1 Travaux précédents

1.1 Modèle Joint

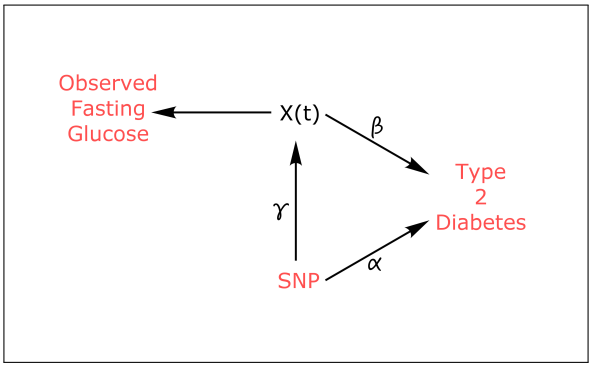

Figure 1: Diagramme général d'un modèle joint pour le T2D (adapté de **?**)

$X(t)$: trajectoire de la glycémie à jeun inférée des données longitudinales observées ; α : effet du SNP sur le diabète ; γ : effet du SNP sur la trajectoire de la glycémie à jeun ; β : effet de la trajectoire de la glycémie à jeun sur le diabète.

En utilisant, l'approche de modèle joint implémentée dans l'extension JM [**?**] du logiciel R (version 3.2.3)[**?**], 124 095 SNPs de la MetaboChip ont été testés simultanément pour leur association avec la glycémie à jeun et le risque de DT2.

La formulation standard du modèle joint implique deux composantes, d'une part, une composante longitudinale pour modéliser la trajectoire de la variable étudiée, et d'autre part, une composante de survie pour modéliser la survenue de l'événement étudié. La composante longitudinale consiste typiquement à l'application d'un modèle linéaire mixte :

$$Y_{ij} = X_{ij} + \epsilon_{ij}, \tag{1}$$

où Y_{ij} est la valeur observée et X_{ij} la vraie (non-observée) valeur de la variable longitudinale. Le terme ϵ_{ij} est le terme d'erreur aléatoire supposé être distribué selon la Loi Normale :

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

La quantité X_{ij} (ou $X(t)$) est la fonction de la trajectoire et est définie usuellement comme une fonction linéaire (ou quadratique) du temps t.

Des covariables peuvent être incluses dans la fonction de la trajectoire, comme l'âge, le sexe ou l'IMC. Dans notre étude, Y_{ij} représente les valeurs mesurées la glycémie à jeun au temps t_{ij}, Z_i désigne le génotype du SNP analysé pour l'individu i et W_i désigne les covariables selon le modèle suivant :

$$Y_{ij} = X_{ij} + \epsilon_{ij} = \theta_{0i} + \theta_{1i} \times t_{ij} + \gamma \times Z_i + \delta \times W_i \epsilon_{ij} \tag{3}$$

Pour simplifier, le terme $\delta \times W_i$ sera omis dans la suite.

Les paramètres θ_{0i} et θ_{1i} sont supposés être distribués selon une distribution normale multivariée :

$$\boldsymbol{\theta} \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4}$$

Le paramètre γ évalue l'effet additif du SNP (Z_i) sur la fonction de la trajectoire. Pour tenir compte éventuellement de pentes variant entre les génotypes, un terme d'interaction entre le SNP et le temps peut être inclus dans la fonction de la trajectoire. Le terme d'interaction n'a pas été inclus dans notre étude.

La composante de survie (survenue du DT2) se compose généralement d'un modèle paramétrique (p.ex. exponentielle ou Weibull) ou semi-paramétrique (p.ex. risques proportionnels de Cox) avec :

$$h_i(t) = h_0(t)exp(\beta X_i(t) + \alpha Z_i), \tag{5}$$

où $h_i(t)$ est la fonction de risque au temps t pour l'individu i et $h_0(t)$ est la fonction de risque de base non spécifiée, supposée être une constante par morceaux avec deux noeuds placés à des temps intermédiaires (c'est-à-dire, à trois et six ans sur les neuf ans du suivi). Le coefficient α mesure l'effet du SNP sur le délai d'apparition du DT2, alors que le coefficient β mesure l'association entre la trajectoire du niveau de la glycémie à jeun et le temps d'apparition du DT2.

1.2 Simulation

Des études de simulation ont été menées pour examiner la puissance statistique et l'erreur de type 1 des SNPs trouvés comme nominalement associé (à 5%) en utilisant le modèle joint, comme implémenté par **?**. Notre objectif principal était de déterminer le gain ou la perte de puissance du modèle joint par rapport aux approches classiques transversales (p.ex. régression logistique ou linéaire, modèle de Cox) pour détecter l'effet d'un SNP, sur la glycémie à jeun et le statut DT2 dans notre étude. Les jeux de données de simulation, qui ont été réalisés avec R, suivent les Equations 1 à 5, avec la fonction de risque de base fixée ($h_0(t) = \lambda$) de façon à obtenir une incidence équivalente à celle de la cohorte D.E.S.I.R (environ 5%), durant la période de suivi de neuf ans. Les temps d'événements ont été générés selon une distribution exponentielle dans le cadre du modèle de Cox à risque proportionnel [**?**].

$$H(T) = \int_0^T \lambda exp(\beta \times X(t) + \alpha \times Z)dt \tag{6}$$

$$T = \frac{1}{\beta\theta_1} log\left(-\frac{\beta\theta_1 \times log(1-u)}{\lambda exp(\beta\theta_0 + (\beta\gamma + \alpha)Z)} + 1\right) \tag{7}$$

1.3 Etudes des estimateurs du modèle joint par simulation

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

Table 1: Paramètres initiaux pour la simulation des données. Caractéristiques basées sur le SNP de TCF7L2 (SNP le plus fortement associé au DT2).

Dans le but d'identifier les avantages et limites des approches de type modèle joint (`JM`), un jeu de données a été simulé sur la base de la cohorte D.E.S.I.R. et les paramètres donnés dans la Table 1.

Plusieurs scénarios de simulation ont été réalisés pour tester la robustesse des estimations des paramètres en présence de données manquantes, en utilisant 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.

D'autres paramètres ont également été étudiés, tels que l'effectif de la population, la fréquence du marqueur génétique et dans le cas plus général des LMM, le nombre de mesures longitudinales. Ces scénarios ont été étudiés avec les paquets `JM`.

Les scénarios adoptés et étudiés sont les suivants :

Scénario 1 Données complètes et variation de la fréquence allélique ;

Scénario 2 Données complètes et variation du nombre de mesures longitudinales ;

Scénario 3 Données complètes et variation de l'effectif ;

Scénario 4 Données manquantes distribuées de façon uniforme (MCAR) ;

Scénario 5 Perte au suivi (MCAR).

La Figure 2 et la Table 2 montrent les résultats obtenus pour 10 000 jeux de données simulées (Equations 1 à 5) en faisant varier la fréquence allélique du marqueur de 5% à 95% (ici les résultats sont symétriques puisque le marqueur est bi-allélique). Pour ce scénario, les estimations du modèle joint se montrent proches des valeurs des paramètres simulés, avec cependant une légère sous-estimation des paramètres relatifs (α et γ) au marqueur génétique testé. Ceci est d'autant plus vrai lorsque la fréquence de l'allèle mineur diminue.

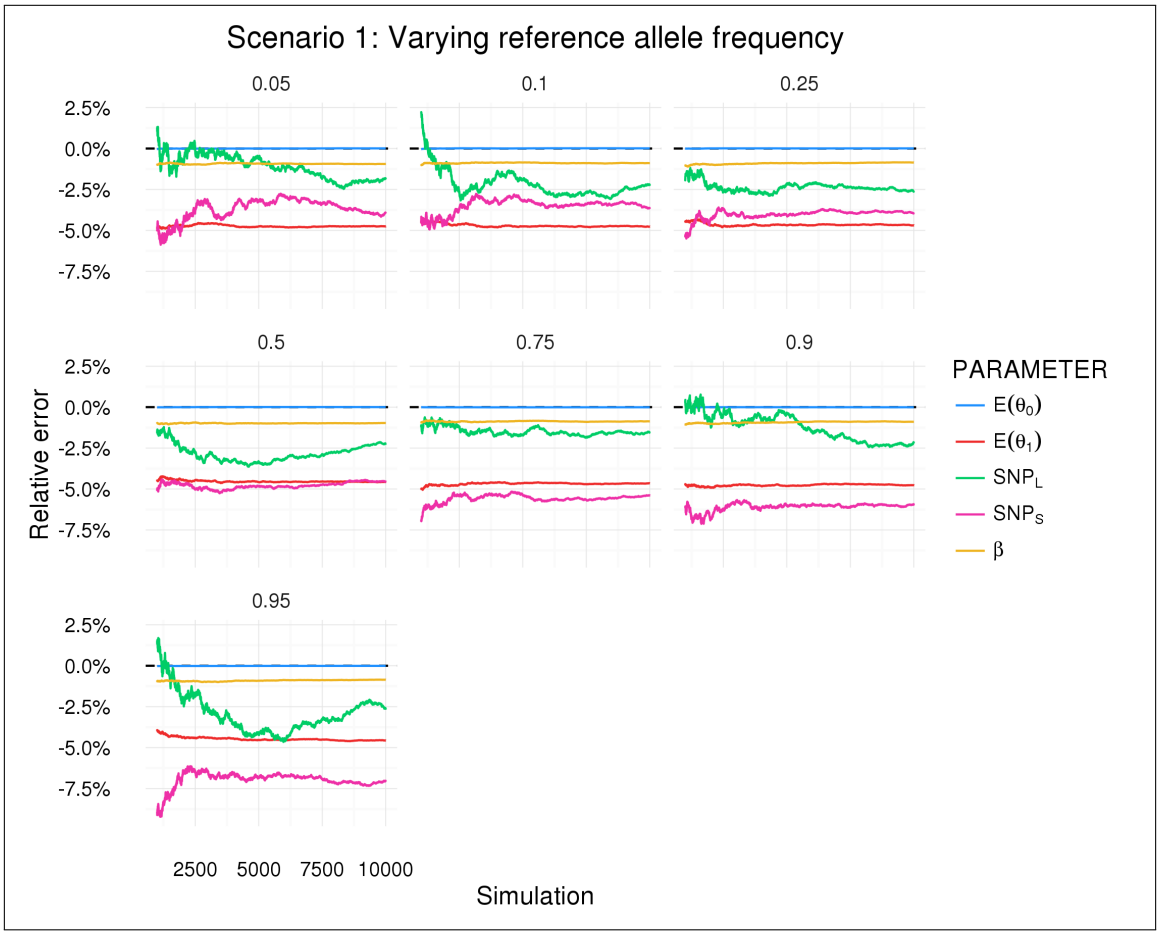

Figure 2: Estimation de l'erreur relative des paramètres du modèle joint en fonction de la fréquence allélique (10 000 simulations).

Le même type de résultats a pu être observé pour les autres scénarios (données non montrées), à savoir :

Scénario 2 sous-estimations (α, β et γ) réduites avec l'augmentation du nombre de mesures, notamment pour β ;

Scénario 3 sous-estimations (α, β et γ) réduites avec l'augmentation du nombre d'individus, le gain reste cependant faible au-delà de 1 000 individus ;

Scénario 4-5 le taux de données manquantes impacte les estimations de façon importante au-delà de 10%, notamment sur la composante longitudinale du modèle joint.

Paramètre	Estimée	Simulée	Fréquence allélique
α	0.215 [-0.0762, 0.551]	0.23	0.05
	0.219 [0.00346, 0.459]		0.10
	0.22 [0.0741, 0.373]		0.25
	0.219 [0.101, 0.343]		0.50
	0.219 [0.0825, 0.349]		0.75
	0.218 [0.0284, 0.398]		0.90
	0.218 [-0.0538, 0.461]		0.95
β	3.56 [3.29, 3.85]	3.60	0.05
	3.57 [3.3, 3.85]		0.10
	3.57 [3.3, 3.86]		0.25
	3.56 [3.29, 3.85]		0.50
	3.57 [3.29, 3.85]		0.75
	3.57 [3.29, 3.85]		0.90
	3.57 [3.29, 3.85]		0.95
γ	0.0196 [-0.0164, 0.0558]	0.02	0.05
	0.0195 [-0.00712, 0.0456]		0.10
	0.0194 [0.00111, 0.038]		0.25
	0.0197 [0.00322, 0.0353]		0.50
	0.0196 [0.00115, 0.0385]		0.75
	0.0196 [-0.00678, 0.0457]		0.90
	0.0195 [-0.0165, 0.0558]		0.95

Table 2: Estimation des paramètres du modèle joint en fonction de la fréquence allélique (10 000 simulations).

1.4 Modèles pour données longitudinales

Nous avons comparé plusieurs approches, d'une part, pour tester l'effet principal d'un SNP (β_g), et d'autre part pour tester l'effet d'interaction entre le SNP et le temps. Pour le premier, nous avons comparé cinq méthodes : les modèles de régression linéaires en utilisant les mesures à l'inclusion dans la cohorte D.E.S.I.R. ou en utilisant la moyenne de l'ensemble des mesures du suivi, l'approche "Two-Step" avec l'ordonnée à l'origine en terme aléatoire (RI), les équations estimantes généralisées (GEE) et les modèles linéaires mixtes (LMM) ; tandis que pour le dernier, nous avons comparé l'approche "Two-Step" avec une pente aléatoire, "Two-Step conditionnelle", GEE et LMM avec terme d'interaction.

Y_i est la variable mesurée et G_i désigne le génotype (SNP) codés 0, 1 ou 2.

Modèle linéaire (à l'inclusion)

$$Y_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ où } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Modèle linéaire (Moyenne des m mesures)

$$\bar{Y}_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ où } \epsilon_i \sim \mathcal{N}(0, \frac{\sigma^2}{m})$$

"Two-Step"

1. $Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \epsilon_{ij}$, où $\epsilon_{ij} \sim \mathcal{N}_m(0, V^\dagger{}_i \equiv Z^\dagger{}_i D^\dagger Z^{\dagger\prime}_i + \sigma^{\dagger 2} I_m)$
2. $\hat{b}_{0i} = \beta_0^* + \beta_g^* G_i + \epsilon_i^*$, où $\epsilon_i^* \sim \mathcal{N}(0, \sigma^{*2})$

Modèle linéaire mixte

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \beta_g G_i + \epsilon_{ij}, \text{ où } \epsilon_{ij} \sim \mathcal{N}_m(0, V_i \equiv Z_i D Z_i' + \sigma^2 I_m)$$

Equations d'Estimation Généralisée (GEE)

$\mathbb{E}(Y_i) = \beta_0 + \beta_1 t_{ij} + \beta_g G_i$ et $\mathbb{V}(Y_i) = V_i$ (Compound Symmetry).

L'approche "Two-Step", consiste dans un premier temps, à l'application d'un modèle linéaire mixte sans le génotype G_i, puis dans un second temps, à l'application d'un modèle linéaire sur l'ordonnée à l'origine (aléatoire) en incluant cette fois, le génotype dans les variables explicatives.

L'erreur de type I et la puissance statistique, de l'effet du SNP, ont été obtenus via des procédures de ré-échantillonnages et de permutations appliquées sur le jeu de données complet, ou par simulations dans le cas d'un où un terme d'interaction était inclus. Dans tous les modèles testés, l'erreur de type I était maintenu au seuil de 5%. En revanche, une puissance statistique accrue de l'approche classique (modèle de régression linéaire) par rapport

aux méthodes prenant en compte des mesures répétées, a pu dans certains cas être mis en évidence, en dépit du fait que la prise en compte de mesures répétées apporte un gain de puissance.

Ce comportement "contre-intuitif" peut s'observer et s'expliquer au niveau du paramètre de décentralité (NCP) de ces modèles, à l'aide des formules closes suivantes, pour tester l'association d'un SNP sous l'approche transversale (CS : Cross-Sectional), avec l'ordonnée à l'origine en aléatoire (RI : Random Intercept) et pente et ordonnée à l'origine aléatoire (RIS : Random Intercept and Slope) :

$$NCP_{CS} = nd^2 \left(\frac{2p(1-p)}{\sigma^2}\right) \tag{8}$$

$$NCP_{RI} = NCP_{CS} \left(\frac{m\sigma^2}{\sigma^2+m\sigma^2_{b0}}\right) \tag{9}$$

$$NCP_{RIS} = NCP_{RI}U \tag{10}$$

$$\text{avec} \quad U = \frac{(\sigma^2+\sigma^2_{b1}\sum_{j=1}^{m}(t_j-\bar{t})^2)(\sigma^2+m\sigma^2_{b0})}{(\sigma^2+\sigma^2_{b1}\sum_{j=1}^{m}(t_j-\bar{t})^2)(\sigma^2+m\sigma^2_{b0})-m\rho^2\sigma^2_{b0}\sigma^2_{b1}\sum_{j=1}^{m}(t_j-\bar{t})^2} \geq 1 \tag{11}$$

Où d est la taille d'effet ;

n est la taille d'échantillon (nombre d'individus) ;

m est le nombre de mesures répétées ;

σ^2 est la variance des résidus ;

σ^2_{b0} et σ^2_{b1} sont les variances des paramètres en effet aléatoire b_{0i} et b_{1i} ;

ρ est le coefficient de corrélation entre b_{0i} et b_{1i}.

Nous pouvons écrire le NCP_{RIS} comme le produit du NCP sous le modèle RI et un facteur U supérieur ou égal à un (Equations 10 à 11). Cela garantit que le NCP sous le modèle RIS est toujours supérieur ou égal au NCP du modèle RI, mais n'implique pas que NCP_{RIS} est supérieur au NCP_{CS} dans l'approche transversale.

$$NCP_{RIS} \geq NCP_{RI} \nRightarrow NCP_{RIS} > NCP_{CS} \tag{12}$$

Les résultats des calculs des NCPs montrés dans la Table 3 concordent avec notre étude de puissance, où une analyse transversale pouvait être plus puissante à détecter un effet qu'une approche prenant en compte des mesures répétées (Figure 3).

SNP	$Gene$	NCP_{CS}		NCP_{RIS} (NCP_{RI})
rs560887	G6PC2	63.33	<	93.08 (92.69)
rs2908289	GCK	17.02	<	26.37 (26.26)
rs16913693	IKBKAP	12.75	>	6.55 (6.53)
rs6072275	TOP1	7.78	>	6.78 (6.76)

Table 3: Calcul du paramètre de décentralité pour une sélection de SNPs associés au glucose sanguin.

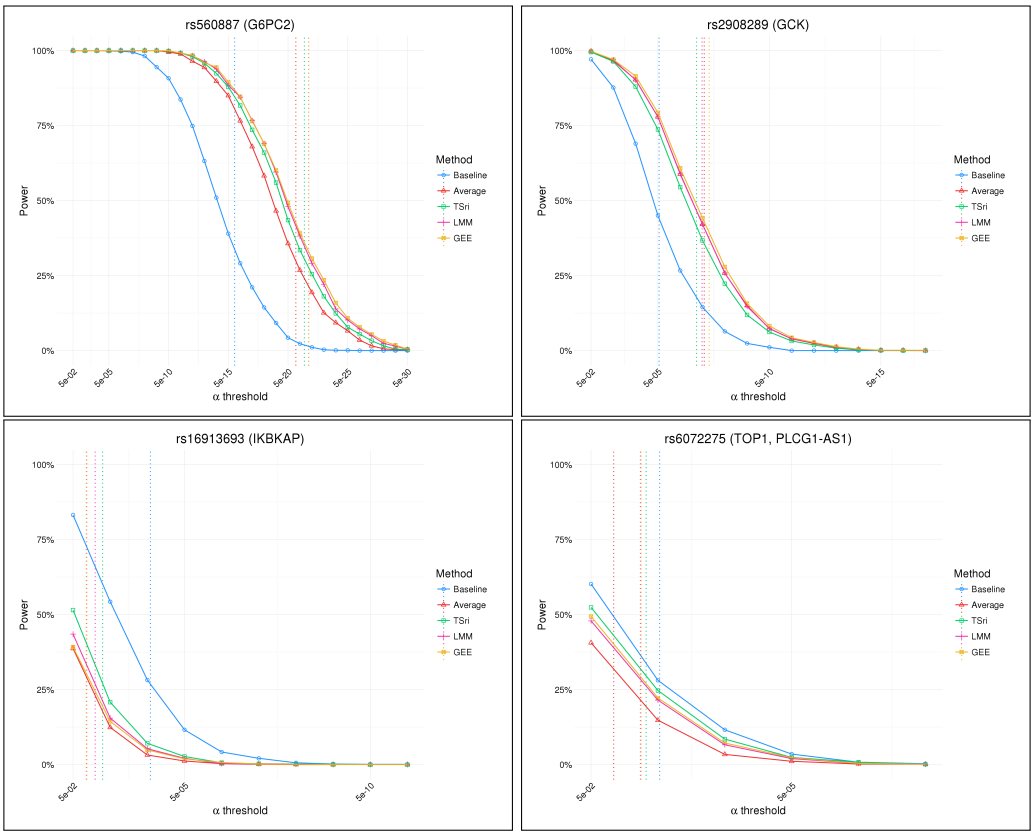

Figure 3: Calcul de puissance pour une sélection de SNPs associés au glucose sanguin.