Impact des données manquantes informatives dans l'estimation d'un modèle linéaire mixte

Moussa NGAMBE, Mahamadou Ousmane Keita et Xaver Wangerpohl

Résumé

Ce projet vise à approfondir la compréhension du modèle linéaire mixte et de son application aux données longitudinales. À travers l'étude d'un échantillon de patients atteints d'un cancer de la prostate, nous avons étudié l'impact des données manquantes informatives sur l'estimation des paramètres et la modélisation des trajectoires individuelles.

Les données, recueillies sur une période maximale de 15 ans après le traitement par chimiothérapie ou chirurgie, comprennent des mesures répétées du marqueur PSA (Prostate Specific Antigen). En raison de la nature des mesures de PSA, une transformation de la variable d'intérêt a été appliquée. Pour le modèle longitudinal, nous avons développé un modèle d'évolution non-linéaire du log(PSA + 0.1) décomposé en 3 parties (Desmée, 2016) : le niveau de PSA après le traitement par chimiothérapie ou chirurgie, l'évolution à court terme dans l'année suivant le traitement et l'évolution à long-terme. Ce modèle capture convenablement la non-linéarité biphasique de la cinétique du PSA. Au vu de la cinétique du PSA, ces modèles mathématiques vont correctement ajuster les données longitudinales de PSA.

Nous avons comparé les résultats obtenus à partir de deux versions du jeu de données : la version complète et la version incomplète, caractérisée par des données manquantes informatives. Différents modèles linéaires mixtes biphasiques ont été ajustés, et les critères AIC ont été utilisés pour sélectionner le modèle optimal.

L'analyse comparative entre le modèle complet et le modèle incomplet a révélé des légères différences dans l'estimation paramètres, mettant en évidence l'influence des données manquantes sur les résultats du modèle.

Ce rapport présente une discussion approfondie sur ces observations, met en lumière les implications des données manquantes et propose des recommandations pour atténuer cet impact dans le contexte spécifique du modèle linéaire mixte biphasique appliqué aux données longitudinales du cancer de la prostate.

1 Introduction

La recherche en modélisation statistique, particulièrement dans le domaine des données longitudinales, joue un rôle crucial dans la compréhension des trajectoires évolutives de divers phénomènes. Notre projet se situe à la convergence du modèle linéaire mixte (MLM) et de l'analyse de données longitudinales, se concentrant sur un échantillon de patients atteints d'un cancer de la prostate. La prostate, une glande de l'appareil génital masculin située sous la vessie, en avant du rectum, joue un rôle essentiel dans la sécrétion du liquide séminal, l'un des constituants du sperme, et de son stockage. Pour son bon fonctionnement, la prostate dépend des androgènes, notamment la testostérone, une hormone masculine prédominante. Le cancer de la prostate se caractérise par une prolifération excessive de cellules formant une tumeur. Ces cellules détournent les ressources des cellules saines, favorisant notamment l'angiogenèse, la création de nouveaux vaisseaux sanguins à partir de vaisseaux préexistants (Desmée, 2016).

Le marqueur PSA (Prostate Specific Antigen) est une protéine produite par les cellules de la prostate. Sa mesure fréquente dans le sang permet d'évaluer la santé de la prostate. L'inconvénient de l'évaluation de la taille de la tumeur par imagerie médicale est qu'elle est coûteuse et ne peut pas être réalisée fréquemment, contrairement au PSA qui nécessite une simple prise de sang. Le PSA apparaît comme un "substitut" à la taille de la tumeur puisqu'il est produit par les cellules prostatiques dont les cellules tumorales font partie. S'il est controversé en tant qu'outil diagnostique, le PSA est un outil pronostique utilisé pour le suivi des patients après un traitement (Stephenson et al., 2006; Roach et al., 2006).

Une augmentation du taux de PSA peut signaler diverses conditions, dont le cancer de la prostate. Cependant, en raison de sa forte variabilité, les mesures de PSA nécessitent une transformation, comme proposé par (Ferrer et al., 2016), qui suggèrent de modéliser la variable transformée Y=log(PSA+0.1).

Le choix du MLM dans notre analyse s'explique par la présence de mesures répétées dans nos données. Il vise à modéliser les trajectoires individuelles, à estimer les caractéristiques de la population (moyenne et variabilité inter-sujet), ainsi que les paramètres propres à chaque sujet. Cette approche robuste permet de comprendre l'évolution du marqueur dans le contexte du cancer de la prostate, et elle est particulièrement adaptée en présence de données manquantes. En effet, notre objectif est d'étudier l'impact des données manquantes informatives sur l'estimation des paramètres et la modélisation des trajectoires individuelles dans le cadre du MLM biphasique.

Cette problématique émerge de la dualité de nos jeux de données, l'un complet et l'autre incomplet, chacun présentant des défis distincts. Cependant, le modèle linéaire de l'évolution de log(PSA + 0.1) est décomposé en 3 parties : le niveau de PSA après le traitement par chimiothérapie ou chirurgie, l'évolution à court terme dans l'année suivant le traitement et l'évolution à long terme, nous utilisons un Modèle Linéaire à Effets Mixtes (MLEM) pour tenir compte de la non-linéarité biphasique de la cinétique du PSA. En considération de cette cinétique spécifique, ce modèle mathématique ajustera précisément les données longitudinales de PSA. Nous comparerons ensuite les résultats obtenus à partir de ces deux ensembles de données pour évaluer l'influence des données manquantes sur nos conclusions, offrant ainsi une compréhension approfondie de l'impact de la variabilité des données sur nos résultats.

2 Méthode

2.1 Les données

Dans le cadre de ce projet, nous avons examiné des données relatives à un échantillon de patients atteints d'un cancer de la prostate et suivis sur une période maximale de 15 ans. Après avoir subi des traitements tels que la chimiothérapie ou la chirurgie, le marqueur PSA (Antigène Spécifique de la Prostate), largement utilisé comme indicateur dans le cancer de la prostate, a été mesuré de manière répétée dans le but d'analyser la progression de la maladie. En effet, le marqueur PSA constitue un indicateur prédictif crucial pour suivre la présence et l'évolution de la maladie. Deux versions du jeu de données sont à notre disposition : une base de données complète et une base de données incomplète caractérisée par l'absence de mesures de PSA pour certains patients à certains moments au cours de l'essai clinique. Ces deux versions distinctes du jeu de données ont été fournies pour cette étude.

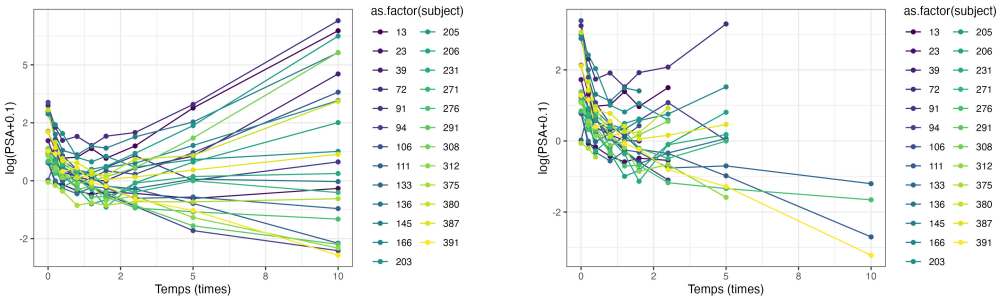
— Version complète (`dataPSA_complete.Rdata`) : Cette version regroupe des informations exhaustives sur 400 patients, offrant une vision détaillée de l'évolution du PSA au fil du temps.

— Version incomplète (`dataPSA_incomplete.Rdata`) : Composée de données manquantes informatives, cette version simule un scénario réaliste où des informations cruciales peuvent faire défaut, ajoutant ainsi une dimension supplémentaire à notre analyse.

Le tableau 2.1 fournit une description des patients présents dans nos deux bases de données.

Descriptions	Jeu de données complètes	Jeu de données incomplètes
Nombre de patients	400	400
Age normalisé médian (min-max)	2(0-4)	2(0-4)
Nombre de patients sous traitement chirurgie	1071	872
Nombre de patients sous traitement chimiothérapie	2529	1994
Temps médian (min-max)	2 (0-10)	1(0-10)
Mesures médian (min-max) du PSA	1(0-35451)	1 (0-70)
Nombre total de mesures de PSA	3600	2866

Table 2.1 : Caractéristiques des patients et des données disponibles, des jeux de données complètes et incomplètes



(a) Evolutaion log(PSA+0.1) en fonction du temps des données complètes

(b) Evolutaion log(PSA+0.1) en fonction du temps des données incomplètes)

Figure 1 – Evolution du log(PSA+0.1) chez 25 patients tirés au sort parmi les patients des jeux de données

2.2 Méthodes statistiques

La mesure du PSA chez les patients atteints du cancer de la prostate présente des observations faites au cours du temps. Ces observations sont répétées, étant donné qu'il existe un ensemble d'observations provenant d'un même patient. Par conséquent, pour modéliser la trajectoire du marqueur longitudinal, nous utilisons un modèle linéaire mixte. Sous des hypothèses gaussiennes, nous supposons que Y_{ij}, la mesure observée du marqueur au moment t_{ij}, est une mesure bruitée du niveau réel $Y_i^*(t_{ij})$. Ce niveau non observé $Y_i^*(t_{ij})$ s'explique en fonction du temps et des covariables avec des effets fixes β au niveau de la population, et des effets aléatoires b_i qui prennent en compte la corrélation entre les mesures répétées du même individu :

$$Y_{ij} = Y_i^*(t_{ij}) + \varepsilon_{ij} = X^T(t_{ij})\beta + Z_i^T(t_{ij})b_i + \varepsilon_{ij}, \tag{1}$$

avec $X(t_{ij})$ et $Z_i(t_{ij})$ les vecteurs de covariables éventuellement dépendantes du temps associées au vecteur de covariables fixes β et au vecteur d'effets aléatoires b_i, $b_i \sim N(0, D)$, respectivement.
Notez que $\varepsilon_i = \begin{bmatrix}\varepsilon_{i1} & \dots & \varepsilon_{ini}\end{bmatrix}^T \sim N(0, \sigma^2 I)$, où I est la matrice identité, D matrice de covariance des effets aléatoires ; ε_i et b_i sont indépendants.

2.2.1 Spécification du modèle linéaire mixte

La cinétique de $\log(\text{PSA} + 0.1)$ présente une évolution non linéaire (voir Figure 2). Elle peut être décomposée en trois parties : le niveau de PSA après le traitement, l'évolution à court terme au cours de l'année suivant le traitement et l'évolution à long terme. Ainsi, nous allons modéliser la trajectoire du $\log(\text{PSA} + 0.1)$ de manière biphasique à l'aide d'un modèle mixte linéaire avec deux fonctions du temps $f_1(t)$ et $f_2(t)$, comme cela a été réalisé dans des travaux antérieurs Proust-Lima et al. (2008).

Les fonctions $f_1(t)$ et $f_2(t)$ sont définies comme suit :

$$f_1(t) = (1+t)^{-\alpha} - 1$$
$$f_2(t) = \frac{t^{1+\nu}}{(1+t)^{\nu}}$$

où ν et α ont été estimés par le profil de vraisemblance ($\alpha = -1.2$, $\nu = 0$). Les fonctions $f_1(t)$ et $f_2(t)$ décrivent la baisse initiale suivie d'une stabilisation ou d'une augmentation linéaire à long terme du $\log(\text{PSA} + 0.1)$.

Par conséquent, l'équation (1) devient :

$$\begin{aligned} Y_{ij} =& Y_i^*(t_{ij}) + \epsilon_{ij} \\ Y_{ij} =& \left(X_i^\top \beta_0 + b_{i0}\right) + \left(X_i^\top \beta_1 + b_{i1}\right) \times f_1(t_{ij}) + \left(X_i^\top \beta_2 + b_{i2}\right) \times f_2(t_{ij}) + \epsilon_{ij}, \end{aligned} \tag{2}$$

Les vecteurs $b_i = (b_{i0}, b_{i1}, b_{i2})^T \sim N(0, D)$ et $X_i^\top = \begin{bmatrix}1 & \text{age}_i & \text{Intervention}_i\end{bmatrix}$ représentent respectivement le vecteur des coefficients aléatoires et le vecteur des covariables. Les vecteurs $\beta_0, \beta_1, \beta_2$ sont des vecteurs de coefficients associés aux covariables de taille (3×1).

2.2.2 Modélisation statistique et choix du modèle

Nous avons utilisé la base de données complète (`dataPSA_complete.Rdata`) pour comparer plusieurs modèles à l'aide du critère AIC, une mesure permettant de sélectionner le

modèle offrant le meilleur compromis entre l'ajustement et la complexité. Le modèle de référence, **Modèle 1**, inclut les covariables de l'âge du patient et de l'intervention.

Où nous avons :

- $Y_{ij} = log(PSA + 0.1)$: La variable dépendante, généralement mesurée ou observée.
- β : L'intercept, représentant la valeur attendue de Y_{ij} lorsque toutes les autres variables explicatives sont nulles.
- β_{time} : Le coefficient associé à la variable explicative times_{ij}, qui mesure la contribution de cette variable à la variation de Y_{ij}.
- β_{age} : Le coefficient associé à la variable explicative age_{ij}, représentant l'effet de l'âge sur la variable dépendante.
- $\beta_{\text{intervention}}$: Le coefficient associé à la variable explicative $X_{\text{intervention}_{ij}}$, représentant l'effet de l'intervention sur la variable dépendante.
- β_{phase1} et β_{phase2} : Les coefficients associés aux fonctions $f_1(t_{ij})$ et $f_2(t_{ij})$, respectivement.
- b_{0i}, b_{1i}, b_{2i} : Les termes d'effet aléatoire pour le sujet i.
- ϵ_{ij} : Le terme d'erreur aléatoire.

Modèle 1 :

$$\begin{aligned} Y_{ij} &= \beta_{01} + \beta_{02} \cdot \text{age}_i + \beta_{03} \cdot X_{\text{intervention}_i} + b_{0i} \\ &\quad + (\beta_{11} + \beta_{12} \cdot \text{age}_i + \beta_{13} \cdot X_{\text{intervention}_i} + b_{1i}) \cdot f_1(t_{ij}) \\ &\quad + (\beta_{21} + \beta_{22} \cdot \text{age}_i + \beta_{23} \cdot X_{\text{intervention}_i} + b_{2i})) \cdot f_2(t_{ij}) + \epsilon_{ij} \end{aligned}$$

Modèle 2 :

$$\begin{aligned} Y_{ij} &= \beta_{01} + \beta_{02} \cdot \text{age}_i + \beta_{03} \cdot X_{\text{intervention}_i} + b_{0i} \\ &\quad + (\beta_{11} + \beta_{12} \cdot \text{age}_i + \beta_{13} \cdot X_{\text{intervention}_i} + b_{1i}) \cdot f_1(t_{ij}) + \beta_{interaction_{\text{phase1}}} \cdot \text{age}_i \cdot f_1(t_{ij}) \\ &\quad + (\beta_{21} + \beta_{22} \cdot \text{age}_i + \beta_{23} \cdot X_{\text{intervention}_i} + b_{2i})) \cdot f_2(t_{ij}) + \beta_{interaction_{\text{phase2}}} \cdot \text{age}_i \cdot f_2(t_{ij}) + \epsilon_{ij} \end{aligned}$$

Modèle 3 :

$$\begin{aligned} Y_{ij} &= \beta_{01} + \beta_{02} \cdot \text{age}_i + \beta_{03} \cdot X_{\text{intervention}_i} + b_{0i} \\ &\quad + (\beta_{11} + \beta_{12} \cdot \text{age}_i + \beta_{13} \cdot X_{\text{intervention}_i} + b_{1i}) \cdot f_1(t_{ij}) + \beta_{interaction_{\text{phase1}}} \cdot \text{intervention}_i \cdot f_1(t_{ij}) \\ &\quad + (\beta_{21} + \beta_{22} \cdot \text{age}_i + \beta_{23} \cdot X_{\text{intervention}_i} + b_{2i})) \cdot f_2(t_{ij}) + \beta_{interaction_{\text{phase2}}} \cdot \text{intervention}_i \cdot f_2(t_{ij}) + \epsilon_{ij} \end{aligned}$$

Modèle 4 :

$$Y_{ij} = \beta_{01} + \beta_{02} \cdot \text{age}_i + b_{0i} + (\beta_{11} + \beta_{12} \cdot \text{age}_i + b_{1i}) \cdot f_1(t_{ij}) + (\beta_{21} + \beta_{22} \cdot \text{age}_i + b_{2i})) \cdot f_2(t_{ij}) + \epsilon_{ij}$$

Modèle 5 :

$$\begin{aligned} Y_{ij} &= \beta_{01} + \beta_{02} \cdot \text{age}_i + b_{0i} \\ &\quad + (\beta_{11} + \beta_{12} \cdot \text{age}_i + b_{1i}) \cdot f_1(t_{ij}) + \beta_{interaction_{\text{phase1}}} \cdot \text{age}_i \cdot f_1(t_{ij}) \\ &\quad + (\beta_{21} + \beta_{22} \cdot \text{age}_i + b_{2i}) \cdot f_2(t_{ij}) + \beta_{interaction_{\text{phase2}}} \cdot \text{age}_i \cdot f_2(t_{ij}) + \epsilon_{ij} \end{aligned}$$

Nous avons implémenté les quatre modèles sous `R` à l'aide du package `nlme`, obtenant ainsi les valeurs AIC pour chaque modèle, comme présenté dans le tableau ci-dessous.

Modèle	AIC
Modèle 1	5621
Modèle 2	4931
Modèle 3	5023
Modèle 4	4998
Modèle 5	4921

Table 2.2 : Critères AIC pour chaque modèle

Le Modèle 5, avec un AIC de 4921, présente le plus bas critère d'information d'Akaike (AIC) parmi les modèles évalués. Cette valeur suggère que le Modèle 5 offre le meilleur compromis entre l'ajustement aux données et la complexité du modèle. Ainsi, selon le critère AIC, le Modèle 5 est préférable aux autres modèles examinés, qui est rappelé ci-dessous :

Modèle 5 :

$$\begin{aligned} Y_{ij} &= \beta_{01} + \beta_{02} \cdot \text{age}_i + b_{0i} \\ &\quad + (\beta_{11} + \beta_{12} \cdot \text{age}_i + b_{1i}) \cdot f_1(t_{ij}) + \beta_{interaction_{\text{phase1}}} \cdot \text{age}_i \cdot f_1(t_{ij}) \\ &\quad + (\beta_{21} + \beta_{22} \cdot \text{age}_i + b_{2i}) \cdot f_2(t_{ij}) + \beta_{interaction_{\text{phase2}}} \cdot \text{age}_i \cdot f_2(t_{ij}) + \epsilon_{ij} \end{aligned}$$

Ce modèle 5 modélise un modèle linéaire mixte modélisant la variable dépendante $Y = \log(PSA + 0.1)$ en fonction de l'âge (age_i) et du temps (t_{ij}), avec des effets aléatoires spécifiques à chaque sujet (b_{0i}, b_{1i}, b_{2i}). Les effets fixes sont modulés par des fonctions spécifiques à chaque phase ($f_1(t_{ij})$ et $f_2(t_{ij})$), avec des termes d'interaction entre l'âge et ces fonctions pour chaque phase.

3 Résultats

Ayant identifié le Modèle 5 comme étant le meilleur choix parmi les modèles en utilisant la base de données complète, nous avons procédé à une évaluation approfondie de l'impact des données manquantes, celle-ci a été réalisée en utilisant *le Modèle 5.*

3.1 Estimation du modèle avec la base de données complète

Nous avons ajusté le **Modèle 5** en utilisant l'ensemble de données complet (`dataPSA_complete.Rdata`) pour établir une référence solide. Les résultats obtenus à partir de cette analyse représentent nos estimations de base, sans l'impact des données manquantes. Ces résultats sont présentés dans les tableaux ci-dessous.

Effets aléatoires			
	StdDv	Corr	
(Intcp)	0.6	(Intcp)	times
times	0.3	0.1	
$I(1 + times)^{-1.2} - 1$	1.3	0.5	0.7
Residual	0.3		

Table 3.3 : Tableau avec effets aléatoires

Variable	Value	Std.Error	DF	t-value	p-value
(Intercept)	-0.3	0.10	3196	-3	7×10^{-3}
I((1 + times)$^{(-1.2)} - 1$)	0.9	0.22	3196	4	5×10^{-5}
times	-0.1	0.05	3196	-2	7×10^{-2}
age_normalise	0.8	0.05	398	17	3×10^{-48}
age_normalise :I((1 + times)$^{(-1.2)} - 1$)	1.0	0.10	3196	10	4×10^{-21}
times :age_normalise	0.2	0.03	3196	9	3×10^{-18}

Table 3.4 : Tableau avec les résultats de la régression linéaire mixte

Les résultats du modèle linéaire mixte, ajusté à l'aide des données complètes provenant de l'étude visant à évaluer les facteurs influençant le niveau de log(PSA + 0.1) au fil du temps, révèlent que le niveau de log(PSA + 0.1) au début de l'étude pour les patients est estimé à -0.3 en moyenne, ce qui est statistiquement significatif (p-value $= 7 \times 10^{-3} < 0.05$) (Tableau 3.4), et présente une variabilité de 0.6 d'un patient à l'autre (Tableau 3.3).

Leur niveau de log(PSA + 0.1) augmente en moyenne de 0.9 avec l'augmentation du temps durant la première phase (Tableau 3.4). Cette augmentation est statistiquement significative (p-value $= 5 \times 10^{-5} < 0.05$), et cette croissance du niveau de log(PSA + 0.1) des patients au cours du temps est variable d'un patient à un autre (1.3) (Tableau 3.3).

De même, les patients ont un niveau de log(PSA + 0.1) qui diminue en moyenne de -0.1 au niveau de la seconde phase en fonction du temps (Tableau 3.4). Cette décroissance n'est pas statistiquement significative (p-value $= 7 \times 10^{-2} > 0.05$). Cette baisse non significative du niveau de log(PSA + 0.1) varie d'un patient à l'autre de 0.3 (Tableau 3.3).

Le niveau de log(PSA + 0.1) augmente de 0.8 lorsque l'âge augmente d'une unité, ce qui est statistiquement significatif (p-value $= 3 \times 10^{-48} < 0.05$, (Tableau 3.4)). De plus, le niveau de log(PSA + 0.1) des patients augmente de 0.1 durant la première phase et de 0.2 pour la seconde phase en fonction de l'âge, ce qui est statistiquement significatif car leurs p-values sont très faibles (voir Tableau 3.4).

3.2 Estimation du modèle avec la base de données incomplète

Dans le but d'évaluer l'impact des données manquantes, nous avons ajusté le **Modèle 5** en utilisant l'ensemble de données incomplètes (`dataPSA_incomplete.Rdata`) pour comparer les estimations avec celles obtenues à partir de la base de données complète (`dataPSA_complete.Rdata`). Les résultats de ces estimations avec la base incomplète sont présentés dans les trois tableaux ci-dessous.

Effets aléatoires			
	StdDv	Corr	
(Intcp)	0.6	(Intcp)	times
times	0.3	0.1	
$I(1+times)^{-1.2}-1$	1.3	0.5	0.7
Residual	0.3		

Table 3.5 : Tableau avec effets aléatoires

	Données incomplètes		**Données complètes**	
Variable	**Value**	**p-value**	**Value**	**p-value**
(Intercept)	-0.26	1×10^{-3}	-0.3	7×10^{-3}
I((1 + times)$^{(-1.2)} - 1$)	0.97	6×10^{-5}	0.9	5×10^{-5}
times	-0.09	2×10^{-1}	-0.1	7×10^{-2}
age_normalise	0.80	1×10^{-46}	0.8	3×10^{-48}
age_normalise :I((1 + times)$^{(-1.2)} - 1$)	0.91	3×10^{-15}	1.0	4×10^{-21}
times :age_normalise	0.21	3×10^{-12}	0.2	3×10^{-18}

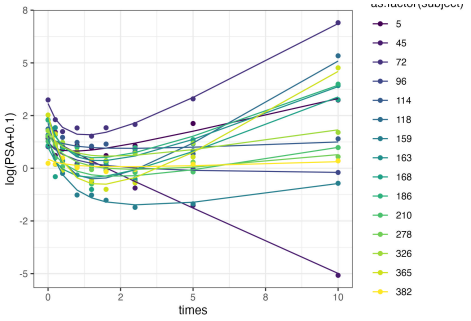
Table 3.6 : Tableau de comparaison des coefficients estimés données complètes et incomplètes

Les résultats du modèle linéaire mixte, ajusté à l'aide des données incomplètes provenant de l'étude visant à évaluer les facteurs influençant le niveau de PSA au fil du temps, indiquent que les estimations des effets aléatoires similaires à celles obtenues avec les données complètes (voir tableau Tableau 3.5). La comparaison suggère que les données manquantes n'ont pas introduit de biais significatif dans les estimations. La trajectoire du PSA est principalement influencée par l'âge, avec une absence d'effet significatif sur l'évolution du niveau de log (PSA+0.1) durant la seconde phase. Les résultats restent cohérents malgré la présence de données manquantes informatives.

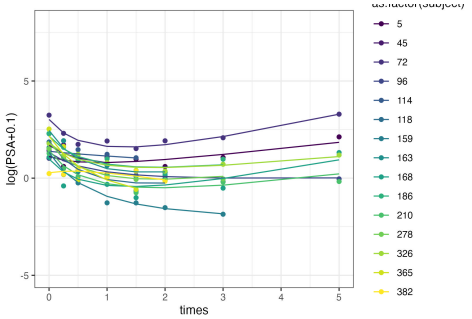
4 Comparaison des trajectoires

Pour évaluer l'impact des données manquantes, cette section compare les trajectoires individuelles prédites pour 15 patients sélectionnés de manière aléatoire et identiques dans les deux bases de données (complète et incomplète). Ensuite, elle examine les trajectoires moyennes (trajectoires fixes) estimées par les bases de données complètes et incomplètes pour l'ensemble des patients.

4.1 Comparaison des trajectoires individuelles prédites



(a) Évolution prédite de log(PSA+0.1) en fonction du temps - Données complètes



(b) Évolution prédite de log(PSA+0.1) en fonction du temps - Données incomplètes)

FIGURE 2 – Comparaison des trajectoires prédites log(PSA+0.1) pour 15 patients sélectionnés au hasard dans les deux bases de données

Les graphiques ci-dessus représentent les trajectoires prédites de log(PSA+0.1) pour 15 patients sélectionnés au hasard, communs aux deux bases de données. L'observation révèle que les trajectoires individuelles prédites dans le modèle complet et le modèle incomplet ne suivent pas une forme générale similaire au fil du temps. Certains patients présentent des tendances différentes dans les deux bases de données, notamment à partir de la cinquième mesure. Cette différence s'explique par le fait que certains de ces patients n'ont pas eu de mesure PSA après une certaine date, impactant ainsi les trajectoires dans la base de données incomplète. De plus, la base de données complète semble offrir un ajustement plus précis que la base de données incomplète. Ces graphiques suggèrent qu'il existe une différence dans la manière dont log(PSA+0.1) évolue au fil du temps en présence de données manquantes informatives.

4.2 Comparaison des trajectoires des effets fixes (données complètes vs incomplètes)

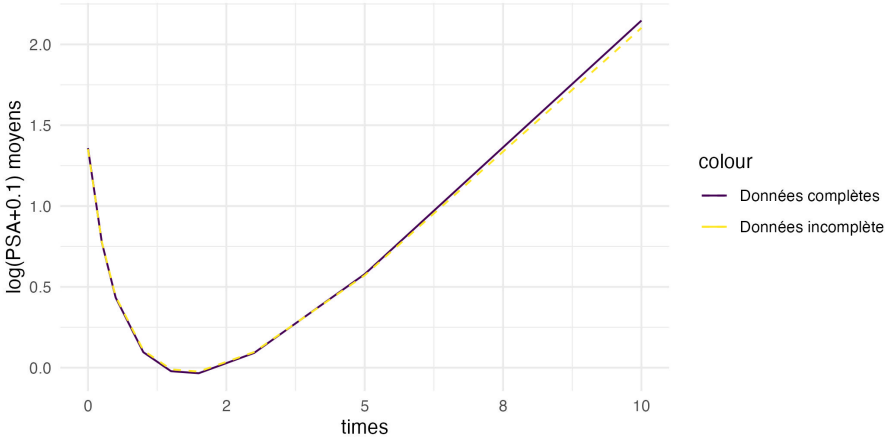


FIGURE 3 – Comparaison des trajectoires des moyennes des données complètes et incomplètes.

L'analyse de ce graphique montre que les trajectoires moyennes ajustées dans nos deux bases de données sont identiques pour les premières mesures du PSA. Cependant, à partir de la cinquième mesure, la trajectoire des données complètes est au-dessus de celle de

la base incomplète. Cette dissymétrie à partir de cette mesure s'explique par la présence de données manquantes informatives à partir de cette date. Ces résultats confirment qu'il existe une différence dans la manière dont log(PSA+0.1) évolue au fil du temps en présence de données manquantes informatives.

4.3 Comparaison de la performance du modèle sur les deux bases de données

Dans cette section, nous avons entrepris une comparaison de la performance des modèles ajustés par les deux bases de données. Nous avons réalisé une analyse de prédiction sur deux ensembles de données distincts : complet et incomplet. Ce processus a été répété 1000 fois pour atténuer les variations liées à l'échantillonnage aléatoire.

À chaque itération, nous avons sélectionné aléatoirement 15 patients pour former l'ensemble d'apprentissage. Ces patients ont été utilisés pour ajuster les modèles sur les données d'apprentissage de la base de données complète et incomplète. Ensuite, nous avons choisi 15 patients distincts pour constituer l'ensemble de test dans la base de données complète et avons effectué des prédictions à l'aide des deux modèles ajustés (modèle complet et incomplet).

La performance de chaque prédiction a été évaluée en calculant l'erreur quadratique moyenne (RMSE) pour chaque groupe de données. Enfin, un **boxplot** a été généré pour visualiser la distribution des erreurs de prédiction entre les deux ensembles de données.

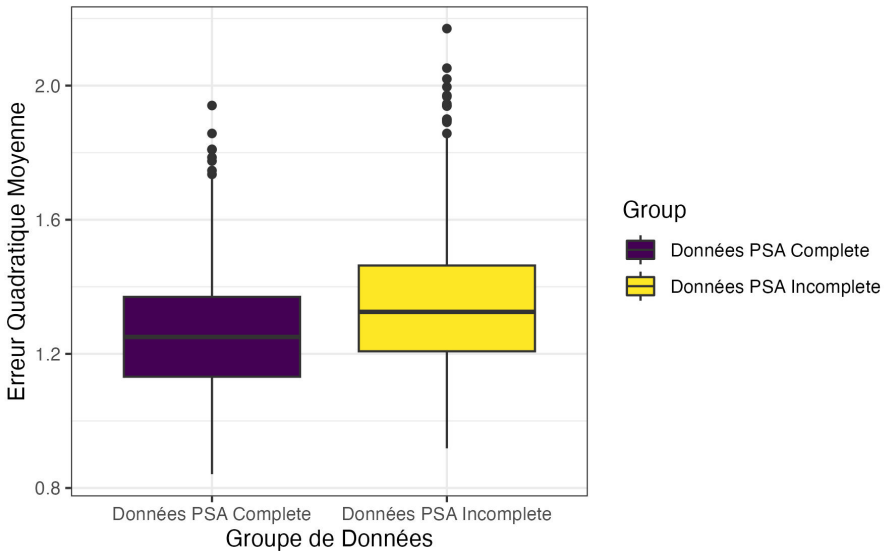


Figure 4 – Boxplot des erreurs quadratiques de prédiction moyennes pour les deux bases de données.

Les résultats de cette analyse (Figure 3) indiquent que l'erreur quadratique moyenne de prédiction est globalement plus faible pour les données complètes par rapport aux données incomplètes. Cette observation suggère que le modèle ajusté sur la base de données complète a une meilleure capacité de prédiction par rapport au modèle ajusté sur la base de données incomplète, ce qui confirme l'impact des données manquantes sur la qualité des prédictions du modèle. En particulier, à partir d'une certaine date, notre modèle ajusté aux données incomplètes aura plus tendance à faire des erreurs de prédiction en raison de l'absence de certaines mesures pour les patients à l'apprentissage du modèle.

5 Discussion

L'objectif principal de ce projet était d'étudier l'impact des données manquantes informatives en comparant l'estimation des paramètres et la modélisation des trajectoires individuelles.

Dans un premier temps, une analyse descriptive approfondie des deux bases de données disponibles a été réalisée. Cette analyse a jeté les bases pour le développement d'un modèle linéaire mixte diphasique. Ce modèle repose sur l'évolution non linéaire du $\log(PSA+0.1)$, décomposé en trois composantes distinctes : le niveau de PSA après le traitement par chimiothérapie ou chirurgie, l'évolution à court terme dans l'année suivant le traitement, et l'évolution à long terme. Le choix de ce modèle s'est appuyé sur le critère AIC, intégrant les variables dépendantes telles que l'âge normalisé ainsi que deux fonctions du temps de mesure.

L'estimation du modèle dans les deux bases de données a révélé que les données manquantes n'ont pas introduit de biais significatif dans les estimations des paramètres. La trajectoire du PSA semble principalement influencée par l'âge. Ces résultats restent cohérents malgré la présence de données manquantes informatives.

La comparaison des trajectoires individuelles entre les deux bases de données, complète et incomplète, a été illustrée graphiquement. Les figures générées pour 15 patients sélectionnés au hasard ont mis en évidence la concordance des prédictions entre les deux ensembles de données. Cependant, certains patients présentent des tendances différentes dans les deux bases de données, notamment à partir de la cinquième mesure. Cette différence s'explique par le fait que certains de ces patients n'ont pas eu de mesure PSA après une certaine date, impactant ainsi les trajectoires dans la base de données incomplète. Les trajectoires moyennes ajustées dans nos deux ensembles de données présentent une similarité pour les premières mesures du PSA. Néanmoins, à partir de la cinquième mesure, la trajectoire des données complètes surpasse celle de la base incomplète. Cette disparité à partir de ce point s'explique par la présence de données manquantes informatives à partir de cette date, renforçant ainsi la fiabilité des résultats malgré la présence de données manquantes.

Enfin, une comparaison des erreurs quadratiques moyennes de prédiction a été effectuée pour évaluer les performances prédictives du modèle dans les deux bases de données. Les résultats révèlent une tendance où l'erreur quadratique moyenne de prédiction est généralement plus faible pour les données complètes par rapport aux données incomplètes. Cette observation suggère que le modèle ajusté sur la base de données complète démontre une meilleure aptitude à la prédiction que le modèle ajusté sur la base de données incomplète. Ces constatations confirment l'impact des données manquantes sur la qualité des prédictions du modèle. Plus spécifiquement, à partir d'une certaine date, le modèle ajusté aux données incomplètes montre une propension accrue à générer des erreurs de prédiction en raison de l'absence de certaines mesures pour les patients pendant la phase d'apprentissage du modèle.

En conclusion, ce projet contribue au développement des connaissances sur les modèles linéaires mixtes et leur application aux données longitudinales, ouvrant de nouvelles perspectives pour leur utilisation et soulignant la nécessité d'une approche prudente lors de la gestion des données manquantes. L'omission de ces données peut conduire à des divergences notables dans les prédictions du modèle, notamment à partir d'une certaine date, soulignant la nécessité d'adopter des approches de modélisation adaptées en présence de données manquantes. Ces résultats soulignent la pertinence d'une gestion proactive des données manquantes pour garantir la robustesse des modèles longitudinaux.

5.1 Répartition du travail

Pour la réalisation du projet, vous avez réparti les tâches de la manière suivante : Moussa NGAMBE et Xaver Wangerpohl ont effectué une revue de littérature sur le modèle linéaire mixte biphasique et son application aux données longitudinales, en plus de rédiger le rapport. Mahamadou Ousmane Keita et Moussa NGAMBE sont responsables de l'analyse descriptive des données, de la modélisation, l'interprétation des résultats ainsi que la rédaction du rapport.

Références

Desmée, S. (2016). Modélisation conjointe de données longitudinales non-linéaires et de données de survie : Application au cancer de la prostate métastatique. *HAL Open Science*, (1.2.1) :21–21.

Ferrer, L., Rondeau, V., Dignam, J., Pickles, T., Jacqmin-Gadda, H., and Proust-Lima, C. (2016). Joint modelling of longitudinal and multi-state processes : application to clinical progressions in prostate cancer. *Statistics in Medicine*, 35(22) :3933–3948.

Proust-Lima, C., Taylor, J., Williams, S., Ankerst, D., Liu, N., Kestin, L., Bae, K., and Sandler, H. (2008). Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. *International Journal of Radiation Oncology* Biology* Physics*, 72(3) :782–791.

Roach, M., Hanks, G., Thames, H., Schellhammer, P., Shipley, W., Sokol, G., and Sandler, H. (2006). Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer : Recommendations of the rtog-astro phoenix consensus conference. *International Journal of Radiation Oncology* Biology* Physics*, 65 :965–974.

Stephenson, A. J., Kattan, M. W., Eastham, J. A., Dotan, Z. A., Bianco, F. J., Lilja, H., and Scardino, P. T. (2006). Defining biochemical recurrence of prostate cancer after radical prostatectomy : A proposal for a standardized definition. *Journal of Clinical Oncology*, 24 :3973–3978.