. Predictions are computed from either the final multimarker joint model (solid line) or a unimarker joint model based on MMSE (short dashed line), IADL4 (dotted line), IST (dashed-dotted line) or BVRT (long dashed line).



Web Figure 4: Estimation of the AUC and the Brier Scores for 5-year dementia prediction from time s on PAQUID (5-fold cross-validation) (A) and 3C (external validation) (B) for individuals with no cognitive impaireement (MMSE>23) and no dependency at study entry (IADL4=0): N=1,908 subjects for PAQUID (937 men, 971 women, with a median follow-up time equal to 9.98 years, and 502 incident dementia cases), and N=3,390 subjects for 3C (1,422 men, 1,968 women, with a median follow-up time equal to 11.3 years, and 450 incident dementia cases). Definitions of AUC n°1 and AUC n°2 are identical to Figure 3 in the main manuscript. Predictions are computed from either the final multimarker joint model (solid line) or a unimarker joint model based on MMSE (short dashed line), IADL4 (dotted line), IST (dashed-dotted line) or BVRT (long dashed line).



Web Figure 5: Cumulative incidence of dementia (with its 95% confidence interval) between $\boldsymbol{s}$ and $\boldsymbol{s + 5}$ estimated by a non-parametric illness-death model accounting for interval censoring for each quartile of the probabilities predicted from the final joint model, on PAQUID (A) and 3C (B). Solid line, very high risk group; dashed line, high risk group; dotted line, medium risk group; dashed-dotted line, low risk group; shaded area, 95% confidence intervals.

Web appendix 3 : R code for the statistical analyses

```
################ Cross-validation analysis ################

library(JMbayes)
library(survival)
library(splines)

######## Description of the datasets ########
###Datasets
# Paquid_trainCV1 : cross-validation training subsamble n°1 (long format)
# Paquid_validCV1 : cross-validation validation subsamble n°1 (long format)
# Paquid_Surv_trainCR1 : cross-validation training subsamble n°1 (wide format)

###Variables
# isa15ST : IST score truncated after 15 seconds (standardized)
# mmsnormST : MMSE score after normalization (standardized)
# bentonnormST : BVRT score after normalization (standardized)
# IADL_4 : IADL score
# SEXE : sex (1: male, 2: female)
# CEP: education level
# AGEENTREST : age at study entry (standardized)
# VISIT10: time (between study entry and visit, in years, divided by 10)
# NUMERO: subject id
# Time10: minimum time between time of event or last visit of follow-up
# dem: event indicator (dementia)
# AGEENTRE: age at study entry (not standardized)

######## Estimation of the joint models (Paquid, cross-validation subsamples)
########
#NB: the codes below correspond to the estimation for training subsamble n°1.
#Steps have been repeated for all 5 training subsamples

###Survival model
JMCoxFit_CR1 <- coxph(Surv(Time10, dem) ~ SEXE + CEP + AGEENTRE, data =
Paquid_Surv_trainCR1, model = TRUE)

###IST (unimarker)
#1) mixed model
MixedModelisa_CR1 <- mvglmer(list(isa15ST ~  SEXE + CEP + AGEENTREST +
ns(VISIT10,2)  + (1 + ns(VISIT10, 2) | NUMERO)), data = Paquid_trainCV1,
                families = list(gaussian))
#2) slope calculation formula
Forms_isa <- list("isa15ST" = "value", "isa15ST" = list(fixed = ~ 0 + dns(VISIT10, 2),
random = ~ 0 + dns(VISIT10, 2), indFixed = 5:6, indRandom = 2:3, name = "slope"))
#3) joint model
Joint.isa_CR1 <- mvJointModelBayes(MixedModelisa_CR1, JMCoxFit_CR1, timeVar =
"VISIT10", Formulas = Forms_isa)
```

```
###MMS (unimarker)
#1) mixed model
MixedModelmms_CR1 <- mvglmer(list(mmsnormST ~  SEXE + CEP + AGEENTREST +
ns(VISIT10,2)  + (1 + ns(VISIT10, 2) | NUMERO)), data = Paquid_trainCV1,
                    families = list(gaussian))
#2) slope calculation formula
Forms_mms <- list("mmsnormST" = "value", "mmsnormST" = list(fixed = ~ 0 +
dns(VISIT10, 2), random = ~ 0 + dns(VISIT10, 2), indFixed = 5:6, indRandom = 2:3,
name = "slope"))
#3) joint model
Joint.mms_CR1 <- mvJointModelBayes(MixedModelmms_CR1, JMCoxFit_CR1, timeVar =
"VISIT10", Formulas = Forms_mms)

###BVRT (unimarker)
#1) mixed model
MixedModelbenton_CR1 <- mvglmer(list(bentonnormST ~  SEXE + CEP + AGEENTREST
+ ns(VISIT10,4)  + (1 + ns(VISIT10, 4) | NUMERO)), data = Paquid_trainCV1,
                     families = list(gaussian))
#2) slope calculation formula
Forms_benton <- list("bentonnormST" = "value", "bentonnormST" = list(fixed = ~ 0 +
dns(VISIT10, 4), random = ~ 0 + dns(VISIT10, 4), indFixed = 5:8, indRandom = 2:5,
name = "slope"))
#3) joint model
Joint.benton_CR1 <- mvJointModelBayes(MixedModelbenton_CR1, JMCoxFit_CR1,
timeVar = "VISIT10", Formulas = Forms_benton)

###IADL (unimarker)
#1) mixed model
MixedModeliadl_CR1 <- mvglmer(list(IADL_4 ~  SEXE + CEP + AGEENTREST + VISIT10 +
(1 + VISIT10 | NUMERO)), data = Xcomplet_trainCR1,
                    families = list(poisson))
#2) slope calculation formula
Forms_IADL <- list("IADL_4" = "value", "IADL_4" = list(fixed = ~ 1, random = ~ 1,
indFixed = 5, indRandom = 2))
#3) joint model
Joint.IADL_CR1 <- mvJointModelBayes(MixedModeliadl_CR1, JMCoxFit_CR1, timeVar =
"VISIT10", Formulas = Forms_IADL)

###Multimarkers (IST + MMSE + IADL)
#1) mixed model
MixedModelisammsiadl_CR1 <- mvglmer(list(isa15ST ~  SEXE + CEP + AGEENTREST +
ns(VISIT10,2)  + (1 + ns(VISIT10, 2) | NUMERO),
                          mmsnormST ~  SEXE + CEP + AGEENTREST + ns(VISIT10,2)  + (1 +
ns(VISIT10, 2) | NUMERO),
                          IADL_4 ~  SEXE + CEP + AGEENTREST + VISIT10  + (1 + VISIT10 |
NUMERO)), data = Paquid_trainCV1,
                       families = list(gaussian,
gaussian,poisson),control=list(n.iter=56000,n.adapt=6000,n.burnin=6000,n.chains=4))
#2) slope calculation formula
```

```
Forms_isammsiadl <- list("isa15ST" = "value", "isa15ST" = list(fixed = ~ 0 + dns(VISIT10,
2), random = ~ 0 + dns(VISIT10, 2),
                                        indFixed = 5:6, indRandom = 2:3, name = "slope"),
                 "mmsnormST" = "value", "mmsnormST" = list(fixed = ~ 0 + dns(VISIT10,
2), random = ~ 0 + dns(VISIT10, 2),
                                          indFixed = 5:6, indRandom = 2:3, name = "slope"),
                 "IADL_4" = "value", "IADL_4" = list(fixed = ~ 1, random = ~ 1,
                                       indFixed = 5, indRandom = 2, name = "slope"))
#3) joint model
Joint.multimarkers_CR1 <- mvJointModelBayes(MixedModelisammsiadl_CR1,
JMCoxFit_CR1, timeVar = "VISIT10", Formulas = Forms_isammsiadl)

######## 5-years Predictions (Paquid, cross-validation subsamples) #######
#NB:
#1) the codes below correspond to the estimation for validation subsamble n°1
#and for the IST unimarker model. The same code was used for all models.
#2) Steps have been repeated for all 5 validation subsamples and predictions were
#then pooled to estimate AUC and BS
#3) Predictions were computed at each time s for all subjects, but for the final
#analysis, we only kept in the final matrix predictions for subjects still at
#risk of dementia at time s
#4) The predictions estimated by the package JMbayes are predictions of survival
without event.


PAQUID_s0.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=0, survTimes = 0.5)
PAQUID_s2.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=0.2, survTimes = 0.7)
PAQUID_s4.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=0.4, survTimes = 0.9)
PAQUID_s6.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=0.6, survTimes = 1.1)
PAQUID_s9.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=0.9, survTimes = 1.4)
PAQUID_s11.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=1.1, survTimes = 1.6)
PAQUID_s14.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=1.4, survTimes = 1.9)
PAQUID_s16.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=1.6, survTimes = 2.1)
PAQUID_s18.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=1.8, survTimes = 2.3)
PAQUID_s21.isa_train1 <- survfitJM(Joint.isa_CR1, newdata = Paquid_validCV1, idVar =
"NUMERO", last.time=2.1, survTimes = 2.6)
```

```
################## External validation analysis ##################

library(JMbayes)
library(survival)
library(splines)

######## Description of the datasets ########
###Datasets
# Paquid : Paquid dataset (long format)
# Paquid_Surv : Paquid dataset (wide format)
# ThreeC : 3C dataset

###Variables
# isa15ST : IST score truncated after 15 seconds (standardized)
# mmsnormST : MMSE score after normalization (standardized)
# bentonnormST : BVRT score after normalization (standardized)
# IADL_4 : IADL score
# SEXE : sex (1: male, 2: female)
# CEP: education level
# AGEENTREST : age at study entry (standardized)
# VISIT10: time (between study entry and visit, in years, divided by 10)
# NUMERO: subject id
# Time10: minimum time between time of event or last visit of follow-up
# dem: event indicator (dementia)
# AGEENTRE: age at study entry (not standardized)

######## Estimation of the joint models (Paquid, total sample) ########
###Survival model
JMCoxFit <- coxph(Surv(Time10, dem) ~ SEXE + CEP + AGEENTRE, data = Paquid_Surv,
model = TRUE)

###IST (unimarker)
#1) mixed model
MixedModelisa <- mvglmer(list(isa15ST ~  SEXE + CEP + AGEENTREST + ns(VISIT10,2)
+ (1 + ns(VISIT10, 2) | NUMERO)), data = Paquid,
                families = list(gaussian))
#2) slope calculation formula
Forms_isa <- list("isa15ST" = "value", "isa15ST" = list(fixed = ~ 0 + dns(VISIT10, 2),
random = ~ 0 + dns(VISIT10, 2), indFixed = 5:6, indRandom = 2:3, name = "slope"))
#3) joint model
Joint.isa<- mvJointModelBayes(MixedModelisa, JMCoxFit, timeVar = "VISIT10", Formulas
= Forms_isa)

###MMSE (unimarker)
#1) mixed model
MixedModelmms <- mvglmer(list(mmsnormST ~  SEXE + CEP + AGEENTREST +
ns(VISIT10,2)  + (1 + ns(VISIT10, 2) | NUMERO)), data = Paquid,
                families = list(gaussian))
#2) slope calculation formula
```

```
Forms_mms <- list("mmsnormST" = "value", "mmsnormST" = list(fixed = ~ 0 +
dns(VISIT10, 2), random = ~ 0 + dns(VISIT10, 2), indFixed = 5:6, indRandom = 2:3,
name = "slope"))
#3) joint model
Joint.mms <- mvJointModelBayes(MixedModelmms, JMCoxFit, timeVar = "VISIT10",
Formulas = Forms_mms_ns2)

###BVRT (unimarker)
#1) mixed model
MixedModelbenton <- mvglmer(list(bentonnormST ~  SEXE + CEP + AGEENTREST +
ns(VISIT10,4)  + (1 + ns(VISIT10, 4) | NUMERO)), data = Paquid,
                  families = list(gaussian))
#2) slope calculation formula
Forms_benton <- list("bentonnormST" = "value", "bentonnormST" = list(fixed = ~ 0 +
dns(VISIT10, 4), random = ~ 0 + dns(VISIT10, 4), indFixed = 5:8, indRandom = 2:5,
name = "slope"))
#3) joint model
Joint.benton <- mvJointModelBayes(MixedModelbenton, JMCoxFit, timeVar = "VISIT10",
Formulas = Forms_benton)

###IADL (unimarker)
#1) mixed model
MixedModeliadl <- mvglmer(list(IADL_4 ~  SEXE + CEP + AGEENTREST + VISIT10 + (1 +
VISIT10 | NUMERO)), data = Paquid,
                  families = list(poisson))
#2) slope calculation formula
Forms_IADL <- list("IADL_4" = "value", "IADL_4" = list(fixed = ~ 1, random = ~ 1,
indFixed = 5, indRandom = 2))
#3) joint model
Joint.iadl <- mvJointModelBayes(MixedModeliadl, JMCoxFit, timeVar = "VISIT10",
Formulas = Forms_IADL)

###Multimarkers (IST + MMSE + IADL)
#1) mixed model
MixedModelisammsiadl <- mvglmer(list(isa15ST ~  SEXE + CEP + AGEENTREST +
ns(VISIT10,2)  + (1 + ns(VISIT10, 2) | NUMERO),
                        mmsnormST ~  SEXE + CEP + AGEENTREST + ns(VISIT10,2)  + (1 +
ns(VISIT10, 2) | NUMERO),
                        IADL_4 ~  SEXE + CEP + AGEENTREST + VISIT10  + (1 + VISIT10 |
NUMERO)), data = Paquid,
                     families = list(gaussian,
gaussian,poisson),control=list(n.iter=56000,n.adapt=6000,n.burnin=6000,n.chains=4))
#2) slope calculation formula
Forms_isammsiadl <- list("isa15ST" = "value", "isa15ST" = list(fixed = ~ 0 + dns(VISIT10,
2), random = ~ 0 + dns(VISIT10, 2),
                                        indFixed = 5:6, indRandom = 2:3, name = "slope"),
                 "mmsnormST" = "value", "mmsnormST" = list(fixed = ~ 0 + dns(VISIT10,
2), random = ~ 0 + dns(VISIT10, 2),
                                          indFixed = 5:6, indRandom = 2:3, name = "slope"),
```

```
                "IADL_4" = "value", "IADL_4" = list(fixed = ~ 1, random = ~ 1,
                                                    indFixed = 5, indRandom = 2, name = "slope"))
#3) joint model
Joint.isammsiadl <- mvJointModelBayes(MixedModelisammsiadl, JMCoxFit, timeVar =
"VISIT10", Formulas = Forms_isammsiadl)

######## 5-years Predictions (3C) ########
#NB:
#1) the codes below correspond to the estimation for for the IST unimarker model.
#The same code was used for all models.
#2) Predictions were computed at each time s for all subjects, but for the final
#analysis, we only kept in the final matrix predictions for subjects still at
#risk of dementia at time s
#3) The predictions estimated by the package JMbayes are predictions of survival
without event.

ThreeC_s0.isa <- survfitJM(Joint.isa, newdata = ThreeC, idVar = "NUMERO", last.time=0,
survTimes = 0.5)
ThreeC_s3.isa <- survfitJM(Joint.isa, newdata = ThreeC, idVar = "NUMERO",
last.time=0.3, survTimes = 0.8)
ThreeC_s5.isa <- survfitJM(Joint.isa, newdata = ThreeC, idVar = "NUMERO",
last.time=0.5, survTimes = 1.0)
ThreeC_s8.isa <- survfitJM(Joint.isa, newdata = ThreeC, idVar = "NUMERO",
last.time=0.8, survTimes = 1.3)
ThreeC_s10.isa <- survfitJM(Joint.isa, newdata = ThreeC, idVar = "NUMERO",
last.time=1.0, survTimes = 1.5)


########## Illness-death models and Nelson-Aalen estimators ###########

######## NB: here we provide an example for the 3C dataset (external validation)
# at time s = 3. The exact same procedure was replicated for each time s,
# both in external validation and in cross-validation

library(SmoothHazard)
library(survival)
library(riskRegression)

######## Description of the datasets ########
###Datasets
# ThreeCsmooth_s3 : ThreeCities dataset that only contains information for
# subjects still at risk of dementia at closest visit to time s, with
# s = 3 (wide format)

###Variables
# ISAMMSIADL.3: prediction (of dementia in the next 5 years) estimated with the
multimarkers (IST + MMSE + IADL) model at time s = 3
# VISITbeforediag : in case of dementia diagnosis, last visit before the diagnosis
# Time : minimum between time of dementia diagnosis or last follow-up
```

```
# TimeDeath : minimum between time of death or last follow-up
# dem: event indicator (dementia)
# death: death indicator (dementia)
# debut: corresponds to time s-1 (here, equal to 2)

######## Computing the cumulative risk of dementia before death ########
### Building the risk categories

ThreeCsmooth_s3$Cat <- NA
ThreeCsmooth_s3$Cat <-
ifelse(ThreeCsmooth_s3$ISAMMSIADL.3>=quantile(ThreeCsmooth_s3$ISAMMSIADL.3,p
rob=0) &
ThreeCsmooth_s3$ISAMMSIADL.3<=quantile(ThreeCsmooth_s3$ISAMMSIADL.3,prob=0
.25),1,ThreeCsmooth_s3$Cat)
ThreeCsmooth_s3$Cat <-
ifelse(ThreeCsmooth_s3$ISAMMSIADL.3>quantile(ThreeCsmooth_s3$ISAMMSIADL.3,pr
ob=0.25) &
ThreeCsmooth_s3$ISAMMSIADL.3<=quantile(ThreeCsmooth_s3$ISAMMSIADL.3,prob=0
.5),2,ThreeCsmooth_s3$Cat)
ThreeCsmooth_s3$Cat <-
ifelse(ThreeCsmooth_s3$ISAMMSIADL.3>quantile(ThreeCsmooth_s3$ISAMMSIADL.3,pr
ob=0.5) &
ThreeCsmooth_s3$ISAMMSIADL.3<=quantile(ThreeCsmooth_s3$ISAMMSIADL.3,prob=0
.75),3,ThreeCsmooth_s3$Cat)
ThreeCsmooth_s3$Cat <-
ifelse(ThreeCsmooth_s3$ISAMMSIADL.3>quantile(ThreeCsmooth_s3$ISAMMSIADL.3,pr
ob=0.75) &
ThreeCsmooth_s3$ISAMMSIADL.3<=quantile(ThreeCsmooth_s3$ISAMMSIADL.3,prob=1
),4,ThreeCsmooth_s3$Cat)
ThreeCsmooth_s3$Catfactor <- as.factor(ThreeCsmooth_s3$Cat)

### Estimating the illness-death models for each category

ThreeC_Smooth_s3_Cat4 <-
idm(formula01=Hist(time=list(VISITbeforediag,Time),event=dem, entry = debut)~1,
                formula02=Hist(time=TimeDeath,event=death)~1,
                formula12=~1,maxiter=1000,

data=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==4),method="splines",
                conf.int=FALSE)

ThreeC_Smooth_s3_Cat3 <-
idm(formula01=Hist(time=list(VISITavtdiagbis,Time),event=dem, entry = debut)~1,
                formula02=Hist(time=TimeDeath,event=death)~1,
                formula12=~1,maxiter=1000,

data=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==3),method="splines",
                conf.int=FALSE)
```

```
ThreeC_Smooth_s3_Cat2 <-
idm(formula01=Hist(time=list(VISITavtdiagbis,Time),event=dem, entry = debut)~1,
                formula02=Hist(time=TimeDeath,event=death)~1,
                formula12=~1,maxiter=1000,

data=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==2),method="splines",
                conf.int=FALSE)

ThreeC_Smooth_s3_Cat1 <-
idm(formula01=Hist(time=list(VISITavtdiagbis,Time),event=dem, entry = debut)~1,
                formula02=Hist(time=TimeDeath,event=death)~1,
                formula12=~1,maxiter=1000,

data=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==1),method="splines",
                conf.int=FALSE)

### Estimating the cumulative risk of dementia before death between s and s+5

predict(ThreeC_Smooth_s3_Cat1,s=4,t=9) # low risk group
predict(ThreeC_Smooth_s3_Cat2,s=4,t=9) # medium risk group
predict(ThreeC_Smooth_s3_Cat3,s=4,t=9) # high risk group
predict(ThreeC_Smooth_s3_Cat4,s=4,t=9) # very high risk group

######## Calibration: computing Nelson-Aalen estimators ########
#1) Estimating the models for each risk categories
fit3C_Cat4_s3 <- coxph(Surv(Time, dem)~1,
data=subset(ThreeC_Smooth_s3,ThreeC_Smooth_s3$Catfactor==4&ThreeC_Smooth_s3$
Time>3),
        x=TRUE)
fit3C_Cat3_s3 <- coxph(Surv(Time, dem)~1,
data=subset(ThreeC_Smooth_s3,ThreeC_Smooth_s3$Catfactor==3&ThreeC_Smooth_s3$
Time>3),
        x=TRUE)
fit3C_Cat2_s3 <- coxph(Surv(Time, dem)~1,
data=subset(ThreeC_Smooth_s3,ThreeC_Smooth_s3$Catfactor==2&ThreeC_Smooth_s3$
Time>3),
        x=TRUE)
fit3C_Cat1_s3 <- coxph(Surv(Time, dem)~1,
data=subset(ThreeC_Smooth_s3,ThreeC_Smooth_s3$Catfactor==1&ThreeC_Smooth_s3$
Time>3),
        x=TRUE)

#2) Computing the 5-year cumulative incidences
fit.pred3Cs3_Cat4 <- predictCox(fit3C_Cat4_s3,
newdata=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==4&ThreeCsmooth_s3
$Time>3), times=8, se = TRUE, band = TRUE)
fit.pred3Cs3_Cat3 <- predictCox(fit3C_Cat3_s3,
newdata=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==3&ThreeCsmooth_s3
$Time>3), times=8, se = TRUE, band = TRUE)
```

```
fit.pred3Cs3_Cat2 <- predictCox(fit3C_Cat2_s3,
newdata=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==2&ThreeCsmooth_s3
$Time>3), times=8, se = TRUE, band = TRUE)
fit.pred3Cs3_Cat1 <- predictCox(fit3C_Cat1_s3,
newdata=subset(ThreeCsmooth_s3,ThreeCsmooth_s3$Catfactor==1&ThreeCsmooth_s3
$Time>3), times=8, se = TRUE, band = TRUE)

#2) Obtaining the mean of the 5-year cumulative incidences of dementia in the next 5
years
1-mean(fit.pred3Cs3_Cat4$survival)
1-mean(fit.pred3Cs3_Cat3$survival)
1-mean(fit.pred3Cs3_Cat2$survival)
1-mean(fit.pred3Cs3_Cat1$survival)
```

References :

1. Rizopoulos D. The R Package JMbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *J. Stat. Softw.* 2016; 1(7) [electronic article]. (https://www.jstatsoft.org/v072/i07)

2. Blanche P, Proust-Lima C, Loubère L, et al. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*. 2015;71(1):102–113.