

Développement et Application de Méthodologies Statistiques pour Etudes Longitudinales d'Association Génétique

Comité de suivi de thèse: première année

Mickaël Canouil
mickael.canouil@cnrs.fr

Direction de thèse

Dr. Ghislain Rocheleau & Pr. Philippe Froguel

28 septembre 2015



Université
de Lille
2 DROIT
ET SANTÉ



Sommaire

1 Introduction

2 Objectifs

3 Matériels

4 Méthodes

5 Résultats préliminaires

6 Simulations

7 Perspectives

8 Références

Introduction

En 2014, la prévalence de diabète de type 2 (T2D) a été estimée à près de 9% chez l'adulte de 18 ans et plus.

Sur la dernière décennie, l'essor des études d'association pangénomiques (GWAS) a permis l'identification de :

- 65 variants associés à la susceptibilité au T2D ;
- 36 variants associés à la glycémie à jeun (FG) chez les normoglycémiques.

Introduction

La grande majorité des **GWAS** a utilisé un design transversal, quand un design longitudinal offre la possibilité :

- de décrire la trajectoire temporelle d'une variable ;
- d'accroître la puissance pour détecter des variants génétiques associés à la trajectoire.

La modélisation de ces trajectoires temporelles optimiserait les tests d'association et l'exploitation des phénotypes disponibles.

Objectifs

Cette thèse s'organise sur deux principaux objectifs :

- 1 Développement et implémentation des approches basées notamment sur les modèles joints ;
- 2 Application à un jeu de données (p.ex. cohorte **DESIR**, **FRAMINGHAM**, etc) ;
- 3 Optimisation du temps de calcul avec R (portage Julia).

Matériels

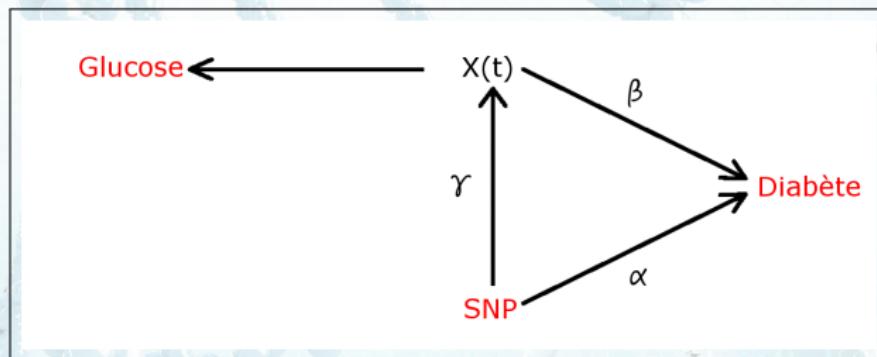
Le laboratoire (**UMR CNRS 8199**) dispose de l'accès à la cohorte prospective **DESIR** (Données Epidémiologiques sur le Syndrome d'Insulino-Résistance), comptant **5 214** individus suivis pendant **9** ans, tous les **3 ans** (0, 3, 6 et 9 ans).

En plus de données phénotypiques (p.ex. **FG**, **hba1c**, etc), des données génotypiques sont également disponibles pour une grande partie de ces individus.

Cette cohorte comporte **187** cas incidents de **T2D**, définis à partir d'une glycémie supérieure à **7 mM/L** ou par la prise d'un traitement anti-diabétique.

Méthodes : modèle joint

L'approche par modèle joint a été décrite par [Tsiatis and Davidian \[2004\]](#) et [Ibrahim et al. \[2010\]](#)



Deux paquets R implémentent cette approche sous deux angles :

- **JM** [[Rizopoulos, 2010](#)] : intérêt sur la composante de survie ;
- **joineR** [[Philipson et al., 2012](#)] : intérêt sur la composante longitudinale.

Méthodes : modèle joint

Le modèle joint se décompose en deux parties :

- Composante longitudinale (Modèle linéaire mixte)

$$Y_{ij} = X_{ij} + \epsilon_{ij} \quad (1)$$

$$X_{ij} = \theta_{0i} + \theta_{1i} \times t_{ij} + \gamma \times SNP_i \quad (2)$$

- Composante de survie (Modèle de Cox)

$$\lambda_i(t) = \lambda_0(t) \exp(\beta X_i(t) + \alpha SNP_i) \quad (3)$$

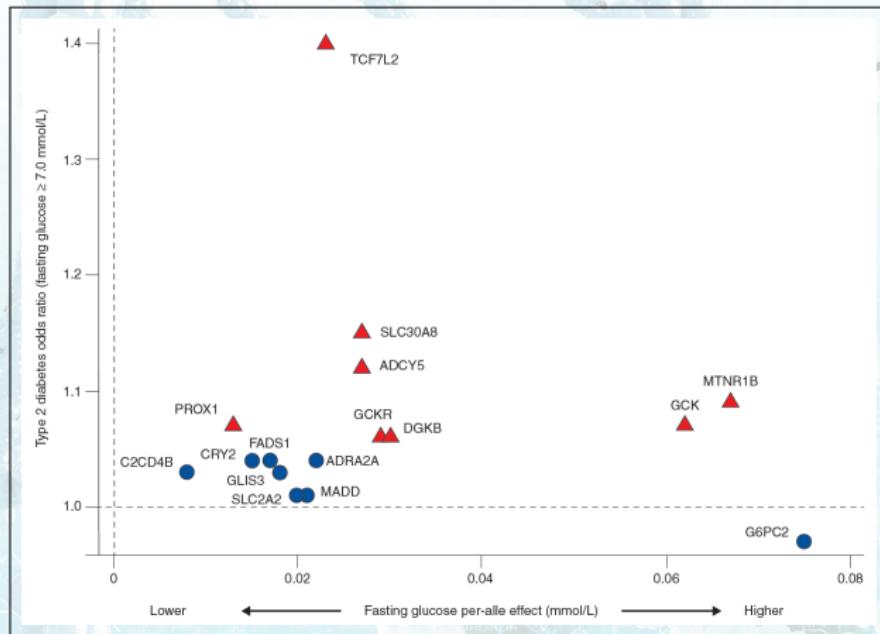
Méthodes : modèle "Two-Stage"

L'approche dite "Two-Stage" (TS) [Self and Pawitan, 1992] consiste en la succession de deux étapes :

- Etape 1** Ajustement d'un modèle linéaire mixte pour estimer la "vraie" trajectoire.
- Etape 2** Ajustement d'un modèle de survie incorporant la trajectoire estimée de l'étape 1 en tant que covariable dépendante du temps.

Résultats préliminaires : modèle joint

L'approche du paquet JM a été utilisé pour analyser une sélection de SNPs, basée sur leur association au T2D ou FG.



Résultats préliminaires : modèle joint

Les résultats d'association entre le **FG** et les **SNPs** sont confirmés avec l'approche **JM** ($\gamma \neq 0$).

	γ	α	β
rs7903146 (TCF7L2)	0.02465	0.2204	3.477
rs3802177 (SLC30A8)	0.038	0.01066	3.542
rs10278336 (GCK)	0.0383	0.09214	3.527
rs560887 (G6PC2)	0.09504	-0.3237	3.568
rs780094 (GCKR)	0.06271	-0.09694	3.568
rs10830963 (MTNR1B)	0.0959	-0.3868	3.611
rs11717195 (ADCY5)	0.02581	-0.1202	3.554

Résultats préliminaires : modèle "Two-Stage"

L'approche TS donne des résultats similaires à ceux obtenus par JM.

	γ	α	β
rs7903146 (TCF7L2)	-	0.317	3.356
rs3802177 (SLC30A8)	-	0.1291	3.375
rs10278336 (GCK)	-	-0.1373	3.376
rs560887 (G6PC2)	-	-0.2725	3.432
rs780094 (GCKR)	-	-0.04382	3.404
rs10830963 (MTNR1B)	-	-0.2683	3.456
rs11717195 (ADCY5)	-	-0.08409	3.405

Simulations : paramètres

Pour identifier les avantages et limites des approches de type modèle joint (**JM** et **joineR**), des simulations ont été réalisées avec R, en suivant les **Equations 1 à 3** et pour jeu de paramètres :

Paramètres	Valeurs
Effectif (N)	5000
Temps de mesures (en années)	0, 3, 6, 9
Incidence à neuf ans (I)	5%
LMM : Trajectoire $\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$	$\mathcal{N}_2 \left(\begin{bmatrix} 4.50 \\ 0.013 \end{bmatrix}, \begin{bmatrix} 0.16 & 0 \\ 0 & 1 \times 10^{-3} \end{bmatrix} \right)$
LMM : Effet du SNP (γ)	0.025
Cox : Effet du SNP (α)	0.2
JM : Effet de la trajectoire (β)	3.50

La fonction de risque de base λ_0 a été fixée pour une incidence de 5%.

Simulations : scénarios

Plusieurs scénarios de simulation ont été réalisés pour tester la robustesse des estimations des paramètres :

- 1 Données complètes et variation de la fréquence allélique ;
- 2 Données complètes et variation du nombre de mesures longitudinales ;
- 3 Données complètes et variation de l'effectif ;

Simulations : scénarios

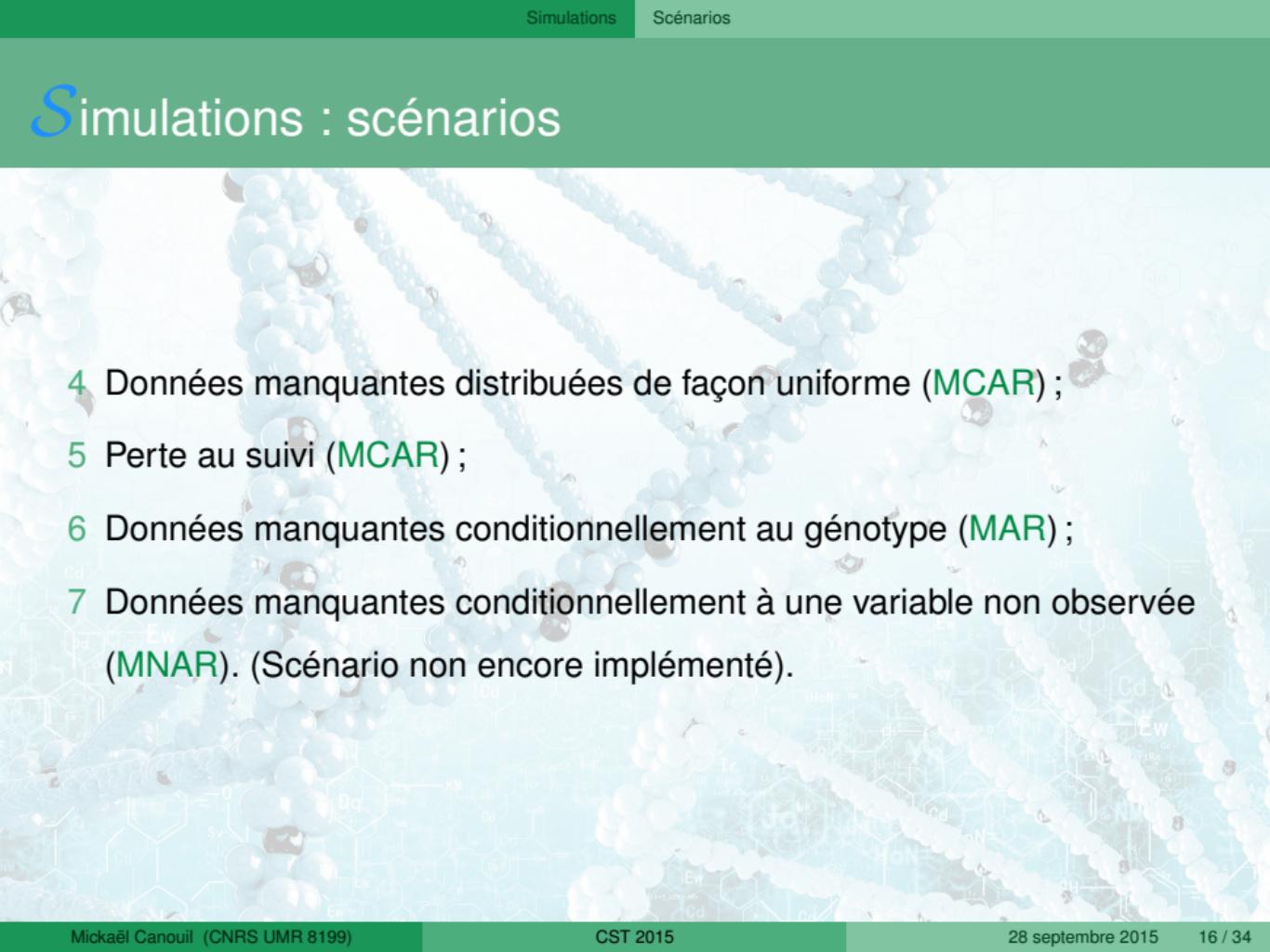
En présence de données manquantes, selon la classification usuelle :

MCAR (missing completely at random) : les données sont manquantes indépendamment des données observées et non observées ;

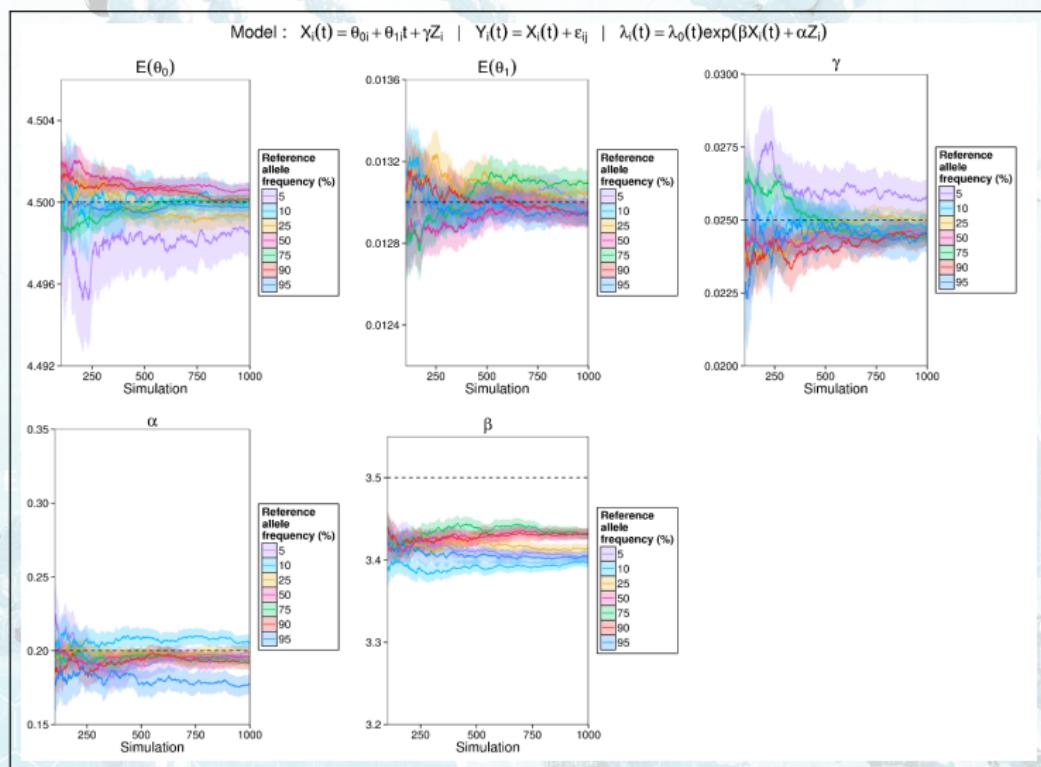
MAR (missing at random) : conditionnellement aux données observées, les données manquantes sont indépendantes des données non observées ;

MNAR (missing not at random) : les données manquantes sont dépendantes de variables non observées.

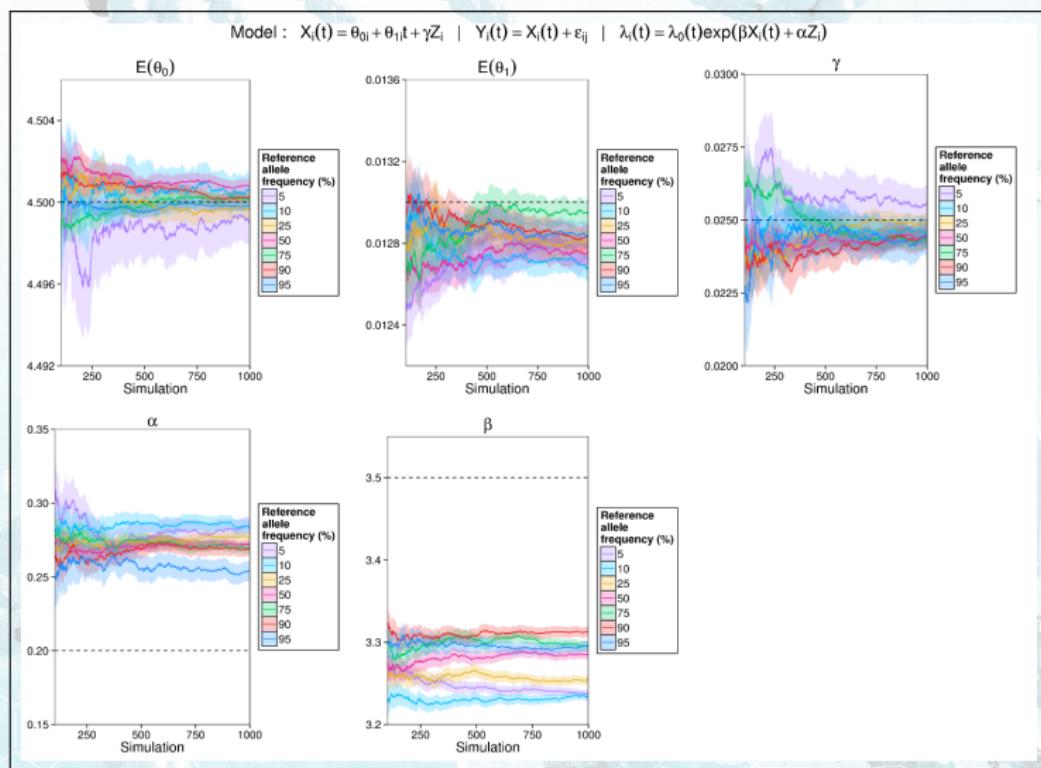
Simulations : scénarios

- 
- 4 Données manquantes distribuées de façon uniforme (**MCAR**) ;
 - 5 Perte au suivi (**MCAR**) ;
 - 6 Données manquantes conditionnellement au génotype (**MAR**) ;
 - 7 Données manquantes conditionnellement à une variable non observée (**MNAR**). (Scénario non encore implémenté).

Scénario 1 : JM et fréquence allélique



Scénario 1 : joineR et fréquence allélique



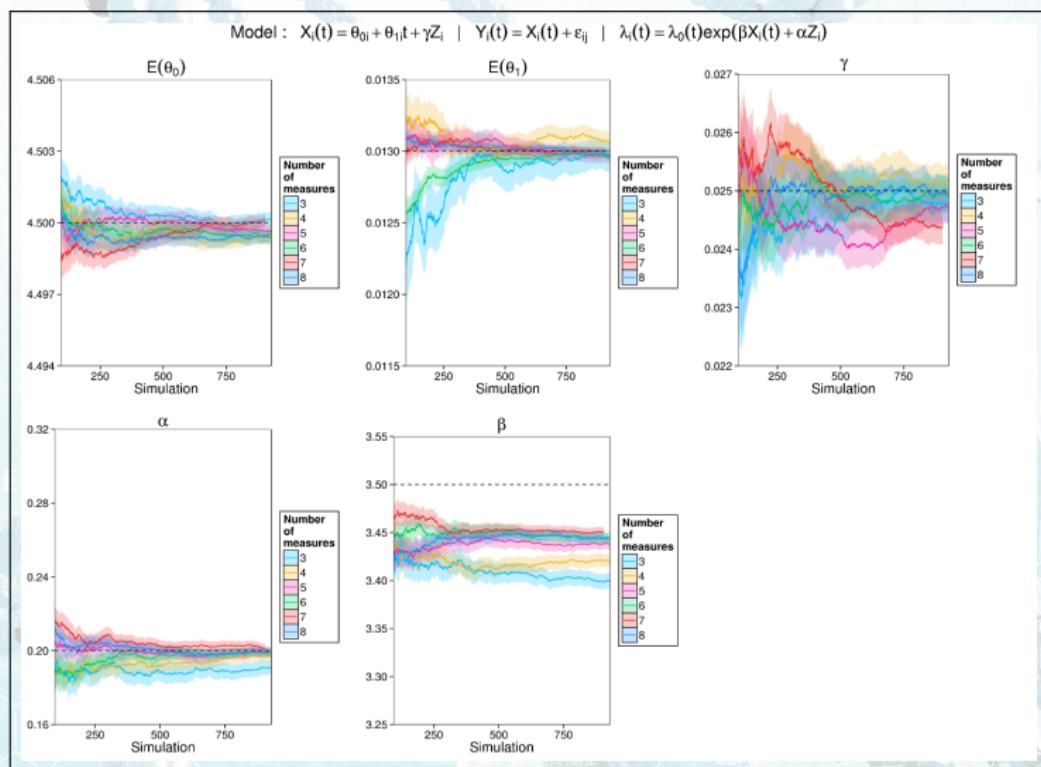
Scénario 1 : JM et joineR

Les résultats des estimations de ce scénario révèlent que l'implémentation **joineR** serait plus biaisée que **JM**. Notamment, pour les paramètres de la composante de survie (α et β).

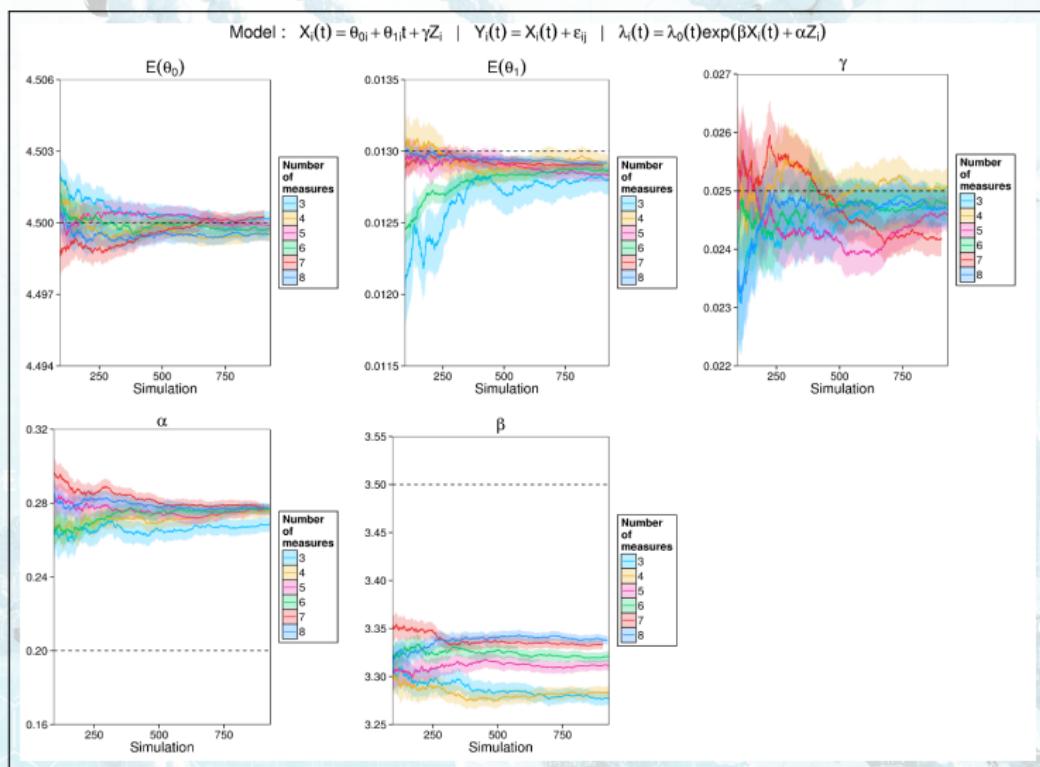
Ces résultats sont à confirmer sur un plus grand nombre de simulations et sur d'autre scénarios.

La variation de la fréquence allélique influe sur les estimations des paramètres, essentiellement pour une fréquence allélique faible (< 5%).

Scénario 2 : JM et nombre de mesures



Scénario 2 : joineR et nombre de mesures

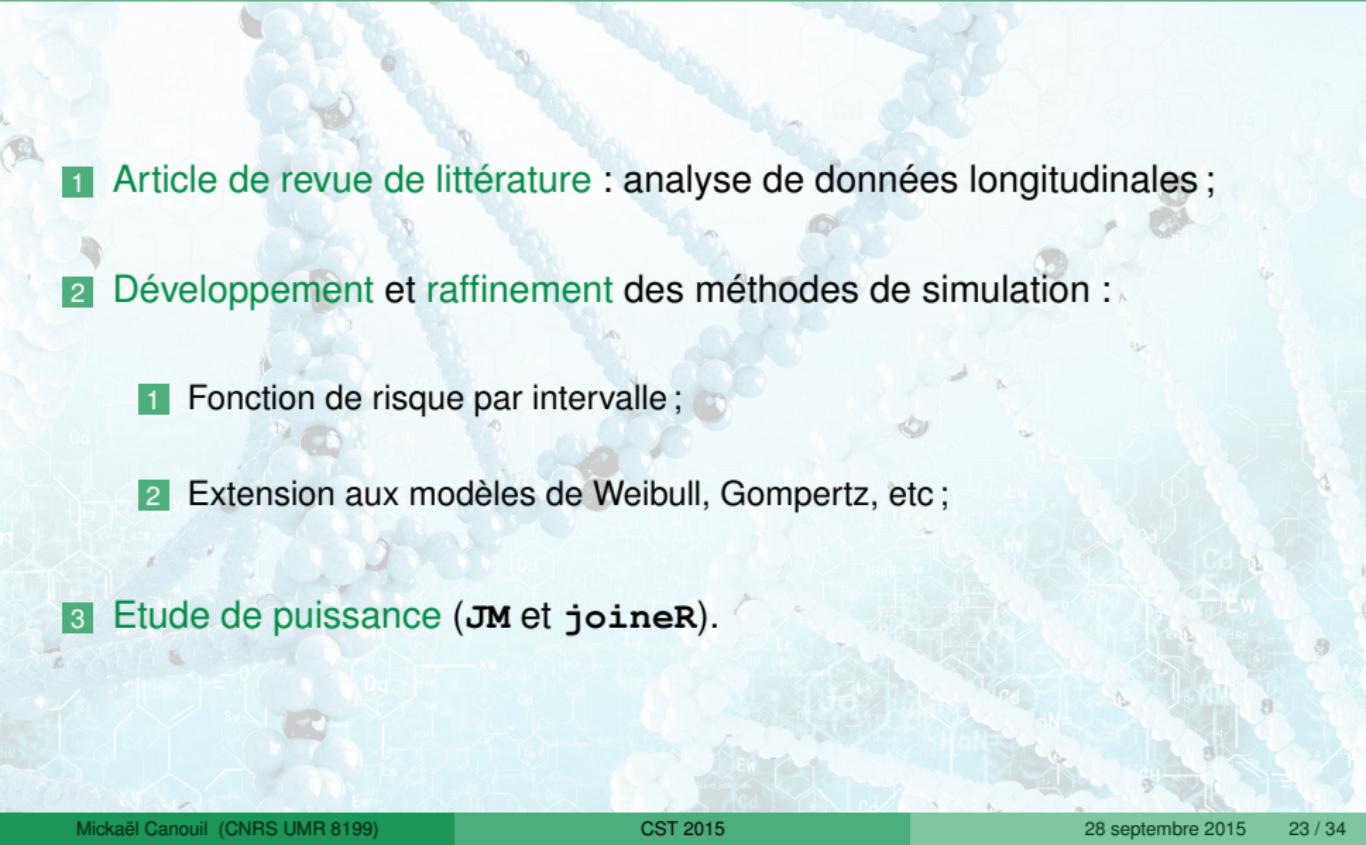


Scénario 2 : JM et joineR

Le biais des estimations de **joineR** reste plus important que celui de **JM**, même si celui-ci se réduit avec l'augmentation du nombre de mesures longitudinales.

Cependant, le biais d'estimation de α varie très peu selon le nombre de mesures pour **joineR**.

Perspectives : en cours / à court terme

- 
- 1 Article de revue de littérature : analyse de données longitudinales ;
 - 2 Développement et raffinement des méthodes de simulation :
 - 1 Fonction de risque par intervalle ;
 - 2 Extension aux modèles de Weibull, Gompertz, etc ;
 - 3 Etude de puissance (**JM** et **joineR**).

Perspectives : à long terme

- 
- 1 Application à la cohorte DESIR.
 - 2 Développement et implémentation de solutions :
 - pour réduire le temps de calcul ;
 - pour réduire le biais des estimations.

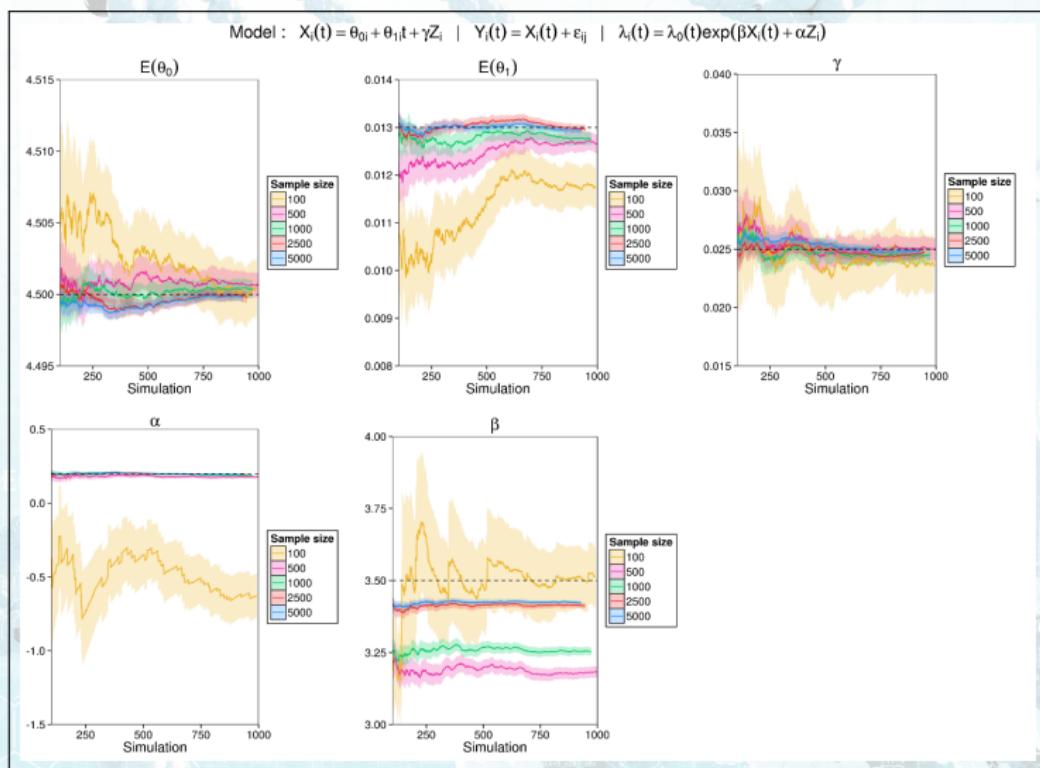
Références I

- Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal of Clinical Oncology*, 28(16) :2796–2801.
- Philipson, P., Diggle, P., Sousa, I., Kolamunnage-Dona, R., Williamson, P., and Henderson, R. (2012). joineR : Joint modelling of repeated measurements and time-to-event data.
- Rizopoulos, D. (2010). JM : An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9) :1–33.

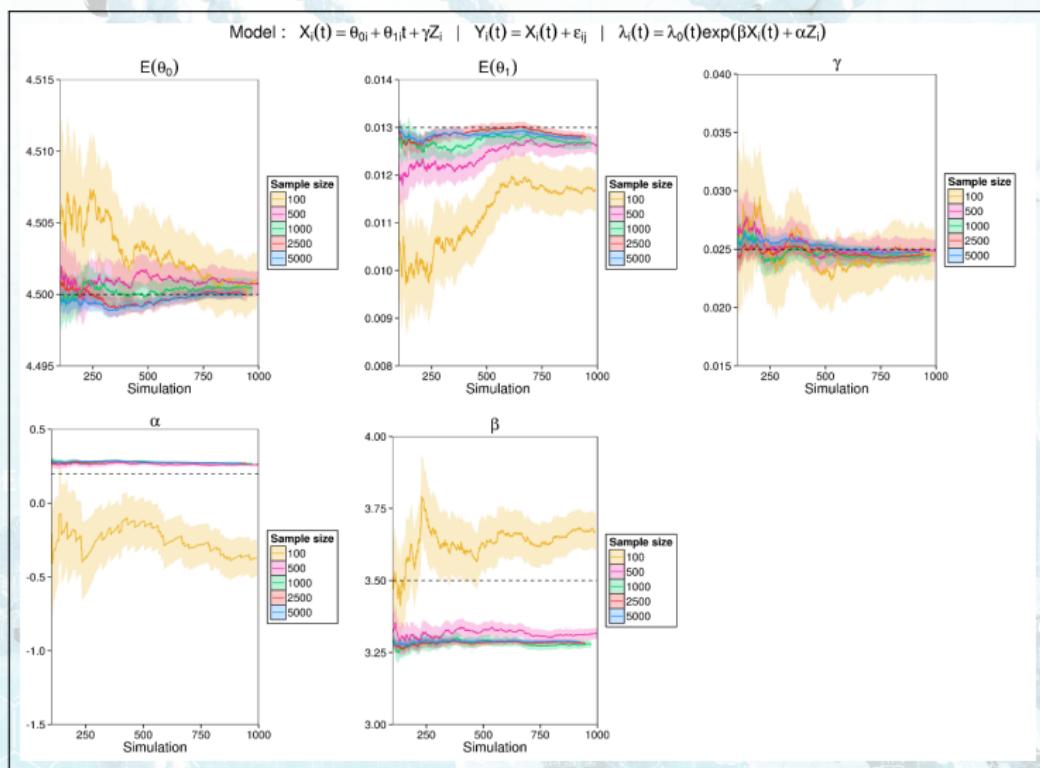
Références II

- Self, S. and Pawitan, Y. (1992). Modeling a Marker of Disease Progression and Onset of Disease. In Jewell, N. P., Dietz, K., and Farewell, V. T., editors, *AIDS Epidemiology*, pages 231–255. Birkhäuser Boston.
- Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data : an overview. *Statistica Sinica*, 14 :809–834.

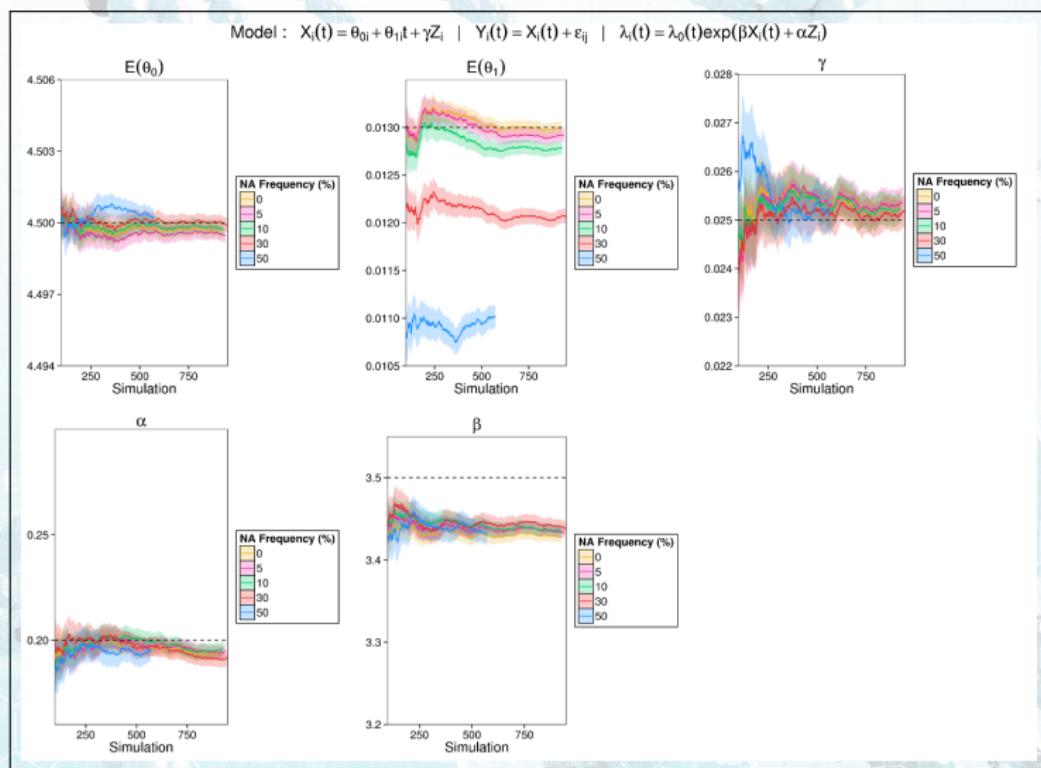
Scénario 3 : JM et effectif



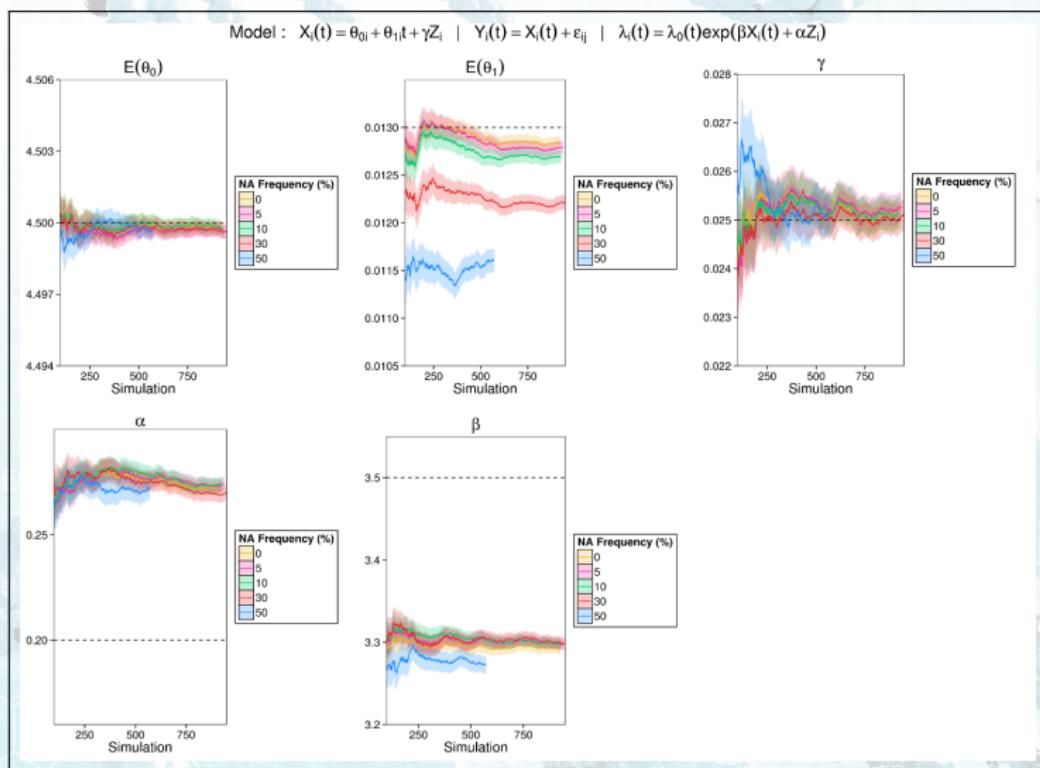
Scénario 3 : joineR et effectif



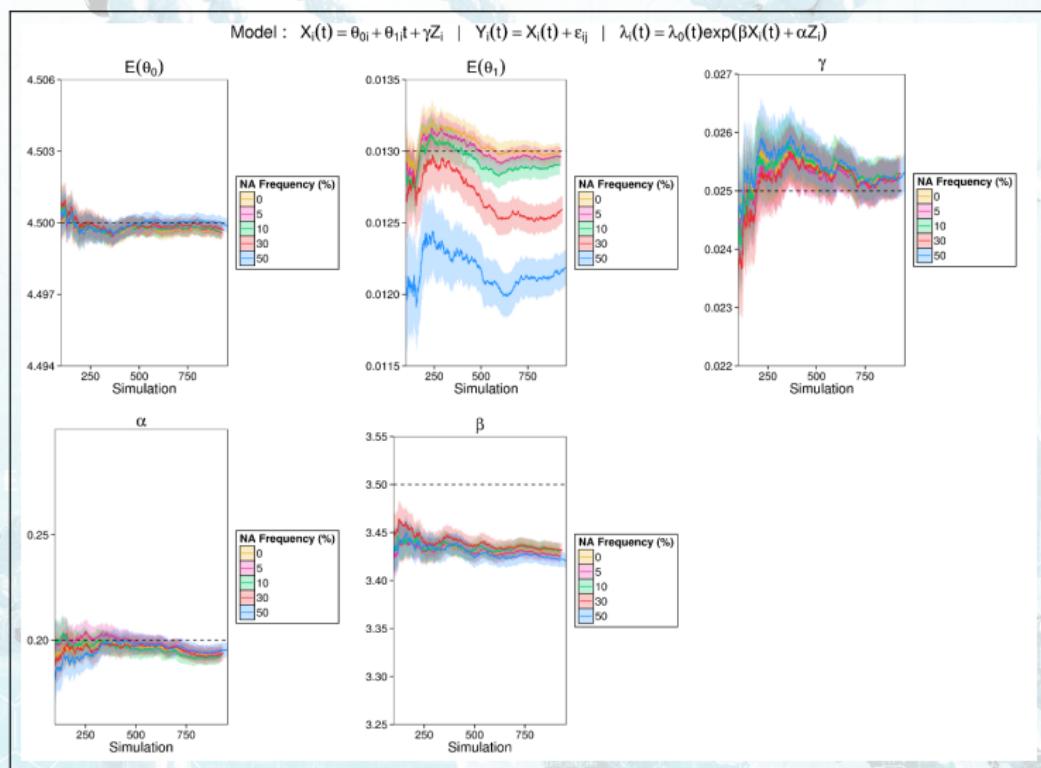
Scénario 4 : JM et NA uniforme



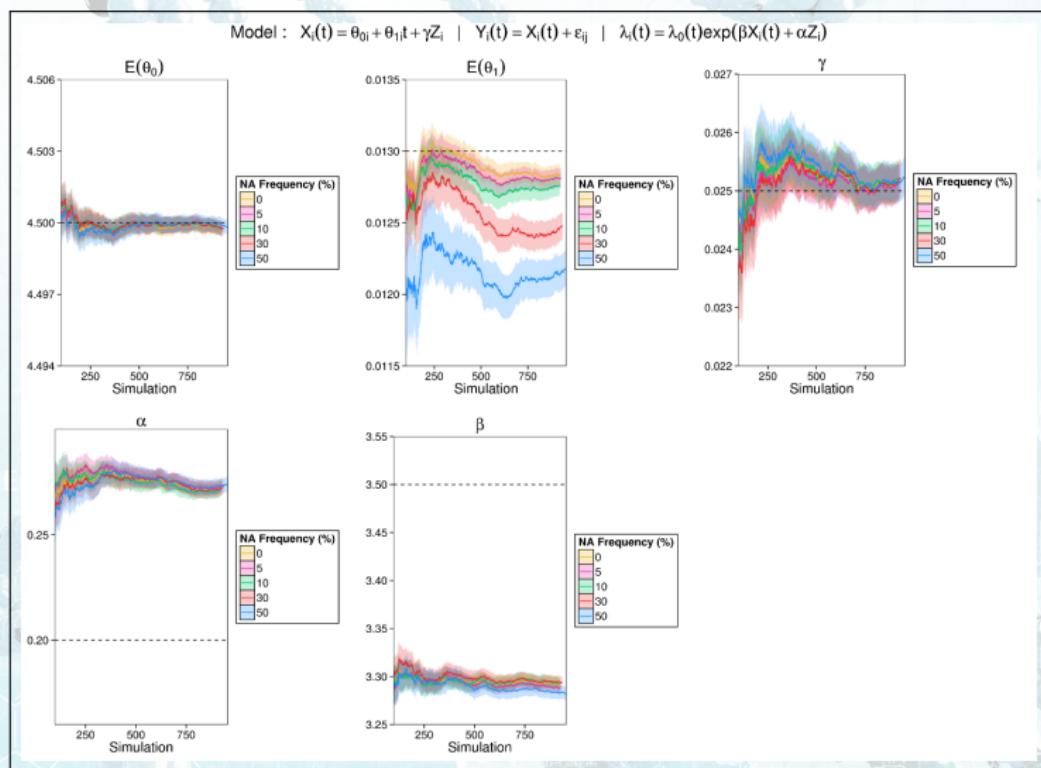
Scénario 4 : joineR et NA uniforme



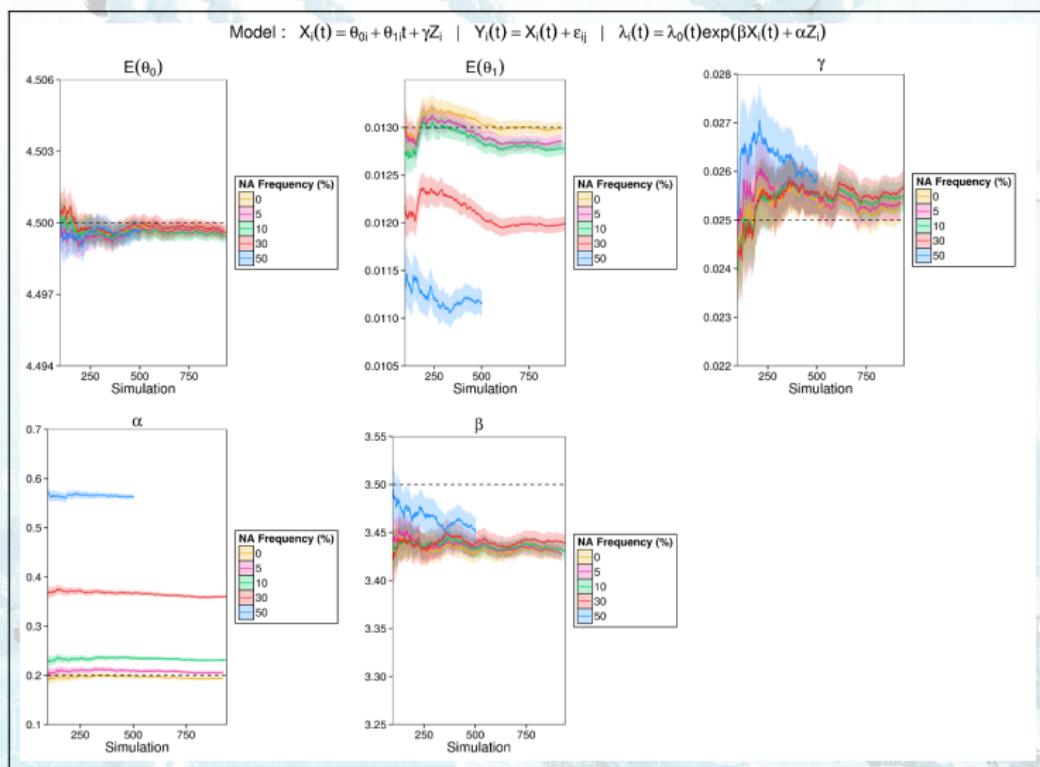
Scénario 5 : JM et perte au suivi



Scénario 5 : joineR et perte au suivi



Scénario 6 : JM et NA/génotypes



Scénario 6 : joineR et NA/génotypes

