

UNIVERSITÉ DE LILLE

UFR3S-MÉDECINE

Année : 2026

THÈSE POUR LE DIPLÔME D'ÉTAT DE DOCTEUR EN MÉDECINE

Analyse de biomarqueurs quantitatifs recueillis en routine : Mise au point et évaluation de la méthode « *Individual Regressions Based on Defined Temporally-Independant Functions* » (IRBODeTIF)

Présentée et soutenue publiquement le 26 juin 2026 à 14h00

au Pôle Formation

par Valentin JOHNSON

JURY

Président :

Monsieur le Professeur Philippe AMOUYEL

Assesseurs :

Monsieur le Docteur Luc DAUCHET

Monsieur le Docteur Raphaël BENTEGEAC

Directeur de thèse :

Monsieur le Professeur Emmanuel CHAZARD

Avertissement

La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Sigles

ACP	Analyse en Composante Principale
AIC	<i>Akaike Information Criterion</i>
ANOVA	Analyse de Variance
BIC	<i>Bayesian Information Criterion</i>
CART	<i>Classification and Regression Tree</i> , ou Arbre de Classification et de Régression
CPER	Contrat de Plan État-Région
FDA	<i>Functional Data Analysis</i> , ou Analyse de Données Fonctionnelles
FPCA	<i>Functional Principal Component Analysis</i> , ou Analyse en Composante Principale Fonctionnelle
HbA1c	Hémoglobine glyquée
HR	Hazard Ratio
IMC	Indice de Masse Corporelle
IRBODeTIF	<i>Individual Regressions Based on DEfined Temporally-Independent Functions</i> , ou Régressions Individuelles Basées sur des Fonctions Temporellement Indépendantes
LMM	Linear Mixed Models, ou Modèles Linéaires Mixtes
MAR	<i>Missing At Random</i>
MCAR	<i>Missing Completely At Random</i>

Sommaire

Avertissement.....	3
Remerciements	4
Sigles.....	9
Sommaire	10
Introduction.....	13
1 Biomarqueurs et données fonctionnelles.....	13
1.1 Généralités sur les biomarqueurs.....	13
1.2 Généralités sur les données fonctionnelles	14
1.3 Difficultés liées aux données fonctionnelles collectées en routine	15
1.4 Enjeux liés aux données fonctionnelles collectées en routine	15
2 Etat de l'art	17
2.1 Analyse de données longitudinales classiques	17
2.1.1 Analyses de survie.....	17
2.1.2 Analyse de variance répétée.....	18
2.1.3 Analyse en composante principale	19
2.1.4 Modèles linéaires mixtes.....	19
2.1.5 Tableau récapitulatif.....	20
2.2 Analyse de données sous forme fonctionnelle	21
2.2.1 Définition et problématique	21
2.2.2 Lissage et représentation fonctionnelle.....	22
2.2.3 Analyse en composantes principales fonctionnelle.....	22
2.2.4 Régression fonctionnelle.....	23
2.2.5 Modèles joints longitudinal-survie	24
2.2.6 Tableau récapitulatif.....	25
3 Régression linéaire.....	26
3.1 Régression linéaire simple	26
3.1.1 Indicateurs	26
3.1.2 Critères de validation	27
3.1.3 Prédiction.....	30
3.2 Régression linéaire multiple	30
3.2.1 Généralités.....	30
3.2.2 Limites du modèle.....	31

3.2.3	Processus itératif	31
3.2.4	Critères propres à la régression linéaire multiple	32
4	Objectif	33
	Matériel et méthodes	34
1	Données de test	34
2	Définition de la méthode	35
2.1	Principe général	35
2.2	Définition des fonctions sous-jacentes	35
2.3	Gestion des données manquantes	36
2.4	Calcul des coefficients pour chaque individu	37
3	Applications pratiques de cette méthode	38
3.1	Modélisations individuelles	38
3.2	Clustering et description de sous-groupes choisis d'individus	38
3.3	Prédiction de la trajectoire d'HbA1c d'après d'autres variables	38
3.4	Prédiction d'autres variables à l'aide de la trajectoire d'HbA1c	38
3.5	Prédiction de la suite d'une trajectoire d'HbA1c	39
3.6	Description de l'évolution conjointe de l'HbA1c et de l'IMC	39
4	Cadre réglementaire	40
4.1	Autorisations	40
4.2	Financement	40
	Résultats	41
1	Diagramme de flux (<i>flowchart</i>)	41
2	Modélisations individuelles	42
3	Clustering et description de sous-groupes choisis d'individus	44
4	Prédiction de la trajectoire d'HbA1c d'après d'autres variables	46
5	Prédiction d'autres variables à l'aide de la trajectoire de l'HbA1c	49
5.1	Mise sous insuline à terme, en binaire	49
5.2	Délai avant mise sous insuline	49
5.3	Développement de complication(s)	51
6	Prédiction de la suite d'une trajectoire d'HbA1c	52
7	Description de l'évolution conjointe de l'IMC et de l'HbA1c	54
	Discussion	56
	Liste des tables	58
	Liste des figures	59
	Références	61

Introduction

Ce travail de thèse consiste à mettre au point et évaluer une méthode d'analyse des biomarqueurs quantitatifs recueillis en routine. L'introduction est structurée comme suit :

La première partie présentera une définition des données fonctionnelles, qui seront utilisées et analysées ultérieurement. Nous soulignerons leur potentiel d'utilisation en recherche biomédicale.

La deuxième partie présentera un état de l'art pour l'analyse de telles données.

La troisième partie proposera une explication des modèles de régression linéaires, sur lesquels la nouvelle méthode sera fondée.

La quatrième partie explicitera cette nouvelle méthode, ses objectifs, ses principes et son fonctionnement.

La cinquième partie exposera les différents résultats de ce travail de thèse, suivi d'une discussion.

1 Biomarqueurs et données fonctionnelles

1.1 Généralités sur les biomarqueurs

Un biomarqueur est une caractéristique biologique mesurable lié à un processus physiologique ou pathologique du corps humain [1,2]. Les biomarqueurs peuvent être utilisés dans le cadre du dépistage, du diagnostic, ou de la réponse à un traitement médical [1,2]. Dans ce travail, nous nous intéresserons uniquement aux biomarqueurs quantitatifs.

Les biomarqueurs les plus utilisés en routine clinique peuvent être des ions, protéines, ou éléments figurés, dosées dans le sang ou dans les urines. On citera par exemple le nombre d'érythrocytes (globules rouges), de leucocytes (globules blancs), de plaquettes, le dosage des enzymes hépatiques ou biliaires, de marqueurs de la fonction rénale (urée et créatinine), de la glycémie, des ions, des lipides, etc.

Les marqueurs utilisés en pratique clinique sont ceux dont, à la fois :

- La mesure est possible, abordable et suffisamment fiable [3,4].
- La distribution statistique est clairement connue chez les individus indemnes de pathologies.
- Des plages de valeurs (généralement les extrémités de distribution) sont connues pour être des tests diagnostiques plus ou moins fiables de pathologies connues.

Les biomarqueurs seraient théoriquement mesurables en continu dans l'organisme, leur quantité variant notamment en fonction de leur production et de leur dégradation par les processus physiologiques et pathologiques. La mesure de ces biomarqueurs est parfois invasive, et présente dans tous les cas un coût financier ou humain [3]. Par conséquent, en pratique, cette mesure est réalisée à des instants choisis et limités en nombre [5,6]. Ainsi, la mesure d'un biomarqueur dans le cadre du soin ou de la

recherche engendre, pour un individu, une courbe de valeurs quantitatives, discrètes en fonction du temps, qui n'est qu'un reflet de la courbe réelle, quantitative continue.

Notre travail portera sur l'analyse des biomarqueurs sous forme de données quantitatives continues dépendantes du temps, dites fonctionnelles.

1.2 Généralités sur les données fonctionnelles

Dans le cadre de ce travail, nous nous limiterons aux valeurs quantitatives continues. Les données fonctionnelles (quantitatives continues) sont des observations répétées faites sur un même individu au cours du temps, produisant une suite de valeurs [7–9]. Le terme « fonctionnel » provient du fait que ces valeurs témoignent d'une fonction sous-jacente du temps [7,10]. Cette fonction est théoriquement définie pour chaque instant, contrairement aux valeurs recueillies.

Les observations ou mesures discrètes, c'est-à-dire effectuées à certaines dates précises, peuvent être représentées pour chaque patient par une courbe brisée du marqueur en fonction du temps. Cette courbe brisée est une approximation de la fonction sous-jacente, d'une part parce qu'elle n'est définie qu'à des instants choisis, et d'autre part parce que les mesures font elles-mêmes l'objet d'erreurs de mesure [7,11]. En tenant compte de ces deux défauts, il est possible d'approcher la fonction continue réelle par plusieurs méthodes d'interpolation et de lissage [7,8].

Les méthodes d'interpolation et de lissage concernent les observations mesurées sur un même individu au cours du temps. Les méthodes appliquées à un ensemble d'individus ne seront pas abordées ici.

En ce qui concerne les méthodes d'interpolation pour un seul individu :

- Le *Last Observation Carried Forward (LOCF)*, technique simple et conservatrice, impute une valeur manquante par la valeur précédente. Elle rend constante la dérivée première entre 2 observations (à l'exception des jonctions) [12].
- L'interpolation linéaire minimise la dérivée première de la pente entre deux observations, notamment aux jonctions [7].
- Le *spline* minimise la dérivée seconde de la pente entre deux observations et offre une courbe plus réaliste pour les phénomènes biologiques [7,8].

Les méthodes de lissage partent du principe que l'erreur de mesure accentue la variabilité dans le temps et qu'il serait raisonnable de considérer que la fonction sous-jacente présente moins de variabilité :

- Le lissage par la moyenne, purement basé sur l'échantillon, consiste à remplacer chaque point par la moyenne des points voisins [7].
- Le lissage par régressions locales consiste à effectuer des régressions de manière itérative sur chaque sous-ensemble de points de la courbe [7,8].

Le progrès des technologies médicales, notamment le développement et la démocratisation des capteurs et autres outils de mesures, a considérablement accru la quantité de données disponibles [13,14]. La réutilisation de données de santé déjà collectées pour des raisons ou objectifs différents, notion dite de « *data re-use* », a rendu nécessaire le développement d'une méthodologie d'analyse fiable et robuste propre aux données fonctionnelles [9,15].

L'analyse de données de santé sous forme de fonctions apporterait plusieurs avantages par rapport aux tests classiques se basant sur des données discrètes, moins nombreuses [8,15,16]. Parmi ces avantages, on retrouve :

- L'intégration de la composante temporelle des données, permettant par exemple de détecter des différences localisées dans le temps, ou d'envisager la prédiction de l'évolution des individus [8,9].
- La possibilité de visualisation graphique (courbes) [7,8].
- L'utilisation d'outils mathématiques propres aux fonctions [7,10].
- L'augmentation de la puissance due à la prise en compte de l'ensemble de la courbe, au lieu de données discrètes [9,15].

Ce champ d'étude offre de nouvelles perspectives pour exploiter les données de santé (issues des domaines médicaux, paramédicaux, d'auto-mesures provenant de dispositifs médicaux ou non-médicaux, ou bien en pharmacie) et fournir des analyses complémentaires aux méthodes statistiques plus traditionnelles [9,13,15].

1.3 Difficultés liées aux données fonctionnelles collectées en routine

Les principales limites dans l'analyse des données fonctionnelles collectées en routines sont liées à leur caractère discret par rapport au temps, que ce soit comme cause ou conséquence. Les données obtenues ne sont qu'un échantillon des données réelles, et reflètent une fonction sous-jacente invisible dont on ne connaît qu'un ensemble limité de points :

- Les temps de mesures sont eux aussi discrets, comme en recherche clinique [5,6].
- Les temps de mesures sont asynchrones et non-protocolisés, à l'inverse de la recherche clinique, et soumis à un biais d'indication, les paramètres étant d'avantage mesurés par le clinicien s'il est vraisemblable qu'ils soient anormaux ou changeants [5,17–19].
- Les erreurs de mesures sont toujours présentes à des degrés divers, comme pour les données collectées en recherche clinique. Cependant, contrairement aux données de recherche clinique, cette erreur est elle-même changeante en fonction du contexte, du matériel et des réactifs utilisés, etc. [3,4].
- Le suivi des patients (l'inclusion et l'exclusion) est lui-même non-protocolisé, ce qui engendre des censures informatives (à gauche et/ou à droite) [18,20].

1.4 Enjeux liés aux données fonctionnelles collectées en routine

Idéalement, les données fonctionnelles collectées en routine présentent un potentiel d'utilisation dans plusieurs cadres statistiques :

- **Description** : ces données devraient permettre de décrire de manière synthétique un ensemble d'individus au sein d'un échantillon [7–9].
Exemple : conclure que dans l'échantillon, les individus ont une HbA1c qui suit globalement une certaine courbe.

- **Clusterisation** : ces données devraient également permettre de regrouper (ou rapprocher) les individus qui paraissent similaires du point de vue de ces données [21,22].
Exemple : conclure que l'échantillon comporte 4 profils-types d'évolution de l'HbA1c sur la durée de suivi.
- **Prédiction d'une variable** : ces données devraient pouvoir être utilisées comme variable explicative (*input*) d'une méthode permettant de prédire une variable réponse (*outcome*), comme une variable binaire, qualitative, quantitative, ou de survie [14].
Exemple : prédire la mise en place d'une insulinothérapie par l'augmentation progressive de l'HbA1c.
- **Prédiction par des variables** : ces données devraient pouvoir être utilisées comme variable de réponse (*outcome*) d'une méthode permettant de les prédire à l'aide d'autres variables plus traditionnelles (binaires, qualitatives, quantitatives) [8,9].
Exemple : prédire l'évolution péjorative de l'HbA1c par le profil pondéral et socio-économique des patients.
- **Prédiction de l'évolution du paramètre** : ces données, au regard de leur profil précoce, devraient permettre de prédire la courbe d'un individu à partir d'une certaine date [8,11] .
Exemple : utiliser les valeurs déjà connues de l'HbA1c d'un individu pour prédire leur évolution à partir d'une certaine date.
- **Evolution jointe** : enfin, il devrait être possible de décrire la corrélation entre l'évolution de deux données fonctionnelles [23,24].
Exemple : montrer que la courbe d'HbA1c et la courbe d'IMC de patients diabétiques évoluent dans le même sens.

De plus, en réutilisation de données, il serait souhaitable de respecter deux contraintes techniques naturellement observées dans les données (et moins souvent retrouvées en recherche clinique protocolisée) :

- **L'asynchronicité des mesures** : la nature arbitraire des temps auxquels sont réalisés les mesures ne doit pas être un obstacle à l'analyse [5,11,25].
- **L'inégalité des durées de suivi** : s'il est sûrement souhaitable de limiter cette inégalité (en imposant par exemple une durée de suivi allant du simple au double au maximum), il doit néanmoins être possible de traiter des suivis de durées différentes, notamment dans une perspective de prédiction de l'évolution [5,9].

Tout ce potentiel n'est malheureusement pas réalisé avec les méthodes actuelles.

2 Etat de l'art

Dans cet état de l'art, nous décrirons d'abord les méthodes classiques pour analyser les données de santé, puis les méthodes d'analyse des données fonctionnelles, au regard des enjeux décrits en introduction.

2.1 Analyse de données longitudinales classiques

Les méthodes statistiques classiques traitent le temps comme une variable explicative discrète et raisonnent point par point (valeur à $t = 0$, valeur à $t = 1$ etc.) [26–28]. Une observation est un vecteur de mesure prise à un temps défini, et l'analyse se fait sur l'ensemble des mesures prises à un même temps t . La courbe ou trajectoire d'un individu n'est pas ici un objet manipulable en tant que tel [8,9].

Plusieurs méthodes existent pour analyser une suite de mesures répétées dans le temps : l'analyse de survie, l'analyse de variance répétée, l'analyse en composante principale et les modèles linéaires mixtes.

2.1.1 Analyses de survie

Les analyses de survie modélisent le temps jusqu'à l'occurrence d'un événement, comme un décès, une rechute ou une hospitalisation [29,30]. Leur objectif est de comparer la vitesse de survenue de cet événement dans plusieurs groupes définis *a priori*.

Par construction, ces méthodes gèrent l'asynchronicité des données ainsi que la censure (durées de suivi différentes) [31,32] :

- Censure à gauche quand, pour un individu, l'événement est survenu avant la période d'observation.
- Censure à droite quand, pour un individu, l'événement n'est pas encore survenu à la fin du suivi (le plus courant).

Cependant, ces méthodes traitent uniquement des *outcomes* de nature binaire (0 ou 1, vivant ou décédé), dont le statut est nécessairement l'absence d'événement au début du suivi. Généralement, l'événement est unique et irréversible pour un individu donné. De plus, les co-variables sont généralement considérées comme constantes tout au long du suivi, hormis le temps [30,33]. Le modèle de Cox à événements répétés et co-variables temps-dépendantes permet de lever ces deux restrictions [33,34]. Les co-variables doivent néanmoins rester constantes sur les segments de suivi des individus (entre une censure à gauche et une censure à droite ou un événement).

Les analyses de survie ne permettent pas de créer des clusters de patients similaires, ni de prédire une variable extérieure ou d'effectuer une matrice de corrélation entre plusieurs variables [29,30].

Il existe trois méthodes principales dans les analyses de survie, présentées ci-après.

2.1.1.1 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est une méthode dite non-paramétrique (c'est-à-dire que l'on ne connaît pas la distribution des données étudiées) qui estime la fonction de survie $S(t)$, soit la probabilité de ne pas observer l'événement au cours du temps, recalculée à chaque nouvel événement [31,32].

Cette méthode produit des courbes « en escalier » qui descendent à chaque événement et permet de visualiser et de comparer les différents groupes [30,32]. La courbe de Kaplan-Meier permet de décrire la probabilité de survie au cours du temps.

2.1.1.2 Test du Log-Rank

Le test du Log-Rank est aussi non-paramétrique, et sert à comparer les courbes de survie de plusieurs groupes [30,31]. Il s'agit de comparer le nombre d'événements observés dans chaque groupe au nombre d'événements attendus sous l'hypothèse H_0 (hypothèse où les groupes ont le même risque de survenue de l'événement) [32,33].

2.1.1.3 Modèle de Cox

Le modèle de Cox, ou régression à risques proportionnels, est un modèle semi-paramétrique qui modélise l'effet de plusieurs variables sur le risque instantané de survenue de l'événement sous la forme de l'équation suivante [31,33,34]:

$$h(t) = h_0(t) \times e^{(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}$$

Avec $h_0(t)$ le risque de base (non spécifié et même non-estimé par le modèle) et les coefficient β_i quantifiant l'effet de chaque variable X_i sur le risque [33,34]. Le Hazard Ratio (HR) de la variable X_i est égal à e^{β_i} [30,31].

2.1.2 Analyse de variance répétée

L'analyse de variance (ANOVA) est un test statistique qui permet de comparer les moyennes de deux groupes ou plus pour déterminer si l'une au moins d'entre elles est significativement différente des autres [35,36]. Elle exige plusieurs conditions d'application : la normalité des résidus dans chaque groupe, l'homoscédasticité et l'indépendance des observations.

Le principe de l'ANOVA est de décomposer la variabilité totale des données en deux parties :

- La variance inter-groupes, c'est-à-dire les différences entre les moyennes des groupes (effet du facteur étudié).
- La variance intra-groupe, c'est-à-dire les différences à l'intérieur de chaque groupe (liée à la variabilité naturelle et aux erreurs de mesure).

Est ensuite calculé le rapport de ces deux variances, appelé statistique F . Une F élevée indique que les groupes diffèrent d'avantage que ce qui est attendu par le hasard [35,36].

$$F_{ANOVA} = \frac{Variance_{inter-groupe}}{Variance_{intra-group}}$$

L'ANOVA à mesures répétées est utilisée quand les mêmes individus font l'objet de mesures à plusieurs reprises [26,28,37]. Elle permet d'obtenir des moyennes par temps, qui approchent la courbe d'observation. Ici, la variabilité totale est décomposée en deux parties :

- La variance temporelle, soit l'effet du facteur temps sur le paramètre mesuré.
- La variance inter-sujet.

$$F_{ANOVA \text{ répétée}} = \frac{Variance_{temporelle}}{Variance_{inter-sujet}}$$

L'ANOVA répétée requiert des temps d'observation communs à tous les sujets (synchronicité) ainsi que la présence de toutes les mesures (pas de données manquantes) [28,35,37], et ne permet pas de prédire une trajectoire individuelle ni d'autres variables. Son principal intérêt est de tester l'effet du temps [26,37].

2.1.3 Analyse en composante principale

L'analyse en Composante Principale (ACP) est une méthode de réduction de dimension [38]. L'ACP permet de « remplacer » les variables d'origine observées dans l'échantillon, par autant de nouvelles variables statistiquement indépendantes entre elles, qu'on appelle « composantes ». La variance totale expliquée est la même, mais la part de variance expliquée est maximisée pour les premières composantes, et diminue au fur et à mesure de la construction des autres composantes.

Lorsque de nombreuses variables sont corrélées entre elles, il est alors intéressant de n'examiner qu'un nombre limité de composantes, qu'on peut alors qualifier de « principales » [38].

L'ACP repose sur une matrice de corrélation entre les variables, mais également sur une matrice de corrélation entre les variables observées et les variables construites. Il est habituel de représenter graphiquement les variables observées sur des « cercles de corrélation », qui rendent compte de leur corrélation linéaire avec les composantes principales [38].

Lorsqu'elle est appliquée à des données longitudinales, par exemple en s'intéressant à un seul biomarqueur dont les différentes mesures à des temps successifs seraient des variables observées, l'ACP permet d'identifier des profils-types (clusters) à partir des composantes principales de chaque individu. Elle requiert cependant des temps d'observation identiques pour tous les individus et est purement descriptive, sans possibilité de prédire d'autres variables [38].

2.1.4 Modèles linéaires mixtes

Les Modèles Linéaires Mixtes (*Linear Mixed Models* ou LMM) sont la méthode de référence pour l'analyse de données longitudinales continues [17,27,28]. Le terme "mixte" résulte de la coexistence de deux types d'effets : la trajectoire moyenne de la population – les effets fixes – et la variabilité inter-individuelle – les effets aléatoires. Les LMM modélisent un paramètre comme suit :

$$Y_{i,t} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}).t + \varepsilon_{i,t}$$

- Le terme $Y_{i,t}$ est la mesure du paramètre chez le patient i au temps t .
- Le terme β_0 est l'intercept moyen de la population étudiée (valeur moyenne de Y quand $t = 0$).
- Le terme b_{0i} est l'intercept dit « aléatoire », c'est-à-dire la différence entre β_0 et l'intercept du patient i . La somme $(\beta_0 + b_{0i})$ représente l'intercept propre du patient i .
- Le terme β_1 est la pente moyenne de la population.
- Le terme b_{1i} est la pente dite « aléatoire », c'est-à-dire la différence entre la pente de la population et celle du patient i .
- Le terme $\varepsilon_{i,t}$ est l'erreur résiduelle au temps t .

Avec les LMM, les mesures peuvent être effectuées à des temps différents pour chaque patient. Ces modèles gèrent les données manquantes et les temps d'observation irréguliers grâce aux hypothèses MAR (*Missing At Random*) ou MCAR (*Missing Completely At Random*) [17,39,40].

Les effets fixes permettent d'inclure autant de variables que nécessaire (continues ou catégorielles), et de tester leurs interactions avec le temps. Grâce aux effets aléatoires, chaque patient a sa propre trajectoire avec pente et intercept individuels qu'il est possible d'utiliser pour créer des clusters de patients similaires [27,28]. On peut en conséquence prédire la trajectoire d'un individu spécifique, conditionnellement à ses mesures passées.

Les LMM acceptent des durées de suivi inégales entre individus, sous réserve que la fin du suivi ne soit pas liée à la valeur du paramètre étudié (censure dite « informative »), ce qui est fréquent dans le domaine médical [17,20].

Un LMM standard ne modélise qu'un seul paramètre à la fois. Mais les modèles linéaires mixtes multivariés permettent de modéliser deux paramètres conjointement et d'estimer leur corrélation au cours du temps [23,24].

2.1.5 Tableau récapitulatif

Le tableau suivant (Tableau 1) résume les différentes méthodes classiques pour analyser les données répétées dans le temps, via leur capacité à décrire l'évolution du paramètre étudié (1), à créer des clusters de patients (2), à prédire une ou plusieurs autre(s) variable(s) par le paramètre étudié (3), à prédire le paramètre étudié par une ou plusieurs autre(s) variable(s) (4), à prédire l'évolution du paramètre (5), à analyser l'évolution jointe de deux paramètres (6), à gérer l'asynchronicité des mesures (7) et l'inégalité des durées de suivi (8).

Tableau 1. Résumé des méthodes d'analyse classique, Y = biomarqueur étudié

Fonctionnalités Méthodes	1. Décrire l' évolution de Y	2. Clustering d' individus	3. Prédire de nouvelle(s) variable(s) par Y	4. Prédire Y par d' autre(s) variable(s)	5. Prédire l' évolution de Y	6. Analyser l' évolution jointe de Y_1 et Y_2	7. Gérer l' asynchronicité des mesures	8. Gérer l' inégalité des durées de suivi
Analyses de survie	✓	✗	✗	✓	✓	✗	✓	✓
Analyse de variance répétée	✓	✗	✗	✗	✗	✗	✗	✗
Analyse en Composante Principale	✓	✓	✗	✗	✗	✓	✗	✗
Modèles Linéaires Mixtes	✓	✗	✗	✓	✓	✓	✓	✗

2.2 Analyse de données sous forme fonctionnelle

2.2.1 Définition et problématique

En biostatistiques, on parle de donnée fonctionnelle lorsque l'unité d'observation est une courbe $x_i(t)$ définie sur un intervalle de temps $[0, T]$ plutôt qu'un simple vecteur reprenant la valeur et le temps de mesure [7–9]. En pratique, on observe un nombre fini de mesures pour chaque patient, mais l'idée est de traiter ces mesures comme des échantillons bruités d'une fonction continue sous-jacente [7,11].

Le principe fondamental de l'analyse de données fonctionnelles dans le domaine de la recherche médicale consiste donc à traiter les observations clinico-biologiques non pas comme des données isolées mais comme des fonctions continues évoluant dans le temps [8,15,16].

Cette approche offre plusieurs avantages par rapport aux méthodes classiques qui traitent les mesures répétées. Elle exploite la continuité de la trajectoire (deux mesures proches dans le temps sont plus similaires que deux mesures éloignées), elle est moins sensible aux mesures manquantes isolées, et elle permet d'inclure autant de données que disponibles pour chaque patient, même si toutes n'ont pas été mesurées aux mêmes temps [9,11,15].

Dans le domaine de la recherche médicale observationnelle, l'analyse des données fonctionnelles se heurte à plusieurs problèmes pratiques :

- Tous les patients ne sont pas suivis pendant la même durée. Certains sont perdus de vue, d'autres décèdent en cours d'étude. Les méthodes qui supposent une durée de suivi identique pour tous les patients sont donc inadaptées à la plupart des données réelles [5,20].
- Les instants de mesure ne coïncident pas toujours exactement en vie réelle, même quand deux paramètres sont censés être mesurés le même jour. Il s'agit de la notion d'asynchronicité [5,6].
- Concernant les données de soin issues des dossiers médicaux informatisés, les patients ne viennent pas en consultation à intervalles fixes. La probabilité de consultation est plus élevée si l'état clinique ou biologique du patient est altéré. Les données manquantes informatives qui en résultent ne permettent pas de respecter les hypothèses de la plupart des modèles classiques [17,18].

Pour y remédier, plusieurs méthodes d'analyse des données fonctionnelles ont été développées dans la littérature [8,9,41], que nous précisons ci-après.

2.2.2 Lissage et représentation fonctionnelle

Avant d'effectuer une analyse de données sous forme fonctionnelle, il est nécessaire de convertir préalablement les données discrètes en fonctions continues via des méthodes de lissage [7,8,10]. Ces dernières ne seront pas abordées ici.

2.2.3 Analyse en composantes principales fonctionnelle

La *Functional Principal Component Analysis* (FPCA) est la méthode la plus utilisée pour l'analyse des données sous forme fonctionnelle [8,9,42]. Elle consiste en une application de l'Analyse en Composante Principale classique (cf. 2.1.3) au cas où les observations sont des courbes. Son principe est de décomposer la variabilité des trajectoires individuelles en un nombre réduit de fonctions propres (qu'on appelle « *eigenfunctions* »), chaque patient étant alors caractérisé par les scores qu'obtient sa courbe d'observation pour ces fonctions [42,43].

Avec la FCPA, la courbe d'un patient est définie comme suit :

$$X_i(t) = \mu(t) + \sum_k \xi_{ik} \cdot \varphi_{ik}(t)$$

- Le terme $X_i(t)$ est la courbe observée du sujet i .
- Le terme $\mu(t)$ est la courbe observée moyenne de la population à chaque instant t .
- Le terme $\varphi_{ik}(t)$ est la composante principale k pour le sujet i , il s'agit d'une fonction, l'*eigenfonction*, représentant un mode de variation. La composante φ_{i1} est la direction de variation la plus importante, φ_{i2} la deuxième direction la plus importante, etc.
- Le terme ξ_{ik} est le coefficient ou score, il représente la quantité du mode de variation φ_{ik} pour le patient i .

La FPCA peut servir à la fois à visualiser des tendances (composante descriptive) et à construire des indicateurs pour un modèle de régression (composante prédictive) [14,42]. Cette méthode gère bien les durées de suivi hétérogènes et les données manquantes, mais elle est fondamentalement univariée : elle ne traite qu'un paramètre à la fois [11,43,44].

En outre, même s'il est possible de créer des clusters de patients à partir des scores issus de la FPCA, ces derniers ne sont pas nécessairement les plus pertinents pour séparer les différents groupes [22,45].

2.2.4 Régression fonctionnelle

Le principe de la régression fonctionnelle consiste à appliquer la régression linéaire classique (cf. 3) aux situations où la variable à expliquer et/ou les variables explicatives sont des fonctions [8,9].

Ces modèles ont un rôle purement prédictif, ils ne permettent pas de décrire l'évolution d'un paramètre ni de créer des clusters de patients. Il existe trois types de régressions fonctionnelles, présentées ci-après.

2.2.4.1 *Scalar-on-function*

Dans la régression scalaire-sur-fonction, la variable explicative est une fonction mais pas la variable à expliquer [8,9]. L'équation de régression s'exprime comme suit :

$$Y_i = \alpha + \int \beta(t).X_i(t) d(t) + \varepsilon_i$$

- Le terme Y_i est la variable à expliquer chez un patient i .
- Le terme α est l'intercept, soit la valeur moyenne de Y lorsque $X(t) = 0$.
- Le terme $\beta(t)$ est une fonction dite « coefficient », représentant le caractère informatif de $X_i(t)$ sur la variable à expliquer. Plus β est élevée à un instant t , plus X_i permet de prédire Y_i à ce même instant t .
- Le terme $X_i(t)$ est la courbe de la variable explicative chez le patient i .
- Le terme ε_i est l'erreur résiduelle.

2.2.4.2 *Function-on-scalar*

Dans la régression fonction-sur-scalaire, la variable à expliquer est une fonction mais pas la variable explicative [8,9]. L'équation de régression s'exprime comme suit :

$$Y_i(t) = \beta_0(t) + \beta_1(t).X_i + \varepsilon_i(t)$$

- Le terme $Y_i(t)$ est la courbe ou fonction représentant la variable à expliquer chez un patient i .
- Le terme $\beta_0(t)$ est l'intercept fonctionnel. Il s'agit de la courbe moyenne $Y(t)$ de tous les patients pour lesquels $X = 0$.
- Le terme $\beta_1(t)$ est une fonction dite « coefficient », représentant l'effet de X_i sur la valeur de Y_i au temps t .
- Le terme X_i est la variable explicative chez le patient i .
- Le terme $\varepsilon_i(t)$ est l'erreur résiduelle, variable dans le temps.

2.2.4.3 Function-on-function

Dans la régression fonction-sur-fonction, la variable à expliquer ainsi que la variable explicative sont toutes deux des fonctions [8,9]. L'équation de régression s'exprime comme suit :

$$Y_i(s) = \alpha(s) + \int \beta(s, t).X_i(t) d(t) + \varepsilon_i(s)$$

- Le terme $Y_i(s)$ est la courbe ou fonction représentant la variable à expliquer chez un patient i au temps s .
- Le terme $\alpha(s)$ est l'intercept fonctionnel. Il s'agit de la courbe moyenne $Y(s)$ de tous les patients pour lesquels $X(t) = 0$.
- Le terme $\beta(s, t)$ est un coefficient sous forme de fonction à deux dimensions, représentant l'effet de X_i au temps t sur la valeur de Y_i au temps s .
- Le terme X_i est la variable explicative chez le patient i .
- Le terme $\varepsilon_i(s)$ est l'erreur résiduelle au temps s .

2.2.5 Modèles joints longitudinal-survie

Les modèles joints longitudinal-survie combinent deux sous-modèles, à savoir un modèle linéaire mixte (cf. 2.1.4) et un modèle de survie (cf. 2.1.1), reliés par un paramètre d'association α qui quantifie l'effet de la trajectoire longitudinale d'une variable (modèle linéaire mixte) sur le risque d'un événement (modèle de survie) [46–48]. On obtient alors l'équation suivante, qui se base sur un modèle de Cox (cf. 2.1.1.3) :

$$h(t) = h_0(t) \times e^{(\gamma.W_i + \alpha.m_i(t))}$$

- Le terme $h(t)$ est le risque à l'instant t .
- Le terme $h_0(t)$ est le risque de base, estimé lorsque W_i et $m_i(t)$ sont nuls, et représentant l'évolution du risque dans la population.
- Le terme γ est le vecteur des coefficients des variables fixes W_i (telles que l'âge à l'inclusion, le sexe ou le groupe de traitement) chez le patient i .
- Le terme α est le paramètre d'association du modèle.
- Le terme $m_i(t)$ est la trajectoire de la variable explicative, définie par le modèle linéaire mixte comme $(\beta_0 + b_{0i}) + (\beta_1 + b_{1i}).t$ (cf. 2.1.4).

Si le paramètre d'association $\alpha > 0$, alors la variable explicative augmente le risque de survenue de l'événement. Si $\alpha < 0$ la variable explicative diminue le risque. Si $\alpha = 0$ la variable et le risque sont indépendants [47,49].

Les modèles joint longitudinal-survie corrigent le biais d'information, lié à la dépendance entre le suivi et l'événement, ainsi que les biais de mesure [20,46,49].

2.2.6 Tableau récapitulatif

Le tableau ci-dessous (Tableau 2) résume les différentes méthodes pour analyser les données sous forme fonctionnelle, via leur capacité à décrire l'évolution du paramètre étudié (1), à créer des clusters de patients (2), à prédire une ou plusieurs autre(s) variable(s) par le paramètre étudié (3), à prédire le paramètre étudié par une ou plusieurs autre(s) variable(s) (4), à prédire l'évolution du paramètre (5), à analyser l'évolution jointe de deux paramètres (6), à gérer l'asynchronicité des mesures (7) et l'inégalité des durées de suivi (8).

Tableau 2. Résumé des méthodes d'analyse de données sous forme fonctionnelle, Y = biomarqueur étudié

Fonctionnalités Méthodes	1. Décrire l' évolution de Y	2. Clustering d' individus	3. Prédire de nouvelle(s) variable(s) par Y	4. Prédire Y par d' autre(s) variable(s)	5. Prédire l' évolution de Y	6. Analyser l' évolution jointe de Y ₁ et Y ₂	7. Gérer l' asynchronicité des mesures	8. Gérer l' inégalité des durées de suivi
Analyse en composante principale fonctionnelle (FCPA)	✓	✓	✓	✗	✗	✗	✗	✗
Régressions fonctionnelles	✗	✗	✓	✓	✓	✓	✗	✗
Modèles joints longitudinal-survie	✓	✗	✓	✗	✓	✗	✓	✓

3 Régression linéaire

3.1 Régression linéaire simple

La régression linéaire simple est une méthode statistique permettant de décrire la relation entre une variable à expliquer Y et une variable explicative X sous la forme $Y = a.X + b + \varepsilon$, où a représente la pente, b l'ordonnée à l'origine et ε l'erreur ou résidu, c'est-à-dire la différence entre les valeurs prédites $\hat{Y}_i$ et les valeurs observées Y_i [50–52].

Cette méthode présente un intérêt explicatif, dans la mesure où il sera ainsi possible de commenter l'association entre X et Y , tant en termes de taille d'effet (valeur du coefficient associé à X) que de significativité (position de l'intervalle de confiance à 95% par rapport à la valeur zéro, ou p-valeur, selon la méthode retenue) [51–53].

La régression linéaire simple présente un intérêt prédictif. Il est possible de prédire une valeur y_0 inconnue, dès lors que l'on connaît les coefficients de cette équation et la valeur de x_0 [50,52].

La méthode est agnostique de la relation sémantique entre les deux variables, pour peu qu'elles soient quantitatives. On peut alors étudier, par exemple, l'association entre des phénomènes différents mesurés chez un même individu, l'association entre des mesures successives d'un même phénomène, ou bien la variabilité de mesure entre deux opérateurs [50,53].

3.1.1 Indicateurs

Les indicateurs suivants contribuent à la mise au point d'un modèle de régression linéaire, ou permettent de l'évaluer a posteriori.

3.1.1.1 Covariance

La covariance entre 2 variables X et Y représente la quantité – positive ou négative – de variation de Y lorsque X augmente [50,51]. Elle est définie dans l'échantillon comme étant la moyenne du produit des écarts à la moyenne :

$$cov_{ech}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n}$$

Son principal défaut est de tenir compte des unités des variables X et Y , rendant son interprétation difficile. Le coefficient de corrélation linéaire de Pearson, présenté ci-après, résulte de la standardisation de la covariance [50,52].

3.1.1.2 Coefficient de corrélation linéaire de Pearson

Le coefficient théorique de corrélation linéaire ρ (rhô), compris entre -1 et 1 , permet de s'affranchir des ordres de grandeurs des variables X et Y [50–52]. Plus sa valeur absolue est élevée, plus les variables X et Y sont corrélées, avec une valeur de 0 lorsqu'il n'existe pas de corrélation, 1 lorsque la corrélation est totale et dans le même sens, et -1 lorsque la corrélation est totale et en sens contraire.

En population, ρ est le quotient de la covariance et du produit des écarts-types :

$$\rho = \frac{cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

L'estimation ponctuelle de cette quantité dans l'échantillon est le coefficient empirique de corrélation linéaire de Pearson, noté r :

$$r = \frac{cov_{ec}(X, Y)}{s_{ec\ x} \cdot s_{ec\ y}}$$

3.1.1.3 Coefficients de l'équation

Grâce aux indicateurs précédents, il est possible de déterminer les coefficients a et b de l'équation de régression $Y = a.X + b + \varepsilon$ comme suit [50,51] :

$$a = r \cdot \frac{s_Y}{s_X} = \frac{cov_{ec}(X, Y)}{s_X^2}$$

$$b = \bar{y} - a \cdot \bar{x}$$

3.1.1.4 Test de la pente

La pente de la régression peut être associée à un test de nullité. Ce test est équivalent au test de nullité du coefficient de corrélation linéaire de Pearson [51,52]. Il permet de déterminer la significativité de la corrélation linéaire entre les variables X et Y . Il est à noter que si l'indépendance entre deux variables implique une corrélation nulle, la réciproque n'est pas vraie : une corrélation nulle peut exister alors que les deux variables ne sont pas indépendantes [50].

3.1.1.5 Définition des résidus

Les coefficients présentés précédemment sont calculés par la méthode des moindres carrés, qui permet de minimiser la somme des carrés des résidus [50–52].

Pour chaque individu, le résidu est la différence entre la valeur prédite par le modèle et la valeur réellement observée :

$$\varepsilon_i = \hat{y}_i - y_i$$

Il correspond à la distance verticale, pour un individu donné, entre le point représentant cet individu et la droite de régression.

On peut également parler du résidu comme d'une nouvelle variable, qui sera étudiée dans le processus de validation :

$$\varepsilon = \hat{Y} - Y$$

3.1.2 Critères de validation

Les critères de validation suivants sont nécessaires pour effectuer une régression linéaire. En pratique, ils sont vérifiés a posteriori, la régression étant un processus itératif qui permet au final de générer le modèle le plus approprié [50–52]. Les critères que nous présentons ci-après répondent fondamentalement à trois questions :

- Celle du pouvoir explicatif, et inversement du risque de surajustement [50,53].
- Celle de la relation entre X et Y , qui devrait être soit linéaire, soit constituée d'une composante linéaire et d'un bruit (dans des proportions plus ou moins importantes), mais en aucun cas refléter une autre forme fonctionnelle [50,51].
- Celle de la généralisation possible du résultat obtenu, qui ne devrait pas avoir subi l'influence excessive de quelques individus [50,52].

3.1.2.1 Coefficient de détermination

Le coefficient de détermination R^2 est un indicateur de performance, exprimant la proportion de la variance de Y expliquée par X [50–52]. Il prend une valeur comprise entre 0, lorsque le modèle de régression linéaire n'explique pas du tout la variance de Y , et 1, lorsque la valeur de X explique totalement celle de Y . R^2 est défini comme suit :

$$R^2 = 1 - \frac{SS_{\text{résiduel}}}{SS_{\text{total}}} = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y} - y_i)^2}$$

Avec $SS_{\text{résiduel}}$ la somme des carrés des résidus et SS_{total} la somme totale des carrés des différences à la moyenne de la variable Y à expliquer.

Il est à noter que, dans une régression linéaire simple, ce coefficient R^2 est strictement le carré du coefficient de corrélation linéaire de Pearson, calculé entre la variable Y et la variable X . Plus généralement, dans une régression linéaire simple ou multiple, le coefficient R^2 est strictement le carré du coefficient de corrélation linéaire de Pearson, calculé non plus entre Y et X mais entre la variable Y observée et la variable $\hat{Y}$ prédite [50,51].

3.1.2.2 La parcimonie

Pour éviter le risque de surajustement entre les variables explicatives, leur nombre doit être le plus faible possible avec un coefficient R^2 le plus élevé possible : il s'agit de la notion de parcimonie (cf. 3.2.4.1), d'emblée vérifiée lors de la régression linéaire simple [50,52,53].

3.1.2.3 La normalité des résidus

Les résidus doivent suivre une loi normale (Figure 1). Ce critère est souvent considéré comme acquis si X et Y suivent une loi normale, même si ce n'est pas toujours nécessaire [50,51]. On notera que, par construction, la moyenne des résidus est toujours nulle.

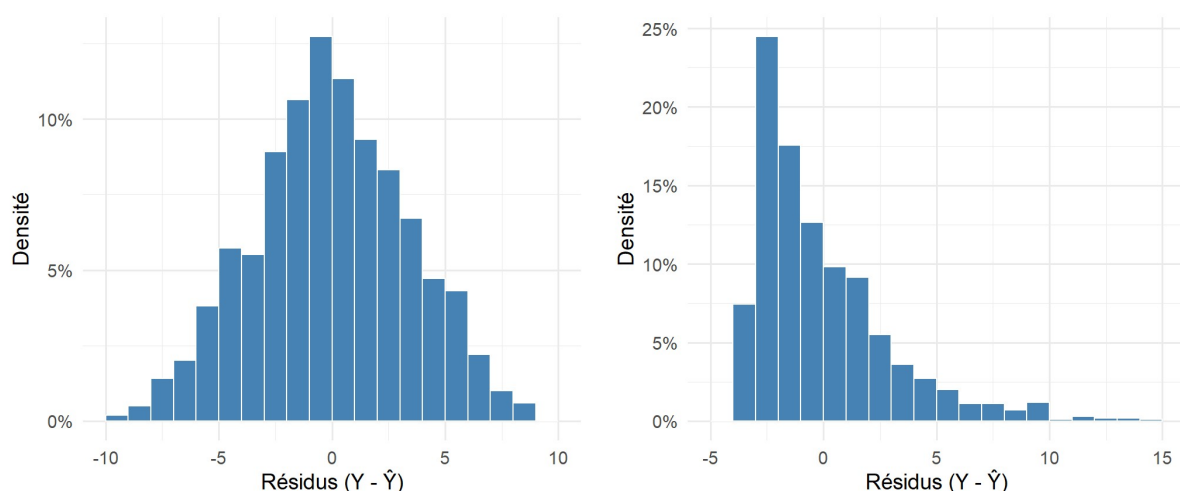


Figure 1. Exemple de distribution normale (à gauche) ou non-normale (à droite) des résidus

3.1.2.4 L'indépendance des résidus

Lorsque la moyenne des résidus n'est pas indépendante de $\hat{y}$, c'est qu'il existe une relation fonctionnelle non encore découverte (Figure 2) [50,51]. Fondamentalement, cela signifie que la relation entre X et Y n'est pas uniquement linéaire.

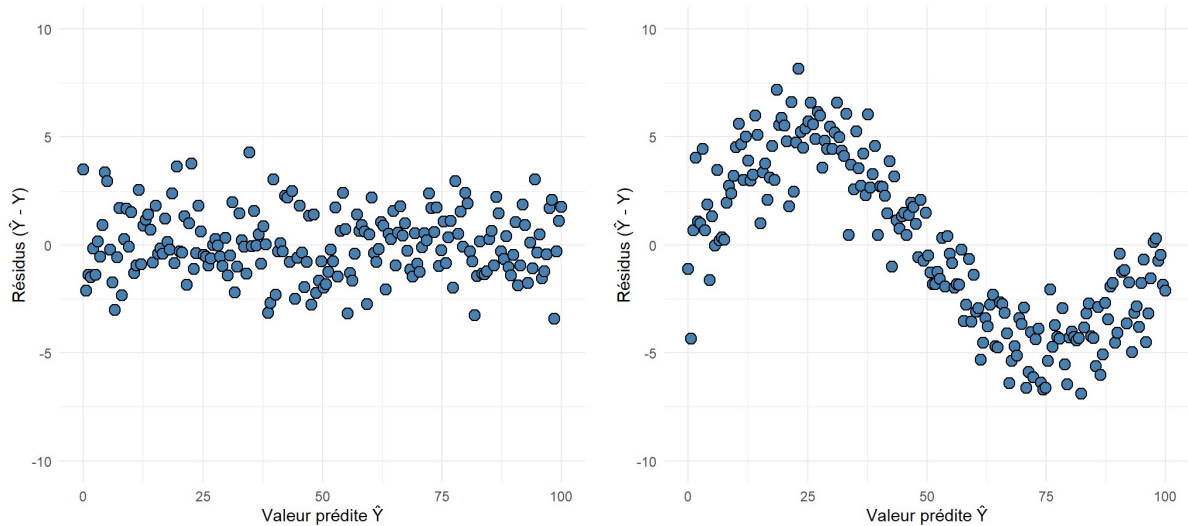


Figure 2. Exemple de distribution des résidus avec moyenne et variance indépendantes de $\hat{Y}$ (à gauche), ou moyenne dépendante de $\hat{Y}$ (à droite)

3.1.2.5 L'homoscédasticité des résidus

Lorsque la variance des résidus ne dépend pas de $\hat{y}$ (Figure 3), on parle d'homoscédasticité des résidus [50–52]. Lorsqu'elle est dépendante de $\hat{y}$, on parle d'hétéroscédasticité.

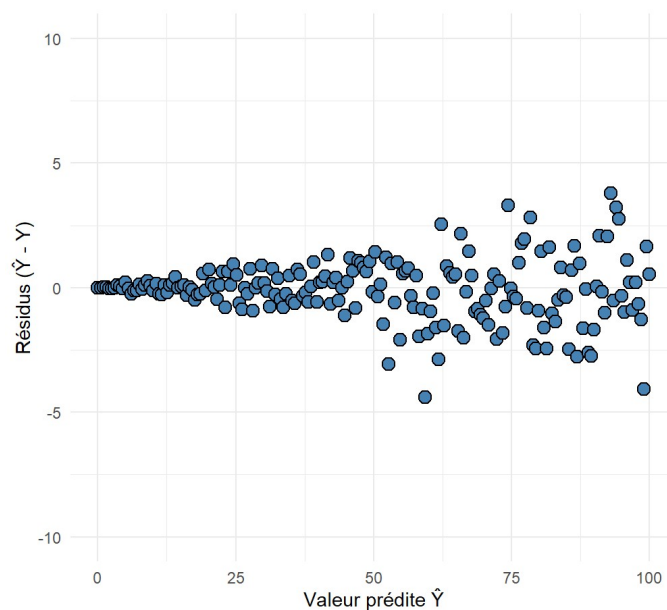


Figure 3. Exemple de distribution des résidus avec variance dépendante de $\hat{Y}$

3.1.2.6 La distance de Cook

La distance de Cook (Figure 4) est un indicateur d'influence des individus [50,51]. Elle est toujours positive et proportionnelle au poids que l'individu exerce sur les coefficients du modèle. Elle doit être la plus faible possible, mais il n'existe pas de seuil d'interprétation : le choix des individus « influents » et la décision de les supprimer se fait empiriquement.

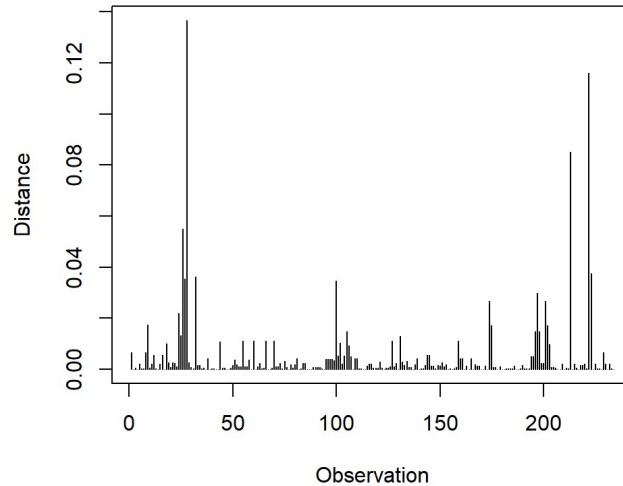


Figure 4. Exemple de distribution des distances de Cook chez 234 individus

3.1.3 Prédiction

Une fois la régression effectuée, l'équation $\hat{y}_0 = a \cdot x_0 + b$ permet de prédire une valeur $\hat{y}_0$ pour un nouvel individu dont la valeur x_0 est connue mais pas la valeur y_0 [50,52]. Cette prédiction est assortie d'un intervalle de confiance à 95% qui n'est valide que si la régression est elle-même valide [50,51] :

$$IC_{\hat{y}_0; 1-\alpha} = \hat{y}_0 \pm u_{\alpha} \cdot s_{YX} = \hat{y}_0 \pm u_{\alpha} \cdot s_Y \cdot \sqrt{1 - r^2}$$

Avec s_{YX} l'écart-type de l'erreur de prédiction.

3.2 Régression linéaire multiple

3.2.1 Généralités

La régression linéaire multiple suit les mêmes principes généraux que la régression linéaire simple, avec plusieurs variables explicatives X_i , quantitatives ou binaires, au lieu d'une seule [50–52]. L'équation de régression prend alors la forme suivante :

$$Y = a_1 \cdot X_1 + a_2 \cdot X_2 + \dots + a_p \cdot X_p + b + \varepsilon$$

Avec p le nombre de variables explicatives, b l'ordonnée à l'origine et ε le résidu. Les coefficients a_i de l'équation s'interprètent directement comme l'augmentation de Y lorsque la variable X_i augmente d'une unité, toutes choses restant égales par ailleurs [50,51].

La régression linéaire multiple est une méthode multivariée, c'est-à-dire qu'elle isole l'effet propre dit « ajusté » de chacune des variables explicatives, en tenant compte de l'effet attribué aux autres variables explicatives [50,52].

Elle repose sur une hypothèse d'additivité des effets des variables X_i qui n'est pas toujours pertinente en santé. Les effets sont généralement conditionnels en vie réelle. Dans les modèles de régression, ces effets conditionnels peuvent être pris en compte à l'aide d'interactions entre variables [50,51]. Il est à noter que certaines familles de méthodes, comme les arbres de régression, offrent une prise en compte plus complète des interactions entre variables explicatives [54].

3.2.2 Limites du modèle

Plusieurs risques doivent être pris en compte lors de la réalisation d'une régression linéaire multiple [50–52] :

- Un faible pouvoir explicatif lorsque Y est peu lié aux variables X_i , que l'on peut estimer en calculant le coefficient de détermination R^2 (cf. 3.1.2.1).
- Un risque de surajustement du modèle lorsque trop de variables X_i sont utilisées, qui peut être étudié en calculant un indice de parcimonie ainsi qu'en effectuant un processus itératif de sélection des variables (cf. 3.2.3).
- Un risque d'instabilité lorsque les variables X_i sont trop corrélées entre elles (on parle de multicollinéarité), que l'on peut résoudre en explorant ces mêmes corrélations (cf. 3.2.4.2).
- Un risque d'inadaptation du modèle lorsque la relation entre Y et les variables X_i n'est pas linéaire, qui peut être détecté par l'analyse de la distribution des résidus (cf. 3.1.1.5).
- Un risque d'erreur lorsque des individus trop influents sont présents, que l'on peut étudier à l'aide de la distance de Cook (cf. 3.1.2).

3.2.3 Processus itératif

En réalisant la régression linéaire de manière itérative, on peut réduire progressivement les risques cités précédemment.

Avant d'effectuer une étape de régression, les variables Y et X_i peuvent être transformées. Aussi, certaines variables X_i peuvent être supprimées, soit lorsqu'elles sont trop corrélées entre elles, soit lorsque la corrélation avec la variable Y en analyse bivariable n'est pas significative, généralement à un seuil élevé comme 20% pour conserver un maximum de variables X_i [50,52].

Une association non significative dans une régression linéaire multiple entre Y et une variable X_i n'implique ni une indépendance, ni même l'absence d'association linéaire entre les deux.

La régression est alors effectuée pour obtenir les coefficients a_i de l'équation avec leurs *p-valeurs*. Le coefficient de détermination R^2 , la distribution des résidus et la présence d'individus marginaux (distance de Cook) doivent être vérifiés à chaque fois [50,51].

On peut ensuite transformer de nouveau les variables X_i , les filtrer selon des critères de significativité ou de parcimonie (indices *AIC* ou *BIC*, cf. 3.2.4.1), mais aussi supprimer certains individus [50,53].

Retirer une variable modifie systématiquement les coefficients de l'équation. Ce processus sera effectué autant de fois que nécessaire pour obtenir un modèle de régression adapté à la question initiale. On rappellera qu'une association significative ne démontre pas le lien de causalité [51,52].

3.2.4 Critères propres à la régression linéaire multiple

Les critères de validation présentés ci-après sont propres à la régression linéaire multiple, du fait de l'existence de plusieurs variables explicatives contre une seule dans la régression linéaire simple.

3.2.4.1 Indices de parcimonie

Le coefficient de détermination R^2 augmente avec le nombre de variable explicatives présentes dans le modèle, jusqu'à ce que ces dernières deviennent des critères d'identification de l'individu et de sa valeur Y_i [50,51].

Pour y remédier, des critères de parcimonie comme l'*Akaike information criterion* (AIC) et le *bayesian information criterion* (BIC) peuvent être utilisés [50–52]. Ils rendent compte à la fois de la valeur de R^2 , que l'on souhaite la plus élevée possible, et du nombre de variables explicatives, que l'on souhaite le plus faible possible. Ils sont calculés à chaque itération de la régression, notamment lorsqu'on ajoute ou supprime des variables explicatives, avec pour objectif leur minimisation. Ils présentent un critère objectif de choix pour conserver ou évincer des variables explicatives, mais ne sont pas les seuls arguments possibles. Ils sont notamment utilisés dans les procédures de sélection pas-à-pas (*stepwise*) de variables explicatives [50,53].

3.2.4.2 Multicolinéarité

La présence de plusieurs variables X_i très corrélées entre elles ne rend pas invalide le modèle de régression, mais peut aboutir à des erreurs d'interprétation voire de prédiction [50,51].

Il est possible de diagnostiquer ce risque en exécutant une matrice de corrélation 2 à 2, consistant à tester la corrélation de chaque variables X_i entre elles. On peut aussi supprimer certaines variables selon des critères experts comme la difficulté de recueil des données, la similarité, etc. [50,52].

4 Objectif

Dans la première partie, nous avons défini les données fonctionnelles, et les enjeux et perspectives qui y sont attachés en recherche biomédicale.

Dans la deuxième partie, nous avons présenté un état de l'art concernant l'analyse de des données fonctionnelles. Ce dernier a permis d'illustrer l'insuffisance des méthodes disponibles, au regard des enjeux listés dans la première partie.

La troisième partie a rappelé les bases théoriques des modèles de régression linéaires, sur lesquels s'appuie la méthode que nous proposons.

L'objectif stratégique de ce travail est de proposer une nouvelle méthode d'analyse de données fonctionnelles.

Après avoir présenté les données utilisées dans notre exemple expérimental, nous définirons une nouvelle méthode provisoirement nommée IRBODETIF (*Individual Regressions Based on Defined Temporally-Independent Functions*).

En analysant les taux d'hémoglobines glyquées de patients, nous montrerons une preuve de concept des applications attendues de cette méthode telles que définies précédemment (voir chapitre 1.4 en page 15), à savoir :

- La description du paramètre.
- La création de clusters.
- La prédiction d'une variable.
- La prédiction par des variables.
- La prédiction de l'évolution du paramètre.
- L'évolution jointe de plusieurs variables.

Matériel et méthodes

1 Données de test

Pour ce travail exploratoire, nous utiliserons un échantillon de données synthétiques générées par la plateforme e-Santé du CPER Tec'Santé.

Il s'agit d'une base de données relationnelle de 349 patients fictifs diabétiques de type 2 suivis durant plusieurs années. Cette base de données est constituée de plusieurs tables, dont on retient :

- Une table « patient », comportant une ligne par patient, et plusieurs colonnes de variables comprenant le sexe, la date de naissance, la date de début de suivi et la date de fin de suivi ;
- Une table de biométrie, comportant notamment 14 428 mesures datées de poids et de taille ;
- Une table de résultats de biologie médicale, comportant notamment 3 375 mesures datées d'HbA1c (hémoglobine glyquée), et reliées chacune à un patient ;
- Une table de médicaments prescrits, comportant notamment les prescriptions datées d'insuline.

Toutes ces données sont asynchrones, d'une part entre patients, et d'autre part entre paramètres mesurés chez les mêmes patients (ex : HbA1c et poids). Ces données sont transformées. En particulier, s'agissant de patients fictifs adultes, la taille est considérée comme étant constante sur toute la durée du suivi, ce qui permet de calculer l'IMC (indice de masse corporelle) à chaque mesure du poids. Nous constituons ainsi formellement le jeu de données suivant :

- Une table des patients comportant :
 - o L'âge à l'inclusion en années.
 - o Le sexe.
 - o L'événement temps-dépendant « mise sous insuline » (avec une censure à droite en l'absence d'observation de cet événement au terme du suivi), décrit sous la forme d'une colonne événement/censure, et une colonne durée de suivi en jours.
 - o L'événement binaire « mise sous insuline à terme ».
 - o L'événement binaire « survenue de complication(s) ».
 - o Le taux d'HbA1c à l'inclusion.
 - o L'IMC à l'inclusion.
- Pour chaque patient, une série datée de valeurs d'HbA1c (valeur et jour relatif de la mesure, en commençant à zéro).
- Pour chaque patient, une série datée de valeurs d'IMC (valeur et jour relatif de la mesure, en commençant à zéro).

Seuls les patients ayant été suivis pendant au moins la moitié du suivi global (soit 1226 jours) et ayant effectué 3 mesures ou plus d'HbA1c avant d'être mis sous insuline ont été inclus, soit 134 patients sur 349.

2 Définition de la méthode

2.1 Principe général

L'idée générale est de décrire la trajectoire de chaque patient (en termes de valeurs datées d'HbA1c) comme étant la combinaison linéaire de plusieurs fonctions dépendantes du temps. La méthode que nous proposons peut s'appuyer sur différents jeux de fonctions, définies théoriquement ou construites par l'observation. Nous utiliserons ici un jeu de fonctions définies *a priori*. Cela permettra de définir, pour chaque patient, un jeu de coefficients complets, c'est-à-dire sans valeur manquante. Ces coefficients seront ensuite utilisés à diverses fins.

2.2 Définition des fonctions sous-jacentes

Nous définissons tout d'abord 4 fonctions primaires. Elles sont définies de manière arbitraire, avec si possible un objectif d'indépendance théorique entre elles par rapport au temps : cela signifie que, pour un individu, un coefficient multiplicateur obtenu vis-à-vis d'une fonction ne détermine pas théoriquement le coefficient vis-à-vis des autres fonctions. Cette indépendance est théorique : rien ne dit qu'elle soit réelle dans un échantillon ou en population (il est possible, par exemple, que les individus qui partent de très haut aient tendance à baisser plus que les autres). Cette indépendance est souhaitable, mais pas impérative : elle simplifie l'interprétation des coefficients multiplicateurs par la suite.

Le jeu de fonctions primaires que nous utiliserons est composé de :

- $f_1(t) = 1$, soit une fonction constante (Figure 5).
- $f_2(t) = t - \frac{t_{max}}{2}$, soit une fonction linéaire croissante (Figure 6). La valeur t_{max} est la durée maximale de suivi dans l'ensemble des individus (2453 jours dans notre exemple).
- $f_3(t) = -\sin(\frac{\pi \cdot t}{t_{max}}) + \frac{2}{\pi}$, soit une fonction f_3 en cloche (Figure 7).
- $f_4(t) = \sin(\frac{3 \cdot \pi \cdot t}{t_{max}} - \frac{\pi}{2})$, soit une fonction f_4 sinusoïdale (Figure 8).

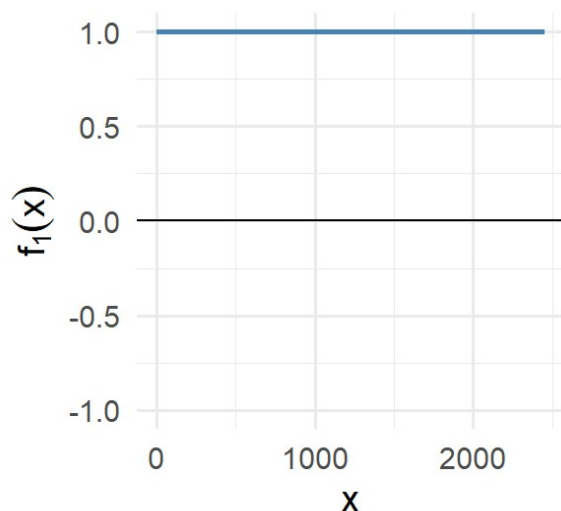


Figure 5. Fonction $f_1(x)$

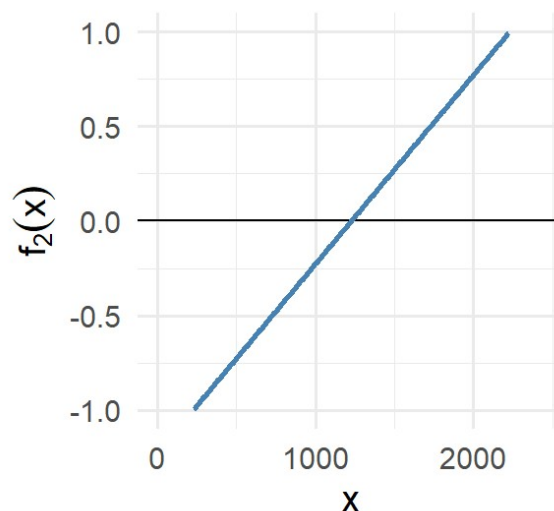


Figure 6. Fonction $f_2(x)$

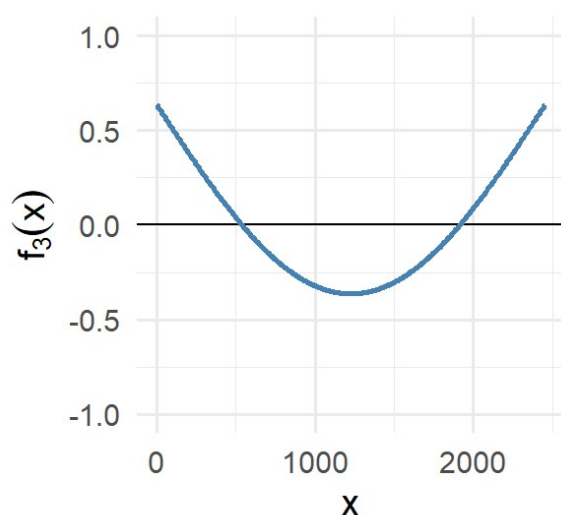


Figure 7. Fonction $f_3(x)$

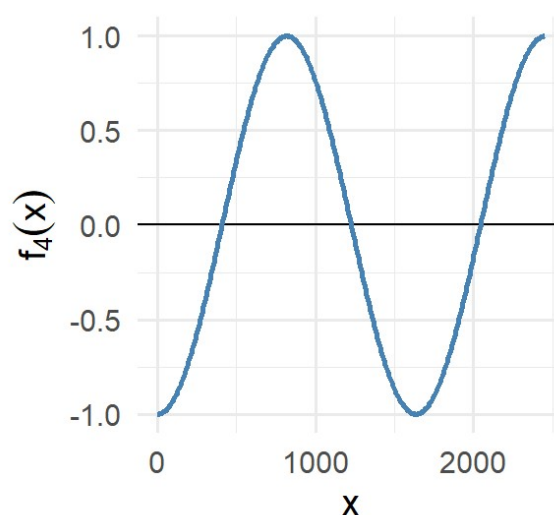


Figure 8. Fonction $f_4(x)$

On notera que ces quatre fonctions ne sont pas toutes théoriquement indépendantes par rapport au temps (t) : il existe une relation croissante entre f_2 et f_4 .

2.3 Gestion des données manquantes

Pour chaque individu, les valeurs manquantes d'hémoglobine glyquée à chaque instant t seront déterminées par interpolation linéaire entre les valeurs précédentes et suivantes réellement observées. Ce procédé ne semble pas techniquement indispensable. Cependant, pour les étapes suivantes, il permet d'obtenir une pondération uniforme des valeurs observées par rapport au temps en cas de mesures réparties de manière hétérogène sur l'axe temporel.

2.4 Calcul des coefficients pour chaque individu

Le principe fondamental de la méthode proposée est d'exprimer, pour chaque individu i séparément des autres, la courbe d'observation f_i de l'hémoglobine glyquée HbA1c (ou bien d'une autre variable observée comme l'IMC) au cours du temps comme étant la combinaison linéaire des quatre fonctions primaires f_1, f_2, f_3 et f_4 , en utilisant une régression linéaire.

Pour chaque individu, cette courbe est donc définie par la fonction suivante :

$$f_i = c_1 \cdot f_1 + c_2 \cdot f_2 + c_3 \cdot f_3 + c_4 \cdot f_4 + \varepsilon$$

Un principe de parcimonie peut être appliqué en imposant un seuil de nombre de mesures disponibles. De tels seuils n'ont pas été imposés dans notre exemple, ayant préalablement limité l'échantillon aux patients présentant au moins 3 mesures.

Lorsque les fonctions sont indépendantes, il est possible de calculer directement les coefficients d'une régression linéaire multiple. Cependant, dans notre exemple et lorsque les fonctions ne sont pas indépendantes, nous proposons de calculer ces coefficients de manière emboîtée : le coefficient c_1 est issu de la régression linéaire simple utilisant uniquement la première fonction :

$$f_i = c_1 \cdot f_1(t) + \varepsilon_1$$

Bien que la fonction $f_1(t)$ soit constante, on notera que c_1 n'est pas nécessairement la moyenne des valeurs, car ces dernières sont généralement distribuées de manière non-uniforme sur l'axe temporel.

Le coefficient c_2 est ensuite obtenu par modélisation du résidu (la partie de gauche de l'équation étant désormais déterminée, seul le coefficient c_2 reste à calculer) :

$$\varepsilon_1 = c_2 \cdot f_2(t) + \varepsilon_2$$

Idem pour les coefficients c_3 et c_4 , qui sont obtenus par modélisation du résidu (la partie de gauche de l'équation étant désormais déterminée, seul le coefficient c_3 reste à calculer, puis le coefficient c_4) :

$$\varepsilon_2 = c_3 \cdot f_3(t) + \varepsilon_3$$

$$\varepsilon_3 = c_4 \cdot f_4(t) + \varepsilon_4$$

En plus des coefficients c_1, c_2, c_3 et c_4 , nous retenons arbitrairement, pour chaque patient, deux variables supplémentaires :

- Le coefficient R^2 de chaque modélisation finalement obtenue.
- Le nombre de mesures d'HbA1c.

Enfin, les variables représentant les coefficients c_1, c_2, c_3 et c_4 sont centrées et réduites.

La méthode reste identique si l'on souhaite utiliser un jeu de fonctions complètement différent.

3 Applications pratiques de cette méthode

Le logiciel « R » avec l'interface graphique « RStudio » est utilisé pour effectuer les calculs et tests présentés dans cette partie, ainsi que pour créer les graphiques.

3.1 Modélisations individuelles

Nous montrerons, pour certains individus, l'aspect de la courbe observée d'HbA1c, et l'aspect modélisable par les coefficients ainsi calculés. Le tout sera résumé dans un tableau présentant la distribution des coefficients et des R^2 .

3.2 Clustering et description de sous-groupes choisis d'individus

Afin d'illustrer les résultats, nous définissons arbitrairement 8 sous-groupes d'individus. Les coefficients c_1 , c_2 et c_3 centrés-réduits sont classés selon leurs signes respectifs (positif ou négatif), afin de créer 8 clusters de patients définis de la manière suivante :

- Cluster 1 : $c_1 \geq 0$; $c_2 \geq 0$; $c_3 \geq 0$
- Cluster 2 : $c_1 \geq 0$; $c_2 < 0$; $c_3 \geq 0$
- Cluster 3 : $c_1 \geq 0$; $c_2 \geq 0$; $c_3 < 0$
- Cluster 4 : $c_1 \geq 0$; $c_2 < 0$; $c_3 < 0$
- Cluster 5 : $c_1 < 0$; $c_2 \geq 0$; $c_3 \geq 0$
- Cluster 6 : $c_1 < 0$; $c_2 < 0$; $c_3 \geq 0$
- Cluster 7 : $c_1 < 0$; $c_2 \geq 0$; $c_3 < 0$
- Cluster 8 : $c_1 < 0$; $c_2 < 0$; $c_3 < 0$

Pour chaque cluster, une courbe sera modélisée à partir de la moyenne des coefficients c_1 à c_4 des patients qui appartiennent à ce cluster. Cette courbe exprime le profil d'évolution moyen de l'HbA1c au sein du cluster.

Sera ensuite testée, en utilisant la méthode CART (*Classification and Regression Trees*, [54]), la capacité à prédire le cluster auquel appartient un patient à partir de 4 variables initialement connues au début du suivi ($J = 0$) :

- L'âge.
- Le sexe.
- La valeur initiale d'HbA1c.
- L'IMC initial.

3.3 Prédiction de la trajectoire d'HbA1c d'après d'autres variables

Des régressions linéaires multiples seront utilisées pour prédire la valeur des coefficients (représentant la trajectoire de l'HbA1c) à partir des 4 variables initialement connues (âge, sexe, HbA1c initiale, IMC initiale).

3.4 Prédiction d'autres variables à l'aide de la trajectoire d'HbA1c

Par la suite, des analyses multivariées seront réalisées pour tester la capacité à prédire la mise sous insuline, la survenue de complication, et le délai avant la mise sous insuline, selon :

- Les coefficients c_1 , c_2 , c_3 et c_4 (représentant la trajectoire de l'HbA1c).
- Les 4 variables initialement connues au début du suivi (âge, sexe, HbA1c initiale, IMC initiale).
- Le nombre d'observations et les coefficients R^2 , obtenus précédemment.

Le modèle de Cox est utilisé pour prédire le délai avant la mise sous insuline. La régression logistique multiple est utilisée pour prédire les 2 autres variables (mise sous insuline et survenue de complication), qui sont binaires.

3.5 Prédiction de la suite d'une trajectoire d'HbA1c

Dans cette partie, la méthode de régression linéaire décrite précédemment (cf. 2.4) est effectuée non plus sur la totalité du suivi, mais sur les trois-quarts initiaux du suivi. Les coefficients ainsi calculés permettront de montrer l'aspect modélisable « précocement » de la courbe d'HbA1c, à côté de la courbe effectivement observée sur l'ensemble le suivi.

3.6 Description de l'évolution conjointe de l'HbA1c et de l'IMC

Enfin, seront effectuées des régressions linéaires individuelles selon la méthode décrite précédemment (cf. 2.4), appliquées non plus à la courbe d'observation de l'HbA1c mais à celle de l'IMC. Seront ainsi obtenus, pour chaque patient, 4 coefficients définissant la trajectoire de l'IMC comme étant la combinaison linéaire des quatre fonctions primaires f_1 , f_2 , f_3 et f_4 .

À l'aide d'une matrice de corrélation, les coefficients issus des régressions sur les courbes d'IMC seront comparés aux coefficients issus des régressions sur les courbes d'HbA1c.

4 Cadre réglementaire

4.1 Autorisations

La présente étude est réalisée sur des données simulées dans le cadre du CPER Tec'Santé. Ne s'agissant pas de données réelles ni de recherche sur les personnes, aucune autorisation n'est nécessaire. Aucune information des patients n'est requise, ni consentement.

4.2 Financement

Le CPER Tec'Santé est financé par la Région Hauts-de-France.

Cette étude a été réalisée dans le cadre d'un travail individuel de thèse de médecine. L'auteur déclare n'avoir aucun conflit d'intérêt financier ou personnel avec des entreprises privées du secteur de la santé en lien avec les thématiques abordées.

Résultats



À des fins didactiques, nous expliciterons la lecture des résultats expérimentaux issus de notre base de données à l'aide d'un paragraphe marqué de cette icône.

1 Diagramme de flux (flowchart)

Sur les 349 patients présents dans la base de données, 215 patients ont eu soit un temps de suivi inférieur à la moitié du suivi global, soit moins de 3 mesures d'HbA1c. Au total, 134 patients ont été inclus dans cette analyse (Figure 9).

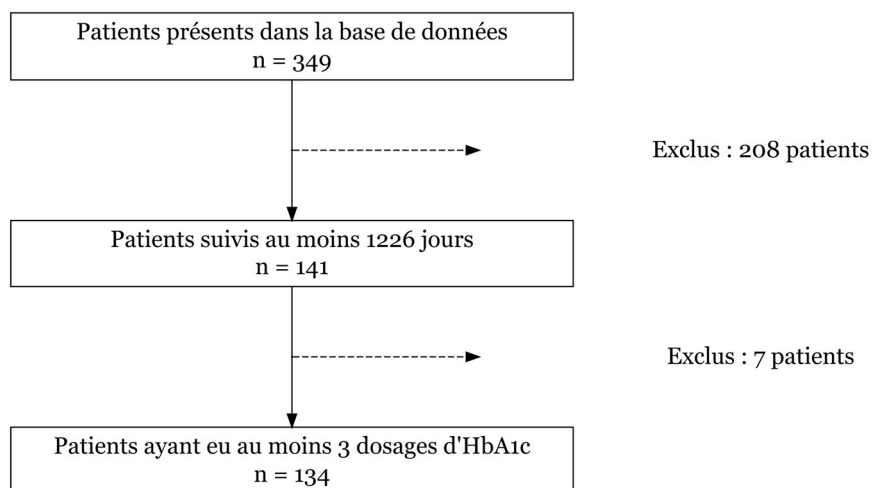


Figure 9. Flowchart des patients fictifs

2 Modélisations individuelles

Les régressions linéaires individuelles permettent d'obtenir 4 coefficients pour chaque patient. Ci-après se trouve la distribution des coefficients sur l'ensemble des patients (Tableau 3, Figure 10, Figure 11, Figure 12, Figure 13), suivie de quelques courbes permettant d'illustrer l'objectif de modélisation individuelle.

Tableau 3. Distribution des coefficients dans la population d'étude

Coefficient	Moyenne	Écart-type	Minimum	Maximum
C_1	6.65	0.92	5.13	11.71
C_2	-0.02	0.66	-2.79	4.40
C_3	0.05	0.29	-1.06	1.46
C_4	-0.02	0.27	-0.98	1.30
R^2	0.53	0.27	0.00	1.00

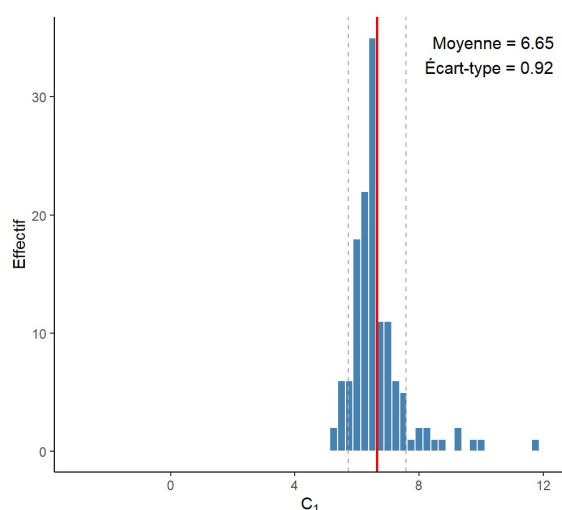


Figure 10. Distribution de C_1

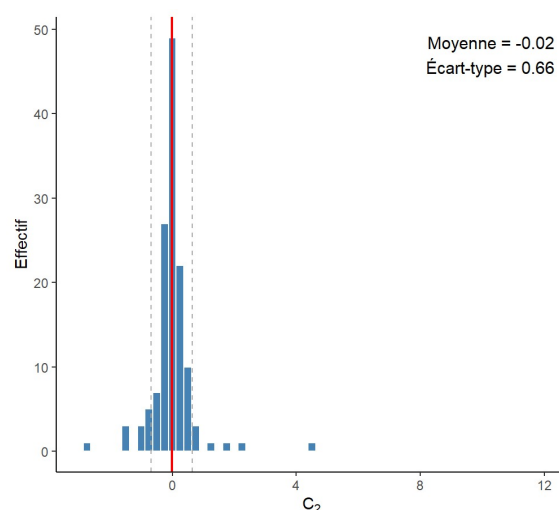


Figure 11. Distribution de C_2

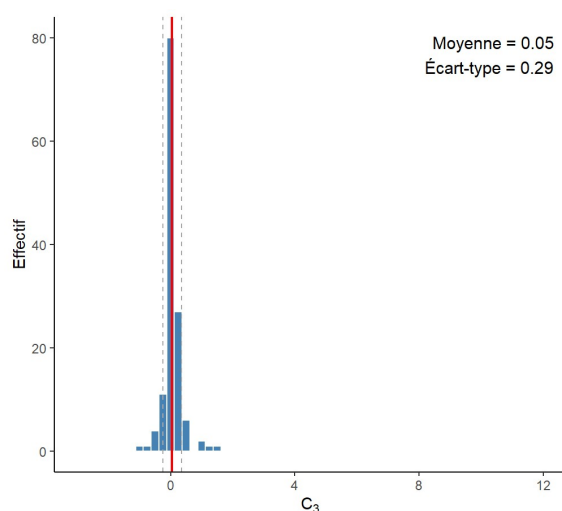


Figure 12. Distribution de C_3

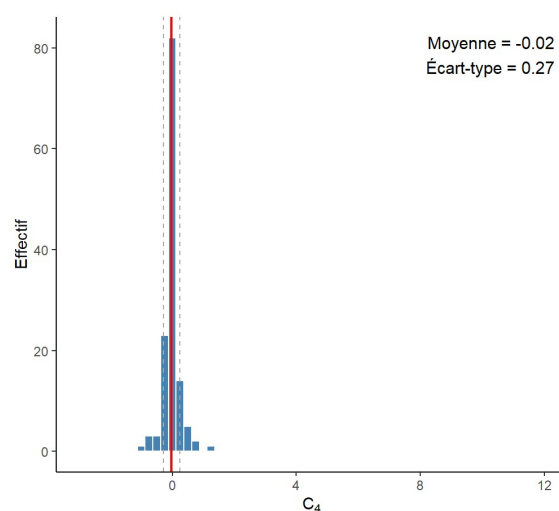


Figure 13. Distribution de C_4

Les courbes modélisées à partir des coefficients issus de la régression linéaire illustrent les trajectoires de l'hémoglobine glyquée réellement observée en fonction du temps. Le coefficient R^2 rend compte de l'adéquation de la courbe modélisée à la courbe observée.

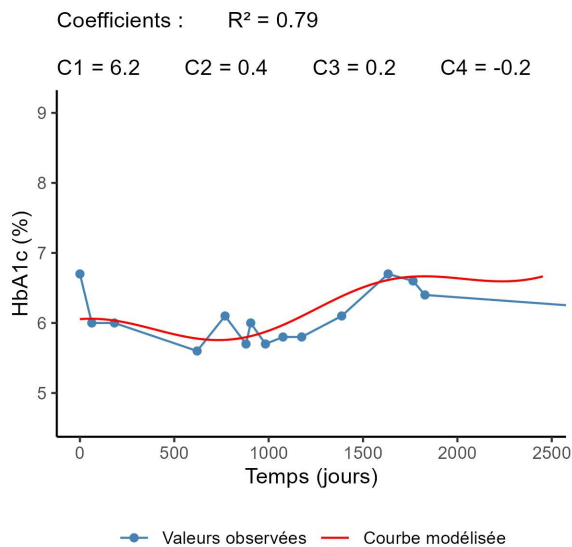


Figure 14. Exemple satisfaisant de modélisation individuelle

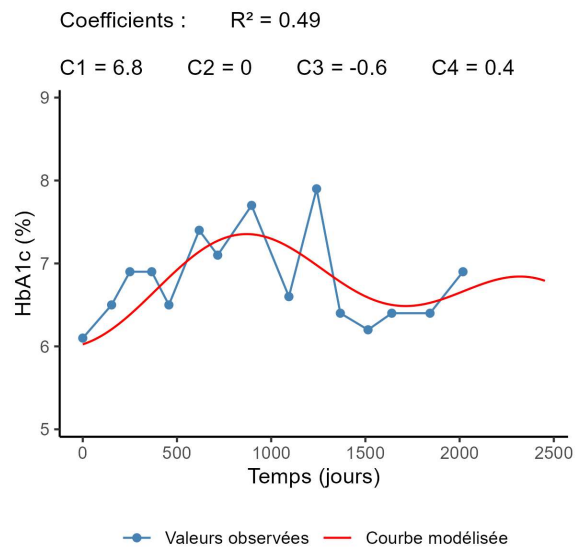


Figure 15. Autre exemple satisfaisant de modélisation individuelle

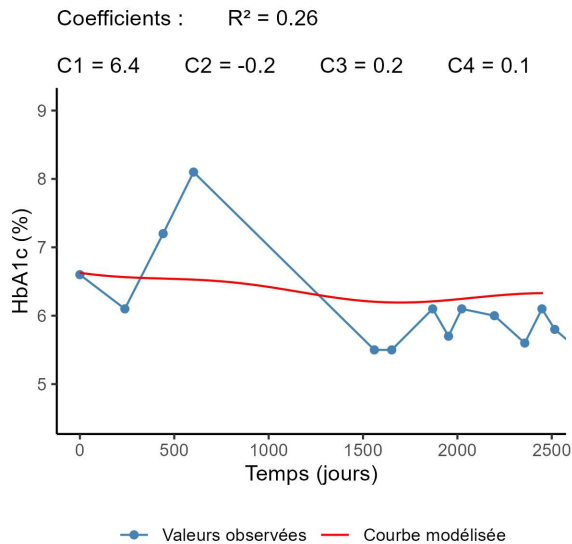


Figure 16. Exemple peu satisfaisant de modélisation individuelle

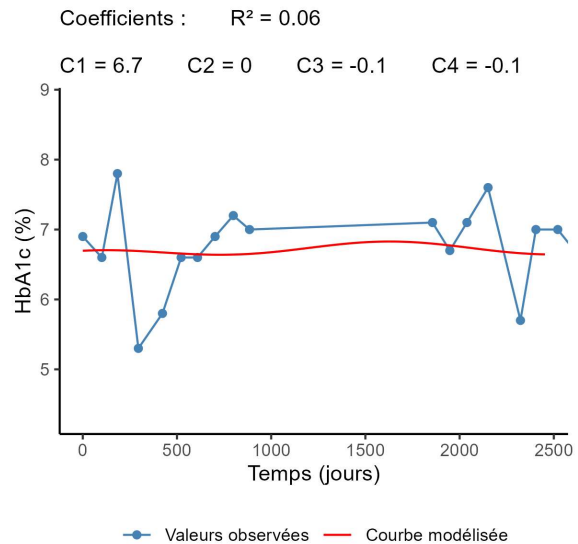


Figure 17. Autre exemple peu satisfaisant de modélisation individuelle



Les deux premiers exemples (Figure 14, Figure 15) montrent des modélisations satisfaisantes, tandis que les exemples suivants (Figure 16, Figure 17) montrent l'inverse. La distribution des R^2 (Tableau 3) obtenus montrent que, globalement, la modélisation est satisfaisante.

3 Clustering et description de sous-groupes choisis d'individus

Sont présentés ci-après la distribution en « nuage de points » (*scatterplot*) des patients au sein du cluster n°1 (c_1 , c_2 et c_3 centrés-réduits supérieurs ou égaux à 0, Figure 18) et du cluster n°8 (c_1 , c_2 et c_3 centrés-réduits inférieurs à 0, Figure 19), ainsi que les courbes d'hémoglobine glyquée des patients qui appartiennent à ces deux clusters (Figure 20, Figure 21).

Pour chaque cluster, la moyenne des coefficients c_1 à c_4 (non centrés-réduits) des patients qui y sont inclus permet de modéliser une courbe (en rouge) représentant la tendance moyenne de l'HbA1c dans le cluster.

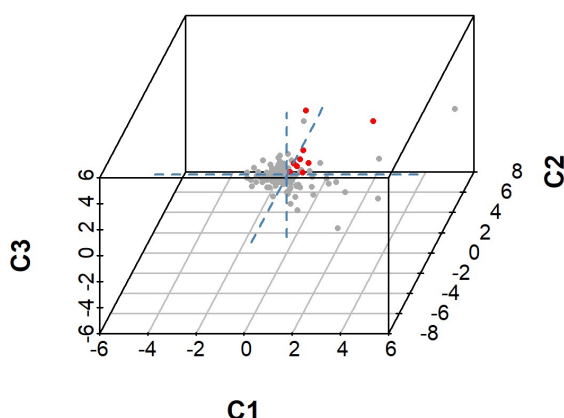


Figure 18. Patients appartenant au cluster n°1

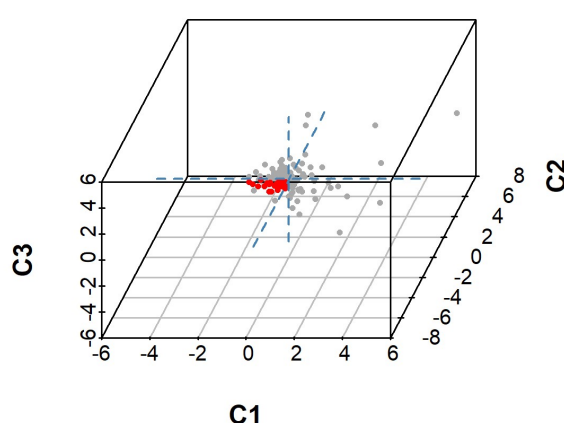


Figure 19. Patients appartenant au cluster n°8

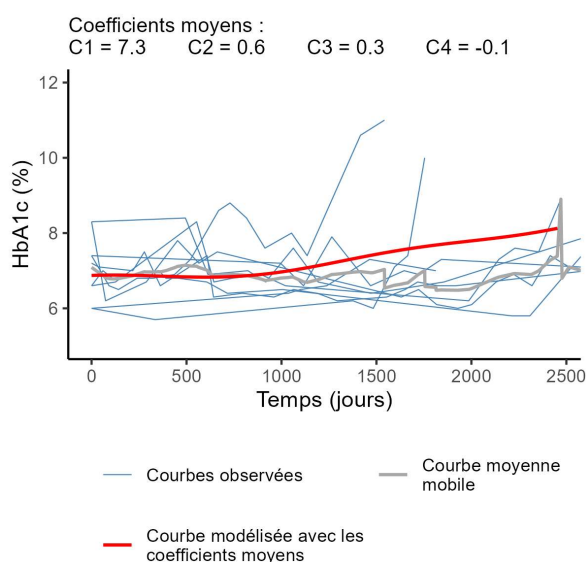


Figure 20. Courbes observées au sein du cluster n°1

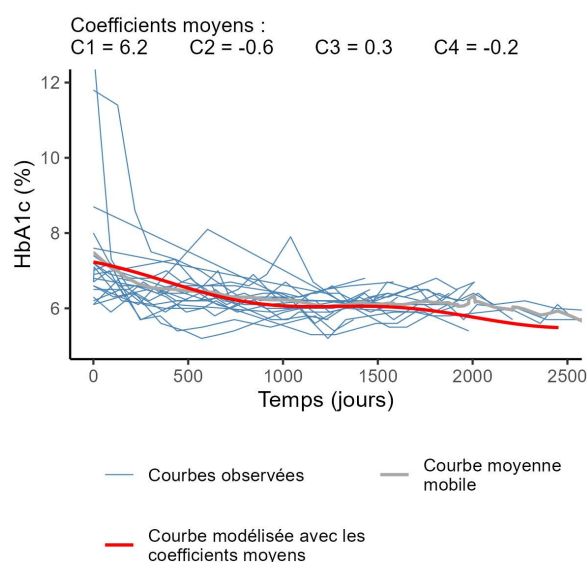


Figure 21. Courbes observées au sein du cluster n°8



Ces exemples illustrent que les clusters reconstituables d'une manière ou d'une autre depuis les coefficients correspondent bien à des modélisations différentes, qui elles-mêmes semblent rendre compte de différentes tendances observées.

Par ailleurs, à l'aide de la méthode CART (*Classification And Regression Tree*), il est possible de déterminer la probabilité d'appartenir à un cluster (Figure 22) uniquement à l'aide des variables initialement connues chez un patient (ici l'âge, le sexe, l'HbA1c et l'IMC initiaux).

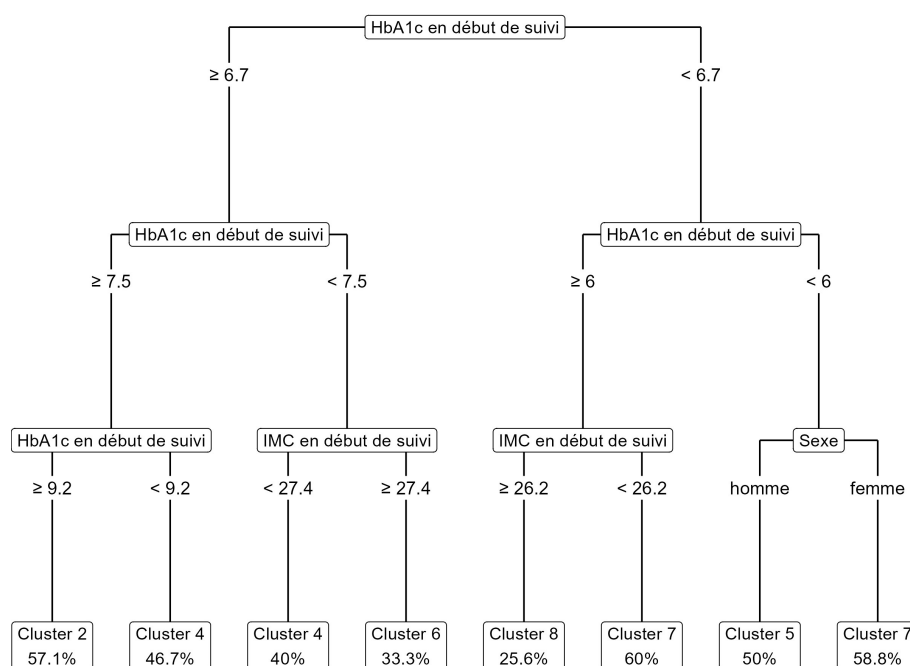


Figure 22. Arbre de classification créé par la méthode CART

4 Prédiction de la trajectoire d'HbA1c d'après d'autres variables

Les coefficients c_1 à c_4 peuvent être prédits à l'aide de nouvelles régressions linéaires, dans lesquelles ils deviennent tour-à-tour la variable à expliquer. Les résultats obtenus sont illustrés dans les graphiques ci-dessous.

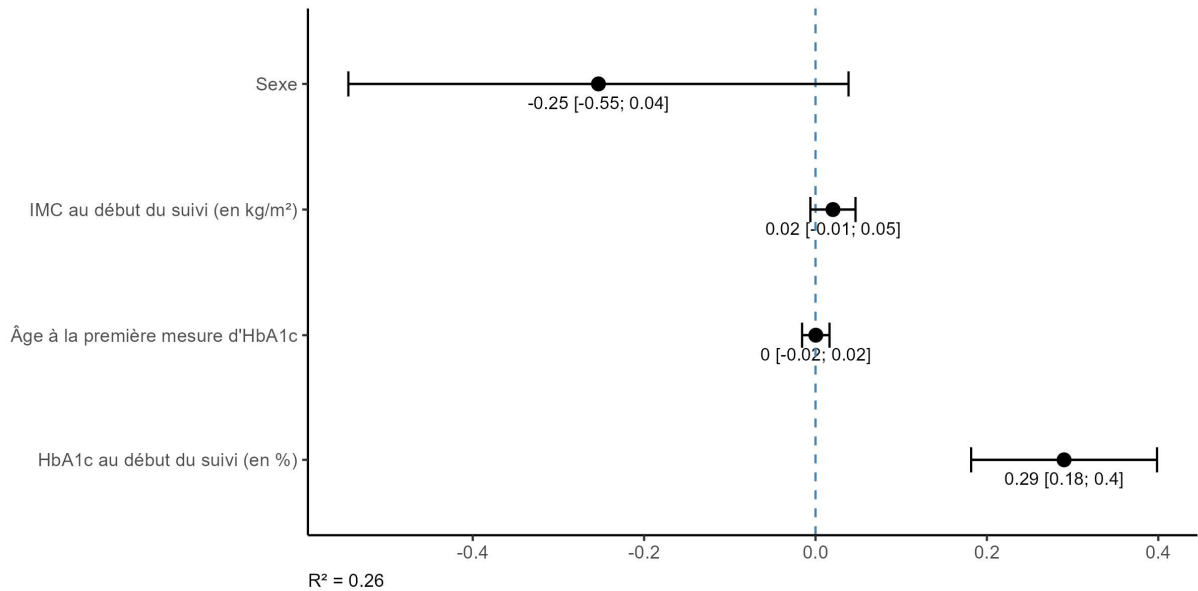


Figure 23. Facteurs associés à C_1



Dans la Figure 23, on observe que la valeur moyenne de l'HbA1c (élevée ou non, représentée par C_1) est positivement associée à la première mesure d'HbA1c, mais pas aux autres variables.

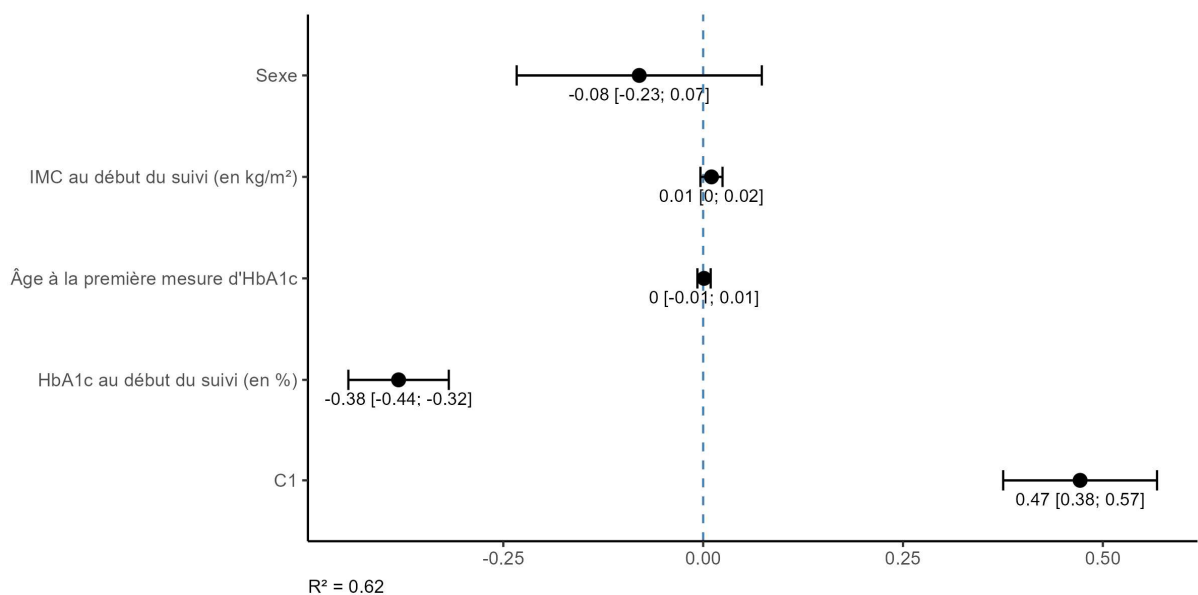


Figure 24. Facteurs associés à C_2



Dans la Figure 24, on observe qu'une tendance croissante d'HbA1c (représentée par C_2) est positivement associée à une première mesure d'HbA1c faible, ainsi qu'à la valeur moyenne de l'HbA1c dans le temps (représentée par C_1).

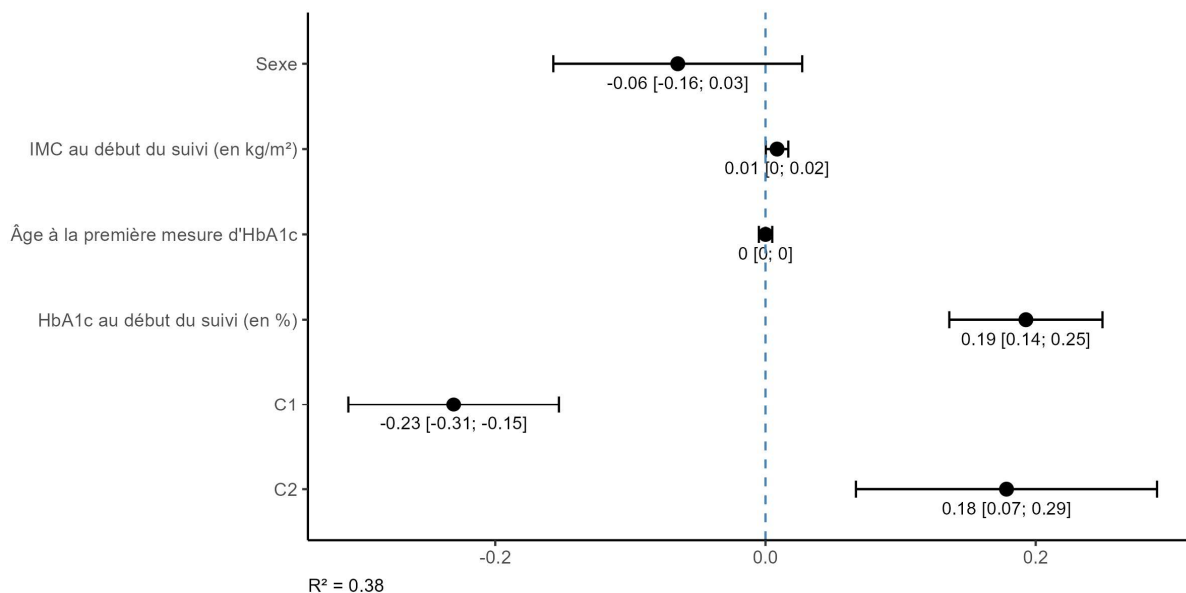


Figure 25. Facteurs associés à C_3



Dans la Figure 25, on observe qu'une tendance d'HbA1c en courbe décroissante puis croissante (représentée par C_3) est positivement associée à la première mesure d'HbA1c, à une valeur moyenne d'HbA1c faible dans le temps (représenté par C_1), ainsi qu'à la tendance croissante de l'HbA1c (représentée par C_2).

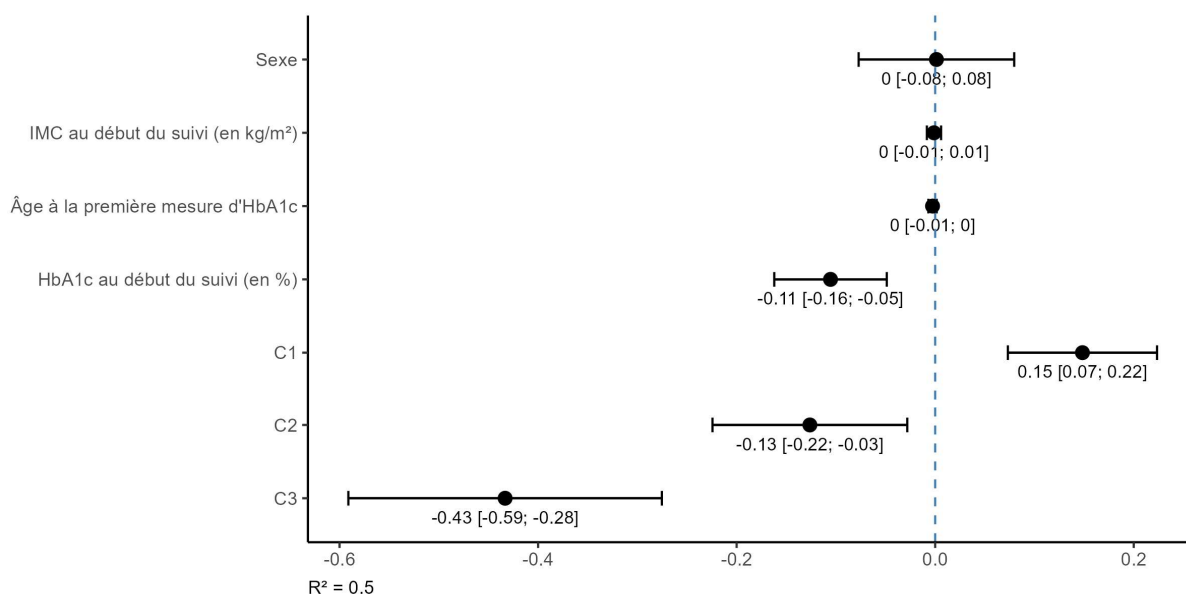


Figure 26. Facteurs associés à C_4



Dans la Figure 26, on observe qu'une tendance sinusoïdale d'HbA1c (représentée par C_4) est positivement associée à une première mesure d'HbA1c faible, à la valeur générale de l'HbA1c dans le temps (représentée par C_1), à la tendance décroissante d'HbA1c (représentée par C_2 , dont le coefficient est négatif), ainsi qu'à une tendance d'HbA1c croissante puis décroissante (représentée par C_3 , dont le coefficient est négatif).

5 Prédiction d'autres variables à l'aide de la trajectoire de l'HbA1c

Les coefficients c_1 à c_4 peuvent aussi être utilisés comme variable explicative dans le cadre d'une régression linéaire ou logistique. Les résultats pour notre base de données sont présentés ici.

5.1 Mise sous insuline à terme, en binaire

La mise sous insuline à terme étant un *outcome* binaire, une régression logistique doit être effectuée pour la prédire.

Variable	N	Odds Ratio	p
Sexe			(Baseline)
homme	59		
femme	60	0.05 [0.00 ; 0.64]	<0.05
IMC au début du suivi (en kg/m ²)	119	1.56 [1.22 ; 2.23]	<0.05
Âge à la première mesure d'HbA1c	119	1.29 [1.09 ; 1.63]	<0.05
Première mesure d'HbA1c (en %)	119	1.26 [0.31 ; 6.54]	0.8
Nombre de mesures d'HbA1c	119	0.91 [0.76 ; 1.04]	0.2
C1	119	52.54 [3.52 ; 1980.40]	<0.05
C2	119	68.94 [2.98 ; 5926.73]	<0.05
C3	119	0.08 [0.00 ; 7.70]	0.3
C4	119	1.56 [0.02 ; 174.17]	0.8
R ²	119	0.00 [0.00 ; 0.08]	<0.05

Figure 27. Facteurs associés à la mise sous insuline à terme



Dans notre exemple (Figure 27), il semble que la mise sous insuline soit influencée par le fait d'être de sexe féminin, d'avoir un IMC et un âge élevés au début du suivi, par des valeurs élevées d'HbA1c (C_1 élevée), par une tendance croissante d'HbA1c (C_2 élevée) et par un comportement bruité (R^2 faible). On notera que la valeur initiale d'HbA1c seule n'est plus significative dans ce modèle.

5.2 Délai avant mise sous insuline

À l'aide d'un modèle de Cox, on peut prédire le délai avant mise sous insuline à partir des coefficients c_1 à c_4 et d'autres variables. Les courbes de survie correspondant à l'événement temps-dépendant « mise sous insuline » sont présentés immédiatement après.

Variable	N	Hazard Ratio	p
Sexe	homme 59		(Baseline)
	femme 60		0.24 [0.04 ; 1.56] 0.1
IMC au début du suivi (en kg/m ²)	119		1.25 [1.07 ; 1.46] <0.05
Âge à la première mesure d'HbA1c	119		1.14 [1.02 ; 1.28] <0.05
Première mesure d'HbA1c (en %)	119		0.90 [0.36 ; 2.22] 0.8
Nombre de mesures d'HbA1c	119		0.83 [0.73 ; 0.96] <0.05
C1	119		8.93 [1.43 ; 55.62] <0.05
C2	119		7.66 [0.86 ; 68.30] 0.1
C3	119		2.45 [0.13 ; 44.64] 0.5
C4	119		0.66 [0.02 ; 18.46] 0.8
R ²	119		0.00 [0.00 ; 0.23] <0.05

Figure 28. Facteurs associés à la durée (en années) avant mise sous insuline - Modèle de Cox



Un peu comme précédemment avec la régression logistique, il semble que le délai avant mise sous insuline (Figure 28) soit influencé par un IMC et un âge au début du suivi élevés, des valeurs élevées d'HbA1c (représentées par C₁) et un comportement bruité (R² faible).

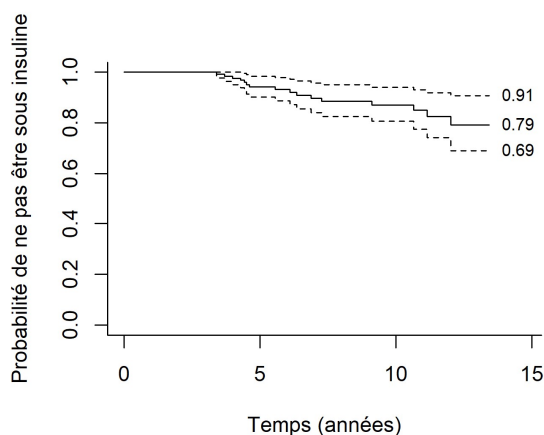


Figure 29. Courbe de survie globale (délai avant mise sous insuline)

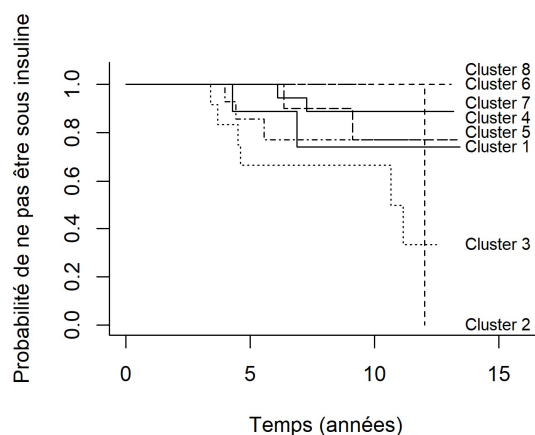


Figure 30. Courbe de survie (délai avant mise sous insuline) selon le cluster



On constate qu'il est aussi techniquement possible de prédire la mise sous insuline à l'aide des clusters issus des coefficients, plutôt que directement par les valeurs de ces coefficients (Figure 29, Figure 30).

5.3 Développement de complication(s)

Grace à une régression logistique, on peut prédire le développement de complication à partir des coefficients c_1 à c_4 et des autres variables connues (Figure 31).

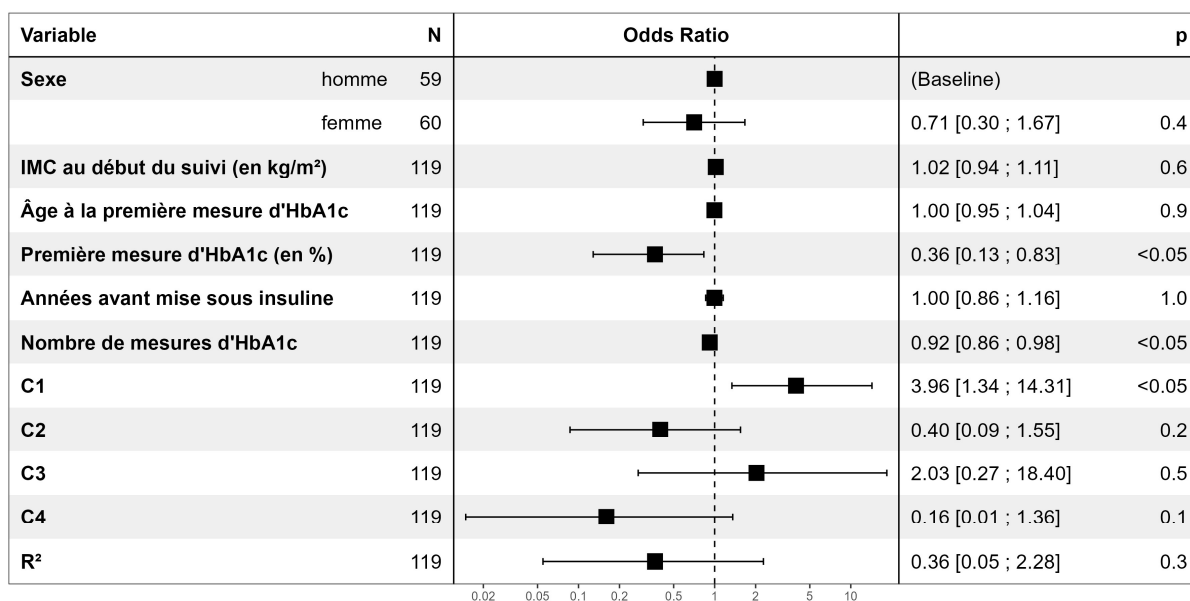


Figure 31. Facteurs associés au développement de complication(s)



On observe que le développement de complication (variable binaire) est corrélé significativement à la valeur moyenne de l'HbA1c au cours du temps (représentée par C_1), à une première mesure faible d'HbA1c et à un faible nombre de mesures d'HbA1c.

6 Prédiction de la suite d'une trajectoire d'HbA1c

Les coefficients c_1 , c_2 , c_3 et c_4 obtenus par régression sur les trois-quarts initiaux du suivi permettent de modéliser « précocement » (c'est-à-dire sans connaître les valeurs observées sur le dernier quart du suivi) la trajectoire de l'HbA1c.

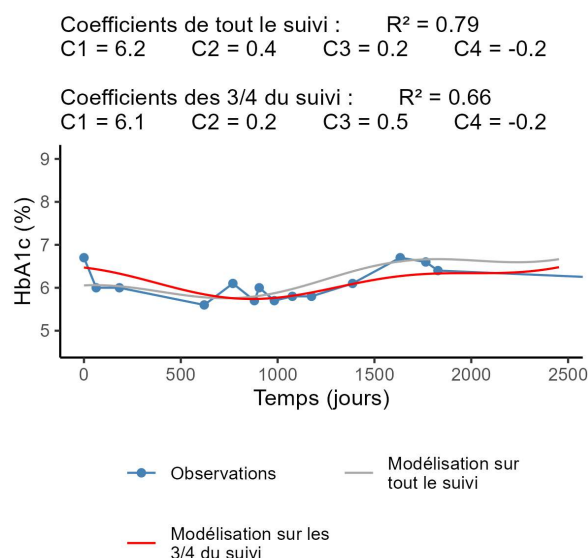


Figure 32. Exemple satisfaisant de prédiction de la trajectoire d'HbA1c

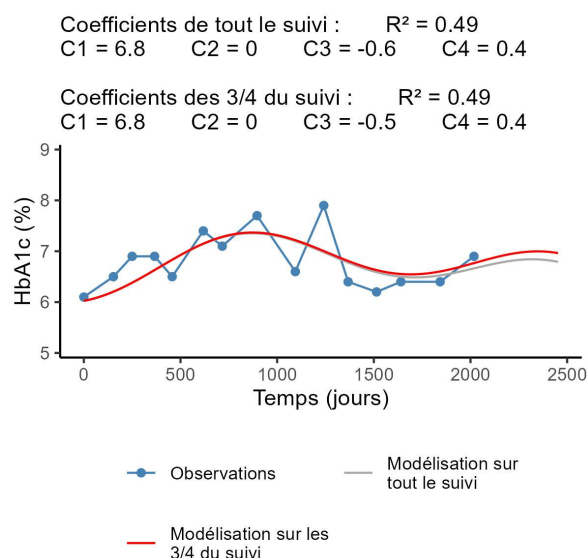


Figure 33. Autre exemple satisfaisant de prédiction de la trajectoire d'HbA1c

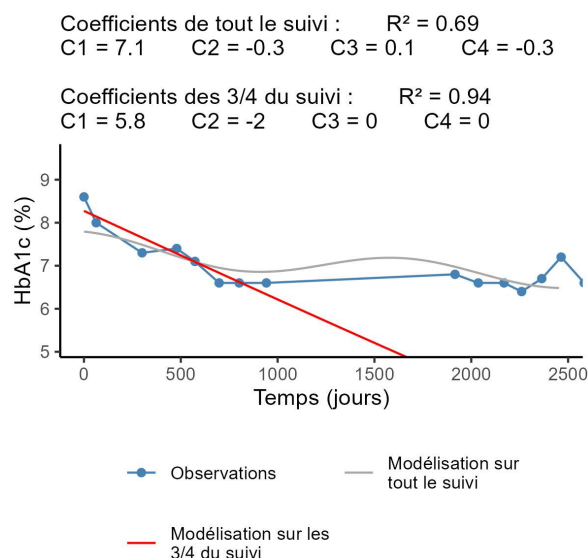


Figure 34. Exemple peu satisfaisant de prédiction de la trajectoire d'HbA1c

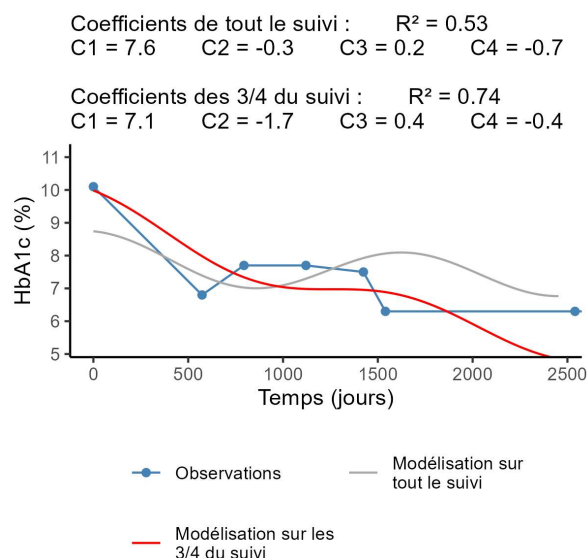


Figure 35. Autre exemple peu satisfaisant de prédiction de la trajectoire d'HbA1c



Dans ces exemples (Figure 32, Figure 33, Figure 34, Figure 35), il est techniquement possible de prédire la suite de l'évolution de l'HbA1c chez ces patients. Des développements ultérieurs devront permettre de vérifier expérimentalement l'adéquation entre cette tendance prédite et la tendance réelle masquée.

7 Description de l'évolution conjointe de l'IMC et de l'HbA1c

Le processus décrit dans la partie Méthodes (voir p. 37) permet de calculer les variables $C_{1,HbA1c}$ à $C_{4,HbA1c}$. De la même manière, en utilisant les mêmes fonctions primaires théoriques de départ, nous calculons en plus les variables $C_{1,IMC}$ à $C_{4,IMC}$ pour décrire l'évolution de l'indice de masse corporelle. La question est désormais de savoir s'il est possible de décrire les covariations des deux variables HbA1c et IMC.

Du fait de données manquantes sur l'une ou l'autre (il faut disposer d'au moins 3 mesures d'HbA1c et d'IMC par individu), cette tentative sera réalisée sur seulement 20 patients.

Les coefficients de régressions issus des courbes observées de l'HbA1c et de l'IMC de chaque patient sont résumés dans le tableau suivant (Tableau 4), suivi d'une matrice de corrélation des deux variables (Figure 36) :

Tableau 4. Coefficients de régression des courbes d'IMC et d'HbA1c

Patient	N _{obs} , IMC	C ₁ , IMC	C ₂ , IMC	C ₃ , IMC	C ₄ , IMC	C ₁ , HbA1c	C ₂ , HbA1c	C ₃ , HbA1c	C ₄ , HbA1c
355	3	30.51	-0.89	-0.09	0.69	6.08	-0.11	0.00	0.03
464	3	30.56	-0.76	-0.15	0.26	6.14	-0.69	0.02	-0.03
489	3	31.06	5.72	-0.02	0.00	7.81	1.63	0.04	-0.01
1297	3	45.56	-2.69	0.10	0.69	6.47	-0.19	0.05	0.35
2301	5	27.02	-1.75	0.03	-0.46	7.42	-0.12	0.01	-0.40
3721	4	32.84	0.40	0.06	-0.12	7.26	-0.36	0.13	-0.15
5337	6	27.79	-0.30	-0.08	-0.48	7.09	-0.05	0.05	0.30
5839	7	33.77	-0.74	1.73	-1.34	6.68	-0.12	0.06	-0.10
5869	4	28.27	-1.28	0.22	-0.10	6.19	-0.29	0.04	-0.01
5961	4	24.59	0.52	0.12	-0.07	6.35	-0.80	0.10	-0.05
5997	3	28.31	-1.96	0.04	-0.02	7.86	0.19	0.50	-0.25
7348	6	27.00	1.98	-0.26	0.27	8.91	3.02	-0.74	1.00
7736	5	33.63	-1.47	-0.05	0.04	6.57	0.09	0.01	0.05
7738	3	24.54	0.17	0.01	0.13	8.58	-0.72	-0.08	-1.14
9603	5	34.68	-2.04	-0.10	0.65	6.34	-0.03	0.01	0.01
9669	4	26.83	-3.16	0.02	0.09	5.18	-0.33	-0.05	0.07
9948	4	20.52	-0.39	0.00	-0.01	7.09	0.01	0.00	0.00
11863	4	40.10	9.79	0.00	-0.02	6.09	0.75	0.00	-0.02
12190	3	26.37	-0.37	-0.07	0.23	6.26	0.30	-0.02	0.06
13996	6	26.96	-6.97	0.00	0.01	4.91	-0.71	0.00	0.01

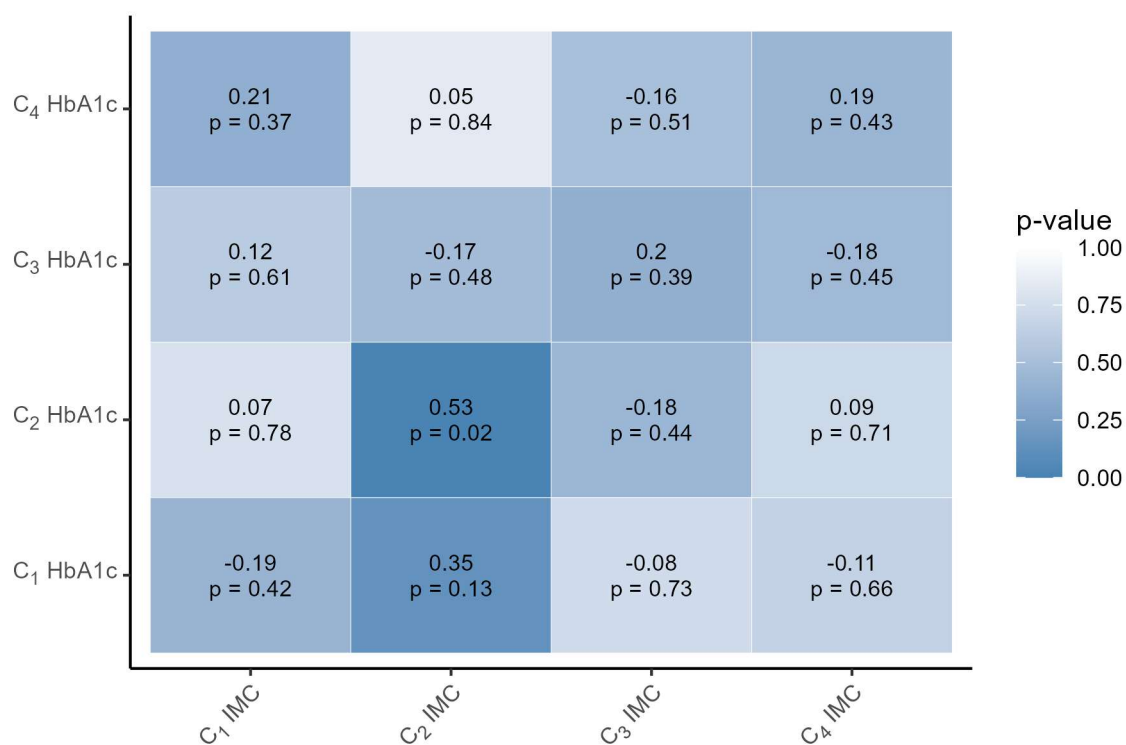


Figure 36. Matrice de corrélation des coefficients d'évolution de l'IMC et de l'HbA1c



Dans cet exemple, les coefficients c_2 de l'HbA1c et de L'IMC sont les seuls à être significativement corrélés. Cela signifie que les tendances croissantes (ou respectivement décroissantes) sur un biomarqueur se retrouvent d'une manière similaire sur l'autre. Ce résultat est cohérent avec la physiopathologie du diabète.

Discussion

Nous avons proposé une méthode qui permet d'analyser des biomarqueurs recueillis en vie réelle, c'est-à-dire notamment avec des mesures asynchrones, et sur une durée variable (raisonnablement au moins la moitié de la durée maximale). En appliquant individuellement la régression linéaire sur les courbes d'observations de l'HbA1c, cette méthode permet d'accomplir les points suivants.

→ Décrire des trajectoires individuelles :

Nous décrivons ainsi des trajectoires individuelles à l'aide de coefficients sans valeur manquante, avec une modélisation satisfaisante. Dans notre exemple, notre méthode a permis de modéliser la courbe d'évolution de l'HbA1c de 134 patients via 4 coefficients de régression.

→ Décrire un groupe d'individu (clustering) :

À l'aide de la distribution de ces coefficients, nous pouvons créer des clusters de patients aux trajectoires similaires. Il est aussi possible de modéliser une tendance « moyenne » au sein de chaque cluster.

→ Prédire une trajectoire à l'aide de variables externes :

Nous pouvons désormais prédire ces coefficients (et donc des trajectoires attendues) à l'aide de variables externes.

→ Utiliser les trajectoires pour prédire des variables externes :

Nous pouvons aussi prédire des variables externes à l'aide de ces coefficients. Nous l'avons montré notamment avec une variable à prédire binaire (régression logistique) ou temps-dépendante (modèle de Cox).

→ Prédire la suite d'une trajectoire, d'après son début :

En appliquant notre méthode sur la partie initiale du suivi, nous pouvons modéliser une courbe d'évolution approchant la trajectoire réelle sur la totalité du suivi.

→ Décrire le lien entre les trajectoires de deux biomarqueurs différents :

Enfin, nous pouvons décrire le lien entre les tendances temporelles de deux variables ou biomarqueurs différents.

Ce travail est une preuve de concept : il montre que la méthode proposée peut techniquement être mise en œuvre, et permet techniquement de mener les analyses attendues. Il ne s'agit pas d'un travail de recherche appliquée mené sur des données réelles à des fins d'amélioration des connaissances médicales. En l'état, les résultats obtenus ne sauraient être interprétés comme des résultats d'une étude clinique ou épidémiologique robuste.

Cette thèse appelle des explorations complémentaires, visant en particulier à approfondir les éléments suivants :

- Améliorer la définition des fonctions de départ, soit empiriquement (d'après les données de l'échantillon), soit à dire d'expert.
- Évaluer de manière neutre et interprétable la qualité de l'ajustement ainsi obtenu.
- Proposer un clustering guidé par les données, qui utilise une méthode communément admise (par exemple la méthode des K-Means).
- Utiliser d'autres méthodes multivariées, tant pour prédire les coefficients que pour utiliser ces coefficients pour une prédiction (exemple : arbres de décision).
- Valider la véracité statistique de la prédiction des trajectoires.
- Imaginer des outils statistiques et graphiques plus pertinents pour décrire la co-variation de deux biomarqueurs.
- Analyser l'impact de la méthode d'interpolation linéaire utilisée d'office pour répartir la pondération des observations sur l'axe temporel.

Liste des tables

Tableau 1. Résumé des méthodes d'analyse classique, Y = biomarqueur étudié....	21
Tableau 2. Résumé des méthodes d'analyse de données sous forme fonctionnelle, Y = biomarqueur étudié.....	25
Tableau 3. Distribution des coefficients dans la population d'étude	42
Tableau 4. Coefficients de régression des courbes d'IMC et d'HbA1c	54

Liste des figures

Figure 1. Exemple de distribution normale (à gauche) ou non-normale (à droite) des résidus	28
Figure 2. Exemple de distribution des résidus avec moyenne et variance indépendantes de $\hat{Y}$ (à gauche), ou moyenne dépendante de $\hat{Y}$ (à droite)	29
Figure 3. Exemple de distribution des résidus avec variance dépendante de $\hat{Y}$	29
Figure 4. Exemple de distribution des distances de Cook chez 234 individus	30
Figure 5. Fonction $f1(x)$	36
Figure 6. Fonction $f2(x)$	36
Figure 7. Fonction $f3(x)$	36
Figure 8. Fonction $f4(x)$	36
Figure 9. Flowchart des patients fictifs	41
Figure 10. Distribution de $C1$	42
Figure 11. Distribution de $C2$	42
Figure 12. Distribution de $C3$	42
Figure 13. Distribution de $C4$	42
Figure 14. Exemple satisfaisant de modélisation individuelle	43
Figure 15. Autre exemple satisfaisant de modélisation individuelle	43
Figure 16. Exemple peu satisfaisant de modélisation individuelle	43
Figure 17. Autre exemple peu satisfaisant de modélisation individuelle	43
Figure 18. Patients appartenant au cluster n°1	44
Figure 19. Patients appartenant au cluster n°8	44
Figure 20. Courbes observées au sein du cluster n°1	44
Figure 21. Courbes observées au sein du cluster n°8	44
Figure 22. Arbre de classification créé par la méthode CART	45
Figure 23. Facteurs associés à $C1$	46
Figure 24. Facteurs associés à $C2$	46
Figure 25. Facteurs associés à $C3$	47
Figure 26. Facteurs associés à $C4$	47
Figure 27. Facteurs associés à la mise sous insuline à terme	49
Figure 28. Facteurs associés à la durée (en années) avant mise sous insuline - Modèle de Cox	50
Figure 29. Courbe de survie globale (délai avant mise sous insuline)	50

Figure 30. Courbe de survie (délai avant mise sous insuline) selon le cluster.....	50
Figure 31. Facteurs associés au développement de complication(s)	51
Figure 32. Exemple satisfaisant de prédiction de la trajectoire d'HbA1c	52
Figure 33. Autre exemple satisfaisant de prédiction de la trajectoire d'HbA1c	52
Figure 34. Exemple peu satisfaisant de prédiction de la trajectoire d'HbA1c	52
Figure 35. Autre exemple peu satisfaisant de prédiction de la trajectoire d'HbA1c ..	52
Figure 36. Matrice de corrélation des coefficients d'évolution de l'IMC et de l'HbA1c	55

Références

- [1] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;69:89–95. <https://doi.org/10.1067/mcp.2001.113989>.
- [2] Strimbu K, Tavel JA. What are Biomarkers? *Curr Opin HIV AIDS* 2010;5:463–6. <https://doi.org/10.1097/COH.0b013e32833ed177>.
- [3] Cemin R, Daves M. Pre-analytic variability in cardiovascular biomarker testing. *J Thorac Dis* 2015;7:E395–401. <https://doi.org/10.3978/j.issn.2072-1439.2015.10.03>.
- [4] Pradhan S, Gautam K, Pant V. Variation in Laboratory Reports: Causes other than Laboratory Error. *JNMA J Nepal Med Assoc* 2022;60:222–4. <https://doi.org/10.31729/jnma.6022>.
- [5] Pullenayegum EM, Lim LS. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Stat Methods Med Res* 2016;25:2992–3014. <https://doi.org/10.1177/0962280214536537>.
- [6] Lokku A, Lim LS, Birken CS, Pullenayegum EM. Summarizing the extent of visit irregularity in longitudinal data. *BMC Med Res Methodol* 2020;20:135. <https://doi.org/10.1186/s12874-020-01023-w>.
- [7] Ramsay JO, Silverman BW. *Functional Data Analysis*. New York, NY: Springer; 2005. <https://doi.org/10.1007/b98888>.
- [8] Wang J-L, Chiou J-M, Müller H-G. *Functional Data Analysis*. *Annu Rev Stat Its Appl* 2016;3:257–95. <https://doi.org/10.1146/annurev-statistics-041715-033624>.
- [9] Gertheiss J, Rügamer D, Liew BXW, Greven S. *Functional Data Analysis: An Introduction and Recent Developments* 2023. <https://doi.org/10.48550/arXiv.2312.05523>.
- [10] Ramsay JO, Dalzell CJ. Some Tools for Functional Data Analysis. *J R Stat Soc Ser B Methodol* 1991;53:539–61. <https://doi.org/10.1111/j.2517-6161.1991.tb01844.x>.
- [11] Yao F, Müller H-G, Wang J-L. Functional Data Analysis for Sparse Longitudinal Data. *J Am Stat Assoc* 2005;100:577–90. <https://doi.org/10.1198/016214504000001745>.
- [12] Lachin JM. Fallacies of Last Observation Carried Forward. *Clin Trials Lond Engl* 2016;13:161–8. <https://doi.org/10.1177/1740774515602688>.
- [13] Acar-Denizli N, Delicado P. Functional Data Analysis on Wearable Sensor Data: A Systematic Review 2024. <https://doi.org/10.48550/arXiv.2410.11562>.
- [14] Yang Q, Jiang M, Li C, Luo S, Crowley MJ, Shaw RJ. Predicting health outcomes with intensive longitudinal data collected by mobile health devices: a functional principal component regression approach. *BMC Med Res Methodol* 2024;24:69. <https://doi.org/10.1186/s12874-024-02193-7>.

- [15] Ullah S, Finch CF. Applications of functional data analysis: A systematic review. *BMC Med Res Methodol* 2013;13:43. <https://doi.org/10.1186/1471-2288-13-43>.
- [16] Sørensen H, Goldsmith J, Sangalli LM. An introduction with medical applications to functional data analysis. *Stat Med* 2013;32:5222–40. <https://doi.org/10.1002/sim.5989>.
- [17] Gasparini A, Abrams KR, Barrett JK, Major RW, Sweeting MJ, Brunskill NJ, et al. Mixed-effects models for health care longitudinal data with an informative visiting process: A Monte Carlo simulation study. *Stat Neerlandica* 2020;74:5–23. <https://doi.org/10.1111/stan.12188>.
- [18] Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *J Am Med Inform Assoc JAMIA* 2022;29:1191–9. <https://doi.org/10.1093/jamia/ocac050>.
- [19] Sang P, Kong D, Yang S. Functional principal component analysis with informative observation times. *Biometrika* 2024;112:asae055. <https://doi.org/10.1093/biomet/asae055>.
- [20] Rudra Gupta T, Schwartz DE, Saha R, Wen PY, Rahman R, Trippa L. Informative censoring in externally controlled clinical trials: a potential source of bias. *ESMO Open* 2025;10:104094. <https://doi.org/10.1016/j.esmoop.2024.104094>.
- [21] Salmanpour MR, Shamsaei M, Hajianfar G, Soltanian-Zadeh H, Rahmim A. Longitudinal clustering analysis and prediction of Parkinson's disease progression using radiomics and hybrid machine learning. *Quant Imaging Med Surg* 2022;12:906–19. <https://doi.org/10.21037/qims-21-425>.
- [22] Mekov EV, Yanev NA, Kurtelova N, Mihalova T, Tsakova A, Yamakova Y, et al. Phenotyping Chronic Obstructive Pulmonary Disease Through Principal Component Analysis: Identification of Clinical Clusters. *Cureus* n.d.;17:e82811. <https://doi.org/10.7759/cureus.82811>.
- [23] Gao F, Philip Miller J, Xiong C, Luo J, Beiser JA, Chen L, et al. Estimating correlation between multivariate longitudinal data in the presence of heterogeneity. *BMC Med Res Methodol* 2017;17:124. <https://doi.org/10.1186/s12874-017-0398-1>.
- [24] Ouko RK, Mukaka M, Ohuma EO. Joint modelling of longitudinal data: a scoping review of methodology and applications for non-time to event data. *BMC Med Res Methodol* 2025;25:40. <https://doi.org/10.1186/s12874-025-02485-6>.
- [25] Weaver C, Xiao L, Lu W. Functional data analysis for longitudinal data with informative observation times. *Biometrics* 2023;79:722–33. <https://doi.org/10.1111/biom.13646>.
- [26] Muhammad LN. Guidelines for repeated measures statistical analysis approaches with basic science research considerations. *J Clin Invest* n.d.;133:e171058. <https://doi.org/10.1172/JCI171058>.
- [27] Gibbons RD, Hedeker D, DuToit S. Advances in Analysis of Longitudinal Data. *Annu Rev Clin Psychol* 2010;6:79–107. <https://doi.org/10.1146/annurev.clinpsy.032408.153550>.
- [28] Ma Y, Mazumdar M, Memtsoudis SG. Beyond Repeated measures ANOVA: advanced statistical methods for the analysis of longitudinal data in anesthesia

- research. *Reg Anesth Pain Med* 2012;37:99–105. <https://doi.org/10.1097/AAP.0b013e31823ebc74>.
- [29] George B, Seals S, Aban I. Survival analysis and regression models. *J Nucl Cardiol Off Publ Am Soc Nucl Cardiol* 2014;21:686–94. <https://doi.org/10.1007/s12350-014-9908-2>.
- [30] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. *Br J Cancer* 2003;89:232–8. <https://doi.org/10.1038/sj.bjc.6601118>.
- [31] Andrade C. Survival Analysis, Kaplan-Meier Curves, and Cox Regression: Basic Concepts. *Indian J Psychol Med* 2023;45:434–5. <https://doi.org/10.1177/02537176231176986>.
- [32] Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J Am Stat Assoc* 1958;53:457–81. <https://doi.org/10.1080/01621459.1958.10501452>.
- [33] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *Br J Cancer* 2003;89:431–6. <https://doi.org/10.1038/sj.bjc.6601119>.
- [34] Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. *Br J Cancer* 2003;89:605–11. <https://doi.org/10.1038/sj.bjc.6601120>.
- [35] Blanca MJ, Arnau J, García-Castro FJ, Alarcón R, Bono R. Repeated measures ANOVA and adjusted F-tests when sphericity is violated: which procedure is best? *Front Psychol* 2023;14:1192453. <https://doi.org/10.3389/fpsyg.2023.1192453>.
- [36] Blanca M, Arnau J, García-Castro F, Alarcón R, Bono R. Non-normal Data in Repeated Measures ANOVA: Impact on Type I Error and Power. *Psicothema* 2023;1:21–9. <https://doi.org/10.7334/psicothema2022.292>.
- [37] Schober P, Vetter TR. Repeated Measures Designs and Analysis of Longitudinal Data: If at First You Do Not Succeed—Try, Try Again. *Anesth Analg* 2018;127:569–75. <https://doi.org/10.1213/ANE.0000000000003511>.
- [38] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Transact A Math Phys Eng Sci* 2016;374:20150202. <https://doi.org/10.1098/rsta.2015.0202>.
- [39] Siddique J, Brown CH, Hedeker D, Duan N, Gibbons RD, Miranda J, et al. Missing Data in Longitudinal Trials - Part B, Analytic Issues. *Psychiatr Ann* 2008;38:793–801. <https://doi.org/10.3928/00485713-20081201-09>.
- [40] Pugh SL, Brown PD, Enserro D. Missing repeated measures data in clinical trials. *Neuro-Oncol Pract* 2021;9:35–42. <https://doi.org/10.1093/nop/npab043>.
- [41] Wang J-L, Chiou J-M, Mueller H-G. Review of Functional Data Analysis 2015. <https://doi.org/10.48550/arXiv.1507.05135>.
- [42] Shang HL. A survey of functional principal component analysis. *AStA Adv Stat Anal* 2014;98:121–42. <https://doi.org/10.1007/s10182-013-0213-1>.
- [43] Ségalas C, Helmer C, Genuer R, Proust-Lima C. Functional principal component analysis as an alternative to mixed-effect models for describing

- sparse repeated measures in presence of missing data 2024. <https://doi.org/10.48550/arXiv.2402.10624>.
- [44] Greven S, Crainiceanu C, Caffo B, Reich D. Longitudinal functional principal component analysis. *Electron J Stat* 2010;4:1022–54. <https://doi.org/10.1214/10-EJS575>.
 - [45] Zhang E, Li D. Contrastive Functional Principal Components Analysis. *Proc AAAI Conf Artif Intell AAAI Conf Artif Intell* 2025;39:22380–8. <https://doi.org/10.1609/aaai.v39i21.34394>.
 - [46] Rizopoulos D. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *J Stat Softw* 2010;35:1–33. <https://doi.org/10.18637/jss.v035.i09>.
 - [47] Tsiatis AA, Davidian M. JOINT MODELING OF LONGITUDINAL AND TIME-TO-EVENT DATA: AN OVERVIEW n.d.
 - [48] Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Med Res Methodol* 2016;16:117. <https://doi.org/10.1186/s12874-016-0212-5>.
 - [49] Sudell M, Kolamunnage-Dona R, Tudur-Smith C. Joint models for longitudinal and time-to-event data: a review of reporting quality with a view to meta-analysis. *BMC Med Res Methodol* 2016;16:168. <https://doi.org/10.1186/s12874-016-0272-6>.
 - [50] Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis n.d.
 - [51] Schneider A, Hommel G, Blettner M. Linear Regression Analysis. *Dtsch Arztebl Int* 2010;107:776–82. <https://doi.org/10.3238/arztebl.2010.0776>.
 - [52] Schober P, Vetter TR. Linear Regression in Medical Research. *Anesth Analg* 2020;132:108–9. <https://doi.org/10.1213/ANE.0000000000005206>.
 - [53] Roustaei N. Application and interpretation of linear-regression analysis. *Med Hypothesis Discov Innov Ophthalmol* 2024;13:151–9. <https://doi.org/10.51329/mehdiophthal1506>.
 - [54] Loh W. Classification and regression trees. *WIREs Data Min Knowl Discov* 2011;1:14–23. <https://doi.org/10.1002/widm.8>.

AUTEUR : Nom : JOHNSON

Prénom : Valentin

Date de Soutenance : 29/06/2026

Titre de la Thèse : Analyse de biomarqueurs quantitatifs recueillis en routine : Mise au point et évaluation de la méthode IRBODeTIF

Thèse - Médecine - Lille 2026

Cadre de classement : Médecine de Spécialité

DES + FST ou option : Santé Publique

Mots-clés : Régression linéaire, données fonctionnelles, méthode

Résumé :

Les méthodes actuelles d'analyse des données de santé répétées dans le temps ne permettent pas d'accomplir simultanément les objectifs suivants : Décrire l'évolution d'un paramètre de santé étudié, créer des clusters ou groupes d'individus similaires, prédire de nouvelle(s) variable(s) à partir du paramètre étudié, prédire le paramètre à partir de nouvelle(s) variable(s), prédire l'évolution du paramètre, analyser l'évolution jointe du paramètre et d'une autre variable, gérer l'asynchronicité des mesures, et gérer l'inégalité des durées de suivi. L'objectif de ce travail est de réaliser la preuve de concept d'une nouvelle méthode d'analyse des données fonctionnelles de santé qui se base sur la régression linéaire, et répondant à ces objectifs.

Un jeu de 4 fonctions primaires théoriquement indépendantes entre elles par rapport au temps est défini. Pour chaque individu, une régression linéaire de la courbe d'hémoglobine glyquée est effectuée sur ces 4 fonctions. Les données manquantes sont imputées par interpolation linéaire entre les données observées. Une base de données synthétique de 349 patients fictifs est utilisée pour effectuer les analyses statistiques.

Résultats : Les trajectoires individuelles d'hémoglobine glyquée peuvent être décrites à l'aide des coefficients issus de la régression. La distribution de ces coefficients permet de créer des groupes de patients aux trajectoires semblables. À l'aide de nouvelles régressions linéaires ou logistiques, ou d'un modèle de survie, ces coefficients peuvent prédire des variables externes, et inversement. En appliquant la méthode de régression sur la partie initiale du suivi, il est possible de modéliser l'évolution de l'hémoglobine glyquée sur l'ensemble du suivi. En appliquant la méthode de régression sur un deuxième paramètre, il est possible d'analyser la corrélation de l'évolution de ces deux paramètres.

La nouvelle méthode permet techniquement de mener les analyses attendues et de répondre aux objectifs présentés en introduction. Il ne s'agit pas d'un travail de recherche appliquée, la véracité statistique des résultats obtenus n'est pas ici validée.

Composition du Jury :

Président : Monsieur le Professeur Philippe AMOUYEL

Assesseurs : Monsieur le Docteur Luc DAUCHET

Monsieur le Docteur Raphaël BENTEGEAC

Directeur : Monsieur le Professeur Emmanuel CHAZARD