Modèles pour l'analyse longitudinale de la qualité de vie en présence de données manquantes non ignorables

Célia Touraine

Séminaire Qualité de Vie 2018 (5ème édition), Montpellier

6-7 septembre 2018

celia.touraine@icm.unicancer.fr

ICM
Institut régional du Cancer
Montpellier | Val d'Aurelle

Contents

Qualité de vie relative à la santé (QdV)

Intérêt grandissant pour la QdV dans les essais cliniques en oncologie

- Critère de jugement
- En complément de la survie

Une notion complexe

- Subjective
- Multidimensionnelle
- Dynamique

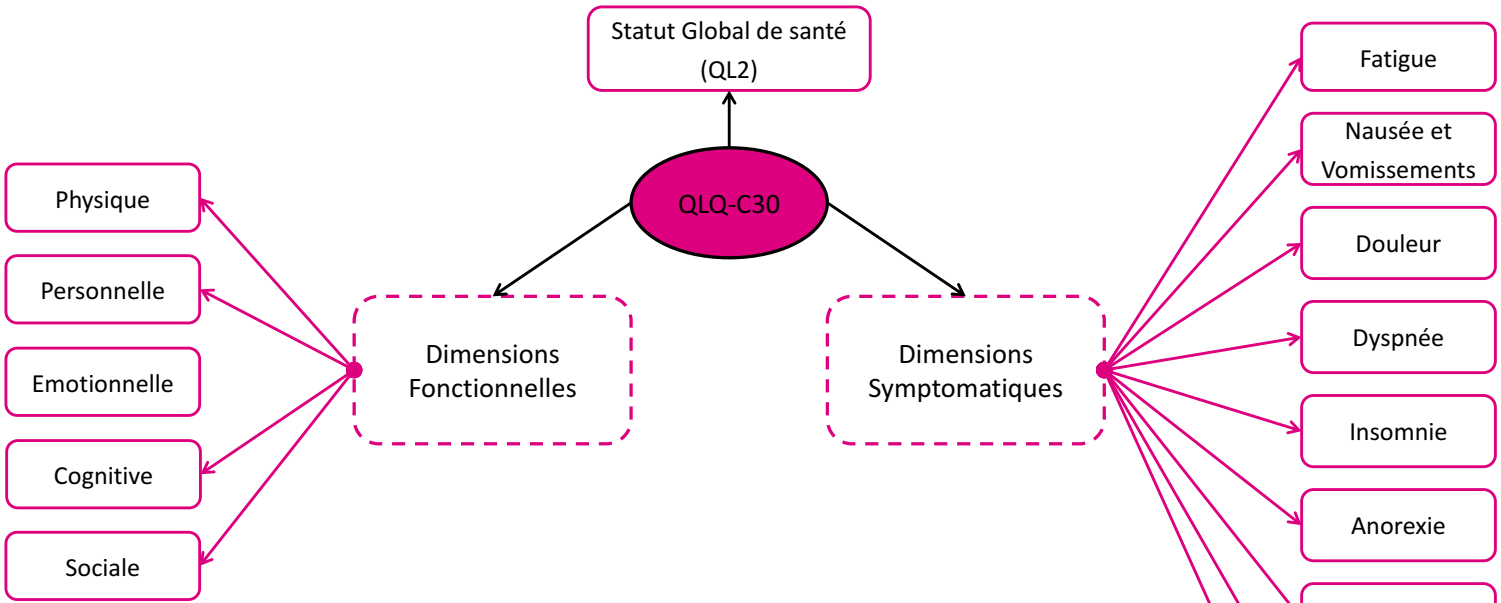
Évaluée à l'aide de questionnaires

- Auto-administrés
- Items regroupés en échelles/dimensions
- Plusieurs évaluation au cours du temps

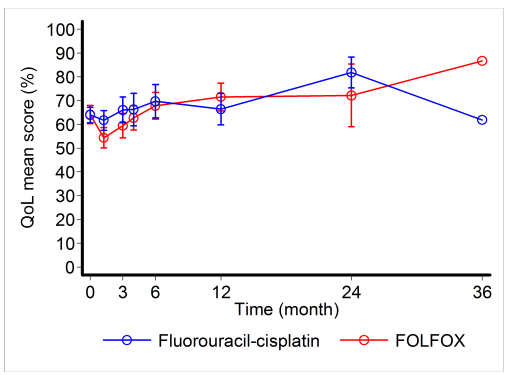
Variable d'intérêt

- Qualitative (ex. : pas du tout / un peu / assez / beaucoup)
- Quantitative (ex. : score de 0 à 100)

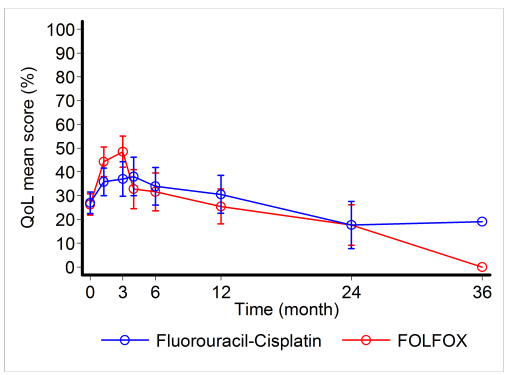
Exemple du questionnaire EORTC QLQ-C30



Statut de santé global



Fatigue



Analyse longitudinale du score de QdV

Modèles linéaires mixtes

Le score de QdV du patient i à l'évaluation j s'écrit :

$$Y_{ij} = \beta^T X_{ij} + b_i{}^T Z_{ij} + \epsilon_{ij}$$

où

- Les effets fixes β caractérisent la trajectoire moyenne du score de QdV
- Les effets aléatoires b_i
  - Permettent de tenir compte de la corrélation des évaluations répétées sur un même patient
  - Correspondent aux déviations spécifiques du patient i par rapport à la trajectoire moyenne

Problèmatique

- Paramètres estimés potentiellement **biaisés en présence de données manquantes**
- Données manquantes fréquentes

Analyse longitudinale du score de QdV

Modèles linéaires mixtes

Le score de QdV du patient i à l'évaluation j s'écrit :

$$Y_{ij} = \beta^T X_{ij} + b_i{}^T Z_{ij} + \epsilon_{ij}$$

où

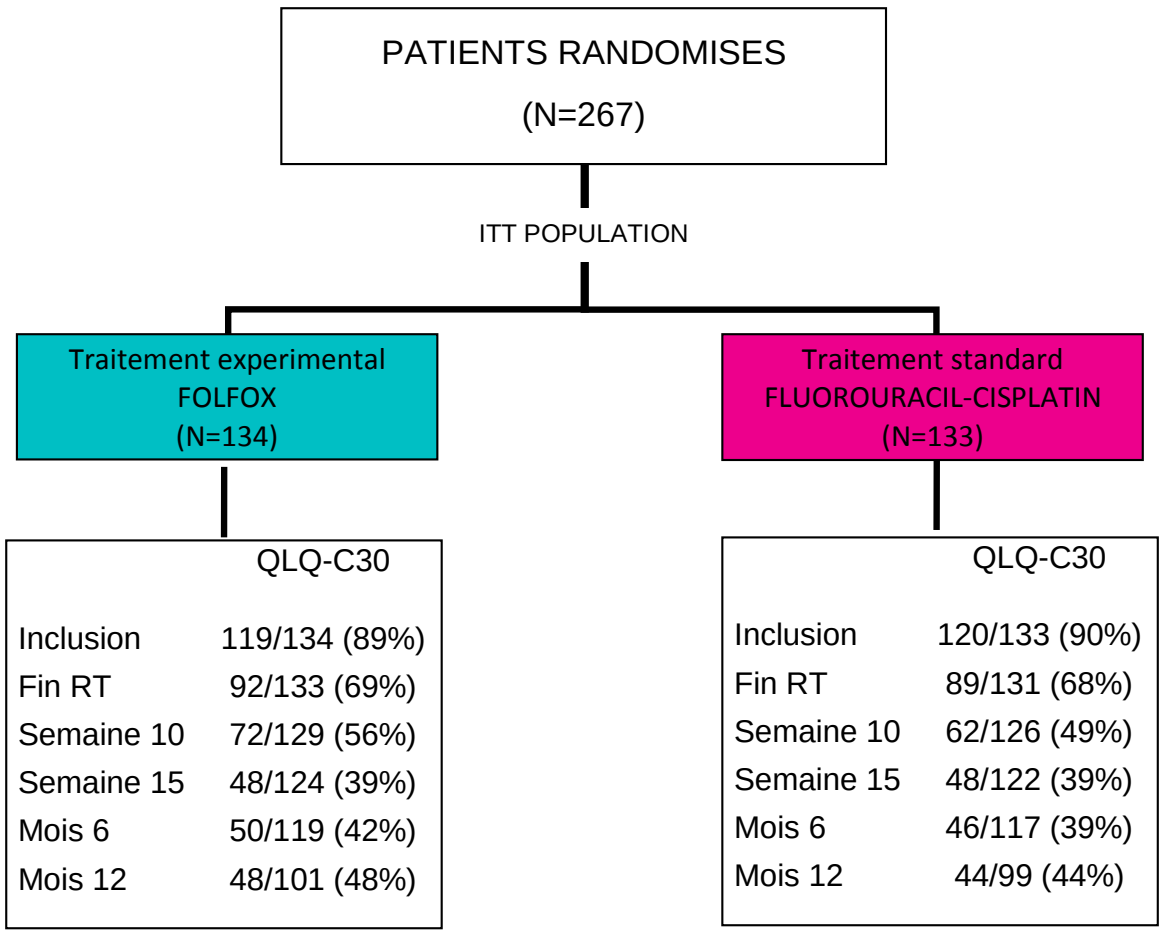
- Les effets fixes β caractérisent la trajectoire moyenne du score de QdV
- Les effets aléatoires b_i
  - Permettent de tenir compte de la corrélation des évaluations répétées sur un même patient
  - Correspondent aux déviations spécifiques du patient i par rapport à la trajectoire moyenne

Problèmatique

- Paramètres estimés potentiellement **biaisés en présence de données manquantes**
- Données manquantes fréquentes

Essai clinique randomisé PRODIGE5/ACCORD17

- Cancer de l'œsophage localement avancé
- Critère principal : Survie sans progression (pas de différence significative)
- Critère secondaire : QdV



Objectif de la présentation

Comment réaliser une analyse longitudinale du score de QdV en présence de données manquantes ?

- Description des 2 types de modèles
- Illustration sur l'essai clinique PRODIGE5/ACCORD17

Contents

Typologie (Rubin, 1987)

Les données manquantes peuvent être :

- MCAR (*missing completely at random*)
  - Manquantes complètement au hasard
  - Mécanisme de non-réponse **indépendant du score de QdV**
  - Pas de problème de biais
- MAR (*missing at random*)
  - Manquantes au hasard
  - Mécanisme de non-réponse **dépendant du score de QdV observé**
  - Biais évitable (si méthode basée sur la vraisemblance)
  - Données manquantes ***ignorables***
- MNAR (*missing not at random*)
  - Manquantes non au hasard
  - Mécanisme de non-réponse **dépendant du score de QdV *non observé***
  - Biais si on ne modélise pas simultanément le score de QdV et le mécanisme de non-réponse
  - Données manquantes ***informatives***

Répartition/Structure

Les données manquantes peuvent être :

- Manquantes par intermittence

Patient	Visite 1	Visite 2	Visite 3	Visite 4	Visite 5
1	75.0	83.3	×	75.0	66.7
2	66.7	×	66.7	50.0	50.0
3	33.3	41.7	×	×	33.3
4	×	75.0	×	83.3	50.0

- De structure **monotone** : sortie d'étude prématurée (*drop-out*)

Patient	Visite 1	Visite 2	Visite 3	Visite 4	Visite 5
1	66.7	66.7	×	×	×
2	83.3	100	66.7	66.7	×
3	50.0	×	×	×	×
4	91.7	25.0	25.0	×	×

Prise en compte des sorties d'études informatives

Prise en compte des données manquantes

- de structure monotone (sorties d'études)
- de type MNAR (informatives)

En présence de sorties d'étude informative, les méthodes d'analyse longitudinale du score de QdV produisent des estimations biaisées si on ne modélise pas simultanément :

$$\begin{cases} \text{le score de QdV} \\ \text{la variable de sortie d'étude} \end{cases}$$

à travers la distribution conjointe de (Y_i, D_i) où :

- Y_i est le score de QdV du patient i
- D_i est la variable de sortie d'étude (*drop-out*) du patient i
  - $D_i = j$ si le patient i est non répondant après la $j^{\text{ème}}$ évaluation
  - $D_i =$ nombre d'évaluation si pas de sortie d'étude

Distribution conjointe de (Y_i, D_i)

La distribution du couple (Y_i, D_i) peut se décomposer selon la distribution marginale de Y_i et la distribution conditionnelle de D_i :

$$P_{\theta\psi}(Y_i, D_i) = P_\theta(Y_i) \times P_\psi(D_i \mid Y_i)$$

où θ et ψ sont les paramètres associés au score de QdV et à la sortie d'étude

La probabilité de sortie d'étude au temps d'évaluation j se décline selon que les données manquantes :

- sont MCAR : $P_\psi(D_i = j \mid Y_i) = \pi$
- sont MAR : $P_\psi(D_i = j \mid Y_i) = P_\psi(D_i = j \mid Y_i^o)$
  où on note $Y_i = (Y_i^o, Y_i^m)$ le vecteur des données complètes avec Y_i^o la composante observée et Y_i^m la composante manquante
- sont MNAR : $P_\psi(D_i = j \mid Y_i)$ ne se simplifie pas

Rq : Les 4 types de données manquantes sont emboîtés

Contents

Définition

Modèle de sélection (*Selection Models*)

Modèle basé sur la décomposition de la distribution de (Y_i, D_i) selon la distribution marginale de Y_i et la distribution conditionnelle de D_i sachant Y_i :

$$P_{\theta\psi}(Y_i, D_i) = P_\theta(Y_i) \times P_\psi(D_i \mid Y_i)$$

où θ est le vecteur de paramètres du modèle pour le score de QdV et ψ est vecteur de paramètres du modèle pour la sortie d'étude

Modèle de Diggle et Kenward [1] :

- Modèle linéaire mixte pour Y_i :

$$Y_{ij} = \beta^T X_{ij} + b_i{}^T Z_{ij} + \epsilon_{ij}$$

avec $Y_i \sim \mathcal{N}(X_i\beta, V_i)$

- Régression logistique pour $D_i \mid Y_i$:

$$\text{logit}(\mathbb{P}(D_i = j \mid Y_i)) = \psi_0 + \psi_1 Y_{ij} + \psi_2 Y_{i,j+1}$$

1. Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied statistics*, 49-93

Estimation

- La distribution conjointe de (Y_i, D_i) s'écrit :

$$P_{\theta\psi}(Y_i, D_i) = P_\theta(Y_i) \times P_\psi(D_i \mid Y_i)$$

où $P_\psi(D_i \mid Y_i)$ dépend des données observées *et* manquantes du score de QdV

- L'écriture de la vraisemblance implique une marginalisation sur les données manquantes Y_i^m :

$$P_{\theta\psi}(Y_i, D_i) = \int P_\theta(Y_i^o, Y_i^m) \times P_\psi(D_i \mid Y_i^o, Y_i^m) \, \mathrm{d}Y_i^m$$

où l'hypothèse sur la distribution de Y_i : « $Y_i \sim \mathcal{N}(X_i\beta, V_i)$ » permet de déduire la distribution de $Y_i^m \mid Y_i^o$.

Illustration : PRODIGE5/ACCORD17 (statut de santé global)

- Modèle linéaire mixte pour le score Y_i :

$$Y_i(t) = \underbrace{\beta_0 + \beta_1 \times t + \beta_2 \times \text{Bras}_i + \beta_3 \text{Bras}_i \times t}_{\textbf{Fixe} : \text{effets moyens}} + \underbrace{b_{0i} + b_{1i} \times t}_{\textbf{Aléatoire} : \text{effets individuels}} + \epsilon_{it}$$

où $\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$ avec $\Sigma = \begin{pmatrix} \sigma^2_{b_0} & \sigma_{b_0 b_1} \\ \sigma_{b_0 b_1} & \sigma^2_{b_1} \end{pmatrix}$ et $\epsilon_{it} \sim \mathcal{N}(0, \sigma^2)$

- Régression logistique pour $D_i \mid Y_i$:

$$\text{logit}(\mathbb{P}(D_i = j \mid Y_i)) = \psi_0 + \psi_1 Y_{ij} + \psi_2 Y_{i,j+1}$$

où $j = 1, 2, 3, 4, 5$ (fin RT, J1C6 CT, M4 suivi, M6 suivi, M12 suivi)

- Résultats :

	Modèle de sélection			Modèle linéaire mixte		
	Estimation	SE	p	Estimation	SE	p
β_1	0.239	0.711	0.74	0.239	0.254	0.35
β_3	0.604	0.635	0.34	0.604	0.360	0.09
ψ_2	-0.001	0.005	0.89	–	–	–
ψ_1	0.000	0.009	1.00	–	–	–

Avantages et inconvénients

Avantage

Il est formellement possible de tester l'hypothèse de non-ignorabilité des données manquantes :

$$\text{logit}(\mathbb{P}(D_i = j \mid Y_i)) = \psi_0 + \psi_1 Y_{ij} + \psi_2 Y_{i,j+1}$$

- Données manquantes MNAR : « $\psi_2 \neq 0$ »
- Données manquantes MAR : « $\psi_2 = 0$ »
- Données manquantes MCAR : « $\psi_1 = \psi_2 = 0$ »

Inconvénient

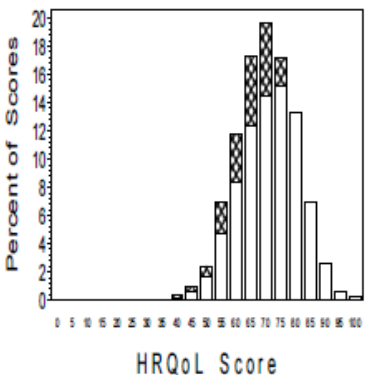
Le modèles de sélection reposent sur une hypothèse de normalité des données complètes Y_i (données observées et manquantes du score de QdV) :

- Hypothèse par définition non vérifiable
- Il a été montré qu'ils y sont particulièrement sensibles

Illustration de l'écart à la normalité

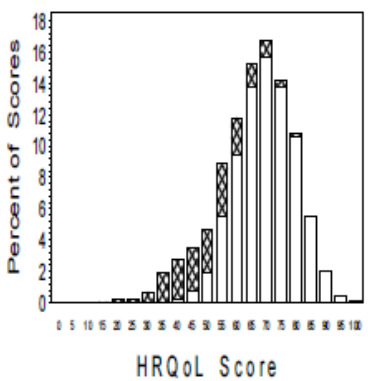
Les 2 figures ci-dessous [2] représentent la distribution du score de QdV

- En hachuré (resp. blanc) : les données manquantes (resp. observées)
- Situation où les sorties d'étude sont corrélées à un score bas de QdV



- Hypothèse de normalité des données complètes : OK
- Pas de normalité des données observées

$\Rightarrow$ Le modèle va bien trouver une dépendance entre score de QdV et sortie d'étude ($\psi_2 \neq 0$)



- Hypothèse de normalité des données complètes : ~~OK~~
- Normalité des données observées

$\Rightarrow$ Le modèle ne va pas détecter la dépendance entre le score de QdV et la sortie d'étude ($\psi_2 = 0$)

2. Source : Fairclough, D. L. (2010). *Design and analysis of quality of life studies in Clinical Trials*

Contents

Définition / Estimation

Il existe une autre décomposition possible de la distribution conjointe du score de QdV et de la variable de sortie d'étude

Modèle de mélange de profils (*Pattern Mixture Models*)

Modèle basé sur la décomposition de la distribution de (Y_i, D_i) selon la distribution marginale de D_i et la distribution conditionnelle de Y_i sachant D_i :

$$P_{\theta\psi}(Y_i, D_i) = P_\psi(D_i) \times P_\theta(Y_i \mid D_i)$$

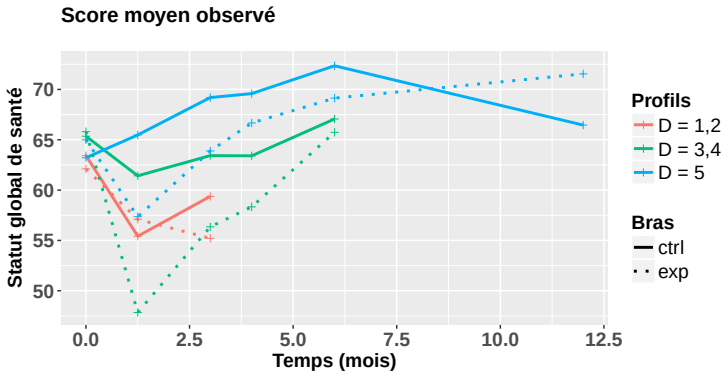
Deux modèles séparés :

- Distribution multinomiale pour D_i : la probabilité d'appartenance au profil « sortie d'étude à la $j^{\text{ème}}$ évaluation » $\mathbb{P}(D_i = j)$ est estimée par la fréquence observée π_j
- Modèles linéaires mixtes pour $Y_i^o \mid D_i$

Remarque : La vraisemblance ne dépend que des données observées

Illustration : PRODIGE5/ACCORD17 (statut de santé global)

- Détermination de 3 profils et estimation des probabilités d'appartenance :



$\mathbb{P}(D = 1, 2)$ estimée par $\pi_{1,2} = 86/253 = 0.34$

$\mathbb{P}(D = 3, 4)$ estimée par $\pi_{3,4} = 75/253 = 0.30$

$\mathbb{P}(D = 5)$ estimée par $\pi_5 = 92/253 = 0.36$

*(1=fin RT, 2=J1C6 CT, 3=M4 suivi, 4=M6 suivi, 5=M12 suivi)

- Modèles linéaires mixtes pour $Y_i \mid D_i$:

$$Y_i(t \mid D_i = 1, 2) = \beta_0 + \beta_1 \times t + \beta_2 \times \text{Bras}_i + \beta_3 \text{Bras}_i \times t + b_{0i} + b_{1i} \times t + \epsilon_{it}$$
$$Y_i(t \mid D_i = 3, 4) = \beta_0 + \beta_1 \times t + \beta_2 \times \text{Bras}_i + \beta_3 \text{Bras}_i \times t + b_{0i} + b_{1i} \times t + \epsilon_{it}$$
$$Y_i(t \mid D_i = 5) = \beta_0 + \beta_1 \times t + \beta_2 \times \text{Bras}_i + \beta_3 \text{Bras}_i \times t + b_{0i} + b_{1i} \times t + \epsilon_{it}$$

- Résultats :

	Modèle de mélange						Modèle linéaire mixte	
	$D = 1, 2$		$D = 3, 4$		$D = 5$			
	Est. (SE)	p	Est. (SE)	p	Est. (SE)	p	Est. (SE)	p
β_1	-2.93 (2.18)	0.180	0.17 (0.74)	0.822	0.17 (0.28)	0.552	0.24 (0.25)	0.347
β_3	-0.15 (2.76)	0.957	-0.97 (1.09)	0.374	0.78 (0.40)	0.052	0.60 (0.36)	0.093

Avantages et inconvénients

Avantages

- Vraisemblance basée sur les données observées
- Facile à mettre en œuvre : Modélisation habituelle sur chaque sous-échantillon correspondant à un profil + Calcul de fréquence

Inconvénient

Si l'on veut faire de l'inférence sur la distribution marginale de Y_i (pas conditionnellement à D_i), on doit extrapoler

- Paramètres non identifiables $\Rightarrow$ Restrictions (ex. : évolution linéaire et $D_i = 1 \Rightarrow$ effet temps le même que lorsque $D_i = 2$)
- Extrapolation de l'évolution du score après sortie d'étude

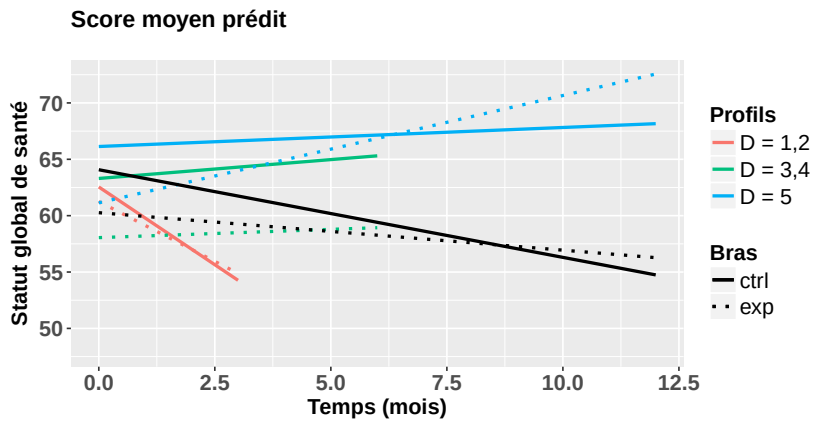
Illustration de l'extrapolation

Estimations relatives à la distribution marginale de Y_i (moyennes pondérées des estimations relatives aux distributions conditionnelles) :

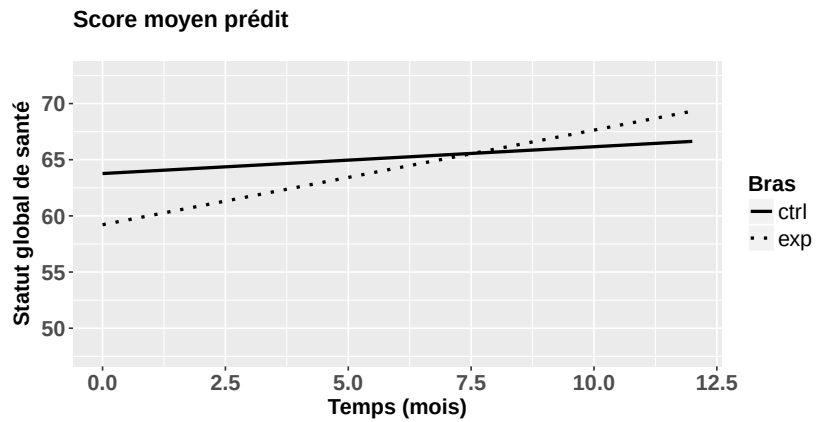
$$\hat{\beta} = \pi_{1,2}\hat{\beta}^{(1,2)} + \pi_{3,4}\hat{\beta}^{(3,4)} + \pi_5\hat{\beta}^{(5)}$$

$\Rightarrow$ Construction d'un modèle pour la distribution marginale de Y_i :

Modèle de mélange de profils



Modèle linéaire mixte



- Extrapolation des valeurs des paramètres après sortie d'étude
- Poids des extrapolations d'autant plus important que les fréquences associées au profils de sortie d'étude précoces sont importantes

Contents

En résumé

Lorsque la sortie d'étude est informative (dépend de la QdV non observée), il est nécessaire de modéliser la distribution conjointe de (Y_i, D_i). Elle peut se décomposer de 2 façons donnant lieu à 2 types de modèles

Modèles de sélection	**Modèles de mélange de profils**
Modélisation directe de distribution marginale du score de QdV	Extrapolation après sortie d'étude (+ restrictions)
Sensibilité à l'hypothèse de normalité des données complètes	Hypothèses fortes (restrictions + extrapolation) mais explicites
Possibilité de tester directement l'hypothèse MNAR	Plus facile de faire des analyses de sensibilité
Intégrale des deux distributions par rapport aux données manquantes (Implémentation dans S-plus : package Oswald, fonction pcmid)	Facile à mettre en œuvre : modèle conditionnel sur les données observées et paramètres des 2 modèles peuvent être estimés séparément

Recommandations générales

- En amont :
  - Éviter les données manquantes
  - Lorsqu'elles surviennent, collecter la raison de sortie d'étude
- Lors de l'analyse :
  - Inclure dans le vecteur des variables explicatives X les variables qui sont associées au score de QdV et à la sortie d'étude
  - Si la sortie d'étude ne dépend de la QdV qu'à travers X, les données manquantes deviennent ignorables conditionnellement à X.
- En présence de données manquantes informatives :
  - Modéliser conjointement le score de QdV et la sortie d'étude
  - Analyses de sensibilité

Une autre approche : les modèles à effets aléatoires partagés

Dans les modèles présentés :

- Il y a des effets aléatoires seulement dans le modèle pour Y_i
- Les paramètres θ et ψ (associés à Y_i et D_i, respectivement) sont disjoints

La distribution conjointe de (Y_i, D_i) peut aussi s'écrire :

$$P_{\theta\psi}(Y_i, D_i) = \int P_{\theta\psi\zeta}(Y_i, D_i, b_i)\, \mathrm{d}b_i$$
$$= \int P_{\theta\psi}(Y_i, D_i|\ b_i) P_\zeta(b_i)\, \mathrm{d}b_i$$

où b_i sont des effets aléatoires et ζ les paramètres associés

Une autre approche : les modèles à effets aléatoires partagés

La distribution conjointe « enrichie » des effets aléatoires peut également se décomposer de 2 façons donnant lieu

- aux modèles de sélection à effets aléatoires (*random-effects selection models*) :

$$P_{\theta\psi}(Y_i, D_i) = \int P_\psi(D_i \mid Y_i, b_i)\, P_\theta(Y_i \mid b_i) P_\zeta(b_i)\, \mathrm{d}b_i$$

- aux modèles de mélange de profils à effets aléatoires (*random-effects pattern mixture models*) :

$$P_{\theta\psi}(Y_i, D_i) = \int P_\psi(Y_i \mid D_i, b_i)\, P_\theta(D_i \mid b_i) P_\zeta(b_i)\, \mathrm{d}b_i$$

Si l'on suppose que Y_i **et** D_i **sont indépendants conditionnellement à** b_i, on obtient les modèles à effets aléatoires partagés :

$$P_{\theta\psi}(Y_i, D_i) = \int P_\psi(Y_i \mid b_i)\, P_\theta(D_i \mid b_i) P_\zeta(b_i)\, \mathrm{d}b_i$$

Une autre approche : les modèles à effets aléatoires partagés

- Plus appropriés lorsque la probabilité de sortie d'étude :
  - ne dépend pas directement de la variable de QdV (par ex. refus de répondre à des items sensibles relatifs à la sexualité dans le QLQ-BR23 pour le cancer du sein)
  - dépend indirectement de la variable de QdV à travers une variable latente sous-jacente telle que la progression du cancer

- Lorsque la sortie d'étude est liée à un tel évènement (progression/décès), il peut être plus intéressant de modéliser le temps jusqu'à progression/décès plutôt que la probabilité de non-réponse à partir de la $j^{\text{ème}}$ évaluation, à travers un modèle conjoint :
  - Modèle linéaire mixte pour Y_i
  - Modèle de survie pour le temps jusqu'à progression/décès T_i :

$$\lambda(t|z_i, b_i) = \lambda_0(t)\mathrm{e}^{\gamma^T z_i + \alpha^T f(b_i,t)}$$

où le paramètre α mesure l'association entre le score de QdV et l'évènement