
Modélisation mathématique et développement de médicaments



CentraleSupélec



Alexandra Augusti
Grégoire Baverez
Jules Dedieu

Client : Laboratoire Servier, représenté
par Sylvain Fouliard
Equipe encadrante : M. Cournède,
Mme. Letort-Le Chevalier

Remerciements

Nous souhaitons tout d'abord remercier Sylvain Fouliard, qui est à l'origine de ce projet. Son aide et ses conseils tout au long du projet ont été très précieux. Nous avons été touchés par le fait que l'objectif principal de M. Fouliard était de nous transmettre sa passion pour son métier et son savoir faire. Nous avons particulièrement apprécié sa bienveillance et sa gentillesse, ce qui a permis à ce projet de se dérouler dans des conditions optimales.

Ensuite, nous tenions à remercier les membres de l'équipe pédagogique de CentraleSupélec qui nous ont aidé dans ce projet. M. Cournède et Mme Letort-Le Chevalier nous ont été d'une grande aide pour comprendre les concepts mathématiques utilisés dans notre modèle. Ils nous ont aussi souvent proposé des méthodes alternatives à celles que nous utilisions, ce qui nous a permis de résoudre certains de nos problèmes mais aussi de beaucoup apprendre.

Nous souhaitons adresser un grand merci à Mme Midy et M. Pietrasz, pour nous avoir conseillés sur l'organisation d'équipe et la communication (qu'elle soit interne ou externe). Leur aide a été précieuse, dans la mesure où elle nous a permis d'avoir une meilleure ambiance de travail au sein de notre équipe.

Enfin, nous remercions Grégoire Martin, qui a travaillé avec nous durant le semestre 7. Nous avons pu tous les quatre saisir les enjeux de ce projet et comprendre ensemble la modélisation biologique du modèle.

Sommaire

1	Abstract	4
2	Introduction	5
2.1	Le groupe Servier	5
2.2	Le diabète	5
2.3	Nos données initiales	5
2.4	Objectif du projet	6
3	Etat de l’art	6
4	Modélisation biologique	6
4.1	Diabète de type 2	6
4.2	Modèle PK/PD	7
4.3	Compartiments d’étude	8
4.4	Démarche expérimentale	9
5	Équations et conditions du modèle	9
5.1	Paramètres du modèle	9
5.2	Équations différentielles du modèle	9
5.2.1	Conditions initiales	9
5.2.2	Équations du système	10
6	Identifiabilité des paramètres	10
6.1	Une introduction à l’identification des modèles	10
6.2	Résultats analytiques	11
6.3	Analyse indicielle	11
6.3.1	Identification par une fonction de transfert du premier ordre	12
6.3.2	Application à notre modèle	12
6.4	Analyse harmonique	13
6.4.1	Procédure expérimentale	13
6.4.2	Application à notre modèle	14
6.5	Identifiabilité empirique	15
6.6	Résumé des résultats obtenus	15
7	Modélisation mathématique	16
7.1	Modèle linéaire	16
7.2	Une approche Machine Learning	17
7.3	Expectation Maximization	18
7.4	Application de l’Expectation-Maximization à notre modèle	20
7.4.1	Cas particulier de la dimension 1	20
7.4.2	Cas général	21
7.5	Régression non linéaire	22

8	Modélisation informatique et estimations dans un cadre simplifié	22
8.1	Obtention des courbes pour les patients	23
8.2	Importance de la variabilité inter-individuelle	24
8.3	Présentation des résultats dans un cadre linéaire	25
9	Résultats dans des conditions expérimentales plus réalistes	26
9.1	Cadre expérimental utilisé	26
9.2	Estimation des paramètres constants	27
9.3	Estimation des paramètres variables	29
9.4	Bruitage des résultats	30
9.4.1	Bruit global	30
9.4.2	Bruit local	30
9.4.3	Résultats sur les paramètres constants	31
9.4.4	Résultats sur les paramètres variables	32
10	Détermination d'un schéma d'administration	32
10.1	Protocole expérimental et résultats	32
10.2	Limitations du modèle	33
11	Rappel des résultats majeurs de l'analyse de sensibilité	34
12	Moyens utilisés par l'équipe	35
12.1	Répartition des tâches	35
12.2	Organisation de l'équipe	35
12.3	Problèmes rencontrés	35
13	Conclusion	35
13.1	Résultats obtenus	35
13.2	Axes de poursuite	35

1 Abstract

L'objectif de ce projet est d'adopter un modèle à effets mixtes PK/PD pour décrire la réponse d'un individu à la rosiglitazone, un médicament anti-diabétique. Le but est d'identifier les paramètres des équations du comportement biologique à partir de données expérimentales.

Les modélisations numériques permettent de comprendre comment évoluent les différentes grandeurs d'intérêt (glucose, hémoglobine glyquée, rosiglitazone) dans le sang au cours du temps, selon les différents types d'injection.

Sur l'ensemble des paramètres, seulement quatre varient d'un individu à l'autre. L'analyse de sensibilité a permis de montrer l'influence assez bien partagée des trois paramètres $Slope_{ins}$, Ka_{glc} et V_4 qui jouent un rôle prépondérant dans la variance des trois plasmas considérés. Un travail supplémentaire serait néanmoins nécessaire pour évaluer l'influence de la clairance CL .

Une étude théorique du modèle et de ses paramètres nous a permis de constater que les paramètres étaient presque tous identifiables en théorie et que ceux qui ne l'étaient pas apparaissaient dans des produits de paramètres que l'on pouvait identifier.

Pratiquement, il s'est avéré que des jeux de paramètres différents pouvaient permettre de réaliser des simulations relativement proches. Le modèle n'est donc pas identifiable en pratique.

Dans un premier temps, nous avons identifié les paramètres dans un cas très idéal (mesures sur les patients fréquentes et sur les huit compartiments d'étude). Les résultats sont satisfaisants, la plupart des erreurs relatives étant en dessous de 1 %.

Nous avons ensuite cherché à rendre notre modèle plus proche des conditions expérimentales réelles en diminuant grandement la fréquence des mesures sur les patients et en les bruitant pour simuler un bruit de mesure. Ces changements nous ont poussé à adopter une approche mathématique différente qui nous a également apporté des résultats satisfaisants.

Après avoir obtenu ces résultats, nous les avons utilisés afin de déterminer un protocole expérimental pour déterminer un traitement adapté pour un patient donné.

2 Introduction

2.1 Le groupe Servier

Servier est un grand groupe pharmaceutique français. Fondé en 1954, il s'est aujourd'hui internationalisé en étant présent dans 149 pays, comptant près de 22 000 employés.

Un des grands axes de développement de l'entreprise est la recherche et développement, pour lequel elle attribue tous les ans environ 25 % de son chiffre d'affaires. Cela leur a permis d'être toujours à la pointe de l'innovation, tout en concevant pas moins de 49 médicaments en 30 ans. Pour rester compétitive dans ce domaine, l'entreprise s'est associée, entre autres, avec le CNRS et l'INSERM. Au sein de centres internationaux de recherche, 2800 personnes sont employées dans ce secteur. L'entreprise se concentre sur plusieurs axes de recherche, notamment les maladies cardiovasculaires, les maladies immuno-inflammatoires, les maladies neuropsychiatriques, les cancers et le diabète.

2.2 Le diabète

Le diabète peut avoir des conséquences graves pour les patients, pouvant parfois même conduire jusqu'à la mort. Le nombre de diabétiques dans le monde a augmenté de 290 % entre 1980 et 2014. Cette augmentation est liée en partie à l'internationalisation du mode de vie occidental. Cette dernière est illustrée par le nombre de cas aux États-Unis ci-dessous. L'OMS estime que le diabète pourrait être la septième cause de décès dans le monde en 2030.

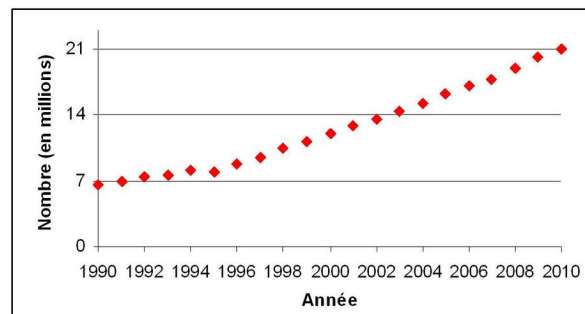


FIGURE 1: Augmentation des patients diabétiques aux USA. Crédits : MST

Face à cette urgence, Servier doit trouver les outils nécessaires pour lutter contre le développement de cette maladie, tout en contrôlant les effets nocifs des médicaments utilisés.

L'intérêt de l'utilisation des simulations numériques pour comprendre comment agit un médicament sur le corps humain est double. Elles permettent en effet de mesurer les conséquences sur les organismes des individus en tenant compte des sources de variabilité interindividuelle, tout en limitant grandement les tests sur des êtres vivants.

2.3 Nos données initiales

Le client nous a fourni un poster scientifique publié en 2017 sur l'application du modèle PK/PD (*c.f* partie 2.3) pour le diabète de type 2 chez des espèces différentes (souris, singe et humain). Le but était de comprendre les mécanismes des organismes communs aux différentes espèces, pour transposer les travaux d'études sur les animaux à l'être humain en seulement quelques modifications.

Néanmoins, les cellules bêta productrices d'insuline dans le pancréas sont différentes chez les souris et les hommes. La réponse cellulaire peut donc différer entre les espèces suite à la sollicitation des récepteurs.

Il est apparu en particulier que les effets de la rosiglitazone chez les patients peuvent être contraires à ceux observés chez la souris, rendant la transposition difficile. Pour proposer un modèle de transposition adéquat pour le développement de ce médicament, il est important de comprendre, pour chaque espèce étudiée, les mécanismes biologiques mis en place ainsi que leurs spécificités. Le but de notre étude est donc de comprendre les effets de la rosiglitazone chez la souris.

A partir d'un code NONMEN fourni par le client, nous devons donc retranscrire les équations différentielles représentant la dynamique d'évolution du médicament dans le métabolisme de la souris.

2.4 Objectif du projet

L'objectif de ce projet est donc d'implémenter en R un modèle PK/PD, pour obtenir des simulations numériques à interpréter et identifier les paramètres, tout en analysant leur sensibilité. Cela s'accompagne par une étude mathématique (notamment statistique) du modèle, pour dégager des tendances spécifiques à la réponse aux médicaments chez la souris.

La rosiglitazone ayant déjà été mise sur le marché, notre étude est particulièrement vouée à être un outil de contrôle et de comparaison à d'autres molécules qui seraient utiles dans le traitement du diabète.

3 Etat de l'art

Le diabète de type 2 étant un problème de santé publique, il n'est pas étonnant que plusieurs études aient été menées sur le sujet, à la fois en physiologie, en biotechnologie et en pharmacologie.

A ce jour, l'état de l'art sur le diabète de type 2 est très avancé et permet de définir des recommandations de bonne pratique (RBP) pour développer des solutions méthodiques afin de rechercher les soins les plus appropriés à chaque patient. Ces RBP sont déterminées à partir de synthèses rigoureuses d'études de fort niveau de preuve fournies par la Haute Autorité de Santé (HAS).

Le but est notamment de prendre en compte les conditions de vie et les antécédents médicaux d'un patient pour proposer des traitements individuels. La HAS définit en particulier les objectifs de glycémie à atteindre, une stratégie médicamenteuse du contrôle de la glycémie et la place de l'autosurveillance glycémique. Ces recommandations s'adressent donc, entre autres, aux laboratoires pharmaceutiques, pour adapter leurs médicaments pour traiter le diabète de type 2. [1]

Les modèles PK/PD ont fait l'objet de nombreuses recherches. Dans un premier temps, les chercheurs utilisaient des modèles minimaux, mais au fur et à mesure, ceux-ci se sont complexifiés afin de tenir compte le plus possible de l'effet d'autres médicaments sur les taux de glucose et d'insuline présents dans le corps. [2]

4 Modélisation biologique

4.1 Diabète de type 2

Le diabète correspond à un excès durable dans le temps de concentration du glucose dans le sang. De manière générale, la glycémie est régulée par différentes hormones parmi lesquelles l'insuline. Seule cette dernière a une action hypoglycémiante, qui permet de faire baisser le taux de sucre dans le sang à la suite des repas. Elle permet l'entrée du sucre dans les cellules pour qu'il soit utilisé par le corps humain, notamment pour le fonctionnement des muscles et du foie.

On distingue 3 types différents de diabète. Le diabète de type 1 est lié à la destruction des cellules du pancréas produisant l'insuline et apparaît le plus souvent chez des personnes peu âgées. Le diabète de type 2 (DT2) est le plus fréquent - il représente en effet 90% des cas de diabète - et sera l'objet de nos propos par la suite. Il arrive en général chez des patients plus âgés et est dû à une résistance à l'insuline. Le diabète gestationnel est un diabète survenant uniquement chez les femmes enceintes. Ces dernières sont alors plus sujettes à souffrir par la suite du diabète de type 2.

Cette maladie est provoquée par la perturbation du métabolisme glucidique. Le corps développe une résistance à l'insuline, empêchant le sucre de rentrer dans les cellules. Cette entrée nécessite l'activation d'un transporteur, activé par un récepteur de l'insuline, qui prend en charge par la suite le glucose. La figure suivante permet d'illustrer ce processus. De plus, pour compenser initialement la résistance à l'insuline, le pancréas se met naturellement à en produire plus. Néanmoins, ce mécanisme est difficile pour l'organe qui finit par s'épuiser et sécréter de moins en moins d'insuline. Cela crée un manque relatif de cette hormone dans l'organisme, qui conduit à une hausse de la glycémie.

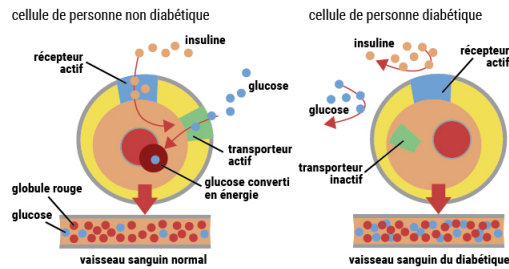


FIGURE 2: Schématisation de l'entrée du glucose dans les cellules. *Crédits : Trust My Science*

Le diabète résulte de la combinaison de facteurs génétiques, environnementaux mais aussi liés au mode de vie. Ainsi, le DT2 apparaît souvent chez des personnes en surpoids, de plus de 40 ans. Bien que les trois points clés pour contrôler cette maladie résident principalement en la perte de poids, l'activité physique et l'alimentation, un traitement médicamenteux doit parfois être prescrit au patient pour diminuer la glycémie ou aider à la contrôler. Plusieurs substances peuvent être administrées par voie orale :

1. **La metformine**, qui permet de diminuer la résistance à l'insuline. L'organisme peut donc utiliser plus efficacement le glucose présent dans le sang, sans risque d'hypoglycémie. Elle est utilisée chez la grande majorité des diabétiques de type 2.
2. **Les médicaments augmentant la sécrétion d'insuline par le pancréas**, grâce à divers mécanismes. Ils comprennent les sulfonylurées et les glinides.
3. **L'acarbose**, inhibiteur de l'alpha-glucosidase, ayant pour effet de diminuer l'absorption des glucides dans l'intestin.
4. **Les incrétines**, famille de médicament augmentant la sécrétion d'insuline après un repas, sans causer d'hypoglycémie.
5. **Les glitazones**, classe de médicaments dont fait partie la rosiglitazone, améliorent la glycémie en diminuant la résistance à l'insuline.

Si un échec est rencontré après l'administration de médicaments oraux, l'insuline par injections est une très bonne alternative. Elle peut également être administrée si les médicaments ne suffisent plus à réguler le taux de glucose dans le sang du patient diabétique.

4.2 Modèle PK/PD

La modélisation PK/PD est une technique qui combine les deux approches pharmacologiques classiques, à savoir la pharmacocinétique (PK) et la pharmacodynamique (PD). Alors que le modèle PK permet de suivre l'évolution de la concentration d'un médicament dans le métabolisme d'un individu, le modèle PD s'intéresse à l'action directe du médicament sur le corps humain et à ses possibles effets secondaires. Ainsi, en combinant ces deux aspects, la modélisation PK/PD permet de décrire les relations entre dosage, concentration et efficacité des médicaments. Correctement décrit, il permet de constituer un outil important à la conception d'expériences et d'essais. Différentes étapes sont à suivre pour mener à bien ce type de modélisation. Elles seront partiellement traitées dans la suite de projet. Elles peuvent être décrites ainsi :

1. Traitement et visualisation des données liées au temps d'absorption du médicament
2. Sélection d'un modèle pharmacocinétique (par exemple à partir d'une bibliothèque)
3. Estimation des paramètres du modèle (à l'aide par exemple d'un modèle à effet mixte)
4. Analyse de sensibilité des paramètres pour étudier dynamiquement le système
5. Simulation des stratégies de dosages à adopter et scénarios possibles

L'objectif d'une telle démarche est notamment de déterminer la fenêtre thérapeutique, afin de déterminer les concentrations pour lesquelles le médicament a une efficacité optimale tout en limitant les effets indésirables possibles. La fenêtre thérapeutique peut être schématisée comme ci-dessous (*cf* figure 2). La particularité de l'insuline est d'avoir une fenêtre thérapeutique très étroite, ne laissant que très peu de possibilité de dosage du médicament pour un patient donné.

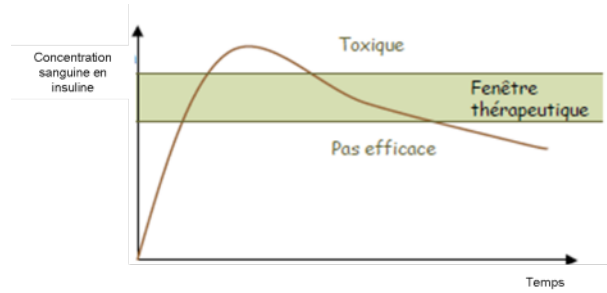


FIGURE 3: Schématisation de la fenêtre thérapeutique pour l'insuline. Crédits : OMéDIT Centre

4.3 Compartiments d'étude

Dans un modèle PK/PD on définit un compartiment comme un espace virtuel où le médicament se répartit de manière homogène (cinétique identique en tout point). Dans notre modélisation nous considérons huit compartiments d'étude :

1. Dépôt rosiglitazone qui décrit les doses de médicament administrées.
2. Plasma rosiglitazone qui décrit l'évolution de la concentration en rosiglitazone dans le sang.
3. Dépôt glucose qui décrit les doses de glucose absorbées par l'individu.
4. Plasma glucose qui caractérise l'évolution de la concentration en glucose dans le sang.
5. Plasma insuline qui modélise l'évolution de la concentration en insuline dans le sang.
6. Plasma *HbA1C* qui décrit l'évolution de la concentration en hémoglobine glyquée dans le sang. Il s'agit d'un marqueur sur le long terme du taux de glycémie dans le sang. Elle permet en effet de donner une estimation facile à obtenir de la glycémie moyenne sur les deux derniers mois.
7. Périphérie rosiglitazone qui modélise la migration du médicament dans les tissus voisins.
8. Compartiment d'effet qui représente le retard de l'effet du médicament par rapport à son ingestion par l'individu.

Le schéma de principe, représentant les relations entre les différents compartiments et introduisant les premiers paramètres du modèle, est donné ci-dessous :

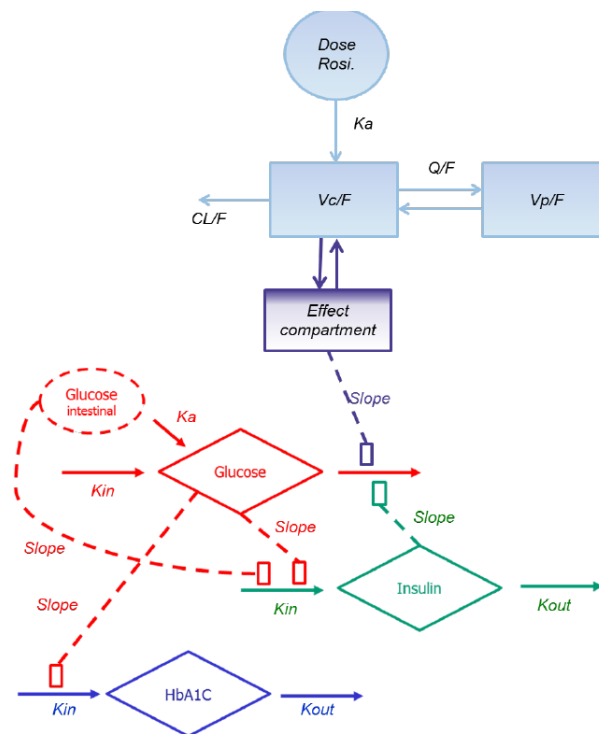


FIGURE 4: Schéma de principe du rosiglitazone. Crédits : Poster PAGE 2017 Servier

4.4 Démarche expérimentale

Pour déterminer si le traitement est efficace, il est important d'avoir une référence à laquelle nous pouvons comparer les résultats. En effet, il est essentiel de vérifier que les diminutions des concentrations de glucose et d'hémoglobine glyquée dans le sang sont dues au traitement et non à d'autres effets (régimes par exemple) que nous pourrions observer sur l'ensemble de la population.

Pour cela, nous disposons d'un groupe témoin auquel nous n'administrons pas de médicament et d'un groupe de patients auquel nous administrons de la rosiglitazone durant toute la durée. Ainsi, si nous observons une baisse des concentrations chez les patients traités mais pas chez le groupe témoin, nous pouvons attribuer ces effets à la rosiglitazone. Cela nous permet de garantir de manière aussi précise que possible que les observations sont dues au traitement.

Par souci de simplification durant notre projet, il nous a parfois été nécessaire de n'utiliser qu'un seul groupe de patients. Dans ces cas-là, nous avons décidé de diviser la durée de l'étude en deux : une première phase où les patients n'étaient pas traités et une deuxième phase sous traitement. Cela nous permettait de comparer les variations avec et sans traitement. Néanmoins, nous avons conscience que cette approche n'est généralement pas fiable car les effets observés sur la deuxième phase pourraient simplement être dus à une stabilisation naturelle au cours du temps des concentrations sanguines.

5 Équations et conditions du modèle

5.1 Paramètres du modèle

Dans la suite, nous désignerons ainsi les paramètres qui nous intéressent :

- CL représente la clairance. Elle correspond à la capacité d'un organisme à se débarrasser d'un liquide biologique. Elle correspond ici au volume de médicament éliminé par unité de temps.
- les volumes sont désignés par des V .
- les valeurs des compartiments d'étude sont représentées par des A , suivi d'un indice i pour spécifier le compartiment d'étude auquel se réfère la grandeur mesurée.
- les coefficients de transfert sont exprimés par des K .

5.2 Équations différentielles du modèle

5.2.1 Conditions initiales

Les conditions initiales de notre système d'équations sont les suivantes :

Pour A_1 (dépôt rosiglitazone) :

$$A_1(0) = 0$$

Pour A_2 (plasma rosiglitazone) :

$$A_2(0) = 0$$

Pour A_3 (dépôt glucose) :

$$A_3(0) = 0$$

Pour A_4 (plasma glucose) :

$$A_4(0) = BSL_{Glc}$$

Pour A_5 (plasma insuline) :

$$A_5(0) = K_{in,Ins} \frac{1 + Slope_{Glc} \frac{BSL_{Glc}}{V_4}}{K_{out,Ins}}$$

Pour A_6 (plasma $HbA1c$) :

$$A_6(0) = K_{in,HbA1C} \frac{BSL_{Glc}}{V_4} \frac{SLGH}{K_{out,HbA1C}}$$

Pour A_7 (périphérique rosiglitazone) :

$$A_7(0) = 0$$

Pour A_8 (effet rosiglitazone) :

$$A_8(0) = 0$$

5.2.2 Équations du système

On peut modéliser le système par les équations suivantes, décrivant ces quatre parties principales :

$$\begin{aligned}
 \text{PK Rosi} \quad & \begin{cases} \frac{dA_1}{dt} = & -K_a A_1 \\ \frac{dA_2}{dt} = & K_a A_1 - K_{27} A_2 + K_{72} A_7 - K_{20} A_2 \\ \frac{dA_7}{dt} = & K_{27} A_2 - K_{72} A_7 \\ \frac{dA_8}{dt} = & K_{e0} \left(\frac{A_2}{V_2} - A_8 \right) \end{cases} \\
 \text{PK Insuline} \quad & \begin{cases} \frac{dA_5}{dt} = & K_{inIns} (1 + S_{Glc} + S_{Gut}) - K_{outIns} A_5 \\ S_{Glc} = & Slope_{Glc} C_4 \\ S_{Gut} = & Slope_{Gut} A_3 \end{cases} \\
 \text{PK HbA1c} \quad & \begin{cases} \frac{dA_6}{dt} = & \frac{A_4}{V_4} K_{inHbA1c} S_{Lgh} - A_6 K_{outHbA1c} \end{cases} \\
 \text{PD} \quad & \begin{cases} S_R = & S_{LR} A_8 \\ S_{Ins} = & Slope_{Ins} C_5 \\ \frac{dA_3}{dt} = & -K_{a,Glc} A_3 \\ \frac{dA_4}{dt} = & K_{a,Glc} A_3 + K_{in,Glc} - K_{out,Glc} S_{Ins} A_4 (1 + S_R) \end{cases}
 \end{aligned}$$

6 Identifiabilité des paramètres

6.1 Une introduction à l'identification des modèles

L'identification de modèles paramétriques [3] à partir d'observations se pose dans de nombreux problèmes. Une question naturelle que nous pouvons nous poser est alors de savoir s'il est possible d'estimer les paramètres inconnus du modèle, à partir de ces observations.

La démarche classique pour identifier un modèle se déroule en 4 étapes [4] :

1. Formaliser les connaissances a priori
2. Recueillir les données expérimentales à disposition
3. Identifier la structure du modèle et les paramètres associés (ainsi que les incertitudes)
4. Évaluer le modèle ainsi obtenu, notamment en le comparant à des données qui n'ont pas servi auparavant

Parfois, il est important d'effectuer cette démarche de façon cyclique, pour pouvoir modifier au fur et à mesure les modèles. La démarche se termine réellement lorsque le modèle répond à l'objectif établi initialement.

Ici, nous considérons un modèle de connaissances (équations mathématiques présentées en 5.2.2). Notre identification a pour seul objectif de déterminer les paramètres (au nombre de 20) qui interviennent dans celui-ci.

L'identifiabilité paramétrique est le problème inverse qui consiste à savoir si les paramètres inconnus d'un modèle peuvent être estimés de façon unique à partir des observations. C'est donc un pré-requis nécessaire à l'identification de paramètres d'un modèle. Ici, nous cherchons donc à savoir si les paramètres intervenant dans les équations précédemment présentées sont identifiables ou non.

Des méthodes d'analyse d'identifiabilité sont connues et satisfaisantes dans le cas des systèmes linéaires. Elles correspondent notamment à l'utilisation de la transformée de Laplace ou du développement en série de Taylor. Par contre, l'étude des systèmes non-linéaires (ce qui est notre cas) s'avère a priori plus difficile. Toutefois, différentes méthodes sont accessibles dans la littérature. L'efficacité de celles-ci est étroitement liée à la nature de la non linéarité ainsi qu'à celle de l'observation. L'identifiabilité peut être analytique (prouvée analytiquement) ou structurelle (prouvée de manière numérique).

Les méthodes d'identification à utiliser dépendent des caractéristiques des modèles et des protocoles expérimentaux applicables au système. Elles peuvent être réunies en deux grandes catégories :

- les méthodes graphiques : elles s'appuient sur les réponses à des signaux usuels (échelon, impulsion ou sinusoïde)

- les méthodes statistiques : elles s'expriment sous la forme d'un problème d'optimisation dont la résolution peut se révéler complexe et faire parfois appel à des algorithmes récursifs

6.2 Résultats analytiques

Les premiers paramètres qui apparaissent simplement identifiables sont K_a et K_{aGlc} . En effet, rappelons qu'ils n'interviennent que dans les équations suivantes :

$$\begin{cases} \frac{dA_1}{dt} = -K_a A_1 \\ \frac{dA_3}{dt} = -K_{aGlc} A_3 \end{cases}$$

Ces équations différentielles sont très simples et peuvent être résolues de manière exacte. Les conditions initiales de A_1 et A_3 peuvent être notées A_{1_0} et A_{3_0} (injection de médicament et de glucose). Dans ces conditions, on a alors :

$$\begin{cases} A_1(t) = A_{1_0} e^{-K_a t} \\ A_3(t) = A_{3_0} e^{-K_{aGlc} t} \end{cases}$$

A tout temps t , les relations suivantes sont vérifiées :

$$\begin{cases} K_a = \frac{1}{t} \ln \left(\frac{A_{1_0}}{A_1(t)} \right) \\ K_{aGlc} = \frac{1}{t} \ln \left(\frac{A_{3_0}}{A_3(t)} \right) \end{cases}$$

Ainsi, ces deux paramètres sont analytiquement identifiables. Numériquement, cette opération est plus complexe. Il n'est possible d'identifier qu'un plateau de valeurs où les paramètres pourraient se situer.

Ainsi, pour les identifier de manière numérique, il est possible de calculer les différentes valeurs de K_a et K_{aGlc} à différents instants t et de prendre la valeur moyenne de tous les K_a et K_{aGlc} calculés. Bien sûr, la précision des résultats obtenus croît avec le nombre de mesure (à différents instants t) choisi.

Par ailleurs, le paramètre BSL_{Glc} est la condition initiale du plasma glucose A_4 . Ce paramètre est donc identifiable.

6.3 Analyse indicielle

La première méthode envisagée pour déterminer si les autres paramètres sont identifiables est d'utiliser les transformées de Laplace. En effet, celle-ci a des propriétés intéressantes du point de vue de la linéarité, mais surtout de la dérivabilité. Cet argument est important compte tenu de notre travail sur des équations différentielles.

On rappelle que la transformée de Laplace est donnée, pour une fonction $x : \mathbb{R}_+ \rightarrow \mathbb{R}$ localement intégrable, par :

$$\mathcal{L}(x)(p) = \int_0^{+\infty} x(t) e^{-pt} dt, \quad p \in \mathbb{C}$$

Il s'agit d'une application linéaire qui vérifie, pour tout x de classe \mathcal{C}^1 :

$$\mathcal{L}(\dot{x})(p) = p\mathcal{L}(x)(p) - x(0^-)$$

Une approche classique revient à déduire un modèle du système physique à partir du relevé de la réponse obtenue lorsqu'on applique un échelon en entrée. La procédure expérimentale consiste à changer instantanément la valeur de l'entrée, alors que le système se trouvait en régime permanent. On enregistre la sortie et on en déduit alors la réponse indicielle. En théorie, la réponse indicielle présente toute l'information nécessaire à la fonction de transfert $H(p)$ du système. Une fois la réponse indicielle obtenue, une analyse graphique rapidement permet de déduire certaines caractéristiques de la fonction de transfert (en s'intéressant à la tangente à l'origine, au retard pur ou encore à la monotonie par exemple).

6.3.1 Identification par une fonction de transfert du premier ordre

Dans le cas où la réponse indicielle a une tangente à l'origine non nulle et est monotone, on peut rechercher un modèle sous la forme d'une fonction de transfert du premier ordre. Dans ce cas, $H(p)$ s'écrit sous la forme :

$$H(p) = K \frac{1}{1 + Tp}$$

En utilisant les transformées de Laplace inverses, on en déduit que la réponse indicielle correspondante est :

$$y(t) = K \left(1 - \exp \left(-\frac{t}{T} \right) \right) \mathbb{1}(t)$$

Il est alors possible d'identifier les paramètres K et T , de la manière suivante, résumée sur l'illustration ci-dessous :

- la réponse tend asymptotiquement vers la droite d'équation $y = K$
- pour $t = T$, la sortie vaut $y(T) \approx 0,63K$
- pour $t = 3T$, la sortie vaut $y(3T) \approx 0,95K$
- pour $t = T$, la tangente à l'origine coupe l'asymptote horizontale

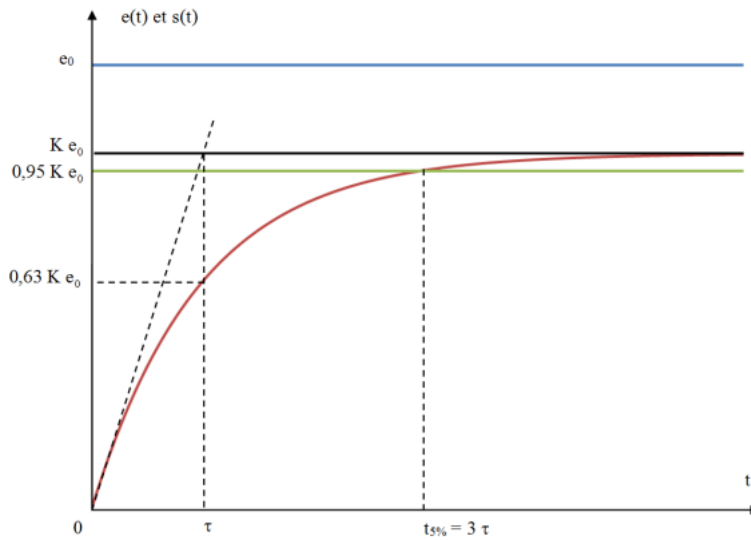


FIGURE 5: Réponse indicielle pour une fonction de transfert du premier ordre. *Crédits : <http://sii-tannarelli.com/>*

Pour déterminer T et limiter les incertitudes, il peut être conseillé de se baser à la fois sur l'asymptote et sur les mesures à différents t et prendre ensuite la valeur moyenne.

6.3.2 Application à notre modèle

Dans notre modèle, nous pouvons passer un certain nombre d'équations dans le domaine de Laplace et celles-ci peuvent s'avérer intéressantes pour identifier nos paramètres. En particulier, les équations faisant intervenir $\frac{dA_8}{dt}$ font apparaître des fonctions de transfert du premier ordre.

La transformée de Laplace de l'équation en $\frac{dA_8}{dt}$ s'écrit :

$$pA_8 = K_{e0} \left(\frac{A_2}{V_2} - A_8 \right)$$

Nous pouvons alors nous intéresser à la fonction de transfert $H_{28}(p) = \frac{A_8(p)}{A_2(p)}$, qui est donné par :

$$H_{28}(p) = \frac{1}{V_2} \frac{1}{1 + \frac{1}{K_{e0}} p}$$

Comme expliqué précédemment, les paramètres V_2 et K_{e0} sont donc identifiables en traçant la réponse indicielle. En effet, $\frac{1}{V_2}$ correspond au gain, tandis que $\frac{1}{K_{e0}}$ correspond à l'instant t où la tangente à l'origine coupe l'asymptote horizontale.

La transformée de Laplace de l'équation en $\frac{dA_6}{dt}$ s'écrit :

$$pA_6 = \frac{A_4}{V_4} K_{inHbA1c} S_{Lgh} - A_6 K_{outHbA1c}$$

Nous pouvons alors nous intéresser à la fonction de transfert $H_{46}(p) = \frac{A_6(p)}{A_4(p)}$, qui est donné par :

$$H_{46}(p) = \frac{K_{inHbA1c} S_{Lgh}}{V_4 K_{outHbA1c}} \frac{1}{1 + \frac{1}{K_{outHbA1c}} p}$$

Ici, les paramètres $K_{outHbA1c}$ et $\frac{K_{inHbA1c} S_{Lgh}}{V_4 K_{outHbA1c}}$ sont donc identifiables en traçant la réponse indicielle. En effet, $\frac{K_{inHbA1c} S_{Lgh}}{V_4 K_{outHbA1c}}$ correspond au gain, tandis que $\frac{1}{K_{outHbA1c}}$ correspond à l'instant t où la tangente à l'origine coupe l'asymptote horizontale.

Nous pouvons directement en déduire que $\frac{K_{inHbA1c} S_{Lgh}}{V_4}$ est lui aussi identifiable. Néanmoins, les paramètres $K_{inHbA1c}$ et S_{Lgh} n'apparaissant dans aucune autre équation et uniquement dans les conditions initiales de A_6 , nous pouvons en déduire que ces paramètres ne sont pas identifiables séparément.

Nous ne pouvons pas encore conclure que le paramètre V_4 (qui, rappelons-le, est égal à V_5) n'est pas identifiable, car il apparaît dans d'autres équations.

La méthode harmonique, présentée ci-après, serait également applicable pour les équations faisant intervenir $\frac{dA_8}{dt}$ et $\frac{dA_6}{dt}$ et donnerait les mêmes résultats d'identifiabilité.

6.4 Analyse harmonique

6.4.1 Procédure expérimentale

La procédure expérimentale de cette méthode consiste à relever le gain $|H(jw)|$ et la phase $\arg(H(jw))$ pour un ensemble de pulsations w , permettant de tracer un graphe suffisamment complet. Cette opération peut se faire avec un oscilloscope ou des logiciels spécialisés par exemple. Les courbes obtenues représentent alors la réponse harmonique du système.

En général, on souhaite obtenir un modèle paramétrique, qui peut s'écrire sous la forme :

$$H(p) = K p^q \frac{\prod_{k=1}^{n_k} \left(1 + \frac{p}{w_k}\right) \prod_{l=1}^{m_l} \left(1 + 2\xi_l \frac{p}{w_l} + \left(\frac{p}{w_l}\right)^2\right)}{\prod_{m=1}^{n_m} \left(1 + \frac{p}{w_m}\right) \prod_{n=1}^{n_n} \left(1 + 2\xi_n \frac{p}{w_n} + \left(\frac{p}{w_n}\right)^2\right)}$$

Les principes généraux pour identifier les paramètres sont donc :

- le comportement asymptotique en basse fréquence permet de déterminer les valeurs du réel K et de l'entier relatif q . En effet, pour $w \rightarrow 0$, $H(jw) \rightarrow K j^q w^q$. L'asymptote du gain est donc une droite de pente $20q$ dB/décade coupant l'axe 0 dB à la pulsation $w = \left|\frac{1}{K}\right|^{\frac{1}{q}}$. La phase permet de déterminer le signe de K car l'asymptote est une droite horizontale d'ordonnée $90q^\circ$ si $K > 0$ ou $90q^\circ + 180^\circ$ si $K < 0$.
- les pulsations w_k , w_l , w_m et w_n correspondent à des pulsations de coupure. Il faut donc utiliser les asymptotes de la courbe de gain tracée dans le plan de Bode, tout en vérifiant que les résultats coïncident avec les courbes de phase. Par exemple, pour $w = |w_k|$ (resp. $w = |w_m|$), le terme à prendre en compte est $1 + \frac{p}{w_k}$ au numérateur (resp. $1 + \frac{p}{w_l}$ au dénominateur). En module, le tracé asymptotique subit un changement de pente de $+20$ dB/décade (resp. -20 dB/décade). En phase, l'asymptote subit un saut de $90^\circ \operatorname{sgn}(w_k)$ (resp. $-90^\circ \operatorname{sgn}(w_m)$). Les termes du second ordre impliquent les mêmes fluctuations (en signe) mais avec des variations de 40 dB/décade et des sauts de 180° (dépendant cette fois des signes des ξ).

- les résonances éventuelles sur la courbe de gain permettent de déterminer les amortissements $\xi_l < 1$. On peut notamment utiliser le dépassement, donné par la formule suivante : $D = \exp\left(-\frac{\pi\xi}{\sqrt{1-\xi^2}}\right)$

6.4.2 Application à notre modèle

En plus des équations faisant intervenir $\frac{dA_8}{dt}$ et $\frac{dA_6}{dt}$, celles en $\frac{dA_2}{dt}$ et $\frac{dA_7}{dt}$ sont également intéressantes dans le domaine de Laplace.

La transformée de Laplace de l'équation en $\frac{dA_2}{dt}$ s'écrit :

$$pA_2(p) = K_a A_1(p) - K_{27}A_2(p) + K_{72}A_7(p) - K_{20}A_2(p)$$

Ce qui peut se réécrire sous la forme de l'équation (1) suivante :

$$(p + K_{27} + K_{20})A_2(p) = K_{72}A_7(p) + K_a A_1(p)$$

Par ailleurs, la transformée de Laplace de l'équation en $\frac{dA_7}{dt}$ s'écrit :

$$(p + K_{72})A_7(p) = K_{27}A_2(p)$$

Cela permet alors d'isoler la valeur de $A_7(p)$:

$$A_7(p) = \frac{K_{27}}{p + K_{72}} A_2(p)$$

En réinjectant cette expression de $A_7(p)$ dans l'équation (1), on obtient :

$$\left(p + K_{72} + K_{20} - \frac{K_{27}K_{72}}{p + K_{72}}\right) A_2 = K_a A_1$$

Nous pouvons alors nous intéresser à la fonction de transfert $H_{12}(p) = \frac{K_a A_1(p)}{A_2(p)}$, qui est donné par :

$$\begin{aligned} H_{12}(p) &= \frac{(p + K_{72})(p + K_{27} + K_{20}) - K_{27}K_{72}}{p + K_{72}} \\ &= \frac{N_{12}(p)}{D_{12}(p)} \end{aligned}$$

avec $N_{12}(p) = \left(1 + \frac{1}{K_{72}}p\right)(K_{27} + K_{20} + p) - K_{27}$ et $D_{12}(p) = 1 + \frac{1}{K_{72}}p$

Finalement, la forme canonique de $H_{12}(p)$ est donnée par :

$$H_{12}(p) = K_{20} \frac{\left(1 + \left(\frac{1}{K_{20}} + \frac{K_{27}+K_{20}}{K_{72}K_{20}}\right)p + \frac{1}{K_{72}K_{20}}p^2\right)}{1 + \frac{1}{K_{72}}p}$$

Pour analyser les résultats, nous gardons les mêmes notations utilisées dans la partie précédente pour la forme générale du modèle.

K_{20} correspond au gain de la fonction, il est donc identifiable en utilisant le comportement asymptotique à basse fréquence.

Nous remarquons que $\sqrt{K_{72}K_{20}}$ correspond à une pulsation de coupure de type w_l . Ainsi, à cette pulsation, le tracé asymptotique subit un changement de pente de +40 dB/décade en gain et un saut de $180^\circ \operatorname{sgn}(\xi_l)$ en phase. Ainsi, $K_{72}K_{20}$ est identifiable. K_{20} ayant été identifié précédemment, K_{72} est lui aussi identifiable.

Enfin, le terme $\frac{1}{K_{20}} + \frac{K_{27}+K_{20}}{K_{72}K_{20}}$ correspond au terme $\frac{2\xi_l}{w_l}$ de la forme générale de la fonction de transfert. La pulsation w_l a été mesurée précédemment et l'amortissement ξ_l peut être calculé grâce à la mesure du dépassement. Ainsi, on en déduit directement que $\frac{1}{K_{20}} + \frac{K_{27}+K_{20}}{K_{72}K_{20}}$ est identifiable. Or, cette expression fait uniquement intervenir les paramètres K_{20} , K_{72} et K_{27} . Comme nous avons déjà prouvé l'identifiabilité des deux premiers paramètres, nous pouvons donc conclure que K_{27} est lui aussi identifiable.

Cette partie a donc pu prouver l'identifiabilité de K_{20} , K_{27} et K_{72} .

6.5 Identifiabilité empirique

Les équations en $\frac{dA_4}{dt}$ et en $\frac{dA_5}{dt}$ ont l'inconvénient de ne pas être linéaire. Or, la transformée d'un produit dans le domaine de Laplace n'est pas simple. Nous ne pouvons donc pas utiliser les méthodes précédemment présentées.

Pour identifier les paramètres restants, nous avons pensé à utiliser un développement en série de Taylor. Néanmoins, cette méthode s'avère difficilement applicable à notre modèle car elle nécessiterait la résolution d'un système 15x15 non linéaire.

Par manque de temps, nous avons décidé de ne pas utiliser cette méthode mais plutôt de regarder si, numériquement, il était possible d'identifier davantage de paramètres.

Pour cela, nous pouvons utiliser les régressions linéaires, simple ou multiple, présentées dans la partie suivante. Cela permet d'identifier numériquement des paramètres supplémentaires

Nous pouvons en effet remarquer que $Slope_{Gut}$ est l'ordonnée à l'origine de la régression linéaire entre $\frac{dA_5}{dt}$ et A_3 tandis que K_{outGlc} correspond à l'ordonnée à l'origine de la régression linéaire entre $\frac{dA_5}{dt}$ et A_5 . Enfin, pour l'équation en $\frac{dA_5}{dt}$, on identifie K_{inIns} grâce à l'ordonnée à l'origine de la régression linéaire multiple entre $\frac{dA_5}{dt}$ et $A_3 + A_4 + A_5$.

De la même façon, l'équation en $\frac{dA_4}{dt}$ permet d'identifier K_{inGlc} (ordonnée à l'origine de la régression linéaire par rapport à A_3) et S_{LR} . Pour ce dernier paramètre, il faut en effet effectuer la régression linéaire par rapport au produit A_4A_5 , puis celle par rapport à $A_4A_5A_8$ et faire le rapport des deux ordonnées à l'origine obtenues.

Ainsi, cette méthode permet d'identifier S_{LR} , K_{inGlc} , K_{inIns} , K_{outGlc} et $Slope_{Gut}$.

6.6 Résumé des résultats obtenus

La figure suivante permet de résumer quels sont les paramètres identifiables de notre modèle. S'ils ne le sont pas, elle donne les produits identifiables dans lesquels les paramètres apparaissent.

Paramètres	Identifiable ?	Produit identifiable
K_a		
K_{27}		
K_{72}		
K_{20}		
K_{e0}		
V_2		
S_{LR}	Régression linéaire multiple	
$Slope_{Ins}$		$\frac{Slope_{Ins}K_{outGlc}}{V_4}$
$V_4 = V_5$		$\frac{Slope_{Ins}K_{outGlc}}{V_4}, \frac{Slope_{Glc}}{V_4}, \frac{K_{inHb1Ac}S_{Lgh}}{V_4}$
K_{aGlc}		
K_{inGlc}	Régression linéaire multiple	
K_{outGlc}		$\frac{Slope_{Ins}K_{outGlc}}{V_4}$
$Slope_{Glc}$		$\frac{Slope_{Glc}}{V_4}$
$Slope_{Gut}$	Régression linéaire multiple	
K_{inIns}	Régression linéaire multiple	
K_{outIns}	Régression linéaire multiple	
$K_{inHb1Ac}$		$\frac{K_{inHb1Ac}S_{Lgh}}{V_4}$
S_{Lgh}		$\frac{K_{inHb1Ac}S_{Lgh}}{V_4}$
$K_{outHb1Ac}$		
BSL_{GLC}		

FIGURE 6: Identifiabilité des paramètres dans notre modèle

Nous remarquons que tous les paramètres non identifiables sont présents dans un produit qui, lui, est identifiable. Par ailleurs, chacun de ses produits fait intervenir la variable V_4 . Finalement, nous pouvons remarquer que fixer V_4 permettrait d'identifier les variables restantes, à l'exclusion de S_{Lgh} et $K_{inHb1ac}$. Or, V_4 correspond à un volume qui est censé représenter l'endroit où se déplace le glucose. Il est en réalité non évident à appréhender mais n'est pas forcément intéressant à connaître de manière précise. En effet, le niveau plasmatique du glucose est le plus intéressant mais les biologistes ne se préoccupent pas nécessairement du glucose qui va ailleurs. Il est donc tout à fait possible de fixer V_4 à un volume arbitraire (par exemple le volume du corps) sans aucune conséquence grave sur l'étude.

7 Modélisation mathématique

7.1 Modèle linéaire

Le modèle statistique que l'on utilise pour étudier les relations entre une variable réponse Y en fonction de variables qualitatives et/ou quantitatives peut s'écrire sous la forme :

$$Y_i = m_i + E_i$$

où i est le numéro de l'unité expérimentale, m_i correspond à l'espérance de Y_i et inclut l'effet des variables explicatives, E_i est une variable résiduelle incluant la variabilité liée au matériel de mesure ou aux variables non incluses dans le modèle.

Selon la nature des variables explicatives, on distingue trois grandes catégories de modèles linéaires :

1. Modèle de régression où les variables explicatives sont **quantitatives**. La régression est dite multiple quand elle est utilisée pour décrire les variations d'une variable endogène associée aux variations de plusieurs variables exogènes. On désignera par p dans la suite ce nombre de variables.
2. Modèle de régression où les variables explicatives sont **qualitatives**.
3. Modèle de régression où les variables explicatives sont **mixtes** (quantitatives et qualitatives).

La régression multiple permet d'expliquer les relations entre Y et un ensemble de variables explicatives X_1, \dots, X_p . Ce modèle est défini par :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + E$$

Les coefficients de régressions β_j sont fixes mais inconnus et E est un terme aléatoire tel que $E \sim \mathcal{N}(0, \sigma^2)$ (modèle à effets fixes). On s'intéresse à des données regroupées pour n individus pour lesquels on dispose de p mesures. On

possède donc n observations des variables Y, X_1, \dots, X_p . Pour un individu i , on note y_i et $x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{pi} \end{pmatrix}$ ces observations.

La matrice de données, notée X devient alors :

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}$$

Pour chaque individu i , on considère que la valeur y_i prise par Y est une réalisation de la variable aléatoire Y_i définie par :

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + E_i$$

où E_i est un terme aléatoire tel que $E_i \sim \mathcal{N}(0, \sigma^2)$. Les E_i sont également supposés indépendants. On obtient alors que $Y_i \sim \mathcal{N}(x_i \beta, \sigma^2)$.

En notant $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ et $E = \begin{pmatrix} E_1 \\ E_2 \\ \vdots \\ E_p \end{pmatrix}$, on obtient alors :

$$Y = X\beta + E$$

L'utilisation de tests d'hypothèses permet de déterminer les propriétés des estimateurs des paramètres du modèle (biais, convergence) et leurs lois de distribution [5].

Quant à lui, le modèle linéaire à effets fixes [6] est une extension du modèle linéaire présenté ci-dessus mais qui prend en compte la variabilité liée aux individus. Il est composé d'une partie fixe, identique pour tous les individus (effet de population), et d'une partie aléatoire, propre à chacun et traduisant la variabilité. Le modèle s'écrit alors :

$$Y_{i,t} = (\alpha + \alpha_i) + \beta t + E_{i,t}$$

où α et β représentent les effets fixes, $E_{i,t} \sim \mathcal{N}(0, \sigma^2)$ représentant le terme d'erreur aléatoire et $\alpha_i \sim \mathcal{N}(0, \mu^2)$ représentant l'effet aléatoire.

La constante aléatoire signifie que, à $t = 0$, il y a un point de départ moyen de α . Cependant, chaque individu peut avoir une variation de $\alpha_i \sim \mathcal{N}(0, \mu^2)$, correspondant à l'inter-individualité qui sera mise en évidence dans la partie 8.2. C'est donc ce modèle que nous utiliserons dans la suite pour des raisons évidentes.

7.2 Une approche Machine Learning

Dans un premier temps, nous avons suivi une approche de type Machine Learning afin d'obtenir les valeurs des paramètres caractérisant les lois log-normales intervenant dans le modèle.

Le principe de l'approche est le suivant :

1. On calcule les valeurs des tirages pour chaque patient en utilisant le modèle linéaire décrit dans la partie précédente.
2. A partir des valeurs des tirages, on détermine des estimateurs des paramètres de la loi log-normale à l'aide du calcul du maximum de vraisemblance.
3. A partir des valeurs des paramètres de la loi, on estime les valeurs des tirages à l'aide du calcul du maximum a posteriori.
4. On répète les opérations 2 et 3 jusqu'à convergence des estimateurs des paramètres.

Commençons par calculer les estimateurs de maximum de vraisemblance qui permettent de réaliser l'étape numéro 2 [7].

On se donne une variable aléatoire $\alpha \sim \text{LogN}(\mu, \sigma^2)$ et un tirage $A = (\alpha_i)_{i \in \llbracket 1, N \rrbracket}$. La vraisemblance s'écrit :

$$\begin{aligned} \mathcal{L}(\mu, \sigma^2 | A) &\propto \prod_{i=1}^N f(\alpha_i | \mu, \sigma^2) \\ &\propto \prod_{i=1}^N \left(\frac{1}{\alpha_i \sigma \sqrt{2\pi}} \exp \left(-\frac{(\ln(\alpha_i) - \mu)^2}{2\sigma^2} \right) \right) \\ &\propto (2\sigma^2 \pi)^{-\frac{N}{2}} \prod_{i=1}^N \left(\frac{1}{\alpha_i} \right) \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (\ln(\alpha_i) - \mu)^2 \right) \end{aligned}$$

On obtient donc la log-vraisemblance :

$$\begin{aligned} -l(\mu, \sigma^2 | A) &= \frac{N}{2} \ln(2\sigma^2 \pi) + \sum_{i=1}^N \ln(\alpha_i) + \frac{1}{2\sigma^2} \sum_{i=1}^N (\ln(\alpha_i) - \mu)^2 + c \text{ où } c \text{ désigne une constante} \\ &= \frac{N}{2} \ln(2\sigma^2 \pi) + \sum_{i=1}^N \ln(\alpha_i) + \frac{1}{2\sigma^2} \sum_{i=1}^N \ln(\alpha_i)^2 - \frac{2}{2\sigma^2} \sum_{i=1}^N \mu \ln(\alpha_i) + \frac{1}{2\sigma^2} N\mu^2 + c \\ &= \frac{N}{2} \ln(2\sigma^2 \pi) + \frac{1}{2\sigma^2} \sum_{i=1}^N \ln(\alpha_i)^2 - \frac{1}{\sigma^2} \sum_{i=1}^N \mu \ln(\alpha_i) + \frac{1}{2\sigma^2} N\mu^2 + \tilde{c} \end{aligned}$$

En dérivant par rapport à μ , on obtient :

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^N \ln(\alpha_i) + \frac{N}{\sigma^2} \mu$$

On a donc l'estimateur de maximum de vraisemblance pour μ :

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \ln(\alpha_i)$$

De même, en dérivant par rapport à σ^2 :

$$\frac{\partial l}{\partial \sigma^2} = \frac{N}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^N (\ln(\alpha_i) - \mu)^2$$

On a donc l'estimateur de maximum de vraisemblance pour σ :

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (\ln(\alpha_i) - \hat{\mu})^2$$

Pour mettre en place l'étape 3, on se place dans le cas où l'on cherche à faire une régression linéaire du type $y = \beta^T x + \epsilon$ avec β_1 suivant une loi log-normale de paramètres (μ, σ^2) connus.

Pour prendre en compte ces deux contraintes, on cherche à minimiser une expression contenant le critère des moindres carrés $\|y - \beta^T x\|^2$ ainsi qu'une pénalité de la forme de la vraisemblance de la loi log-normale.

On a donc :

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|y - \beta^T x\|^2 - \ln(p(\beta_1))$$

Avec

$$\ln(p(\beta_1)) = \ln \left(\frac{1}{\beta_1 \sigma \sqrt{2\pi}} \exp \left(-\frac{(\ln(\beta_1) - \mu)^2}{2\sigma^2} \right) \right)$$

D'où

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left(\|y - \beta^T x\|^2 + \ln(\beta_1) + \frac{(\ln(\beta_1) - \mu)^2}{2\sigma^2} \right)$$

7.3 Expectation Maximization

En statistiques, l'algorithme d'expectation maximization [8] (algorithme EM) permet d'estimer les paramètres du maximum de vraisemblance d'un modèle probabiliste lorsque ce dernier dépend de variables latentes non observables. On appelle variable latente une variable qui n'est pas directement observée mais qui se déduit d'autres variables observées. Plus précisément, ces variables peuvent être mesurées en théorie, mais ne le sont pas pour des raisons pratiques. Par exemple, dans notre modèle, seuls les plasmas sont en réalité mesurables (en réalisant des prises de sang sur le patient par exemple) alors que les autres variables ne peuvent se mesurer en pratique. Elles peuvent néanmoins être simulées et mesurées en théorie. L'expectation maximization est un algorithme très souvent utilisé pour l'analyse des paramètres d'une population dans un modèle PK/PD [9], c'est pourquoi nous avons décidé de nous y intéresser.

L'algorithme d'expectation maximization (appelé espérance-maximisation en français) comporte deux grandes étapes :

- une étape d'évaluation de l'espérance : celle-ci revient à calculer l'espérance de la vraisemblance en tenant compte des dernières variables observées.
- une étape de maximisation : celle-ci estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape précédente.

Le résultat trouvé à l'étape de maximisation sert alors à initialiser une nouvelle étape d'évaluation de l'espérance. On itère ainsi l'algorithme.

Nous détaillons ci-dessous le principe de fonctionnement de l'algorithme.

Considérons donc $x = (x_1, \dots, x_n)$ un échantillon de n individus suivant chacun une loi $f(x_i, \theta)$. Le but de l'algorithme est donc de déterminer θ en maximisant la log-vraisemblance. Pour rappel, la log-vraisemblance a pour formule :

$$L(x, \theta) = \sum_{i=1}^n \log(f(x_i, \theta))$$

De manière générale, on utilise cet algorithme lorsque la maximisation de log-vraisemblance est difficile mais que l'on peut déterminer de manière très simple θ à partir de la connaissance de quelques données judicieusement choisies. Dans ce cas, on complète les données par un vecteur $z = (z_1, \dots, z_n)$. On désigne alors par $f(z_i|x_i, \theta)$ la probabilité de z_i sachant x_i et le paramètre θ . La log-vraisemblance complétée est alors désignée par la quantité suivante :

$$L((x, z), \theta) = \sum_{i=1}^n (\log(f(z_i|x_i, \theta)) + \log(f(x_i, \theta)))$$

Cela permet de réécrire la log-vraisemblance comme :

$$L(x, \theta) = L((x, z), \theta) - \sum_{i=1}^n \log(f(z_i|x_i, \theta))$$

L'algorithme EM est alors une procédure itérative basée sur l'espérance des données complétées conditionnellement au paramètre courant. En désignant par $\theta^{(c)}$ ce paramètre, on peut alors écrire, en prenant l'espérance sous z :

$$\mathbb{E}(L(x, \theta)|\theta^{(c)}) = \mathbb{E}(L((x, z), \theta)|\theta^{(c)}) - \mathbb{E}\left(\sum_{i=1}^n \log(f(z_i|x_i, \theta)) \mid \theta^{(c)}\right)$$

Comme $L(x, \theta)$ ne dépend pas de z , cela se réécrit :

$$L(x, \theta) = \mathbb{E}(L((x, z), \theta)|\theta^{(c)}) - \mathbb{E}\left(\sum_{i=1}^n \log(f(z_i|x_i, \theta)) \mid \theta^{(c)}\right)$$

Pour plus de lisibilité, nous désignerons dans la suite ainsi les quantités suivantes :

$$g(\theta, \theta^{(c)}) = \mathbb{E}(L((x, z), \theta)|\theta^{(c)})$$

$$h(\theta, \theta^{(c)}) = \mathbb{E}\left(\sum_{i=1}^n \log(f(z_i|x_i, \theta)) \mid \theta^{(c)}\right)$$

Ainsi, la suite définie par $\theta^{(c+1)} = \arg \max_{\theta} (g(\theta, \theta^{(c)}))$ fait converger $L(x, \theta^{c+1})$ vers un maximum local.

L'algorithme EM peut alors être défini comme [10] :

1. Initialisation au hasard de θ^c
2. Poser $c = 0$
3. Tant que l'algorithme ne converge pas, faire :
 - Évaluation de l'espérance : $g(\theta, \theta^{(c)}) = \mathbb{E}(L((x, z), \theta)|\theta^{(c)})$
 - Maximisation : $\theta^{(c+1)} = \arg \max_{\theta} (g(\theta, \theta^{(c)}))$
 - $c = c + 1$
4. Renvoyer $L(x, \theta^c)$

Si l'on veut s'affranchir du caractère local du maximum renvoyé, on peut faire tourner un grand nombre de fois l'algorithme en initialisant à des valeurs différentes θ^c .

7.4 Application de l'Expectation-Maximization à notre modèle

7.4.1 Cas particulier de la dimension 1

Comme on sait que tous les paramètres intervenant dans les équations suivent des lois log-normales, et donc que leur logarithme suit une loi normale, on va pouvoir simplifier l'implémentation de cet algorithme.

On se donne une variable aléatoire X suivant une loi normale de densité $f(x, \theta) = \frac{1}{\theta_2 \sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{x-\theta_1}{\theta_2})^2)$.

Soit $(x_i)_{1 \leq i \leq n}$ des tirages de cette variable aléatoire.

La log-vraisemblance associée s'écrit

$$l(x, \theta) = -\frac{n}{2} \ln(2\pi) - n \ln(\theta_2) - \frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \theta_1}{\theta_2} \right)^2$$

On introduit une variable aléatoire Z de densité sachant X $f(Z|X, \theta)$ et la log-vraisemblance complétée

$$l((x, z), \theta) = l(x, \theta) + \sum_{i=1}^n \ln(f(z_i|x_i, \theta)) = -n \ln(2\pi) - 2n \ln(\theta_2) - \frac{1}{2\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)^2 + (z_i - \theta_1)^2$$

Calculons ensuite l'espérance conditionnelle sachant θ^c de cette quantité selon Z .

$$\mathbb{E}[l((x, z), \theta)|\theta^c] = -n \ln(2\pi) - 2n \ln(\theta_2) - \frac{1}{2\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)^2 - \frac{1}{2\theta_2^2} \sum_{i=1}^n \mathbb{E}[(z_i - \theta_1)^2|\theta^c]$$

Pour $1 \leq i \leq n$, $\mathbb{E}[(z_i - \theta_1)^2|\theta^c] = \mathbb{E}[z_i^2|\theta^c] - 2\theta_1 \mathbb{E}[z_i|\theta^c] + \theta_1^2 = (\theta_1^{c^2} + \theta_2^{c^2}) - 2\theta_1 \theta_1^c + \theta_1^2$

On obtient donc l'expression suivante :

$$\mathbb{E}[l((x, z), \theta)|\theta^c] = -n \ln(2\pi) - 2n \ln(\theta_2) - \frac{1}{2\theta_2^2} \sum_{i=1}^n (x_i - \theta_1)^2 - \frac{n}{2\theta_2^2} (\theta_1^{c^2} + \theta_2^{c^2} - 2\theta_1 \theta_1^c + \theta_1^2)$$

On cherche θ permettant d'annuler le gradient de cette fonction. Les dérivées partielles s'écrivent :

$$\partial_{\theta_1} \mathbb{E}[l((x, z), \theta)|\theta^c] = \frac{1}{\theta_2^2} \sum_{i=1}^n x_i - \frac{n}{\theta_2^2} \theta_1 - \frac{n}{\theta_2^2} (\theta_1 - \theta_1^c)$$

Cette quantité s'annule pour $\tilde{\theta}_1 = \frac{1}{2n} (\sum_{i=1}^n x_i + n\theta_1^c)$.

La suite arithmético-géométrique définie ainsi est convergente car elle est de raison $\frac{1}{2}$. On note θ_1 sa limite.

On a

$$\begin{aligned} \theta_1 &= \frac{1}{2n} \left(\sum_{i=1}^n x_i + n\theta_1 \right) \\ \theta_1 &= \frac{1}{n} \sum_{i=1}^n x_i \end{aligned}$$

On retrouve donc l'estimateur de la moyenne empirique.

Par ailleurs,

$$\partial_{\theta_2} \mathbb{E}[l((x, z), \theta)|\theta^c] = -\frac{2n}{\theta_2^3} + \frac{1}{\theta_2^3} \sum_{i=1}^n (x_i - \theta_1)^2 + \frac{n}{\theta_2^3} (\theta_1^{c^2} + \theta_2^{c^2} - 2\theta_1 \theta_1^c + \theta_1^2)$$

Avec $\theta_1 = \tilde{\theta}_1$, cette quantité s'annule pour $\tilde{\theta}_2^2 = \frac{1}{2n} \left(\sum_{i=1}^n (x_i - \tilde{\theta}_1)^2 + n(\theta_1^{c^2} + \theta_2^{c^2} - 2\tilde{\theta}_1 \theta_1^c + \tilde{\theta}_1^2) \right)$

La convergence de la suite en θ_1 nous donne celle de la suite en θ_2 . On note θ_2 sa limite. On a

$$\begin{aligned} \theta_2^2 &= \frac{1}{2n} \left(\sum_{i=1}^n (x_i - \theta_1)^2 + n(\theta_1^2 + \theta_2^2 - 2\theta_1^2 + \theta_1^2) \right) \\ \theta_2^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \theta_1)^2 \end{aligned}$$

On retrouve donc l'estimateur de la variance empirique.

7.4.2 Cas général

Comme précédemment, on sait que tous les paramètres intervenant dans les équations suivent des lois log-normales, et donc que leur logarithme suit une loi normale. On peut alors simplifier l'implémentation de cet algorithme en dimension d .

On se donne un vecteur gaussien $X = {}^t(X_1, \dots, X_d)$ tel que $X \sim \mathcal{N}(\mu, \Sigma)$, où $\mu = \mathbb{E}(X)$ et $\Sigma = \text{Var}(X)$. Sa fonction de densité est donnée par

$$f(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(-\frac{{}^t(x-m)\Sigma^{-1}(x-m)}{2}\right)$$

Soit $(x_i)_{1 \leq i \leq n} \in (\mathbb{R}^d)^n$ des tirages de cette variable aléatoire.

La log-vraisemblance associée s'écrit

$$l(x, \theta) = -\frac{1}{2} \left(nd \ln(2\pi) + n \ln(\det(\Sigma)) + \sum_{i=1}^n {}^t(x_i - m)\Sigma^{-1}(x_i - m) \right)$$

On introduit une variable aléatoire Z de densité sachant X $f(Z|X, \theta)$. La log-vraisemblance complétée s'écrit alors :

$$\begin{aligned} l((x, z), \theta) &= l(x, \theta) + \sum_{i=1}^n \ln(f(z_i|x_i, \theta)) \\ &= - \left(nd \ln(2\pi) + n \ln(\det(\Sigma)) + \frac{1}{2} \sum_{i=1}^n {}^t(x_i - m)\Sigma^{-1}(x_i - m) + {}^t(z_i - m)\Sigma^{-1}(z_i - m) \right) \end{aligned}$$

Calculons ensuite l'espérance conditionnelle sachant θ^c de cette quantité selon Z .

$$\mathbb{E}[l((x, z), \theta)|\theta^c] = -nd \ln(2\pi) - n \ln(\det(\Sigma)) - \frac{1}{2} \sum_{i=1}^n {}^t(x_i - m)\Sigma^{-1}(x_i - m) - \frac{1}{2} \sum_{i=1}^n \mathbb{E}[{}^t(z_i - m)\Sigma^{-1}(z_i - m)|\theta^c]$$

Pour $1 \leq i \leq n$, on a :

$$\mathbb{E}[{}^t(z_i - m)\Sigma^{-1}(z_i - m)|\theta^c] = \mathbb{E}[{}^t z_i \Sigma^{-1} z_i | \theta^c] - 2\mathbb{E}[{}^t m \Sigma^{-1} z_i | \theta^c] + {}^t m \Sigma^{-1} m$$

Nous pouvons calculer chacune des expressions suivantes en développant les produits matriciels et en utilisant la matrice de covariance. En effet, on a en particulier $\mathbb{E}(z_{i,j} z_{i,k} | \theta^c) = \text{Cov}(z_{i,j}, z_{i,k}) - \mathbb{E}(z_{i,j})\mathbb{E}(z_{i,k})$.

Ainsi, en notant $\Sigma^{-1} = (a_{i,j})_{1 \leq i \leq n}$ et $\Sigma^c = (a_{i,j}^c)_{1 \leq i \leq n}$, on obtient alors :

$$\begin{aligned} \mathbb{E}({}^t z_i \Sigma^{-1} z_i | \theta^c) &= \sum_{j=1}^d \sum_{k=1}^d a_{j,k} a_{j,k}^c - \sum_{j=1}^d \sum_{k=1}^d m_j^c a_{j,k} m_k^c \\ &= \text{tr}(\Sigma^c \Sigma^{-1}) - {}^t m^c \Sigma^{-1} m^c \end{aligned}$$

Par ailleurs, nous avons de la même manière $\mathbb{E}({}^t m \Sigma^{-1} z_i | \theta^c) = {}^t m \Sigma^{-1} m^c$

On obtient donc l'expression suivante :

$$\begin{aligned} \mathbb{E}[l((x, z), \theta)|\theta^c] &= -nd \ln(2\pi) - n \ln(\det(\Sigma)) - \frac{1}{2} \sum_{i=1}^n {}^t(x_i - m)\Sigma^{-1}(x_i - m) \\ &\quad - \frac{1}{2} n \text{tr}(\Sigma^c \Sigma^{-1}) + \frac{1}{2} n {}^t m^c \Sigma^{-1} m^c - n {}^t m \Sigma^{-1} m^c + \frac{1}{2} n {}^t m \Sigma^{-1} m \end{aligned}$$

On cherche θ permettant de maximiser cette expression. Les différents gradients s'écrivent :

$$\nabla_m \mathbb{E}[l((x, z), \theta) | \theta^c] = - \sum_{i=1}^n \Sigma^{-1} (m - x_i) - n \Sigma^{-1} m + n \Sigma^{-1} m^c$$

Cette quantité s'annule pour $\hat{m} = \frac{1}{2n} (\sum_{i=1}^n x_i + n m^c)$

Par un raisonnement similaire à celui du cas en dimension 1, on montre que la suite ainsi définie converge vers l'estimateur de la moyenne empirique.

Par ailleurs, posons $\Lambda = \Sigma^{-1}$ et considérons le gradient de l'espérance par rapport à cette variable.

Pour rappel, nous avons

$$\begin{cases} \nabla_{\Lambda} \ln(\det(\Sigma)) = - \frac{\text{com}(\Lambda)}{\det(\Lambda)} \\ \nabla_{\Lambda} \text{tr}(\Sigma^c \Lambda) = \Sigma^c \\ \nabla_{\Lambda} {}^t m^c \Lambda^c = m^c {}^t m^c \end{cases}$$

Finalement, nous avons :

$$\nabla_{\Lambda} \mathbb{E}[l((x, z), \theta) | \theta^c] = \frac{n}{\det(\Lambda)} \text{com}(\Lambda) - \frac{1}{2} n m {}^t m + n m^c {}^t m - \frac{1}{2} n m^c {}^t m^c - \frac{1}{2} n \Sigma^c - \frac{1}{2} \sum_{i=1}^n (x_i - m)^t (x_i - m)$$

Par ailleurs, rappelons que ${}^t \left(\frac{\text{com}(\Lambda)}{\det(\Lambda)} \right) = \Lambda^{-1} = \Sigma$ et que Σ^c est symétrique. Par conséquent, cette quantité, évalué en \hat{m} , s'annule pour :

$$\hat{\Sigma} = \frac{1}{2n} \sum_{i=1}^n (x_i - \hat{m})^t (x_i - \hat{m}) + \frac{1}{2} (m^c {}^t m^c + \hat{m} {}^t \hat{m} - 2 m^c {}^t \hat{m} + \Sigma^c)$$

Par un raisonnement similaire à celui du cas en dimension 1, on montre que la suite ainsi définie converge vers l'estimateur de la matrice de covariance empirique.

Ces raisonnements nous ont permis d'implémenter directement à la main nos expectation maximization sous R dans notre cas particulier, mais nous aurions pu aussi utiliser la documentation déjà existante [11].

7.5 Régression non linéaire

Une régression non linéaire consiste à ajuster un modèle (en général non linéaire, mais pas nécessairement) de la forme $y = f_{a_1, \dots, a_m}(x)$ pour un ensemble de variables $(x_i, y_i)_{1 \leq i \leq m}$ pouvant être des scalaires ou des vecteurs.

Ainsi, une régression non linéaire revient à déterminer les paramètres a_1, \dots, a_m afin de minimiser $R = \|r\|$ où $r_i = y_i - f_{a_1, \dots, a_m}(x_i)$. La norme utilisée est souvent la norme euclidienne ou la norme ℓ_2 (méthode des moindres carrés dans ce cas).

La base de la démarche est similaire à celle d'une régression linéaire. En effet, R est une fonction des paramètres a_1, \dots, a_m pour un jeu de paramètres $(x_i, y_i)_{1 \leq i \leq m}$ fixé. Ainsi, si les dérivées existent, un minimum est donné par :

$$\forall j \in \llbracket 1, m \rrbracket, \frac{\partial R}{\partial a_j} = 0$$

Cela fournit donc un système de m équations non linéaires, qui est a priori difficile à résoudre analytiquement. Ainsi, plusieurs algorithmes peuvent être utilisés pour résoudre ce système.

Dans la suite, nous utilisons le package `nls` sous R pour l'ensemble des régressions non linéaires.

8 Modélisation informatique et estimations dans un cadre simplifié

Cette partie, qui correspond à nos premiers résultats, a été réalisée dans un cadre très simplifié. Elle se déroule donc des conditions expérimentales très éloignées de la réalité (pas de bruit, échantillonnage de l'ordre de la minute, ...). Nous avons néanmoins décidé de laisser les résultats obtenus pour témoigner de notre avancement chronologique et méthodologique dans ce projet.

8.1 Obtention des courbes pour les patients

Dans un premier temps, nous avons implémenté le système d'équations différentielles qui décrit notre modèle en langage R. Les simulations ont été réalisées pour une ou plusieurs injections de glucose et/ou de médicament selon les cas.

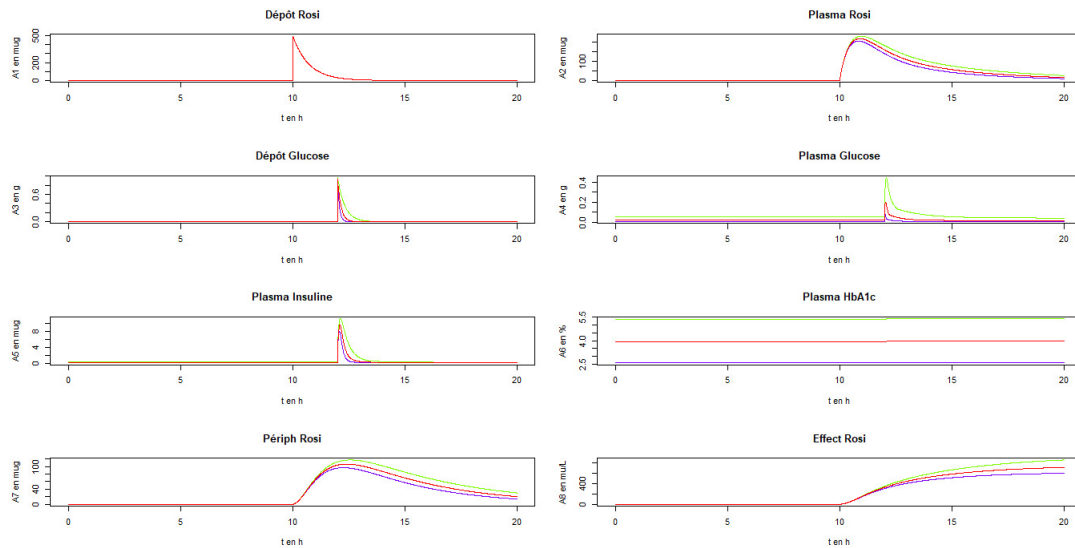


FIGURE 7: Injections de rosiglitazone et de glucose séparées de 2 heures

La figure 7 ci-dessus montre l'évolution de nos 8 compartiments d'étude en fonction du temps à la suite d'une injection de rosiglitazone et d'une injection de glucose toutes deux séparées de 2h pour 100 individus. Les courbes rouges représentent la médiane sur la population, les courbes bleues le minimum et les courbes vertes le maximum. Cela permet d'observer la tendance globale de l'évolution des compartiments d'effet sur la population considérée.

On observe une décroissance exponentielle du taux de glucose et d'insuline dans le sang sur un temps caractéristique avoisinant les 0.4 heures. Le plasma Rosi connaît quant à lui une décroissance beaucoup plus lente, de l'ordre de 5 heures.

L'hémoglobine glyquée a un comportement à première vue étonnant puisque l'on pourrait s'attendre à une décroissance de sa part. En théorie, sa valeur augmente lorsque les périodes d'hyperglycémie ont été fréquentes sur les derniers mois alors qu'elle baisse lorsque la glycémie a été correctement contrôlée sur cette période. Ici, l'écart inter-individuel est trop important et masque donc les variations intra-individuelles. Il serait préférable de tracer une courbe par individu pour observer la décroissance, plutôt que d'essayer d'observer une tendance globale.

La courbe Périph Rosi caractérisant les effets de la rosiglitazone sur les tissus environnants connaît la même forme de décroissance que le Plasma Rosi. Cependant, on observe un premier maximum beaucoup plus tard et une décroissance plus longue pour le Périph Rosi. Ce comportement s'explique par la lenteur des phénomènes mis en jeu. Il s'agit ici en effet de diffusion à travers des parois ce qui est beaucoup plus long que la simple ingestion de rosiglitazone.

Enfin la courbe Effect Rosi est assez difficile à interpréter. Elle est supposée être croissante lorsque le temps d'attente entre la présence de rosiglitazone et la chute de glycémie augmente. On pourrait donc justifier ici la croissance de la courbe sur la première partie.

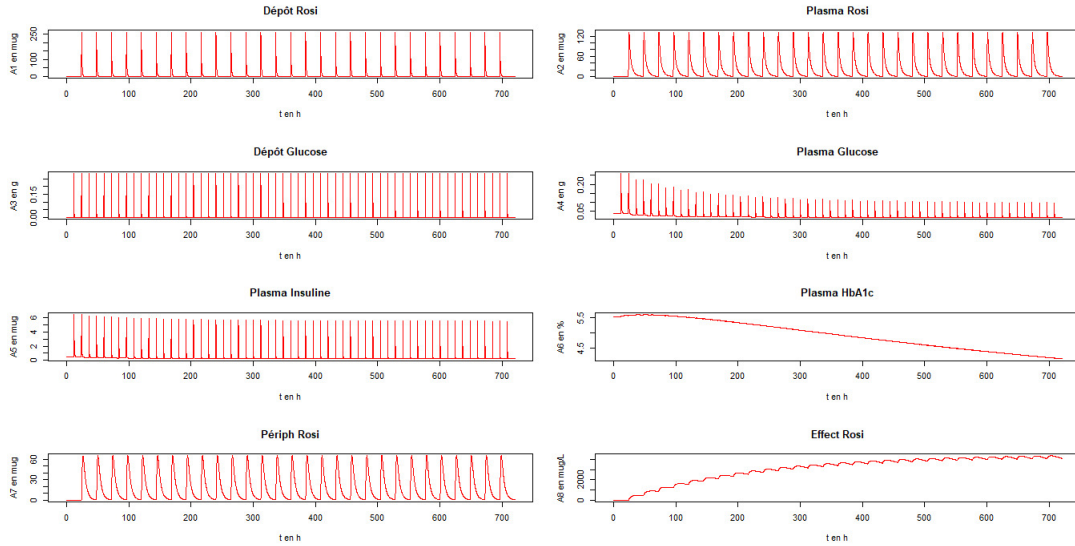


FIGURE 8: Succession d'injections de médicament toutes les 24 heures et de glucose toutes les 12 heures

La figure 8 représente toujours l'évolution temporelle de nos 8 compartiments mais cette fois-ci après une succession d'injections toutes identiques. Les injections de médicaments ont été séparées de 24h et celles de glucose de 12h. Ces choix de séparer ainsi les injections de glucose et de médicament ont été faits afin de se rapprocher au mieux du mode de vie de la plupart des individus. Les injections de médicaments sont, de plus, réglées de manière à ce que l'intégralité de la rosiglitazone injectée ait été éliminée du sang avant une nouvelle prise de celle-ci. Nous avons décidé ici de montrer les résultats uniquement pour un individu, sur un temps beaucoup plus long que précédemment (notamment pour voir les variations de l'hémoglobine glyquée).

On observe un comportement similaire sur les courbes de plasma Rosi et de Plasma Insuline par rapport au cas précédent. En revanche, si la forme de décroissance du plasma glucose reste la même après chaque injection (exponentielle), on observe une diminution significative des extremas de glucose après chaque injection malgré un taux d'insuline similaire. Cette diminution trouve son origine dans la présence de rosiglitazone dans le sang à la suite de chacune des injections. On observe donc le régime transitoire de l'organisme d'un régime sans médicament vers un régime avec traitement. La courbe de l'hémoglobine glyquée illustre bien ce phénomène puisqu'elle décroît lentement au fur à mesure de l'installation du nouveau régime permanent, la glycémie étant bien régulée. Enfin, la courbe de l'effect Rosi modélise cette fois-ci un retard qui semble saturer au fur à mesure que le nouveau régime permanent s'installe.

8.2 Importance de la variabilité inter-individuelle

A partir du modèle implémenté précédemment, nous avons pu simuler la réponse d'une cohorte de 100 patients, choisis aléatoirement, à une injection de glucose et de rosiglitazone selon les conditions expérimentales représentées dans la figure 7. Nous avons stocké la valeur des différents paramètres, constants et variables, de la population.

Dans un second temps, en supposant inconnus ces paramètres, nous avons simulé un échantillonnage sur les 8 compartiments d'effet de chaque patient. Nous avons ensuite pu procéder à l'estimation paramétrique.

Il est intéressant de noter que nous avons deux types de paramètres : des paramètres constants au sein de la population et des paramètres variables (à savoir $K_{a_{glc}}$, $Slope_{ins}$, V_4 et CL). Ces derniers paramètres correspondent à l'inter-individualité. Ils représentent en effet les différences entre individus, liées à l'âge ou au poids par exemple.

Il peut être intéressant de visualiser l'impact de ces paramètres variables sur les patients [12]. Par exemple, nous pouvons nous intéresser au temps d'élimination (ou dose maximale et minimale) de différentes substances dans le sang pour différentes personnes. Pour les mêmes conditions que citées précédemment, nous avons décidé de tracer la concentration dans le sang de glucose et d'insuline pour 12 patients, comme illustré ci-dessous :

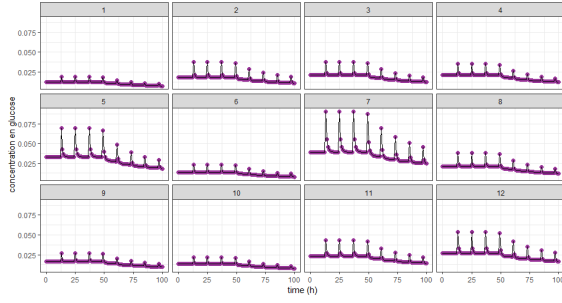


FIGURE 9: Évolution de la concentration de glucose dans le sang selon les patients

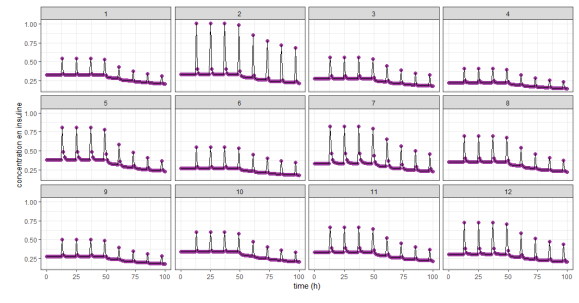


FIGURE 10: Évolution de la concentration d'insuline dans le sang selon les patients

Ainsi, nous pouvons remarquer que les concentrations suivent globalement les mêmes évolutions pour les 12 patients (augmentation pendant les phases d'ingestion et diminution ensuite). Néanmoins, on peut noter une grande diversité de réponses selon les patients. En particulier, les valeurs maximales atteintes sont très différentes d'une personne à l'autre tout comme les temps caractéristiques des variations des compartiments d'effet.

8.3 Présentation des résultats dans un cadre linéaire

Dans un premier temps, nous avons appliqué la méthode de régression linéaires multiples, décrite dans la partie 6.1, afin d'estimer les paramètres constants du modèle PK/PD. Nous avons appliqué cette méthode à chacun des 100 patients de la cohorte avant de moyenner les estimateurs des paramètres constants ainsi obtenus.

La précision relative obtenue dépendait grandement de la période d'échantillonnage. Notre approche, basée sur l'approximation des dérivées par méthode d'Euler, donnait des résultats dans l'ensemble satisfaisants (*cf.* tableau de la figure 11) mais perdait grandement en précision pour une période d'échantillonnage supérieure à quelques minutes (correspondant à des conditions réelles d'échantillonnage).

	Paramètres_estimés	Paramètres_réels	Erreurs_relatives
Ka	1.3900000	1.390000	0.000001636
K72	0.6409823	0.640898	0.013196190
K27	0.5039212	0.503922	0.000078665
Ke0	0.0047247	0.004720	0.098625014
V2	0.0051050	0.005100	0.097282933
KinGlc	0.0122478	0.017000	27.954262025
Slr	0.0005465	0.001032	47.048839502
KoutIns	13.6338721	13.700000	0.482685332
KinIns	0.5565951	0.558000	0.251773283
Slope_gut	378.7126961	381.000000	0.600342243
KoutHbA1C	0.0011999	0.001200	0.004583621

FIGURE 11: Évaluation et erreurs relatives (en %) des paramètres constants, pour une période d'échantillonnage d'une seconde

Notre méthode était ainsi très précise pour tous les paramètres, à l'exception de deux d'entre eux : K_{inGlc} et S_{LR} . Cela était vraisemblablement dû au fait que ces deux paramètres soient estimés par les ordonnées à l'origine des régressions linéaires, qui sont plus fortement impactées que les coefficients directeurs par l'approximation des dérivées par la méthode d'Euler.

Nous avons ensuite cherché à évaluer les moyennes et écart-types des paramètres variables du modèle, à partir de la méthode décrite en 6.2. La précision de nos estimations diminuait, là aussi, grandement avec le pas d'échantillonnage et augmentait avec la taille de la cohorte. Certains paramètres n'étaient pas directement identifiables, comme montré dans l'ensemble de la partie 6. Nous avons donc introduit des nouveaux paramètres, identifiables, définis à partir de produits des paramètres initiaux.

	Paramètres_estimés	Paramètres_réels	Erreurs_relatives
Ka_Glc_Theta	7.287225	6.95000	4.8522
Ka_Glc_var	0.086987	0.09233	5.7868
CL_Theta	0.003040	0.00301	0.9962
CL_var	0.008064	0.00849	5.0170
SlopeGlcV4_mu	6.084612	5.77425	5.3750
SlopeGlcV4_var	0.096875	0.11122	12.8970
SlopeInsKoutGlcV5_mu	0.718205	0.78906	8.9797
SlopeInsKoutGlcV5_var	0.197198	0.20877	5.5416
KinHbA1cSLGHV4_mu	-2.267426	-1.58641	42.9281
KinHbA1cSLGHV4_var	0.094703	0.11122	14.8502

FIGURE 12: Évaluation et erreurs relatives (en %) sur les constantes caractéristiques des paramètres variables, pour une période d'échantillonnage d'une seconde

Bien que les erreurs soient plus élevées pour ces paramètres (figure 12) que pour les paramètres constants (figure 11), notre modèle demeure globalement assez précis. Une partie de l'erreur peut-être expliquée par la taille de la cohorte qui était de 100 patients uniquement. Il serait préférable de prendre en compte plus de patients, mais cela rendrait la complexité temporelle trop élevée.

9 Résultats dans des conditions expérimentales plus réalistes

9.1 Cadre expérimental utilisé

Dans la situation idéale décrite dans la partie précédente, le modèle linéaire que nous avons implémenté était particulièrement efficace. Cependant, il ne permet d'obtenir des estimations satisfaisantes que dans ce cadre peu réaliste. En effet, les conditions de prélèvement ne sont pas réalisables. En particulier, nous effectuons des prélèvements sur les patients toutes les secondes, pendant une durée longue (plusieurs heures), ce qui n'est pas applicable à un individu. De plus, la méthode d'estimation linéaire ne résistait pas à un bruit important (au delà de 0,2 %). Enfin, nous considérons que tous les compartiments d'étude étaient mesurables chez un individu.

L'utilisation du modèle idéal n'est donc pas compatible avec les conditions expérimentales réelles que nous souhaitons à terme modéliser. Dans un second temps, nous avons donc essayé de nous rapprocher le plus possible des conditions expérimentales réelles. Celles-ci correspondent à une période d'échantillonnage d'une heure, et d'un bruit de l'ordre de 5 % en n'utilisant que les 4 compartiments d'études mesurables : les concentrations plasmatiques en glucose, insuline, hémoglobine glyquée et rosiglitazone.

L'utilisation d'uniquement 4 des 8 compartiments d'études implique l'impossibilité d'estimer certains paramètres du modèle (par exemple, K_a n'apparaît que dans l'équation de A_1). De plus, cela rend l'estimation des autres paramètres encore plus difficile. Pour ces raisons, nous avons pour l'instant décidé de conserver l'hypothèse selon laquelle les 8 compartiments sont mesurables.

Dans la suite, nous considérons une cohorte constituée de 200 patients, parmi lesquels 100 vont suivre le traitement à base de rosiglitazone et les 100 autres serviront de témoins à l'expérience. Nous considérerons aussi un temps d'échantillonnage plus long, avec un prélèvement par heure. Nous nous plaçons de plus dans les conditions illustrées par la figure suivante qui correspondent au cadre de l'OGTT. Nous injectons donc du glucose toutes les 12h à l'ensemble des patients, et des injections de rosiglitazone toutes les 12h aux patients traités.

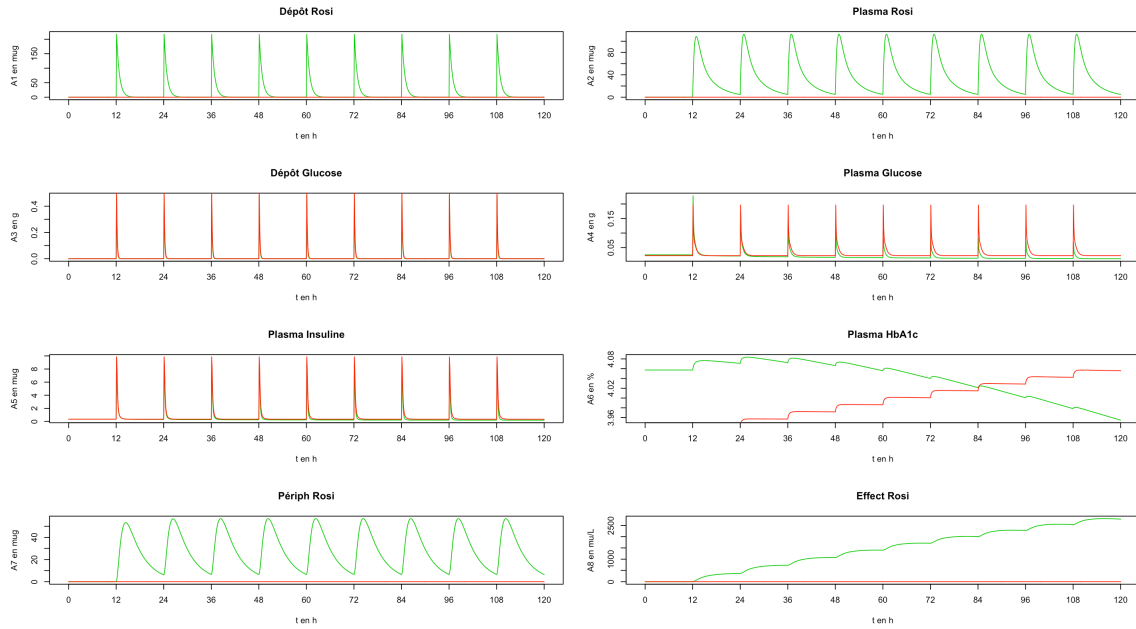


FIGURE 13: Conditions expérimentales

Sur le graphique précédent, les courbes vertes correspondent à un patient traité à la rosiglitazone et les courbes rouges à un patient témoin. On remarque que la concentration en hémoglobine glyquée du patient traité à la rosiglitazone tend à diminuer, contrairement à celle patient témoin. Cela souligne le rôle de la rosiglitazone dans la diminution de la concentration en hémoglobine glyquée dans notre modèle d'équations.

9.2 Estimation des paramètres constants

Afin d'estimer les paramètres constants dans ces nouvelles conditions expérimentales, nous avons eu recours à une méthode d'optimisation non-linéaire. Nous devons tout d'abord initialiser chaque paramètre du modèle. Nous avons pour cela utilisé deux méthodes, la première consistant à initialiser chaque paramètre avec son ordre de grandeur, et la seconde à appliquer l'algorithme de régression linéaire présenté dans les parties précédentes, qui ne nécessite pas d'initialisation.

Nous construisons ensuite une fonction qui, à un jeu de paramètres estimés, associe $\hat{A}_1, \dots, \hat{A}_8$ la valeurs estimées des compartiments d'effet.

Le but de notre algorithme est alors de minimiser la distance entre les valeurs réelles des compartiments d'effet $A_1 \dots A_8$ et les valeurs estimées $\hat{A}_1, \dots, \hat{A}_8$ en jouant sur les paramètres estimés. Cette méthode est déjà implémentée en R, dans le package `nls`.

En appliquant l'algorithme à un seul patient, nous avons obtenu les résultats suivants :

	Paramètres_estimés	Paramètres_réels	Erreurs_relatives
Ka	1.390003e+00	1.39000000	2.369236e-04
K72	6.408982e-01	0.64089776	6.689597e-05
K27	5.039219e-01	0.50392157	7.260035e-05
K20	6.326255e-01	0.63262546	9.877072e-07
Ke0	4.719996e-03	0.00472000	8.549294e-05
V2	5.099996e-03	0.00510000	7.214705e-05
est_K_a_Glc	7.542875e+00	7.54286639	1.125251e-04
S_LR	1.032557e-03	0.00103200	5.394386e-02
Slope_ins_K_out_glc_sur_V5	1.249823e+00	1.25045229	5.034861e-02
K_in_glc	1.699118e-02	0.01700000	5.190945e-02
Slope_glc_sur_V4	2.413186e+02	248.80782098	3.010048e+00
Slope_gut	3.638271e+02	381.00000000	4.507330e+00
K_inIns	5.562449e-01	0.55800000	3.145409e-01
K_outIns	1.329179e+01	13.70000000	2.979611e+00
BSLGLC	3.467257e-02	0.03467257	0.000000e+00
K_inHbA1C_SLGH_sur_V_4	1.563288e-01	0.12348449	2.659794e+01
est_K_outHbA1C	1.185903e-03	0.00120000	1.174751e+00

FIGURE 14: Paramètres estimés par la méthode non linéaire, pour un pas d'échantillonnage de 1h, pour un seul patient.

Ces résultats sont globalement satisfaisants. Les erreurs plus élevées sur certains paramètres sont dues à la propagation d'incertitudes de proche en proche, l'estimation de certains paramètres nécessitant de simuler des équations à partir de paramètres précédemment estimés.

De manière plus qualitative, nous avons aussi tracé, pour les différents paramètres, les courbes estimées que nous avons confrontées aux mesures réelles. Nous voyons que les erreurs sur les paramètres ne semblent pas compromettre l'interprétabilité des résultats. Cela se confirme en étudiant les erreurs quadratiques moyenne entre les \hat{A}_i estimés et les A_i réels, qui est très faible. Cela montre les limites de l'identifiabilité empirique du modèle.

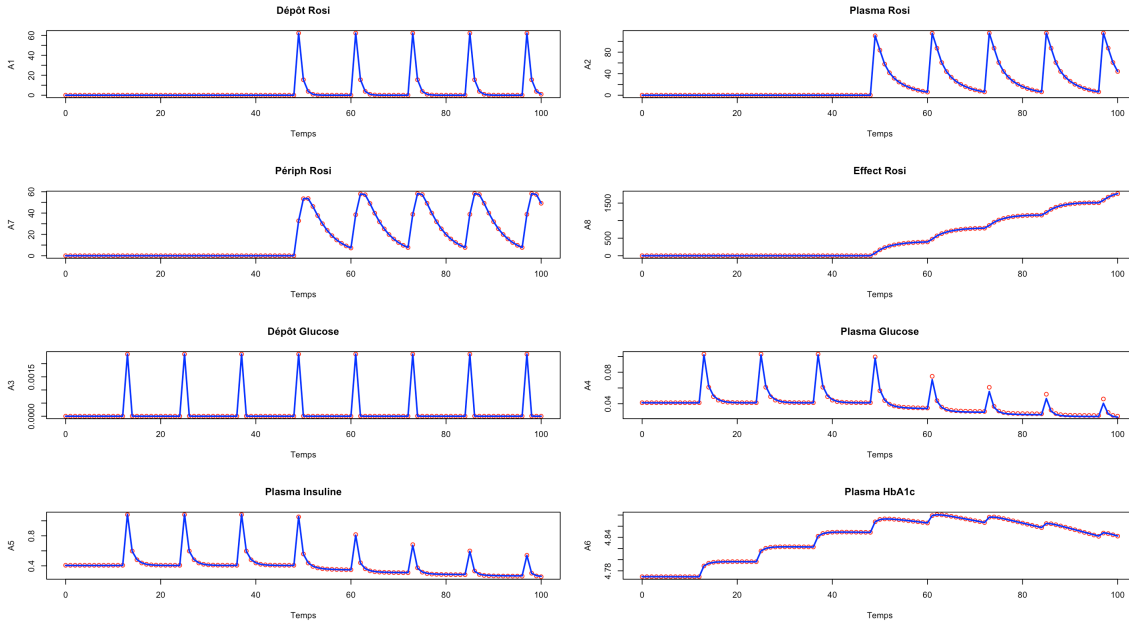


FIGURE 15: Paramètres estimés par la méthode non linéaire, pour un pas d'échantillonnage de 1h

Afin de renforcer la robustesse de notre modèle, nous avons par la suite implémenté plusieurs étapes de régression non-linéaires successives, en actualisant à chaque étape les valeurs initiales des paramètres. Nous avons essayé plusieurs schémas pour ces régressions, et nous avons finalement retenu le schéma suivant :

1. Estimation non-linéaire des paramètres sur un seul patient. Cela permet d'obtenir une première estimation des paramètres, donc de bonnes valeurs initiales pour l'étape suivante, qui vont permettre la convergence de l'algorithme.

2. Estimation non linéaire sur l'ensemble des patients des paramètres, supposés tous constants. Pour cela nous avons découplé le système d'équation en estimant tout d'abord les paramètres liés à A_1 , puis à A_2 et A_7 , puis A_8 , A_3 , puis A_4 et A_5 et enfin A_6 . Cela permet d'estimer les paramètres même quand leur valeurs initiales sont éloignées de leurs valeurs réelles. L'estimation sur l'ensemble des patients simultanément permet de moyenner certaines erreurs liées à l'échantillonnage ou, comme nous le verrons plus tard, au bruitage.
3. Estimation non linéaire des paramètres sur l'ensemble des patients sur les huit équations simultanément. Cet algorithme converge grâce à l'étape précédente qui permet d'affiner les valeurs initiales des paramètres. Néanmoins, cette étape dégrade parfois les résultats, ce qui nous a conduit à l'abandonner, en particulier dans le cas non-bruité.
4. Estimation des paramètres variables et constants, patient par patient, pour les huit équations simultanément. Cette dernière étape permet de prendre en compte la variabilité inter-individuelle, et fournit une estimation relativement robuste des valeurs des paramètres variables pour chaque patient

Cette méthode améliore considérablement les résultats et permet de réduire la propagation d'incertitude de proche en proche entre les paramètres. Nous avons ainsi obtenu les résultats suivants :

	Paramètres_estimés	Paramètres_réels	Erreurs_relatives
Ka	1.390003	1.390000	0.00023645
K72	0.640898	0.640898	0.00006527
K27	0.503922	0.503922	0.00007503
Ke0	0.004720	0.004720	0.00002514
V2	0.005100	0.005100	0.00001624
S_LR	0.001039	0.001032	0.63298560
K_in_glc	0.017147	0.017000	0.86624088
Slope_gut	354.795762	381.000000	6.87775271
K_inIns	0.583702	0.558000	4.60605302
K_outIns	13.805343	13.700000	0.76892902
est_K_outHbA1C	0.001221	0.001200	1.72313234

FIGURE 16: Estimation finale des paramètres constants après les 4 étapes de régression non linéaire.

9.3 Estimation des paramètres variables

Pour l'estimation des paramètres variables, nous avons eu recours à plusieurs méthodes.

Dans un premier temps, nous avons implémenté les estimateurs empiriques des moyennes et des variances d'une loi log-normale, précédemment calculés par maximum de vraisemblance, à partir des valeurs estimées pour chaque patient des paramètres variables.

Par la suite, nous avons implémenté la méthode d'expectation maximization décrite dans les parties précédentes. Comme nous l'avons étudié lors de l'étude théorique de cette méthode dans le cadre de notre problème, les estimateurs obtenus convergent vers les estimateurs empiriques.

	Paramètres_estimés	Paramètres_réels	Erreurs_relatives
Ka_Glc_Theta	6.844932	6.95000	1.5118
CL_Theta	0.002991	0.00301	0.6254
SlopeGlcV4_mu	5.710723	5.77425	1.1002
SlopeInsKoutGLcV5_mu	0.727530	0.78906	7.7979
KinHbA1cSLGHV4_mu	-1.605751	-1.58641	1.2192
Ka_Glc_var	0.113708	0.09233	23.1538
CL_var	0.010606	0.00849	24.9221
SlopeGlcV4_var	0.091033	0.11122	18.1502
SlopeInsKoutGLcV5_var	0.204731	0.20877	1.9332
KinHbA1cSLGHV4_var	0.100813	0.11122	9.3564

FIGURE 17: Estimation finale des paramètres variables par la méthode d'expectation maximization.

Les valeurs estimées sont plutôt satisfaisantes, les paramètres de moyennes sont globalement bien estimés. En raison de la plus grande incertitude associée au calcul de la variance empirique, les valeurs de ces dernières sont moins satisfaisantes, mais demeurent correctes.

9.4 Bruitage des résultats

Les résultats obtenus dans la partie précédente sont idéaux, ils ne représentent pas les résultats de mesure réels que l'on pourrait obtenir. Afin de traduire les imprécisions dues à la mesure, on ajoute aux données un bruit gaussien.

9.4.1 Bruit global

Dans un premier temps, on a ajouté un bruit suivant une loi $\mathcal{N}(0, \sigma^2)$ aux différents compartiments avec un écart-type σ fixé.

Cela correspond à l'utilisation d'un appareil de mesure unique pour mesurer les valeurs. Le paramètre σ étant exprimé comme un certain pourcentage de la moyenne.

Cependant, en raison de l'étendue des valeurs, ce bruit rendait les calculs trop imprécis et a été abandonné.

9.4.2 Bruit local

Dans un second temps, nous avons considéré que les appareils de mesure pouvaient être calibrés en fonction de la grandeur de la mesure et que le modèle de l'erreur de mesure unique n'était pas assez réaliste.

Nous avons donc choisi que le bruit de mesure devait dépendre uniquement de la valeur locale que l'on cherche à bruite pour rendre compte de la possibilité de calibrer. On a donc remplacé chaque valeur y par la valeur $\hat{y} = y(1 + \epsilon)$ avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$ où σ est ici un pourcentage.

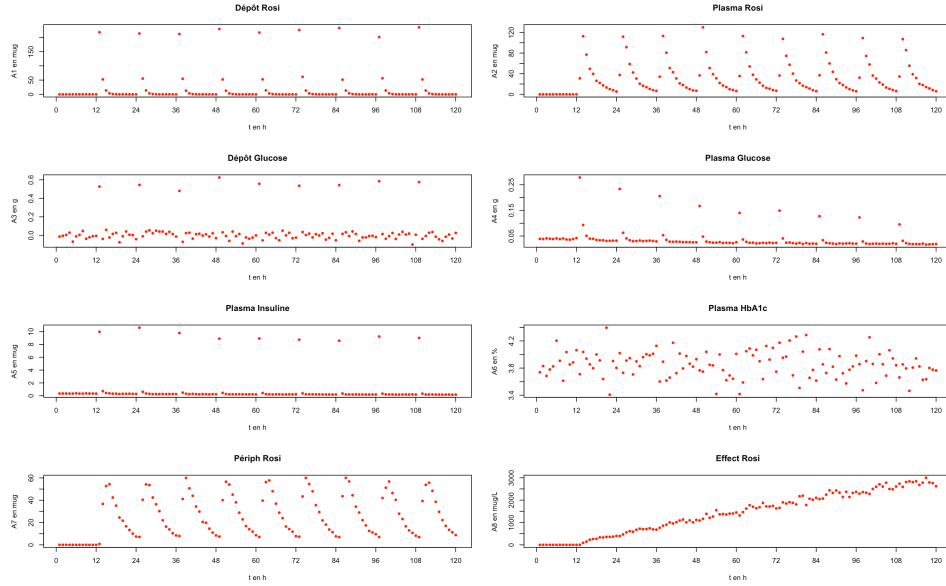


FIGURE 18: Représentation des courbes obtenues après bruitage local des données

Les résultats obtenus en bruitant de cette manière sont bien plus précis.

Pour implémenter un bruit de mesure plus proche de la réalité, il faudrait utiliser un modèle hybride. La valeur de l'écart-type serait imposée en fonction de la plage de valeur à laquelle appartient la grandeur que l'on cherche à mesurer.

9.4.3 Résultats sur les paramètres constants

Nous nous sommes placés dans le cadre d'un bruit local. Grâce à la méthode décrite dans la partie précédente, nous avons pu obtenir une bonne robustesse à des valeurs de bruits de l'ordre de 5%. On obtient ainsi les estimations suivantes :

	Paramètres_estimés	Paramètres_réels	Erreurs_relatives
Ka	5.840314e-01	1.3900000	57.983352
K72	6.690452e-01	0.6408978	4.391878
K27	4.894421e-01	0.5039216	2.873356
Ke0	4.919604e-03	0.0047200	4.228908
V2	5.653633e-03	0.0051000	10.855547
S_LR	1.867162e-03	0.0010320	80.926559
K_in_glc	6.398118e-02	0.0170000	276.359904
Slope_gut	1.786990e+03	381.0000000	369.026377
K_inIns	1.284999e-01	0.5580000	76.971341
K_outIns	1.128353e+01	13.7000000	17.638475
est_K_outHbA1C	-1.731272e-04	0.0012000	114.427264

FIGURE 19: Estimation des paramètres constants après bruitage local à 5% des données

On remarque que la plupart des estimations sont dégradées par rapport au modèle sans bruit, néanmoins ces dernières restent correctes compte tenu de l'importance du bruitage. Les paramètres dont les estimations sont les moins bonnes sont K_{inGlc} et $Slope_{Gut}$, cela est sans doute lié au fait qu'ils apparaissent dans le terme non linéaire des équations de A_5 et de A_6 et sont par conséquent plus difficiles à estimer.

9.4.4 Résultats sur les paramètres variables

Notre modèle d'estimation des paramètres variables étant moins robuste, nous nous attendons à avoir une plus forte dégradation des estimations. Nous avons obtenu les résultats suivants :

	Paramètres_estimés	Paramètres_réels	Erreurs_relatives
Ka_Glc_Theta	1.022694	6.95000	85.285
CL_Theta	0.003092	0.00301	2.730
SlopeGlcV4_mu	5.133076	5.77425	11.104
SlopeInsKoutGLcV5_mu	1.068387	0.78906	35.400
KinHbA1cSLGHV4_mu	-0.297111	-1.58641	81.271
Ka_Glc_var	0.019026	0.09233	79.393
CL_var	0.007275	0.00849	14.305
SlopeGlcV4_var	0.104339	0.11122	6.186
SlopeInsKoutGLcV5_var	0.308784	0.20877	47.908
KinHbA1cSLGHV4_var	0.629315	0.11122	465.834

FIGURE 20: Estimation des paramètres variables après bruitage local à 5% des données

Nous remarquons là aussi que les résultats restent satisfaisants. En dépit du bruitage. La seule valeur aberrante est la variance du dernier paramètre dont la valeur est très éloignée de la valeur théorique, sans doute à cause d'une propagation des incertitudes.

10 Détermination d'un schéma d'administration

10.1 Protocole expérimental et résultats

On fait l'hypothèse qu'un traitement contre le diabète est efficace lorsqu'il conduit à une baisse importante de la concentration d'hémoglobine glyquée dans le sang.

En se basant sur cette hypothèse, nous avons mis en place le protocole suivant que nous avons ensuite testé sur notre modèle.

Pour un patient donné, nous simulons pendant un certain temps une prise régulière de glucose sans rosiglitazone. Ainsi, on constate que la concentration d'hémoglobine glyquée croît jusqu'à atteindre un seuil. Cette valeur de seuil nous servira à simuler une condition initiale plus réaliste pour le patient.

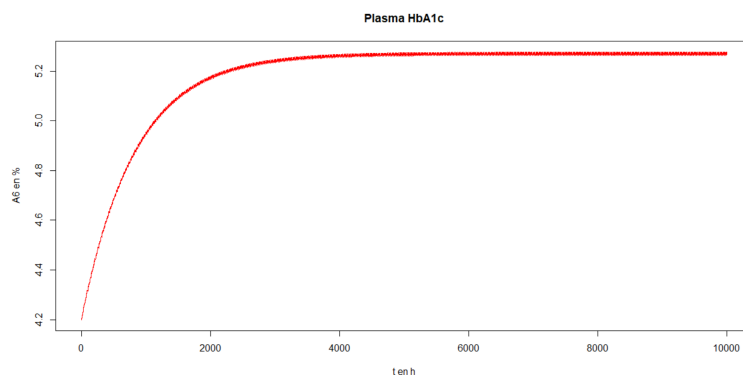


FIGURE 21: Simulation du comportement quotidien d'un patient

Connaissant la dose de rosiglitazone à prendre en une journée, nous avons décidé de la diviser en deux sous-doses à prendre à deux moments pouvant éventuellement être égaux.

Nous simulons ensuite 250 jours de traitement pour tous les différents types de traitement possibles (temps nécessaire pour atteindre un nouveau seuil) et regardons ensuite quel est celui qui a le plus fait descendre la concentration moyenne d'hémoglobine glyquée. On stocke les résultats dans une matrice M tel que $M_{i,j}$ est la concentration moyenne d'hémoglobine glyquée dans le sang à la fin de l'expérience si le patient prenait une demi-dose de rosiglitazone quotidiennement aux heures i et j . Les premières lignes et les premières colonnes de cette matrice sont représentées sur la figure ci-dessous.

	V1	V2	V3	V4	V5	V6	V7
1	2.135455	2.135654	2.136176	2.136879	2.136720	2.136380	2.136089
2	2.135654	2.135868	2.136406	2.137122	2.136965	2.136623	2.136330
3	2.136176	2.136406	2.136967	2.137704	2.137553	2.137211	2.136916
4	2.136879	2.137122	2.137704	2.138470	2.138329	2.137989	2.137693
5	2.136720	2.136965	2.137553	2.138329	2.138202	2.137872	2.137581
6	2.136380	2.136623	2.137211	2.137989	2.137872	2.137555	2.137274
7	2.136089	2.136330	2.136916	2.137693	2.137581	2.137274	2.137007

FIGURE 22: Concentration d'hémoglobine glyquée à l'issue de l'expérience

D'un point de vue statistique, les simulations nous permettent d'obtenir que le traitement optimal pour un patient revient à prendre une seule dose de rosiglitazone trois heures avant le petit-déjeuner.

Cependant, d'un point de vue biologique, les valeurs maximale et minimale obtenues présentent assez peu de différence (dans l'exemple du patient illustré ici, il s'agit de 2.133247 et 2.138483 là où la concentration seuil sans traitement est de 5.269522). La figure ci-dessous présente la décroissance de la concentration d'hémoglobine glyquée dans le sang dans le cas du "meilleur" et du "pire" traitement.

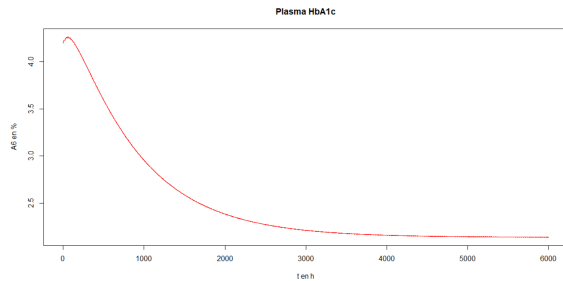


FIGURE 23: Décroissance de la concentration en hémoglobine glyquée dans le sang avec le traitement le moins performant

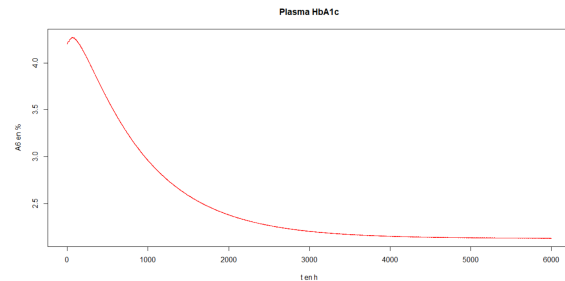


FIGURE 24: Décroissance de la concentration en hémoglobine glyquée dans le sang avec le traitement le plus performant

10.2 Limitations du modèle

Rappelons que notre étude portait sur des souris et non sur des êtres humains. Or, les travaux biologiques sont réalisés sur des modèles d'animaux qui sont génétiquement modifiés pour reproduire une pathologie (sachant que ce n'est pas réellement la pathologie). Ici, les souris ont des prédispositions pour l'obésité et le diabète liées à des modifications génétiques. De plus, la manière dont une souris est alimentée n'est pas la même que l'homme. Pour toutes ces raisons, il y a des restrictions importantes à la transposition du modèle chez la souris à celui chez l'homme. Bien sûr, tout un ensemble de données est facilement transposable chez l'homme (notamment sur la pharmacocinétique). En effet, l'effet du médicament est équivalent chez la souris et chez l'homme. Néanmoins, la question des repas ne peut pas se transposer raisonnablement de la souris à l'homme.

Ainsi, habituellement, l'étude est réalisée directement chez l'homme. Dans le schéma expérimental réel, on caractérise notre pharmacocinétique chez l'homme puis on compare les schémas d'administration à d'autres en extrapolant les résultats. Ainsi, en réalité, les recommandations de posologie chez l'homme ne sont pas tirés d'un modèle de souris diabétique.

Nous avons cependant décidé de conserver notre étude car elle est conceptuellement proche de ce qui pourrait se faire chez l'homme.

Il faudrait à ce stade essayer de comprendre les différences entre souris/homme et voir ce qui est pertinent à mettre en place comme expérience (en accord avec des diabétologue, pharmaciens, experts des projets,...) pour pouvoir proposer des modèles construits sur les meilleures données.

11 Rappel des résultats majeurs de l'analyse de sensibilité

Dans certains modèles complexes, comme dans le notre, l'influence des paramètres d'entrées sur la variable de sortie n'est pas simple à déterminer. Cela rend difficile le tri des paramètres entre ceux qui sont indispensables et ceux qui sont moins utiles.

Grégoire Martin, avant de partir en S8, avait pu travailler sur l'analyse de sensibilité sur notre modèle, notamment pour les 4 paramètres variables Ka_{glc} , V_4 , CL et $Slope_{ins}$. Nous rappelons ici les résultats principaux, directement tirés de son rapport final.

Dans notre modèle, il existe 3 causes principales d'incertitudes au niveau de nos entrées :

- *Les incertitudes de mesure* : Elles caractérisent la précision avec laquelle on parvient à mesurer les paramètres biologiques de notre modèle. Pour modéliser cette incertitude, nous avons décidé de bruitez les résultats en partie 9.4
- *Les incertitudes de population* : Chaque individu possède des paramètres biologiques qui lui sont propres, comme illustré en partie 8.2. Ainsi, on modélise cette différence par une variable aléatoire suivant une loi log-normale dont les paramètres ont été fixés empiriquement.
- *Les incertitudes sur le modèle* : Elles sont liées à une compréhension partielle des phénomènes complexes qui agissent chez un individu. Elles n'ont pas été prises en compte dans ce projet.

Son analyse de sensibilité, à partir des indices de Sobol, a permis de montrer l'influence assez bien partagée des 3 paramètres $Slope_{ins}$, Ka_{glc} et V_4 . Ces trois derniers jouent chacun un rôle prédominant dans la variance des 3 plasmas considérés (glucose, insuline et $HbA1c$).

Le paramètre CL n'a en revanche pas pu bien être analysé. Il nécessite en effet une étude sur un intervalle de temps beaucoup plus grand au vu de la lenteur des phénomènes physiques mis en jeu, que ni Grégoire Martin ni nous n'avons pu mettre en place, pour des raisons de temps ou de complexité temporelle sous R.

Les paramètres Ka_{glc} et V_4 jouent des rôles quasi-symétriques pour les plasmas glucose et insuline, il est donc difficile de juger lequel est le plus important. $Slope_{ins}$ se détache un peu du lot puisqu'il joue un rôle prépondérant dans les 3 plasmas bien qu'il ne s'agisse pas toujours du paramètre le plus influent.

Ainsi, pour une analyse au repos, les paramètres à prioriser seraient :

- Plasma glucose : V_4 et $Slope_{ins}$.
- Plasma insuline : Ka_{glc} et $Slope_{ins}$.
- Plasma hémoglobine glyquée : $Slope_{ins}$.

On pourrait éventuellement simplifier le modèle en considérant CL comme constant si l'on s'intéresse à des échelles de temps courtes.

12 Moyens utilisés par l'équipe

12.1 Répartition des tâches

Nous avons fait le choix de travailler la plupart du temps tous les 3 et de répartir les tâches selon les besoins et les compétences de chacun. Nous travaillions sur tous les créneaux de projet alloués dans notre emploi du temps et nous retrouvions parfois dans la semaine s'il nous était impossible de faire cela à distance. Pour chaque grande étape de notre projet, chacun d'entre nous était responsable d'une sous-étape de celle-ci.

12.2 Organisation de l'équipe

Pour communiquer sur l'avancée du projet, nous utilisions majoritairement WhatsApp. Les codes étaient, eux, envoyés par mail. Nous avons utilisé un document partagé Overleaf pour pouvoir écrire notre rapport.

Dans la mesure du possible et compte tenu des besoins rencontrés, nous essayions de communiquer avec le client au minimum toutes les 3 semaines sur Skype Entreprise, en utilisant des slides partagées pour illustrer nos propos.

Nous avons également essayé de conserver un lien fréquent avec les professeurs, en communiquant par mail ou en faisant des réunions.

12.3 Problèmes rencontrés

Nous avons globalement rencontré trois problèmes majeurs :

- Difficultés mathématiques. Notamment, nous avons eu beaucoup de mal à comprendre comment fonctionner l'algorithme d'expectation maximization et surtout saisir comment nous pouvions l'adapter à notre modèle. Nous avons essayé de pallier ce problème en communiquant avec les professeurs encadrants et en faisant un grand nombre de recherches.
- Difficultés liées à la complexité temporelle de nos algorithmes. Il nous est souvent arrivé qu'un programme tourne pendant plusieurs heures pour obtenir des résultats, nous faisant perdre du temps précieux. Nous essayons le plus possible de faire tourner nos programmes pendant des heures de pauses pour répondre à ce problème. Cela n'était pour autant pas toujours envisageable, compte tenu de l'importance des tests.
- Difficultés liées au langage R, que nous ne maîtrisions pas avant ce projet. Pour palier ce problème, nous avons grandement utilisé la documentation R [13].

13 Conclusion

13.1 Résultats obtenus

Au cours de ce projet, nous avons pu mettre en évidence le fait que le modèle PK/PD qui nous a été donné est identifiable en théorie, mais ne l'est plus forcément dans la pratique.

Nous avons mis en place différentes méthodes pour déterminer les paramètres du modèle en modifiant les données auxquelles on se donnait accès afin de se rapprocher de conditions réelles. Les différents codes implémentant ces méthodes sont donnés en annexe.

Nous avons également établi un protocole afin de déterminer pour chaque patient quel serait un traitement optimal.

13.2 Axes de poursuite

Pour poursuivre ce projet, on pourrait dans un premier temps reprendre les modèles d'estimation des paramètres en limitant les données que l'on prend en compte aux seules données observables (celles des quatre plasmas).

Il faudrait ensuite améliorer le protocole de détermination du schéma d'administration pour le rendre plus conforme aux conditions expérimentales réelles.

Enfin, lorsque ces modèles théoriques seront bien en place et adaptés aux conditions réelles, on pourrait utiliser des données expérimentales réelles afin de voir si l'on retrouve les paramètres initiaux du modèle.

Références

- [1] HAS et ANSM. Stratégie médicamenteuse du contrôle glycémique du diabète de type 2, janvier 2013. https://www.has-sante.fr/upload/docs/application/pdf/2013-02/10irp04_argu_diabete_type2.pdf.
- [2] Cornelia Landersdorfer and William Jusko. Pharmacokinetic/pharmacodynamic modelling in diabetes mellitus. *School of Pharmacy and Pharmaceutical Sciences, University at Buffalo*, 2008.
- [3] Antoine Perasso. Identifiabilité de paramètres pour des systèmes décrits par des équations aux dérivées partielles. application à la dynamique des populations. *Université Paris Sud - Paris XI*, 2009.
- [4] Veronique Letort Le Chevalier and al. *Modélisation : Représentations et analyse des modèles*. Cours de 1ère année à CentraleSupélec, 2019.
- [5] Arthur Tenenhaus. *ST4 - Big Data et Santé*. Cours de ST4 à CentraleSupélec, 2019.
- [6] J Gaudart, R Giorgi, JC Thalabard, D Thiam, and S Whegang. Modèles linéaires à effets mixtes, Septembre 2010. <https://docplayer.fr/60771944-Modeles-lineaires-a-effets-mixtes-j-gaudart-r-giorgi-jc-thalabard-d-thiam-s-whegang.html>.
- [7] Fragkiskos Malliaros and Maria Vakalopoulou. *SG6 - Machine Learning*. Cours de SG6 à CentraleSupélec, 2020.
- [8] Andrew NG. Cs229 lecture notes - expectation maximization, 2011. <http://cs229.stanford.edu/notes/cs229-notes8.pdf>.
- [9] Guzy S. Bauer R.J. *Monte Carlo Parametric Expectation Maximization (MC-PEM) Method for Analyzing Population Pharmacokinetic/Pharmacodynamic Data*. Springer, Boston, MA, 2004.
- [10] Wikipedia. Définition de l'algorithme expectation maximization, 2020. https://fr.wikipedia.org/wiki/Algorithme_expectation_maximization.
- [11] Marc Lavielle. *Statistics in Action with R*. Inria Saclay (Xpop) Ecole Polytechnique (CMAP), 2017.
- [12] Marc Lavielle. Visualisation de l'inter-individualité, 2017. <http://sia.webpopix.org/nlme.html>.
- [13] Documentation du langage R.