UNIVERSITE DE MONTPELLIER 1

Rapport de DEA Biostatistique

Modèles semi–Markoviens Place dans la gestion des maladies chroniques

par

Yohann FOUCHER

sous la direction de

Jean–Pierre DAURES

Organisme d'accueil :
L'**I**nstitut **U**niversitaire de **R**echerche **C**linique

Soutenu le 25 juin 2004

Remerciements

Mes premiers remerciements reviennent au Professeur DAURES pour m'avoir permis de réaliser ce stage. Merci pour ses conseils et son encadrement.

Merci à toute l'équipe de l'IURC, en particulier à Eve et Philippe pour leur collaboration à ce projet.

Enfin, je ne veux pas oublier ma famille et mes proches qui m'ont toujours soutenus et suivis dans mes choix.

Table des matières

Introduction **3**

1 Modèles Markoviens à temps continu **5**
1.1 Définition du modèle 5
1.2 Homogénéité et temps de séjour dans l'état 7

2 Modèles semi–Markoviens **9**
2.1 Définitions 9
2.2 Probabilités de transition du processus semi – Markovien 12
2.3 Fonction de Vraisemblance 13
2.4 Introduction de covariables 15
2.4.1 Modèle à risques semi–proportionnels 15
2.4.2 Probabilité de survie dans l'état 17
2.5 Loi de Weibull comme loi de séjour dans l'état 18
2.6 Généralisations du modèle 18
2.6.1 Loi de Weibull généralisée 18
2.6.2 Choix d'une loi de temps de séjour spécifique à chaque transition . 19

3 Application au VIH **22**
3.1 Définition du modèle 22
3.2 Recueil des données 23
3.3 Stratégie de modélisation 23
3.4 Résultats 25
3.4.1 Description des données 25
3.4.2 Modèle semi–Markovien de type Weibull 26
3.4.3 Modèle semi–Markovien de type Weibull généralisé 28

Conclusions et perspectives **30**

Annexes **32**
Annexe 1 : Vraisemblance basée sur les fonctions de risque 32
Annexe 2 : Fonctions associées à la loi exponentielle 33
Annexe 3 : Modèle semi–Markovien stratifié de type Weibull 34
Annexe 4 : Modèle semi–Markovien stratifié de type Weibull généralisé 42
Annexe 5 : Modèle semi–Markovien multivarié de type Weibull 50
Annexe 6 : Modélisation semi–Markovienne de type Weibull généralisé 51

Bibliographie **53**

Introduction

En médecine, et plus particulièrement pour les pathologies chroniques, les modèles Markoviens connaissent un intérêt grandissant. Ils permettent d'étendre les modèles de survie classiques et d'analyser les processus multi–états. En effet, l'évolution clinique d'un patient ne se résume pas forcément à deux états, vivant et mort par exemple. En cancérologie [1], une fois le malade pris en charge, il peut rester en rémission ou bien rechuter. A partir de ces deux états cliniques, il existe une possibilité de décès. Ce modèle stochastique correspond à une réalité clinique et permet donc une approche détaillée de l'évolution de la maladie. Pour le VIH (Virus de l'Immunodéficience Humaine) [2, 3, 4, 5, 6] ou pour l'asthme [7], ce type de modèle a aussi été récemment appliqué avec succès. De plus, ces méthodes permettent d'étudier des dynamiques complexes et possèdent donc un intérêt accentué dans l'exploration de bases de données observationnelles.

Cependant, la majorité des applications utilisent les outils Markoviens homogènes ayant l'inconvénient d'être sans mémoire, où l'évolution du processus est indépendante du temps déjà passé dans l'état actuel. Nous aborderons cette propriété dans la première partie de ce document, consacrée aux caractéristiques des processus markoviens à temps continu et à espace d'états discret.

Dans le domaine clinique, cette contrainte est souvent trop forte. Les modèles Markoviens à renouvellement ou modèles semi–Markoviens constituent dans ce contexte un outil intéressant, puisqu'ils intègrent la notion de temps de séjour dans le calcul des forces de transition. La loi du temps de séjour dans l'état est alors explicite. La seconde partie sera consacrée à cette généralisation de la propriété Markovienne. Nous nous baserons sur les articles de Perez [8] et de Dabrowska [9].

Le troisième chapitre consistera en l'application et l'interprétation de ce type de modèle à l'évolution des patients séropositifs. Cette dernière étape permettra de mieux comprendre l'intérêt des modélisations semi–Markoviennes, grâce aux interprétations cliniques qui en sont issues. Nous pourrons alors mesurer concrètement l'apport des généralisations précédentes. Cette application est d'autant plus intéressante qu'il s'agit d'une problématique complexe, où l'hypothèse de forces de transition constantes au cours du temps est a priori abusive. En effet, il est probable qu'un individu stable, c'est à dire qui a déjà passé un temps important dans l'état, soit d'autant plus stable dans son évolution future. Nous

utiliserons pour cette application, la cohorte NADIS, suivie au CHU de Nice.

Ce mémoire a pour objectif principal d'expliciter l'approche semi–Markovienne, en développant la théorie et en l'appliquant à des données réelles. Cependant, nous tenterons d'offrir une approche plus générale et plus parcimonieuse que celle définie par Perez [8]. Pour ce faire, nous suivrons trois voix différentes, ne modifiant en rien la théorie générale, mais offrant une originalité à ce travail. Tout d'abord, la taille du vecteur de covariables influençant les vitesses de transition pourra être propre à chaque transition. Autrement dit, les covariables pourront avoir un effet sur certaines transitions, mais pas forcément sur toutes. Ensuite, nous définirons une modélisation plus générale en introduisant une distribution des temps de séjour de type Weibull généralisé. Enfin, nous tenterons d'adopter une approche plus flexible en laissant la possibilité de définir différentes formes de distribution par transition.

Chapitre 1

Modèles Markoviens à temps continu

Dans ce chapitre, nous nous intéressons aux processus Markoviens à temps continu et à espace d'états discret, tels que ceux développés dans le livre de Karlin et Taylor [10]. L'objectif de cette première partie est d'introduire la notion de processus et de mettre en évidence l'apport des modèles semi–Markoviens.

1.1 Définition du modèle

Formellement, nous étudions une famille de variables aléatoires : $\{X(t); 0 \leq t < \infty\}$ où les valeurs prises par $X(t)$ sont des entiers positifs appartenant à l'ensemble $E = \{1, 2, ..., r\}$. La propriété Markovienne résume le passé du processus à l'état précédent, autrement dit pour $t_0 < t_1 < t_1 < ... < t_n < t_{n+1}$ et $\forall i, j, k, h \in E$, nous avons :

$$P(X(t_{n+1}) = j | X(t_n) = i, X(t_{n-1}) = k, ..., X(t_0) = h) = P(X(t_{n+1}) = j | X(t_n) = i) \tag{1.1}$$

Pour simplifier, nous adopterons l'écriture suivante :

$$P_{ij}(t, t+s) = P(X(t+s) = j | X(t) = i)$$

Classiquement, la propriété suivante doit être respectée :

$$\sum_j P_{ij}(t, t+s) = 1$$

d'où

$$P_{ii}(t, t+s) = 1 - \sum_{j \neq i} P_{ij}(t, t+s) \tag{1.2}$$

Autrement dit, soit le processus reste dans le même état, soit il transite vers un autre état. Sous forme matricielle, ces probabilités de transition peuvent être notées :

$$\mathbf{P}(t, t+s) = \begin{pmatrix} P_{11}(t, t+s) & P_{12}(t, t+s) & \cdots & P_{1r}(t, t+s) \\ P_{21}(t, t+s) & P_{22}(t, t+s) & \cdots & P_{2r}(t, t+s) \\ \vdots & \vdots & \ddots & \vdots \\ P_{r1}(t, t+s) & P_{r2}(t, t+s) & \cdots & P_{rr}(t, t+s) \end{pmatrix} = \Big(P_{ij}(t+s)\Big)_{i,j=1,...,r}$$

A partir de la propriété Markovienne (1.1), nous pouvons écrire, $\forall t, s > 0$:

$$\begin{aligned} P_{ij}(0, t+s) &= P(X(t+s) = j | X(0) = i) \\ &= \sum_k P(X(t+s) = j, X(t) = k | X(0) = i) \\ &= \sum_k P(X(t+s) = j | X(t) = k, X(0) = i) P(X(t) = k | X(0) = i) \\ &= \sum_k P(X(t+s) = j | X(t) = k) P(X(t) = k | X(0) = i) \\ &= \sum_k P_{ik}(0, t) P_{kj}(t, t+s) \end{aligned}$$

En reprenant la notation matricielle précédente, il est équivalent d'écrire :

$$\mathbf{P}(0, t+s) = \mathbf{P}(0, t)\mathbf{P}(t, t+s) \tag{1.3}$$

Cette relation est appelée *équation de Chapman–Kolmogorov.*

Le paramètre d'intérêt en analyse de survie est la force de transition (ou fonction de risque instantané), α_{ij}, $i, j \in E$. Celle–ci peut être définie, pour $i \neq j$, comme suit :

$$\begin{aligned} \alpha_{ij}(t) &= lim_{h \to 0} \frac{P(X(t+h) = j | X(t) = i)}{h} \\ &= lim_{h \to 0} \frac{P_{ij}(t, t+h)}{h} \end{aligned} \tag{1.4}$$

Notons que $\alpha_{ij}(t) \times h$ représente la "probabilité" que le processus passe dans l'état j entre t et $t+h$, conditionnellement au fait que ce processus soit dans l'état i en t. $\alpha_{ij}(t)$ constitue donc la vitesse de transition de i vers j au temps t. Pour $i = j$, $\alpha_{ii}(t)$ est défini à partir de la contrainte (1.2) :

$$\sum_{j \neq i} P(X(t+h) = j | X(t) = i) = 1 - P(X(t+h) = i | X(t) = i)$$

d'où

$$\sum_{j \neq i} \frac{P(X(t+h) = j | X(t) = i)}{h} = \frac{1 - P(X(t+h) = i | X(t) = i)}{h}$$

En définissant :

$$lim_{h \to 0} \frac{1 - P(X(t+h) = i | X(t) = i)}{h} = -\alpha_{ii}(t) \tag{1.5}$$

Nous obtenons alors :

$$\sum_{j \neq i} \alpha_{ij}(t) = -\alpha_{ii}(t)$$

et

$$\sum_j \alpha_{ij}(t) = 0$$

1.2 Homogénéité et temps de séjour dans l'état

Dans les applications, le processus Markovien est, le plus souvent, considéré *homogène*. Les probabilités de transition sont alors définies par :

$$P_{ij}(t, t+s) = P_{ij}(0, s) = P_{ij}(s) \tag{1.6}$$

$P_{ij}(s)$ est indépendant de t, $\forall t \geq 0$. L'équation de Chapman–Kolmogorov (1.3) peut alors s'écrire :

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$$

donc

$$\begin{aligned}
\frac{d\mathbf{P}(s)}{ds} &= lim_{h\to 0}\frac{\mathbf{P}(s+h) - \mathbf{P}(s)}{h} \\
&= lim_{h\to 0}\frac{\mathbf{P}(s)\mathbf{P}(h) - \mathbf{P}(s)}{h} \\
&= lim_{h\to 0}\frac{\mathbf{P}(s)(\mathbf{P}(h) - \mathbf{I})}{h} \\
&= \mathbf{P}(s)\ lim_{h\to 0}(\frac{\mathbf{P}(h) - I}{h}) \\
&= \mathbf{P}(s)\mathbf{Q}
\end{aligned} \tag{1.7}$$

avec $\mathbf{I}$, la matrice identité, et où d'après les définitions (1.4) et (1.5), la matrice $\mathbf{Q}$ s'écrit :

$$\mathbf{Q} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1r} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{rr} \end{pmatrix} = \Big(Q_{ij}(t+s)\Big)_{i,j=1,\ldots,r}$$

Remarquons que les forces de transition ne dépendent pas du temps. L'équation différentielle (1.7) admet la solution :

$$\mathbf{P}(t) = exp(\mathbf{Q}t) \tag{1.8}$$

avec comme contraintes $P_{ii}(0)$ égale à 1 et $P_{ij}(0)$ égale à 0. Autrement dit, $\mathbf{P}(0)$ est la matrice identité. Le calcul des termes diagonaux de la matrice $\mathbf{P}(t)$ est assez direct. En effet, d'après la définition (1.6) :

$$\begin{aligned}
P_{ii}(t+u) &= P_{ii}(t)P_{ii}(u) \\
&= P_{ii}(u)(1 - \sum_{j\neq i} P_{ij}(t))
\end{aligned}$$

Développons cette propriété :

$$\begin{aligned}
&P_{ii}(t+u) - P_{ii}(u) = P_{ii}(u)(1 - \sum_{j\neq i} P_{ij}(t)) - P_{ii}(u) \\
&\iff P_{ii}(t+u) - P_{ii}(u) = P_{ii}(u)((1 - \sum_{j\neq i} P_{ij}(t)) - 1) \\
&\iff \frac{P_{ii}(t+u) - P_{ii}(u)}{t} = -P_{ii}(u)\frac{\sum_{j\neq i} P_{ij}(t)}{t}
\end{aligned}$$

Or d'après la relation (1.5), nous obtenons :

$$lim_{t\to 0}\frac{P_{ii}(t+u)-P_{ii}(u)}{t} = -\alpha_{ii}P_{ii}(u)$$
$$\Longleftrightarrow \frac{dP_{ii}(u)}{du} = -\alpha_{ii}P_{ii}(u) \tag{1.9}$$

où $\frac{dP_{ii}(u)}{du}$ est la dérivée de $P_{ii}(u)$ par rapport à u. La probabilité $P_{ii}(u)$, que le processus reste dans l'état i dans $[0, u]$, doit donc satisfaire l'équation différentielle (1.9). Une solution est facilement identifiable : $P_{ii}(u) = c\, exp(-\alpha_{ii}u)$, c étant une constante. Or, le processus ne peut pas changer d'état pendant un intervalle de temps nul. Il faut donc prendre en compte la contrainte $P_{ii}(0) = 1$. Ceci implique directement $c = 1$. La distribution du temps d'attente dans l'état i est donc définie par la loi exponentielle (détails en annexe 2) :

$$\begin{aligned} P_{ii}(u) &= exp(-\alpha_{ii}u) \\ &= exp(-Q_{ii}u) \end{aligned} \tag{1.10}$$

où α_{ii} ne dépend pas du temps. Ces distributions des temps de séjour, données par la diagonale de la matrice $\mathbf{P}(t)$, sont dites sans mémoire (la force de mortalité est constante au cours du temps). Dans l'étude du vivant, cette hypothèse ne correspond pas souvent à la réalité.

Beaucoup de situations nécessitent une fonction de risque évoluant avec le temps de séjour dans l'état. Classiquement, un phénomène d'usure est illustré par une augmentation de la force de mortalité au cours du temps passé dans l'état. Choisissons par exemple la mortalité en fonction de l'âge en Île de France (figure 1.1). Cette problématique sanitaire, où le temps est défini comme l'âge, est encore plus complexe. En effet, le risque de mortalité diminue puis augmente. Cette forme est dite en U (*U-shape* en anglais).

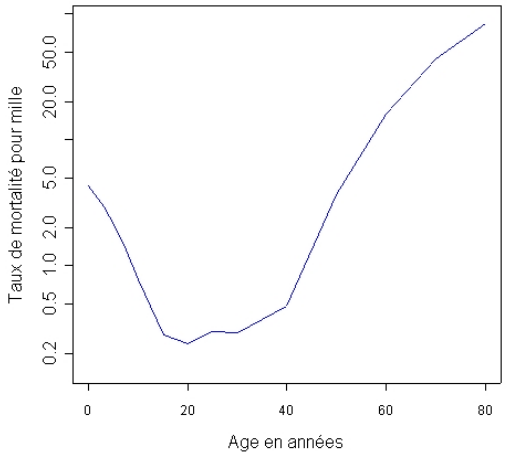

FIG. 1.1 – Taux de mortalité en Île de France entre 1993–1995.

Chapitre 2

Modèles semi–Markoviens

L'intérêt de ce type de modèles est contenu dans le choix explicite de la distribution du temps de séjour dans l'état. La probabilité de rester dans un état peut alors dépendre de la durée déjà passée dans cet état. Ce modèle est présenté dans deux articles dont s'inspire ce document : Dabrowska [9] et Perez [8]. Nous nous fonderons sur les notations issues des processus de comptage, définies par Gill [12]. Elles ont l'avantage d'être explicites et d'être transversales à de nombreuses problématiques Markoviennes.

2.1 Définitions

A des temps différents, le processus occupe des états bien définis. En l'absence de covariable, on observe pour chaque individu le couple $(T, X) = \{(T_n, X_n) : n \geq 0\}$, où $0 = T_0 < T_1 < ... < T_n$ sont les temps consécutifs d'entrée dans les états $X_0, X_1, ..., X_n \in E$, avec $X_{p+1} \neq X_p, \forall p \geq 0$. n représente le numéro de la transition. Pour faire le lien avec les processus de comptage, nous noterons pour une seule réalisation du processus (un individu) :

$$\widetilde{N}_{ij}(t) = \sum_{n \geq 1} I\{T_n \leq t, X_n = j, X_{n-1} = i\} \qquad \forall i \neq j$$

où $\widetilde{N}_{ij}(t)$ représente le nombre de transitions directes $i \to j$ observées dans l'intervalle de temps $[0, t]$. Naturellement, $\widetilde{N}_{ij}(0) = 0$. $\widetilde{N}_{ij}(t)$ est fini, composé de valeurs continues à droite avec des sauts de $+1$ (impossibilité que deux processus sautent au même temps). Ce processus est dit *cadlag*. De plus, nous posons :

$$\widetilde{N}(t) = \sum_{i,j} \widetilde{N}_{ij}(t)$$

$\widetilde{N}(t)$ représente le nombre total de transitions observées dans $[0, t]$. Ainsi, l'état occupé par le processus au temps t, noté $X(t)$ dans la première partie, sera maintenant noté $X_{\widetilde{N}(t)}$. Les séquences $X = \{X_n, n \geq 0\}$ forment une chaîne de Markov sous–jacente. Les probabilités de transition $i \to j$ associées à cette chaîne, notées P_{ij}, sont définies par :

$$P_{ij} = P(X_{n+1} = j | X_n = i)$$

– Si l'état i n'est pas un état absorbant, alors :

$$\begin{cases} P_{ij} \geq 0 \text{ si } i \neq j \\ P_{ij} = 0 \text{ si } i = j \end{cases} \tag{2.1}$$

– Sinon, si l'état i est un état absorbant, alors :

$$\begin{cases} P_{ij} = 0 \text{ si } i \neq j \\ P_{ij} = 1 \text{ si } i = j \end{cases} \tag{2.2}$$

Pour les développements qui vont suivre, nous supposerons que, pour toute transition $i \to j$, l'état i n'est pas absorbant. Cette chaîne de Markov sous–jacente ne gère pas le temps, mais les séquences des états indicées par le numéro de la transition. Les temps d'attente dans les états (ou temps de séjour) sont définis explicitement. Le processus (T, X) est dit semi–Markovien si la distribution des temps de séjour $(T_{n+1} - T_n)$ satisfait la condition suivante :

$$P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_0, T_0, X_1, ..., X_n, T_n) = P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n) \tag{2.3}$$

Sachant la séquence des états X, les temps de séjour $T_1, T_2 - T_1, T_3 - T_2, ...$ sont indépendants et leurs distributions dépendent uniquement des états contigus. Remarquons que le numéro de la transition n'a pas d'importance dans la définition des lois des temps de séjour. Le processus est donc homogène sur le temps chronologique. Afin de faire le lien avec l'analyse de survie, nous noterons :

– la fonction de répartition :

$$F_{ij}(x) = P(T_{n+1} - T_n \leq x | X_{n+1} = j, X_n = i) \tag{2.4}$$

– la fonction de survie :

$$S_{ij}(x) = 1 - F_{ij}(x) = P(T_{n+1} - T_n > x | X_{n+1} = j, X_n = i) \tag{2.5}$$

– la fonction de densité :

$$f_{ij}(x) = lim_{t \to 0^+} \frac{P(x < T_{n+1} - T_n < x + t | X_{n+1} = j, X_n = i)}{t} \tag{2.6}$$

– la fonction de risque :

$$\lambda_{ij}(x) = lim_{h \to 0^+} \frac{P(x < T_{n+1} - T_n < x + h | T_{n+1} - T_n \geq x, X_{n+1} = j, X_n = i)}{h} \tag{2.7}$$

D'après le théorème de Bayes et les définitions (2.2) (2.4), nous pouvons préciser la condition (2.3) définissant les modèles semi–Markoviens. Pour $i \neq j$:

$$\begin{aligned} &P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n = i) \\ &\quad = P(T_{n+1} - T_n \leq x | X_{n+1} = j, X_n = i)\ P(X_{n+1} = j | X_n = i) \\ &\quad = F_{ij}(x)\ P_{ij} \end{aligned} \tag{2.8}$$

Dans la pratique, à un temps d'observation quelconque du processus, seul l'historique de celui–ci est connu. L'état dans lequel va passer le processus est incertain. Il est donc intéressant de définir un temps d'attente marginal, c'est à dire moyenné sur l'état suivant. Par le théorème des probabilités totales, et en reprenant la relation (2.8) :

$$\begin{aligned} F_{i.}(x) &= P(T_{n+1} - T_n \leq x | X_n = i) \\ &= \sum_j P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n = i) \\ &= \sum_j F_{ij}(x) P_{ij} \\ &= \sum_{j \neq i} F_{ij}(x) P_{ij} \ \text{ (puisque } P_{ii} = 0) \end{aligned}$$

Il en découle directement les relations suivantes. La fonction de survie marginale :

$$S_{i.}(x) = 1 - F_{i.}(x) = \sum_{j \neq i} S_{ij}(x) P_{ij} \tag{2.9}$$

La fonction de densité marginale :

$$\begin{aligned} f_{i.}(x) &= \frac{dF_{i.}(x)}{dx} \\ &= \frac{d(\sum_{j \neq i} F_{ij}(x)\ P_{ij})}{dx} \\ &= \sum_{j \neq i} P_{ij} \frac{dF_{ij}(x)}{dx} \\ &= \sum_{j \neq i} P_{ij} f_{ij}(x) \end{aligned}$$

Par définition, la fonction de risque instantané de i vers j, du processus semi–Markovien, correspond à la probabilité du processus à transiter instantanément vers l'état j, sachant qu'il est dans l'état i depuis une durée x :

$$\begin{aligned} \alpha_{ij}(x) &= \lim_{h \to 0} \frac{P[x \leq T_{n+1} - T_n < x + h, X_{n+1} = j | T_{n+1} - T_n \geq x, X_n = i]}{h} \\ &= \lim_{h \to 0} \frac{P[x \leq T_{n+1} - T_n < x + h, X_{n+1} = j | X_n = i]}{h\ P[T_{n+1} - T_n \geq x | X_n = i]} \\ &= \frac{1}{P[T_{n+1} - T_n \geq x | X_n = i]} \\ &\quad \times \lim_{h \to 0} \frac{P[x \leq T_{n+1} - T_n < x + h, X_{n+1} = j | X_n = i]}{h} \\ &= \frac{1}{P[T_{n+1} - T_n \geq x | X_n = i]} \\ &\quad \times \lim_{h \to 0} \frac{P[x \leq T_{n+1} - T_n < x + h | X_{n+1} = j, X_n = i] P[X_{n+1} = j | X_n = i]}{h} \\ &= \frac{P[X_{n+1} = j | X_n = i]}{P[T_{n+1} - T_n \geq x | X_n = i]} \\ &\quad \times \lim_{h \to 0} \frac{P[x \leq T_{n+1} - T_n < x + h | X_{n+1} = j, X_n = i]}{h} \end{aligned}$$

d'où

$$\alpha_{ij}(x) = \frac{P_{ij} f_{ij}(x)}{S_{i.}(x)} \text{ avec } \begin{cases} i \neq j \\ i, j \in E \\ \alpha_{ii}(x) = -\sum_{j \neq i} \alpha_{ij}(x) \end{cases} \tag{2.10}$$

Cette formulation du risque instantané est importante pour comprendre l'apport de la théorie semi–Markovienne. La force de changement d'état, $\alpha_{ij}(x)$, est d'autant plus grande que :

- la probabilité de transition entre i et j de la chaîne de Markov sous–jacente, P_{ij}, est grande ;
- la fonction de densité, $f_{ij}(x)$, est grande ;

– la fonction de survie marginale dans l'état i, $S_{i.}(x)$, est faible. Ceci équivaut à un temps déjà passé dans l'état i grand.

Cette fonction de risque du processus semi–Markovien, $\alpha_{ij}(x)$, ne doit pas être confondue avec la fonction de risque de la loi explicite des temps de séjour, $\lambda_{ij}(x)$, définie en (2.7). A partir de (2.10), nous pouvons écrire :

$$\begin{aligned}\sum_{j\neq i}\alpha_{ij}(t) &= \sum_{j\neq i}\frac{P_{ij}f_{ij}(t)}{S_{i.}(t)}\\ &= \frac{1}{S_{i.}(t)}\sum_{j\neq i}P_{ij}f_{ij}(t)\\ &= \frac{1}{S_{i.}(t)}f_{i.}(t)\\ &= \alpha_{i.}(t)\end{aligned} \tag{2.11}$$

où $\alpha_{i.}(t)$ représente bien la fonction de risque marginale sur j (j étant l'état contigu à droite).

2.2 Probabilités de transition du processus semi – Markovien

Dans la section précédente, nous avons défini P_{ij}, $i \neq j$, comme la probabilité de transition de i vers j du processus de Markov sous–jacent X. X est indicé par le numéro de la transition, mais ne gère pas le temps T d'apparition des transitions. Le processus semi–Markovien étant défini par le couple (X, T), nous avons choisi de noter Z le processus semi–Markovien, tel que $Z(t) = X_{\widetilde{N}(t)}$. Intéressons nous maintenant aux probabilités de transition du processus semi–Markovien Z en reprenant la notion d'homogénéité précédemment abordée :

$$\begin{aligned}p_{ij}(l, l+t) &= P[Z(l+t) = j | Z(l) = i]\\ &= P[X_{\widetilde{N}(l+t)} = j | X_{\widetilde{N}(l)} = i]\\ &= P[X_{\widetilde{N}(t)} = j | X_{\widetilde{N}(0)} = i]\\ &= P[Z(t) = j | Z(0) = i]\\ &= p_{ij}(t) \qquad i, j \in E \text{ et } x \geq 0\end{aligned} \tag{2.12}$$

La propriété d'homogénéité contenue dans le processus Markovien sous–jacent ($P_{ij} = cte$) est donc transmise au processus semi–Markovien sous la forme définie ci–dessus.

Pour calculer ces probabilités, considérons tout d'abord qu'au moins une transition se produise dans l'intervalle $[0, t]$. Ce conditionnement sur le premier événement est particulièrement utilisé dans la théorie du renouvellement [10]. Supposons, de plus, que cette première transition soit de i vers k, $k \in E$. Alors, la probabilité que cette transition ait lieu au temps de séjour x, s'écrit d'après la définition (2.10) :

$$\alpha_{ik}(x)S_{i.}(x) = P_{ik}f_{ik}(x) \tag{2.13}$$

En respectant la définition d'homogénéité sur le temps chronologique (2.12) et en notant que le premier état k est apparu au temps x, la probabilité que le processus soit égal à j au temps t $(t > x)$, s'écrit directement :

$$\begin{aligned} P[Z(t) = j | Z(x) = k] &= P[Z(t-x) = j | Z(0) = k] \\ &= p_{kj}(t-x) \end{aligned} \tag{2.14}$$

Le produit de convolution des deux probabilités (2.13) et (2.14), permet de déterminer la probabilité jointe que $Z(t) = j$ sachant que le premier nouvel état est k et que l'état initial est i :

$$P[Z(t) = j | X_0 = i, X_1 = k, (T_1 - T_0) \leq t] = \int_0^t P_{ik} f_{ik}(x) p_{kj}(t-x) dx$$

Il s'ensuit que, par sommation sur l'espace d'état, la probabilité que $Z(t) = j$, sachant que $Z(0) = i$ et qu'au moins une transition ait lieu dans $[0, t]$, s'écrit :

$$P[Z(t) = j | X_0 = i, (T_1 - T_0) \leq t] = \sum_{k=1}^{r} \int_0^t P_{ik} f_{ik}(x) p_{kj}(t-x) dx \tag{2.15}$$

Enfin, pour pouvoir déterminer $P[Z(t) = j | Z(0) = i]$, il faut noter que dans le cas où $i = j$, nous devons ajouter à la relation (2.15) la possibilité qu'aucun événement ne se produise dans $[0, t]$. En reprenant le résultat (2.9), cette probabilité est égale à :

$$\begin{aligned} P[(T_1 - T_0) > t | X_0 = i] &= S_{i.}(t) \\ &= \sum_{l \neq i}^{r} P_{il} S_{il}(t) \end{aligned} \tag{2.16}$$

En posant $\delta_{ij} = 0$ si $i \neq j$ et 1 sinon, et à partir des résultats (2.15) et (2.16), nous obtenons finalement :

$$p_{ij}(t) = \sum_{k=1}^{r} \int_0^t P_{ik} f_{ik}(x) p_{kj}(t-x) dx \; + \; \delta_{ij} \sum_{l \neq i}^{r} P_{il} S_{il}(t) \tag{2.17}$$

La résolution de cette équation permet de trouver la loi des $p_{ij}(t)$, $i, j \in E$.

2.3 Fonction de Vraisemblance

Reprenons le modèle semi–Markovien comme défini précédemment. Considérons maintenant un échantillon de taille n, où chaque individu est repéré par l'indice h $(h = 1, 2, ..., n)$. Le sujet h change $m_h - 1$ fois d'état aux temps $T_{h,1} < T_{h,2} < ... < T_{h,m_h-1}$. A ces différents temps, les sujets ont successivement occupé les états $X_1^h, X_2^h, ..., X_{m_h-1}^h$, avec $X_p^h \neq X_{p+1}^h$. Etudions plus particulièrement le dernier temps d'observation de l'individu h, noté T_{h,m_h}. Il peut correspondre à une nouvelle transition, ou alors à une censure. Ces deux cas sont classiques en analyse de données de survie :

(i) La transition $i \to j, \forall i \neq j$, est observée après un temps de séjour x. La contribution de cette observation à la vraisemblance est :

$$\alpha_{ij}(x) S_{i.}(x) = P_{ij} f_{ij}(x)$$

(ii) L'observation est censurée à droite, autrement dit le processus reste dans l'état i jusqu'au temps de séjour x, mais nous ne possédons aucune information par la suite. Sa contribution s'exprime donc en terme de survie :

$$S_{i.}(x)$$

En considérant ces deux types d'individus, selon le statut de leur dernière transition, censurée (c) ou non–censurée (nc), la vraisemblance peut alors s'écrire comme le produit de toutes les contributions :

$$L = \prod_{h \in nc} \left[\prod_{r=1}^{m_h} \left\{ P_{X^h_{r-1}, X^h_r} f_{X^h_{r-1}, X^h_r}(T_{h,r} - T_{h,r-1}) \right\} \right] \times \prod_{h \in c} \left[\prod_{r=1}^{m_h - 1} \left\{ P_{X^h_{r-1}, X^h_r} f_{X^h_{r-1}, X^h_r}(T_{h,r} - T_{h,r-1}) \right\} \times S_{X^h_{m_h - 1}.}(T_{h,m_h} - T_{h,m_h-1}) \right]$$

La notion de prédiction pour un processus à censure aléatoire [12] simplifie cette formule. Un processus est dit prédictible lorsqu'il est observable. Les temps auxquels le processus est observé, sont en partie déterminés par un processus prédictible booléen :

$$K(t) = \sum_n I\{T_n < t < V_n\} \tag{2.18}$$

où V_n est une variable aléatoire appelée temps d'arrêt, avec $V_n \in [T_n, T_{n+1}]$. Si $K(t) = 1$, le processus est dit prédictible au temps t, sinon si $K(t) = 0$, il est dit non–prédictible. La figure (2.1) permet de comprendre comment est ainsi construit une fenêtre de prédictibilité. Si $T_n = V_n$ aucune information n'est disponible, que ce soit le temps de séjour dans l'état

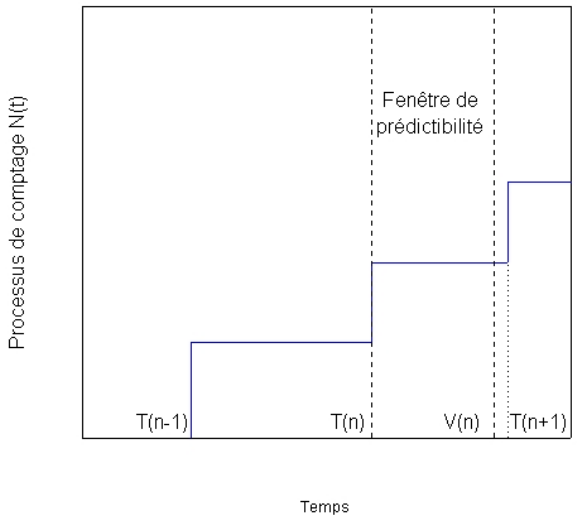

FIG. 2.1 – Prédictibilité d'un processus de comptage

$T_{n+1} - T_n$ ou les états X_n et X_{n+1}. A l'inverse, si $T_{n+1} = V_n$, le temps de séjour $T_{n+1} - T_n$ ainsi que les états adjacents X_n et X_{n+1} sont observables. Enfin, si $T_n < V_n < T_{n+1}$, alors

X_n est prédictible et le temps de séjour $T_{n+1} - T_n$ est connu pour être supérieur à $V_n - T_n$. A partir de ces notations, la vraisemblance est égale à :

$$\begin{aligned} L &= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ P_{X^h_{r-1},X^h_r} f_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \\ &\quad \times \Big\{ S_{X^h_{r-1}}.(V_{h,r-1} - T_{h,r-1}) \Big\}^{1-I\{T_{h,r}=V_{h,r-1}\}} \Big] \end{aligned} \tag{2.19}$$

où $I\{T_{h,r} = V_{h,r-1}\}$ est égal à 1 si $T_{h,r} = V_{h,r-1}$ (la transition est observée) et est égal à 0 sinon. D'après la définition (2.10), la vraisemblance peut aussi s'écrire sans perte d'information (démonstration en annexe 1) :

$$\begin{aligned} L &= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ \alpha_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \Big] \\ &\quad \times exp\Big(- \sum_{r=1}^{m_h} \sum_{k \neq X^h_{r-1}} \int_{T_{h,r-1}}^{min(T_{h,r},V_{h,r-1})} \alpha_{X^h_{r-1},k}(u - T_{r-1}) du \Big) \end{aligned} \tag{2.20}$$

2.4 Introduction de covariables

2.4.1 Modèle à risques semi–proportionnels

Pour prendre en compte d'éventuelles covariables dans le modèle, nous reprendrons le même principe de proportionnalité des risque que Cox [13] ou Andersen [14]. L'hypothèse supplémentaire est que le vecteur de covariables agit sur les fonctions de risque des temps d'attente dans les états, $\lambda_{ij}(x)$. Indirectement cet effet se répercute sur les fonctions de risque du modèle semi–Markovien global, $\alpha_{ij}(x)$.

Définissons $(\widetilde{X}_n, \widetilde{T}_n)$ les valeurs observées du processus et les temps de séjour correspondants. D'après la définition de la fonction de risque, si $T_{n+1} - \widetilde{T}_n < x$, alors la transition $X_n \to X_{n+1}$ est impossible à la durée de séjour x. Formellement :

$$\begin{aligned} \widetilde{\lambda}_{ij}(x) &= \lim_{h \to 0^+} \frac{P[x \leq T_{n+1} - \widetilde{T}_n < x + h | T_{n+1} - \widetilde{T}_n \geq x, \widetilde{X}_n = i, X_{n+1} = j]}{h} \\ &= \begin{cases} \lambda_{ij}(x) \text{ si } \widetilde{X}_n = i \text{ et } (T_{n+1} - \widetilde{T}_n \geq x) \\ 0 \text{ si } \widetilde{X}_n = i \text{ et } (T_{n+1} - \widetilde{T}_n < x) \end{cases} \quad \forall n \geq 0 \end{aligned}$$

où $\widetilde{\lambda}_{ij}(x)$ est appelée fonction d'intensité. Cette fonction peut être exprimée plus simplement en définissant $I\{X_{\widetilde{N}(t^-)} = i\}$. Cette indicatrice est égale à 1 si le processus est dans l'état i juste avant t. Nous pouvons alors écrire :

$$\widetilde{\lambda}_{ij}(t - T_{\widetilde{N}(t^-)}) = I\{X_{\widetilde{N}(t^-)} = i\} \lambda_{ij}(t - T_{\widetilde{N}(t^-)}) \tag{2.21}$$

Remarquons que cette notation possède un réel intérêt lorsque le concept est généralisé à une population. Pour un échantillon de taille n, où chaque individu est repéré par l'indice h $(h = 1, 2, ..., n)$:

$$Y_i(t^-) = \sum_h I\{X_{\widetilde{N}_h(t^-)} = i\}$$

où $X_{\widetilde{N}_h(t-)}$ est la valeur du processus juste avant t pour le sujet h. $Y_i(t-)$ représente ainsi l'effectif à risque d'une transition $i \to j$ au temps t. Nous avons alors :

$$\widetilde{\lambda}_{ij}(t - T_{\widetilde{N}(t^-)}) = Y_i(t^-)\lambda_{ij}(t - T_{\widetilde{N}(t^-)})$$

Nous comprenons alors mieux d'où vient le terme *intensité*, puisque pour une même fonction de risque, la chance d'observer une transition augmente avec le nombre de sujets à risque.

Soit $z_{ij}(x) = \Big(z_{ij}^1(x), z_{ij}^2(x), ..., z_{ij}^{n_{ij}}(x)\Big)$, le vecteur des n_{ij} covariables au temps de séjour x et propre à la transition $i \to j$. Celui–ci modifie la fonction d'intensité $\widetilde{\lambda}_{ij}(x)$, permettant ainsi d'analyser l'hétérogénéité de la population. D'après la définition de la prédictibilité (2.18), les valeurs du vecteur de covariables $z(t - T_{\widetilde{N}(t^-)})$ sont observables pour $t \in [T_{\widetilde{N}(t^-)}, V_{\widetilde{N}(t^-)}]$. Cependant, pour simplifier les calculs, nous supposerons les covariables fixes au cours du temps d'attente : $z_{ij}(x) = z_{ij}$. A partir de (2.21) et de l'hypothèse de proportionnalité des risques, l'introduction des covariables est réalisée comme suit :

$$\begin{aligned} \widetilde{\lambda}_{ij}(t - T_{\widetilde{N}(t^-)}, z) &= \widetilde{\lambda}_{0,ij}(t - T_{\widetilde{N}(t^-)})\ \eta(z) \\ &= I\{X_{\widetilde{N}(t^-)} = i\}\lambda_{0,ij}(t - T_{\widetilde{N}(t^-)})\ \eta(z) \end{aligned} \tag{2.22}$$

où $\eta(z)$ est une fonction quelconque des covariables. Comme dans la grande majorité des modèles (Cox par exemple), nous posons :

$$\eta(z) = exp\Big(\beta_{ij}^T z_{ij}\Big)$$

où $\beta_{ij} = \Big(\beta_{ij}^1, \beta_{ij}^2, ..., \beta_{ij}^{n_{ij}}\Big)$ est le vecteur des coefficients de régression associés à z_{ij}. $\lambda_{0,ij}(x)$ est appelée fonction de risque de base. Il s'agit de la fonction de risque propre à la population de référence, pour laquelle toutes les covariables sont codées 0. L'intérêt de la prise en compte des covariables à travers une forme exponentielle permet à la fonction de risque d'être définie positive et une interprétation sous forme de risque relatif. Les fonctions de survie et de densité correspondantes peuvent alors être facilement calculées :

$$\begin{aligned} S_{ij}(t - T_{\widetilde{N}(t^-)}, z) &= exp\Big(-\int_0^{t-T_{\widetilde{N}(t^-)}} \lambda_{ij}(u, z)du\Big) \\ &= exp\Big(-\eta(z)\int_0^{t-T_{\widetilde{N}(t^-)}} \lambda_{0,ij}(u)du\Big) \\ &= exp\Big(-\int_0^{t-T_{\widetilde{N}(t^-)}} \lambda_{0,ij}(u)du\Big)^{\eta(z)} \\ &= S_{0,ij}(t - T_{\widetilde{N}(t^-)})^{\eta(z)} \end{aligned} \tag{2.23}$$

et

$$\begin{aligned} f_{ij}(t - T_{\widetilde{N}(t^-)}, z) &= S_{ij}(t - T_{\widetilde{N}(t^-)}, z)\ \lambda_{ij}(t - T_{\widetilde{N}(t^-)}, z) \\ &= \lambda_{0,ij}(t - T_{\widetilde{N}(t^-)})\ \eta(z)\ S_{0,ij}(t - T_{\widetilde{N}(t^-)})^{\eta(z)} \end{aligned} \tag{2.24}$$

Comme pour le traitement des processus Markoviens par Andersen [14], ce modèle est dit semi–proportionnel, car la proportionnalité des risques n'est supposée qu'au sein

d'une même classe de transition, aucune contrainte n'est imposée entre classe. De plus, les individus sont comparés à temps de séjour fixé, cette proportionnalité des risques s'applique donc à des individus passant la même période de temps dans un état. Enfin, ce modèle est plus parcimonieux que celui définit par Perez [8]. En effet, dans ce dernier chaque covariable agissait spécifiquement sur chaque transition, alors que les définitions précédentes permettent un vecteur de covariables de taille différente par transition.

Pour estimer les paramètres, nous reprendrons les vraisemblances (2.19) et (2.20) en substituant les termes $f_{ij}(x)$, $S_{ij}(x)$, $F_{ij}(x)$ et $\alpha_{ij}(x)$, par leur forme avec covariables : $f_{ij}(x,z)$, $S_{ij}(x,z)$, $F_{ij}(x,z)$ et $\alpha_{ij}(x,z)$.

2.4.2 Probabilité de survie dans l'état

Cette partie est importante pour retrouver, à partir d'un modèle semi–Markovien, les indicateurs courants, interprétés et utilisés par les cliniciens. La représentation de la survie, en fonction du temps et des covariables, constitue un point majeur. Pour ce calcul, partons de la relation (2.11) et des définitions classiques de l'analyse de survie [15] :

$$S_{i.}(x) = exp\Big(- \int_0^x \alpha_{i.}(u)du\Big)$$

$$A_{i.}(x) = \int_0^x \alpha_{i.}(u)du$$

où $A(x)$ est la fonction de risque cumulée. En supposant les covariables indépendantes du temps pour simplifier le calcul, nous pouvons alors écrire :

$$\begin{aligned} S_{i.}(x,z) &= exp\Big(- \int_0^x \alpha_{i.}(u,z)du\Big) \\ &= exp\Big(- \int_0^x \sum_{j\neq i} \alpha_{ij}(u,z)du\Big) \\ &= exp\Big(- \sum_{j\neq i} A_{0,ij}(x,z)\Big) \end{aligned} \tag{2.25}$$

Ainsi, si $X_0, X_1, ..., X_n$ est la séquence des états visités aux temps $T_0, T_1, ..., T_n$, alors la probabilité qu'un individu survive dans l'état $X_n = i$ jusqu'au temps $T_n + x$ est égale à (2.25).

Pour présenter le modèle et interpréter les résultats, il est intéressant de calculer la probabilité que, sachant que $X_n = i$, un sujet saute dans l'état j au temps $T_{n+1} = T_n + x$:

$$\begin{aligned} &\lim_{h\to 0} P[X_{n+1} = j | x < T_{n+1} - T_n \leq x + h, (T_{n+1} - T_n) \geq x, X_n = i] \\ &\quad = \lim_{h\to 0} \frac{P[X_{n+1} = j, x < T_{n+1} - T_n \leq x + h | (T_{n+1} - T_n) \geq x, X_n = i]}{P[x < T_{n+1} - T_n \leq x + h | (T_{n+1} - T_n) \geq x, X_n = i]} \\ &\quad = \frac{\alpha_{ij}(x,z)}{\alpha_{i.}(x,z)} \\ &\quad = \frac{\alpha_{ij}(x,z)}{\sum_{k\neq i} \alpha_{ik}(x,z)} \end{aligned}$$

2.5 Loi de Weibull comme loi de séjour dans l'état

La loi de Weibull, $W(\nu_{ij}, \sigma_{ij})$, possède de bonnes propriétés pour la modélisation des données de survie. Elle permet de prendre en compte une évolution monotone, notamment croissante, du risque instantané au cours du temps. Elle est d'ailleurs utilisée dans l'article de Perez [8]. Sans covariable, elle est donnée par :

$$\lambda_{ij}(x) = \nu_{ij}\Big(\frac{1}{\sigma_{ij}}\Big)^{\nu_{ij}} x^{\nu_{ij}-1} \qquad \forall x \geq 0,\ \forall \nu_{ij} > 0 \text{ et } \forall \sigma_{ij} > 0$$

La figure (2.2) rend compte des formes possibles de cette fonction. Dans le cas particulier où $\nu_{ij} = 1$, nous retrouvons une distribution exponentielle, sans mémoire (annexe 2). Les modèles semi–Markoviens utilisant une loi de Weibull constituent donc une généralisation des modèles Markoviens homogènes. En respectant les définitions (2.22), (2.23) et (2.24), nous pouvons alors définir la fonction de risque avec covariables :

$$\lambda_{ij}(x, z) = \nu_{ij}\Big(\frac{1}{\sigma_{ij}}\Big)^{\nu_{ij}} x^{\nu_{ij}-1} exp(\beta_{ij}^T z_{ij})$$

La fonction de survie associée :

$$\begin{aligned} S_{ij}(x, z) &= S_{ij}(x)^{exp(\beta_{ij}^T z_{ij})} \\ &= exp\Big(- \int_0^x \nu_{ij}\Big(\frac{1}{\sigma_{ij}}\Big)^{\nu_{ij}} x^{\nu_{ij}-1}\Big)^{exp(\beta_{ij}^T z_{ij})} \\ &= exp\Big(- (\frac{1}{\sigma_{ij}})^{\nu_{ij}}\Big)^{exp(\beta_{ij}^T z_{ij})} \end{aligned}$$

Le calcul de la densité correspondante est direct :

$$\begin{aligned} f_{ij}(x, z) &= S_{ij}(x, z)\ \lambda_{ij}(x, z) \\ &= \nu_{ij}\Big(\frac{1}{\sigma_{ij}}\Big)^{\nu_{ij}} x^{\nu_{ij}-1} exp(\beta_{ij}^T z_{ij}) exp\Big(- (\frac{1}{\sigma_{ij}})^{\nu_{ij}}\Big)^{exp(\beta_{ij}^T z_{ij})} \end{aligned}$$

Même si la loi de Weibull permet de mieux modéliser la dynamique du processus, elle ne donne pas la possibilité de prendre en compte une forme en U du risque instantané. La loi de Weibull généralisée permet ce type de modélisation.

2.6 Généralisations du modèle

2.6.1 Loi de Weibull généralisée

La distribution utilisée, notée $WG(\nu_{ij}, \sigma_{ij}, \theta_{ij})$, permet de modéliser une fonction de risque en forme de U (figure (2.3) et de généraliser à son tour la loi de Weibull. Sa fonction de survie est donnée par :

$$S_{ij}(x) = exp\Big(1 - (1 + (\frac{x}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}}\Big)$$

Le calcul de la fonction de risque est alors direct :

$$log\Big(S_{ij}(x)\Big) = 1 - (1 + (\frac{x}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}}$$

$$\frac{d\Big[log\Big(S_{ij}(x)\Big)\Big]}{dx} = -\frac{1}{\theta_{ij}}\Big(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}}\Big)^{\frac{1}{\theta_{ij}}-1}\nu_{ij}\Big(\frac{x}{\sigma_{ij}}\Big)^{\nu_{ij}-1}\frac{1}{\sigma_{ij}}$$

d'où

$$\begin{aligned}\lambda_{ij}(x) &= -\frac{d\Big[log\Big(S_{ij}(x)\Big)\Big]}{dt} \\ &= \frac{1}{\theta_{ij}}\Big(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}}\Big)^{\frac{1}{\theta_{ij}}-1}\frac{\nu_{ij}}{\sigma_{ij}}\Big(\frac{x}{\sigma_{ij}}\Big)^{\nu_{ij}-1}\end{aligned}$$

Enfin, le densité s'écrit :

$$\begin{aligned}f_{ij}(x) &= \lambda_{ij}(x)\ S_{ij}(x) \\ &= \frac{1}{\theta_{ij}}\Big(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}}\Big)^{\frac{1}{\theta_{ij}}-1}\frac{\nu_{ij}}{\sigma_{ij}}\Big(\frac{x}{\sigma_{ij}}\Big)^{\nu_{ij}-1}exp\Big(1-(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}}\Big)\end{aligned}$$

D'après les définitions (2.22), (2.23) et (2.24), nous obtenons directement :

$$\lambda_{ij}(x,z) = \frac{1}{\theta_{ij}}\Big(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}}\Big)^{\frac{1}{\theta_{ij}}-1}\frac{\nu_{ij}}{\sigma_{ij}}\Big(\frac{x}{\sigma_{ij}}\Big)^{\nu_{ij}-1}\ exp\Big(\beta_{ij}^T z_{ij}\Big)$$

$$S_{ij}(x,z) = exp\Big(1-(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}}\Big)^{exp\Big(\beta_{ij}^T z_{ij}\Big)}$$

$$f_{ij}(x) = \frac{1}{\theta_{ij}}\Big(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}}\Big)^{\frac{1}{\theta_{ij}}-1}\frac{\nu_{ij}}{\sigma_{ij}}\Big(\frac{x}{\sigma_{ij}}\Big)^{\nu_{ij}-1}\ exp\Big(\beta_{ij}^T z_{ij}\Big)\ exp\Big(1-(1+(\frac{x}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}}\Big)^{exp\Big(\beta_{ij}^T z_{ij}\Big)}$$

Il est intéressant, d'un point de vue théorique, de généraliser les lois des temps de séjour dans les états. Cependant, il est évident que la multiplication des paramètres à estimer constitue une difficulté majeure. Pour identifier les distributions adéquates, nous explorerons a priori chaque fonction de survie $S_{ij}(x)$ par une méthode non–paramétrique de type Kaplan–Meier [17]. De plus, nous testerons, a posteriori, l'égalité des paramètres σ_{ij} et θ_{ij} à 1. Néanmoins, le modèle défini précédemment nécessite le choix d'une distribution commune à toutes les transitions, mêmes si certains paramètres apparaissent comme inutiles.

2.6.2 Choix d'une loi de temps de séjour spécifique à chaque transition

Précédemment, $f_{ij}(x)$ a été défini comme une fonction de densité dépendante de x et d'un vecteur de paramètres. Ce vecteur, pour une distribution de Weibull, est par exemple égal à (ν_{ij}, σ_{ij}). Dans le but de généraliser le modèle, notons explicitement :

$$f_{ij}(x) = f(x, \nu_{ij}, \sigma_{ij})$$

En d'autres termes, la forme des distributions des temps de séjour est identique pour chaque transition, seuls les paramètres sont modifiés. Ce choix de modélisation est contraignant et peu parcimonieux. En effet, certaines transitions peuvent nécessiter de nombreux

paramètres, alors que d'autres transitions peuvent être modélisées plus simplement. Dans ce rapport, nous avons abordé trois distributions intéressantes en analyse de survie : exponentielle, Weibull, et Weibull généralisée. Elles nécessitent respectivement un, deux et trois paramètres. Pour permettre un modèle plus adéquate, adoptons plutôt la notation suivante qui ne change rien au reste de la théorie abordée dans ce chapitre :

$$f_{ij}(x) = f^{(ij)}(x, \nu_{ij}, \sigma_{ij})$$

La forme et les paramètres des lois de distribution peuvent alors être spécifiques à chaque transition. La flexibilité du modèle est accrue.

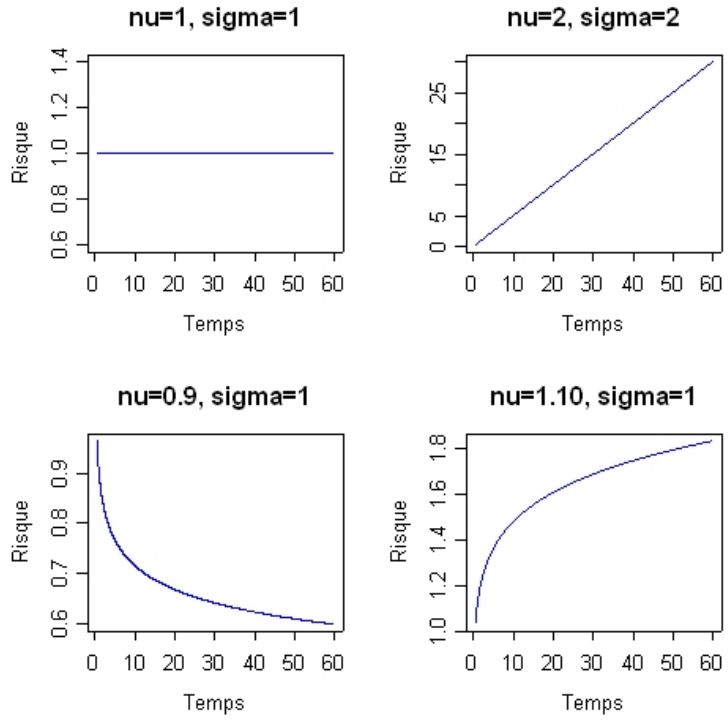

FIG. 2.2 – Exemples de fonctions de risque d'une loi de Weibull

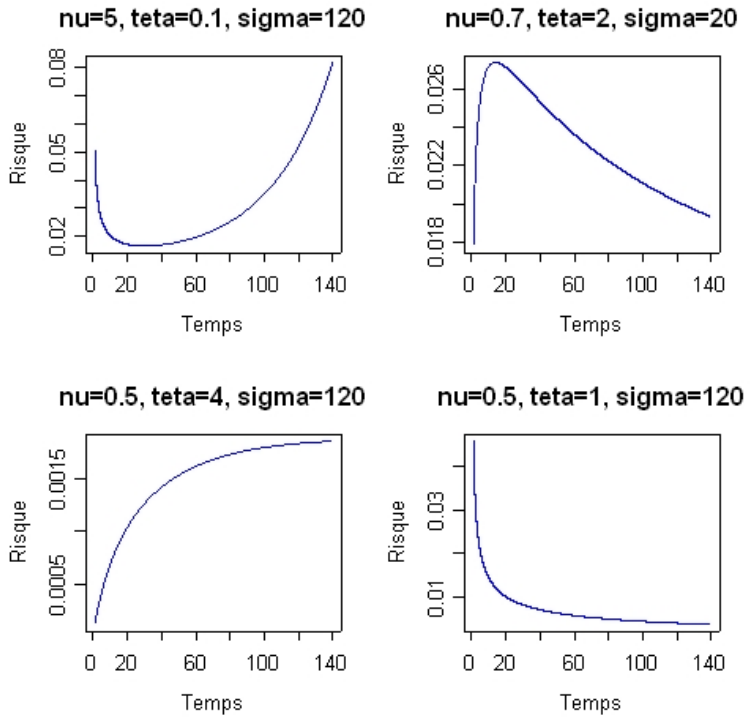

FIG. 2.3 – Exemples de fonctions de risque d'une loi de Weibull généralisée

Chapitre 3

Application au VIH

3.1 Définition du modèle

La notion de processus a un intérêt particulier dans l'étude de la dynamique d'une pathologie. Le VIH (Virus de l'Immuno–déficience Humaine) étant une maladie chronique, l'évolution des patients atteints de cette infection, répond bien à une telle problématique. Le VIH est un virus attaquant les défenses immunitaires. Le stade final de la maladie est le Syndrome de l'Immuno–Déficience acquise (SIDA), où le malade peut décéder de maladies opportunistes.

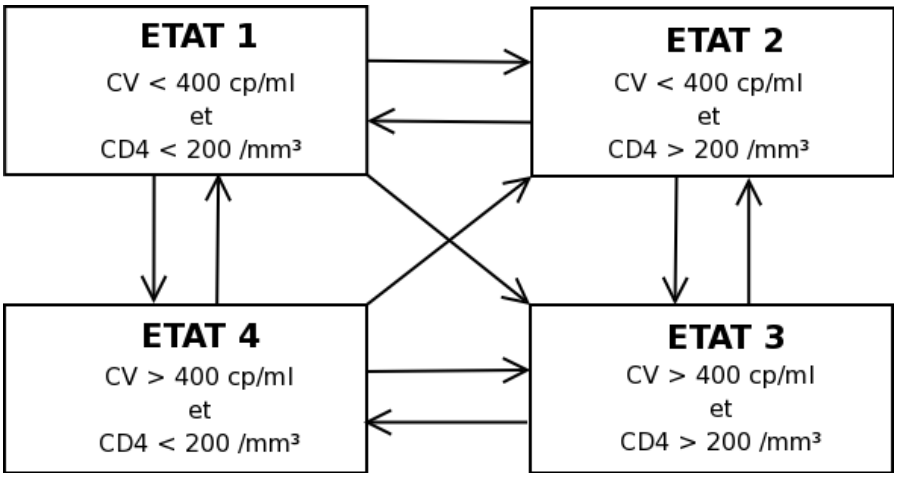

Fig. 3.1 – Graphique des transitions possibles

Deux marqueurs de l'avancement de la maladie sont importants : la charge virale (CV) et la concentration de lymphocytes CD4. La charge virale représente l'activité du virus. Plus sa valeur est importante, moins bon est le pronostic. A l'inverse, le nombre de lymphocytes CD4 représente la capacité immunologique de l'individu. Plus sa valeur est faible, moins bon est le pronostic de la maladie. Ces deux marqueurs permettent aux cliniciens de définir quatre états essentiels dans l'analyse de l'avancement de la maladie, comme l'indique la figure (3.1). Les transitions possibles entre états y sont aussi représentées.

Elles ont été validées par les spécialistes et confirmées par la représentativité de la base de données (tableau (3.1)). Les états 1 et 2 sont des états à CV basse, alors que les états 3 et 4 sont à CV haute. L'état 2 est l'état de meilleur pronostic et l'état 4 est le plus grave. L'objectif de cette étude est de modéliser l'évolution du patient à travers ces états et l'effet de certaines covariables. Elles sont au nombre de 8 : le sexe, l'âge (supérieur ou non à 40 ans), la coinfection par une hépatite C, la coinfection par par une hépatite B et le mode de contamination (hétérosexuelle, homosexuelle, par toxicomanie, ou autres).

3.2 Recueil des données

Le CISIH (Centre d'Information et de Soins de l'Immunodéficience Humaine) du CHU de Nice a développé le logiciel ADDIS permettant la saisie en temps réel des données par les médecins au cours des consultations et l'envoi des informations vers les bases de données internes et externes (PMSI, DMI2). Ce logiciel est fonctionnel depuis juin 1994. Une nouvelle version du logiciel, NADIS, a été développée en partenariat avec Alliance Medica, filiale de Glaxo–Wellcome. Depuis 1998, le logiciel a été distribué dans 6 autres centres hospitaliers français. Nous utiliserons les données issues du CHU de Nice.

La base de données NADIS regroupe tous les patients VIH positifs suivis au moins une fois en consultation au CHU de Nice. Nous choisissons de limiter notre analyse aux données des patients disponibles depuis le 1er janvier 1996, ceci afin de conserver la comparabilité des observations et minimiser le biais période (1996 est l'année d'apparition des tests génotypiques et des inhibiteurs de la protéase). De plus, nous nous limitons aux individus ayant au moins deux séries d'examens biologiques afin d'exclure les patients n'ayant pas une réelle prise en charge thérapeutique. Enfin, pour étudier une population homogène, seuls les adultes ont été conservés (plus de 18 ans). La date de point a été fixée au 30 avril 2004. Le temps chronologique du suivi est mesuré à partir de la première date de mesure biologique. D'après les données, nous avons supposé que des temps de séjour supérieurs à 6 ans dans un même état sans consultation, constituent des valeurs aberrantes qui seront supprimées. Ce seuil est arbitraire permet d'éliminer certaines valeurs anormales, à partir d'une hypothèse clinique acceptable.

3.3 Stratégie de modélisation

L'objectif de notre modèle est d'expliquer au mieux la dynamique du processus à l'aide d'un modèle le plus parcimonieux possible. La sélection d'un tel modèle nécessite de tester l'apport d'éventuels paramètres supplémentaires. Nous utiliserons deux tests. Le test de Wald (sélection des covariables en univarié) et le test du rapport de Vraisemblance (sélection des covariables en multivarié et choix des lois de distribution). Nous pouvons distinguer quatre étapes dans la modélisation. Pour chaque distribution utilisée (Weibull et Weibull généralisé), nous procéderons dans cet ordre. L'intérêt des cette double analyse est de mesurer l'apport du Weibull généralisé et d'évaluer la robustesse des facteurs de risque au choix des lois de temps de séjour.

(i) Analyse Stratifiée – Nous mettrons estimerons un modèle différent par modalité des covariables (analyse en sous–groupe). Cette étape possède plusieurs intérêts. Tout

d'abord, nous pourrons vérifier que la loi utilisée est adéquate par rapport à une loi exponentielle. Ensuite, nous identifierons les covariables qui semblent avoir un effet sur les vitesses de transition. Enfin, nous évaluerons la validité de l'hypothèse de proportionnalité des risques, propre à chaque covariable et à chaque transition. Les résultats seront présentés sous forme graphique pour une meilleure interprétation.

(ii) Analyse univariée – Après cette première étape, plus exploratoire qu'analytique, nous mettrons en place un modèle pour chaque covariable. Nous appelons ces modèles "univariés" au sens où une seule covariable est en présence, même si elle a un effet sur plusieurs transitions. Nous obtenons ainsi 8 modèles différents (correspondants aux 8 covariables). Cette étape permet d'identifier les covariables qui semblent avoir un effet spécifique sur chaque transition. Etant donné le nombre important de facteurs potentiellement influents ($8 \times 10 = 80$), nous avons choisi de retenir, pour l'analyse multivariée, les facteurs dont la p–value est inférieure à 0,05 (test de Wald). Par soucis de claireté, ces résultats ne seront pas présentés exhaustivement dans ce document. Seuls les principaux points seront explicités.

(iii) Analyse univariée – Dans un troisième temps, toujours en supposant une distribution de type Weibull, toutes les variables précédemment retenues seront incluses dans le même modèle. Le vecteur de covariables sera spécifique à chaque transition. Les covariables les moins significatives ($p > 0,05$) seront éliminées une à une du modèle (test du rapport de Vraisemblance), jusqu'à ce que tous les coefficients de régression aient un risque de première espèce inférieur au seuil. A chaque étape de cette stratégie descendante, la constance des autres coefficients de régression sera évaluée (variation relative inférieure à 30%). Cette vérification possède un double intérêt : identifier la présence d'éventuels facteurs de confusion ou d'interaction, et mesurer la stabilité de l'estimation.

(iv) Choix du modèle le plus parcimonieux – Sous la contrainte d'une forme de distribution des temps de séjour invariante selon les transitions, le modèle ainsi obtenu peut être considéré comme le plus parcimonieux. La dernière étape consiste alors à identifier, par le test du rapport de Vraisemblance, si certains des paramètres (ν_{ij} pour la loi de Weibull et θ_{ij} pour la loi de Weibull généralisée) ne diffèrent pas significativement de la valeur théorique 1.

L'ensemble des analyses et des représentations graphiques ont été réalisées à partir du logiciel *R*. Nous avons utilisé la fonction *optim*() pour maximiser la Vraisemblance et estimer la valeur des paramètres ainsi que leur matrice Hessienne. Cette fonction utilise l'algorithme de *quasi–Newton*. Pour lancer l'optimisation des modèles stratifiés et univariés, nous avons initialisé les paramètres propres aux distributions en utilisant la méthode des moindres carrés, à partir des survies estimées par la méthode non–paramétrique de Kaplan–Meier [17]. La chaîne de Markov sous–jacente a été initialisée à partir de simples proportions. Deux tests, équivalents asymptotiquement, ont été cités dans notre stratégie de sélection de modèle. Le test de Wald étant en partie basé sur la matrice Hessienne

dont nous ne possédons qu'une approximation, nous préférerons le test du Rapport de Vraisemblance (LRS) lorsque le nombre de tests à effectuer est peu élevé. C'est pour cette raison que le test de Wald n'est utilisé que pour la stratégie univariée de sélection des covariables potentiellement influentes.

3.4 Résultats

3.4.1 Description des données

La base de données est constituée de 1244 individus, ce qui représente 4804 observations. En moyenne, les patients ont donc un peu moins de quatre mesures de CV et de CD4. Bien sur, selon leur date d'entrée dans la cohorte, ce nombre est plus moins important.

Le tableau (3.1) décrit la représentativité des transitions. Les passages $2 \rightarrow 3$ et $3 \rightarrow 2$ sont les plus observés. Nous pouvons, de plus, remarquer de nombreuses transitions de l'état 4 vers un état de meilleur pronostic. Ceci montre bien l'effort du clinicien à diminuer le réservoir virologique (CV) et à augmenter le réservoir immunologique (CD4).

Transition	Effectif	Pourcentage	Médiane[1]
$1 \rightarrow 1$ [2]	31	0,6 %	0,82
$1 \rightarrow 2$	282	5,9 %	0,52
$1 \rightarrow 3$	58	1,2 %	0,48
$1 \rightarrow 4$	174	3,6 %	0,51
$2 \rightarrow 1$	152	3,2 %	0,63
$2 \rightarrow 2$ [2]	605	12,6 %	1,75
$2 \rightarrow 3$	994	20,7 %	0,81
$3 \rightarrow 2$	1340	27,9 %	0,78
$3 \rightarrow 3$ [2]	231	4,8 %	1,31
$3 \rightarrow 4$	212	4,4 %	0,85
$4 \rightarrow 1$	283	5,9 %	0,76
$4 \rightarrow 2$	109	2,3 %	0,56
$4 \rightarrow 3$	268	5,6 %	0,50
$4 \rightarrow 4$ [2]	65	1,4 %	1,18

Tab. 3.1 – Représentativité des transitions

Comme l'indique le tableau (3.2), notre échantillon est composé pour un tiers de femmes. 32 % des transitions sont relatives à des patients âgés de plus de 40 ans. Le mode de contamination est réparti également entre les 4 catégories définies. Enfin, respectivement 9,7 % et 19,5 % sont coinfectés par une hépatite B et C. Ces données sont comparables à la population cible des patients VIH positifs.

[1] Temps de séjour médian en mois
[2] Censures à droite

Covariables	Effectif	Pourcentage
Femmes	381	30,6 %
Age > 40 ans	395	31,8 %
Coinfection VHB	121	9,7 %
Coinfection VHC	242	19,5 %
Contamination hétérosexuelle	359	28,9 %
Contamination homosexuelle	251	20,2 %
Contamination par toxicomanie	337	27,1 %
Contamination autre (accidents)	297	23,9 %

TAB. 3.2 – Descriptif de la population d'étude

3.4.2 Modèle semi–Markovien de type Weibull

Modèles stratifiés et univariés

Les graphiques présentant les modèles estimés sont en annexe 3. Les écarts entre les courbes pour certaines transitions et pour certaines covariables soulignent l'intérêt de prendre en compte ces facteurs dans l'étude des forces de transition. Cependant, compte tenu des graphiques, l'hypothèse de proportionnalité des risques n'est que rarement vérifiée. En effet, de nombreux croisements ou divergences entre les fonctions de risque peuvent être observés. Ce constat est d'autant plus pénalisant qu'il semble que selon les modalités d'une covariable, la forme de la loi de distribution diffère pour une même transition. C'est par exemple le cas pour la transition $4 \rightarrow 3$ entre hommes et femmes. La force de transition des hommes semble plutôt constante (loi exponentielle sans mémoire), alors qu'elle diminue pour les femmes (loi de Weibull). Dans de telles situations, seule la stratification peut permettre ce type d'approche. Cette méthode ne laissant pas la possibilité de tester l'effet de la variable de stratification, on préférera éliminer l'effet de la covariable lorsque son effet n'est pas risque proportionnel.

Le tableau (3.3) représente ainsi les covariables sélectionnées pour la stratégie stratifiée (signalées par le symbole ×). Cette première sélection, même si elle est subjective, possède l'intérêt d'éviter certains problèmes d'estimation et de ne pas aboutir à un modèle dont les interprétations seraient abusives. La stratégie univariée élimine à son tour les covariables qui paraissent inutiles par transition (voir tableau (3.3), symbole O). Sur les 80 facteurs possibles, 10 ont été sélectionnés pour l'analyse multivariée.

Modèle multivarié

Après sélection, le modèle final repose sur 5 covariables, soit 31 paramètres au total. Leurs estimations ainsi que celles de leur variance sont présentées dans le tableau (3.6) en annexe 5. La logVraisemblance est égale à -6124,0, ce qui correspond à un critère d'AIC[1] égal à 12310. A partir de l'état 3, les patients coinfectés par une hépatite C et ceux dont le mode de contamination est accidentel transitent plus rapidement vers l'état 2.

[1]La minimisation du Critère d'Information d'Akaike (AIC) permet la sélection de modèles non–emboîtés. $AIC = -2 \times LogV + 2 \times$ Nombre de paramètres

Transition	Sexe	Age	VHB	VHC	Co.Hétéro.	Co.Homo.	Co.Toxico.	Co.autre
$1 \rightarrow 2$	$\times$		$\times$	$\times$				$\times$
$1 \rightarrow 3$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$
$1 \rightarrow 4$	$\times$		$\times$	$\times$		$\times$	$\times$	$\times$ O
$2 \rightarrow 1$	$\times$	$\times$ O	$\times$ O	$\times$ O	$\times$		$\times$ O	
$2 \rightarrow 3$	$\times$			$\times$				$\times$
$3 \rightarrow 2$			$\times$	$\times$ O				$\times$ O
$3 \rightarrow 4$	$\times$	$\times$ O	$\times$ O					
$4 \rightarrow 1$	$\times$	$\times$	$\times$			$\times$ O	$\times$	
$4 \rightarrow 2$		$\times$				$\times$	$\times$	
$4 \rightarrow 3$		$\times$			$\times$			$\times$ O

TAB. 3.3 – Covariables retenues après les stratégies stratifiées ($\times$) et univariées (O)

Les malades âgés de plus de 40 ans et ceux coinfectés par une hépatite B, passent plus rapidement de l'état 3 à l'état 4, ce qui va dans le sens d'une aggravation de la maladie. Un patient dont le mode de contamination est accidentel semble transiter plus rapidement de 4 à 3.

Remarquons que certains paramètres ν_{ij} sont proches de la valeur 1, ce qui peut laisser à supposer que certaines transitions obéissent à une loi exponentielle. La prochaine étape consiste donc à identifier ces transitions sans mémoire.

Modèle multivarié avec distributions spécifiques

L'élimination successive, des paramètres ν_{ij} par le test du rapport de Vraisemblance (LRS), diminue le nombre de paramètres total de 31 à 27. En effet, les temps de séjour, de 4 des 10 transitions, semblent suivre des lois exponentielles, sans perte d'information ($LogV = -6126,96$ et $AIC = 12307,92$). Il s'agit des transitions $4 \rightarrow 3$, $3 \rightarrow 2$, $2 \rightarrow 1$ et $1 \rightarrow 4$ (tableau (3.4)). Le modèle final, expliquant le maximum d'information à partir d'un minimum de paramètres, est ainsi représenté dans la tableau (3.5). Les autres coefficients restant dans le modèle ne varient que très peu. Ceci souligne l'intérêt d'une telle simplification et la robustesse de la méthode d'estimation. Les mêmes facteurs influençant l'évolution de la pathologie sont donc à noter : les hépatites B et C, l'âge et les modes de contamination autres que sexuel et par toxicomanie.

Modèles	LogV	LRS	ddl	p–value
Modèle 1 : $\nu_{ij} \neq 1 \; \forall i \neq j \; \in \{1,2,3,4\}$	-6124,00			
Modèle 2 : Modèle 1 avec $\nu_{43} = 1$	-6124,01	0,02	1	0,89
Modèle 3 : Modèle 2 avec $\nu_{32} = 1$	-6124,70	1,38	1	0,24
Modèle 4 : Modèle 3 avec $\nu_{21} = 1$	-6125,81	2,22	1	0,14
Modèle 5 : Modèle 4 avec $\nu_{14} = 1$	-6126,96	1,15	1	0,28

TAB. 3.4 – Sélection du modèle le plus adéquate à partir de lois de Weibull et exponentielles

Paramètres	Coefficients	Écart–types
ν_{12}	1,12	0,05
ν_{13}	1,45	0,14
ν_{23}	0,87	0,02
ν_{34}	0,79	0,04
ν_{41}	0,91	0,04
ν_{42}	1,19	0,08
σ_{12}	0,59	0,03
σ_{13}	0,55	0,05
σ_{14}	0,57	0,05
σ_{21}	0,78	0,08
σ_{23}	1,88	0,07
σ_{32}	0,98	0,04
σ_{34}	1,78	0,26
σ_{41}	0,95	0,07
σ_{42}	0,63	0,06
σ_{43}	0,57	0,04
P_{12}	0,55	0,02
P_{13}	0,11	0,01
P_{21}	0,10	0,01
P_{32}	0,85	0,01
P_{41}	0,45	0,02
P_{42}	0,16	0,01
β_{32}^{VHC}	0,17	0,07
$\beta_{32}^{co.autre}$	0,18	0,07
β_{34}^{age}	0,50	0,17
β_{34}^{VHB}	0,33	0,18
$\beta_{43}^{co.autre}$	0,28	0,16

Tab. 3.5 – Modèle semi–Markovien multivarié final de type Weibull et exponentiel

3.4.3 Modèle semi–Markovien de type Weibull généralisé

En utilisant les stratégies stratifiées et univariées similaires à la partie précédente, 11 covariables ont été sélectionnées (voir tableau (3.7), annexe 6). Remarquons, de plus, la pertinence de prendre en compte une fonction de risque en U puisque la majeure partie des transitions semble répondre à cette problématique. Peu de temps après l'entrée dans un état, le risque de transition est élevé et croissant. Ce phénomène reflète bien le caractère d'instabilité du patient qui vient de changer d'état. Cependant, après un délai variable selon la transition, ce risque diminue, reflet d'une stabilisation. Plus la personne reste de temps dans un état, moins elle a de chance d'en sortir.

Après la stratégie multivariée descendante, 9 covariables sont définies comme influentes (tableau (3.8), annexe 6). Les femmes ont tendance à transiter plus vite de l'état 1 à 3. De la même manière, être âgé de plus de 40 ans, être coinfecté par une hépatite B ou

avoir été contaminé par toxicomanie, semblent accélérer la transition $1 \rightarrow 2$. A l'inverse, les patients contaminés par un rapport homosexuel passent moins rapidement de l'état 1 à l'état 2. Enfin, le mode de contamination par accident, hétérosexuelle, par toxicomanie et le sexe ralentissent respectivement les transitions $2 \rightarrow 3$, $3 \rightarrow 2$, $4 \rightarrow 1$ et $3 \rightarrow 4$.

Aucun paramètre θ_{ij}, $\forall i \neq j \ \in \{1, 2, 3, 4\}$, n'est statistiquement différent de 1. Autrement dit, l'utilisation d'une loi de Weibull généralisée semble justifiée quelle que soit la transition. Ce modèle semble donc le plus parcimonieux, avec une LogVraisemblance égale à -5704,08. Avec 45 paramètres, l'AIC vaut 11498,16. Ce dernier critère est largement inférieur à celui obtenu pour le modèle final fondé sur des distributions Weibull et exponentielles (12307,92).

Conclusions et perspectives

Il est évident, au vue des résultats, que les modélisations Markoviennes homogènes ont un intérêt restreint puisque les forces de transition ne sont pas constantes. Il convient donc de modéliser un phénomène d'usure ou de récupération, à travers des fonctions de risque monotones, qu'elles soient croissantes ou décroissantes. C'est ce que nous permet la loi de Weibull, utilisée par Perez [8] pour la modélisation du cancer.

L'évolution des patients séropositifs nécessite une loi plus complexe en forme de U, comme la loi de Weibull généralisée. Outre cette meilleure qualité d'ajustement du modèle, l'effet des covariables n'est pas robuste au changement de loi de distribution. Le choix de la bonne forme de la fonction de risque de base est donc essentielle pour mettre en évidence des facteurs prédictifs de l'évolution d'un processus. Ce choix est d'autant plus important que la proportionnalité des risques est aussi radicalement changée. Ces résultats justifient donc pleinement notre première généralisation du modèle.

Les résultats obtenus permettent aussi de mesurer l'intérêt de la prise en compte d'un vecteur de covariables spécifique à chaque transition. Cette approche permet de diminuer le nombre de paramètres inutiles. Dans un même modèle multivarié, le nombre de paramètres peut alors être plus important. Cette prise en compte plus complète d'éventuels facteurs de confusion ou d'interaction est essentielle dans la construction d'un modèle prédictif abouti.

Le troisième apport de cette étude est le choix de forme d'une distribution spécifique à chaque transition $i \rightarrow j$, $\forall i \neq j$. Ainsi, nous avons pu faire un mélange de type Weibull et exponentiel, où 4 des 10 transitions répondent à une loi exponentielle. Cette simplification permet, de plus, une meilleure interprétation des résultats par les cliniciens, comme l'amélioration ou la dégradation du pronostic de la maladie au cours du temps de séjour dans un état.

Ce travail laisse place à de nombreuses perspectives. Tout d'abord, il serait intéressant de comparer ce type de modèle entièrement paramétrique à des approches semi–paramétriques où aucune hypothèse n'est faite sur la fonction de risque de base. Ces méthodes, fondées sur la vraisemblance partielle, ne permettent cependant pas de calculer les fonc-

tions de risque, seul l'effet des facteurs est interprétable.

Pour les modèles paramétriques, le choix de la fonction de risque constitue une étape majeure. Pour ce travail, nous avons bénéficié d'une base de données conséquente permettant d'estimer de nombreux paramètres, et ainsi de tester l'efficience de modèles complexes. Cette stratégie n'aurait pas été si aisée avec un échantillon de petite taille. Le choix de distribution a priori est une question importante sur un plan plus méthodologique. L'estimation non–paramétrique des fonctions de risque pourrait répondre en partie à une telle problématique.

Nous avons aussi abordé à plusieurs reprise l'hypothèse de proportionnalité des risques. Celle–ci nous a contraint à ne pas prendre en compte certaines covariables pour des transitions spécifiques. D'autres stratégies de prise en compte des covariables existent. En effet, nous avons toujours supposé :

$$S_{ij}(t,z) = \Big(S_{ij}(t)\Big)^{\eta(z)} \text{ avec } \eta(z) = exp(\beta z)$$

Cette hypothèse de la forme donnée à $\eta(z)$ pourrait être différente de la fonction exponentielle. De même, la relation entre la survie et la fonction de covariable peut être changée. Citons par exemple les formes Odds–Ratio proportionnelles du type :

$$S_{ij}(t,z) = \frac{1}{\frac{\eta(z)}{S_{ij}(t)} + 1 - \eta(z)}$$

Annexes

Annexe 1 : Vraisemblance basée sur les fonctions de risque

$$
\begin{aligned}
L &= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ S_{X^h_{r-1}}.(T_{h,r} - T_{h,r-1}) \alpha_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \\
&\quad \times \Big\{ S_{X^h_{r-1}}.(V_{h,r-1} - T_{h,r-1}) \Big\}^{1-I\{T_{h,r}=V_{h,r-1}\}} \Big] \\
&= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ \alpha_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \\
&\quad \times S_{X^h_{r-1}}.(min(T_{h,r}, V_{h,r-1}) - T_{h,r-1}) \Big] \\
&= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ \alpha_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \\
&\quad \times exp\Big(- \int_{T_{h,r-1}}^{min(T_{h,r},V_{h,r-1})} \alpha_{X^h_{r-1}}.(u - T_{r-1}) du \Big) \Big] \\
&= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ \alpha_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \Big] \\
&\quad \times exp\Big(- \sum_{r=1}^{m_h} \int_{T_{h,r-1}}^{min(T_{h,r},V_{h,r-1})} \alpha_{X^h_{r-1}}.(u - T_{r-1}) du \Big) \\
&= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ \alpha_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \Big] \\
&\quad \times exp\Big(- \sum_{r=1}^{m_h} \int_{T_{h,r-1}}^{min(T_{h,r},V_{h,r-1})} \sum_{k \neq X^h_{r-1}} \alpha_{X^h_{r-1},k}(u - T_{r-1}) du \Big) \\
&= \prod_h \Big[\prod_{r=1}^{m_h} \Big\{ \alpha_{X^h_{r-1},X^h_r}(T_{h,r} - T_{h,r-1}) \Big\}^{I\{T_{h,r}=V_{h,r-1}\}} \Big] \\
&\quad \times exp\Big(- \sum_{r=1}^{m_h} \sum_{k \neq X^h_{r-1}} \int_{T_{h,r-1}}^{min(T_{h,r},V_{h,r-1})} \alpha_{X^h_{r-1},k}(u - T_{r-1}) du \Big)
\end{aligned}
$$

Annexe 2 : Fonctions associées à la loi exponentielle

Comme nous l'avons déjà défini, la fonction de risque d'une loi exponentielle est constante. Pour retrouver directement sa forme à partir de la loi de Weibull, elle sera notée :

$$\lambda_{ij}(t) = \frac{1}{\sigma_{ij}}$$

En respectant la proportionnalité des risques, nous avons :

$$\lambda_{ij}(t,z) = \frac{1}{\sigma_{ij}} exp\Big(\beta_{ij}^T z_{ij}\Big)$$

Il en découle la fonction de survie :

$$\begin{aligned} S_{ij}(t,z) &= exp\Big(-\int_0^t \lambda_{ij}(u,z)du\Big) \\ &= exp\Big(-\int_0^t \frac{1}{\sigma_{ij}} exp\Big(\beta_{ij}^T z_{ij}\Big)du\Big) \\ &= exp\Big(-\frac{1}{\sigma_{ij}} exp\Big(\beta_{ij}^T z_{ij}\Big)\int_0^t du\Big) \\ &= exp\Big(-\frac{t}{\sigma_{ij}} exp\Big(\beta_{ij}^T z_{ij}\Big)\Big) \\ &= exp\Big(-\frac{t}{\sigma_{ij}}\Big)^{exp\Big(\beta_{ij}^T z_{ij}\Big)} \end{aligned}$$

Enfin, la densité est donnée par :

$$\begin{aligned} f_{ij}(t,z) &= \lambda_{ij}(t,z)\ S_{ij}(t,z) \\ &= \frac{1}{\sigma_{ij}} exp\Big(\beta_{ij}^T z_{ij}\Big)\ exp\Big(-\frac{t}{\sigma_{ij}}\Big)^{exp\Big(\beta_{ij}^T z_{ij}\Big)} \end{aligned}$$

Annexe 3 : Modèle semi–Markovien stratifié de type Weibull

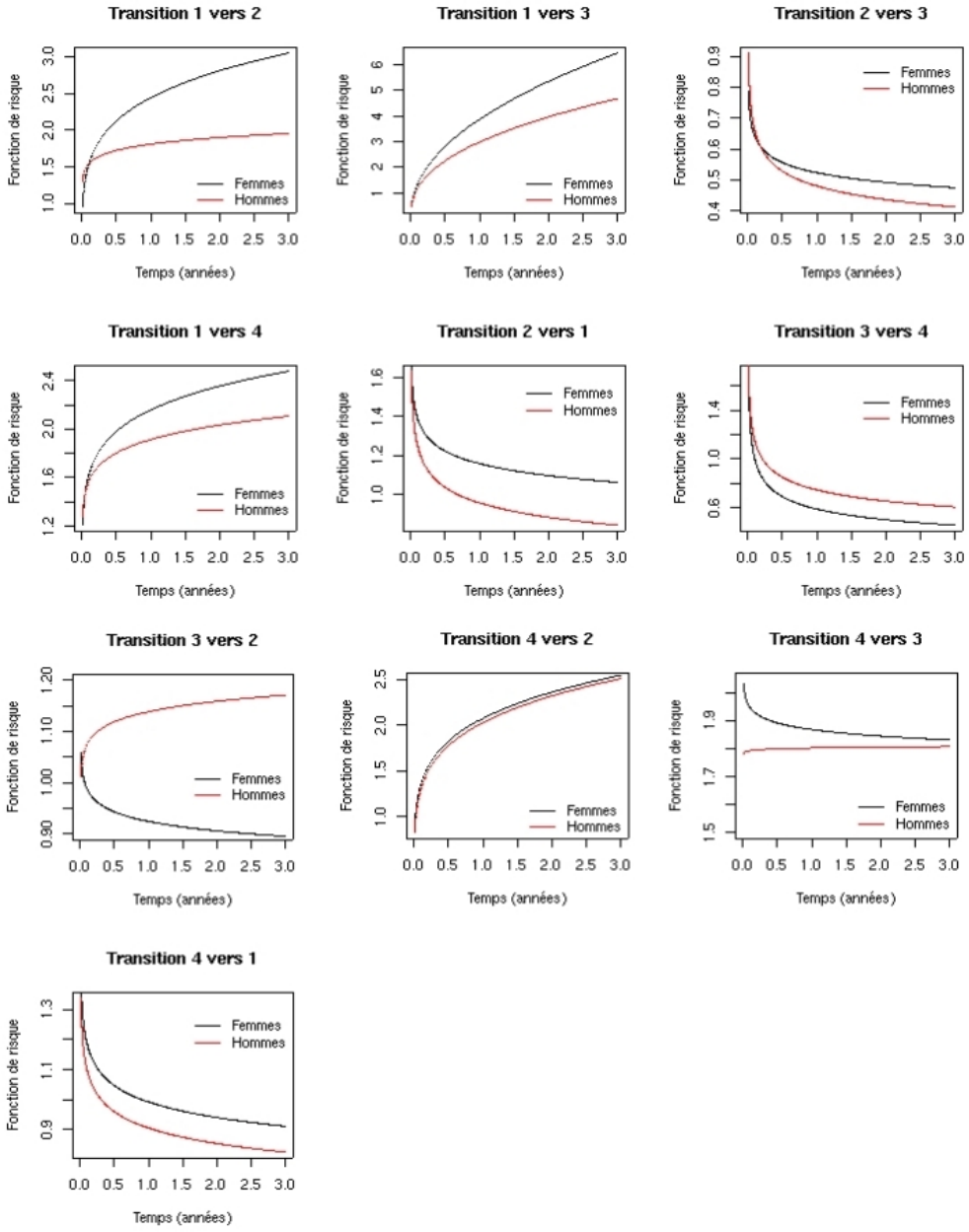

FIG. 3.2 – Fonctions de risque de type Weibull par transition et selon le sexe

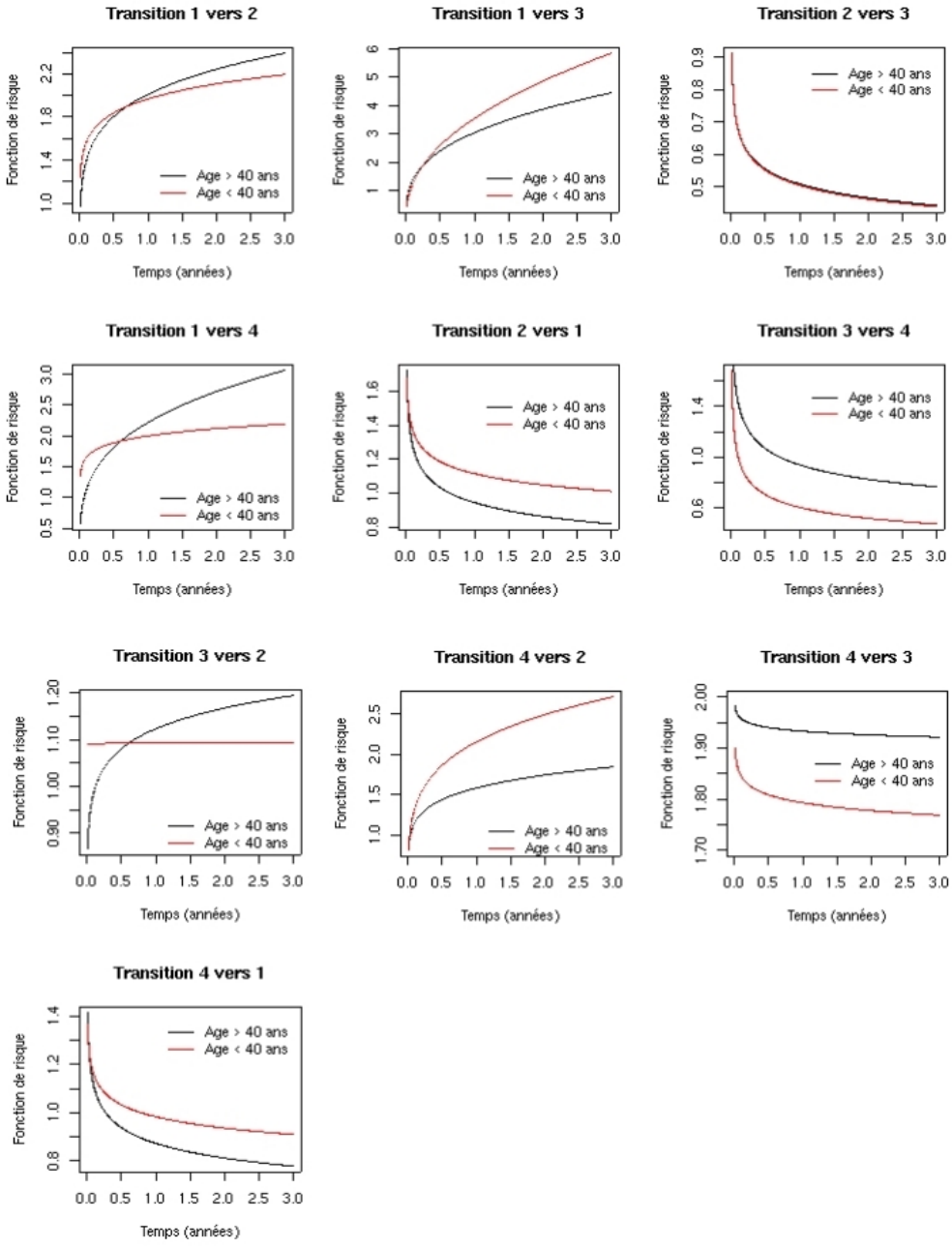

FIG. 3.3 – Fonctions de risque de type Weibull par transition et selon l'âge

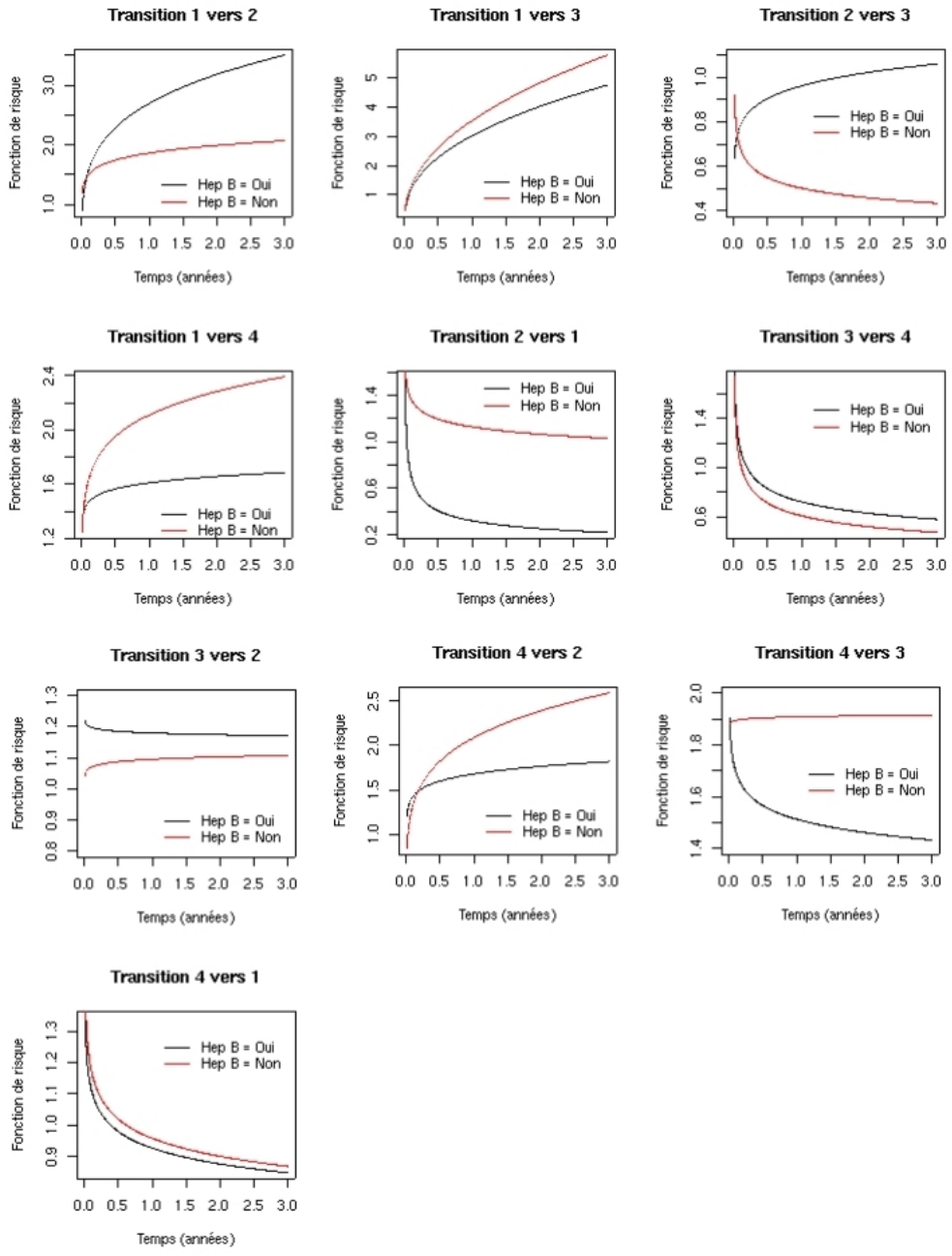

FIG. 3.4 – Fonctions de risque de type Weibull par transition et selon la coinfection par hépatite B

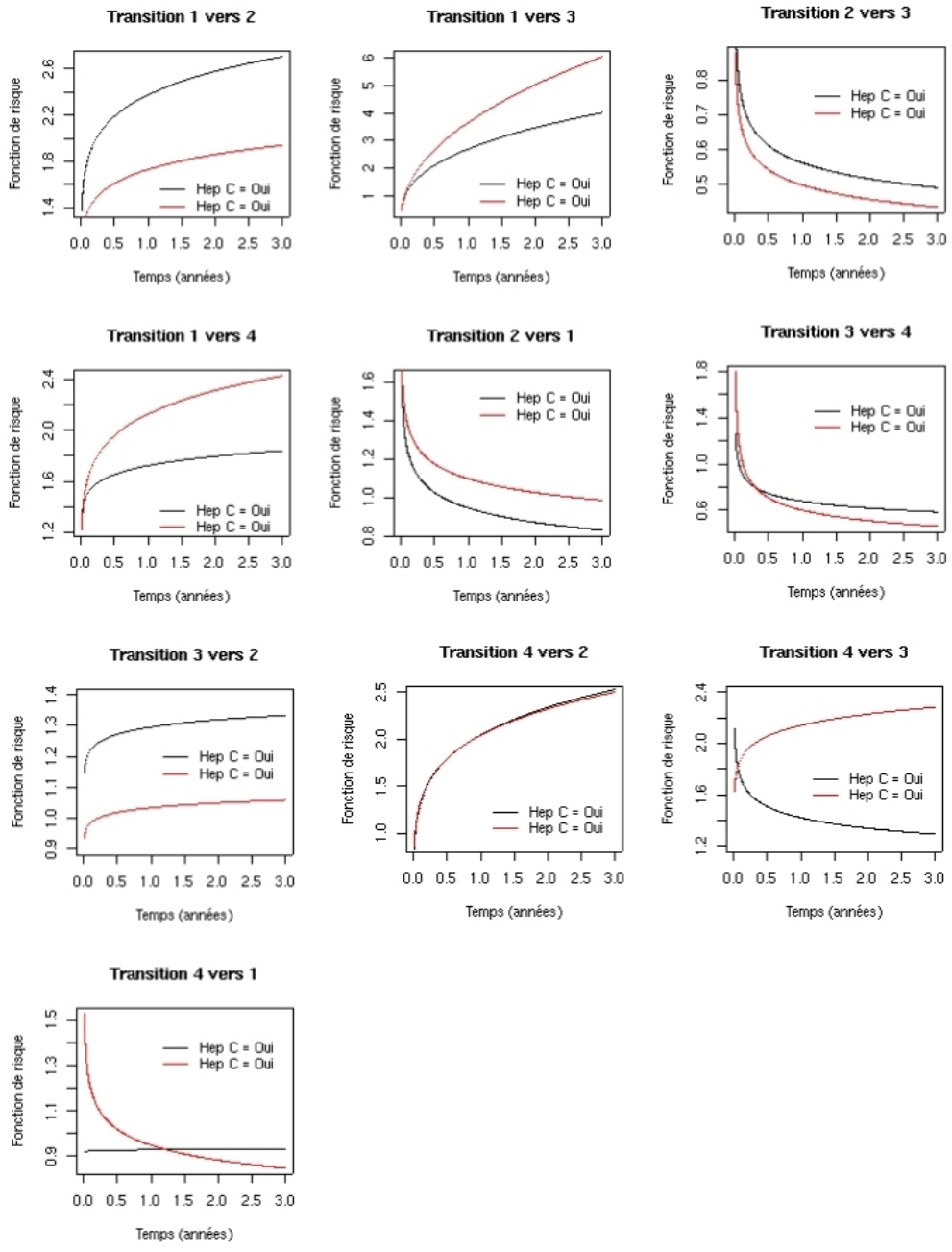

FIG. 3.5 – Fonctions de risque de type Weibull par transition et selon la coinfection par hépatite C

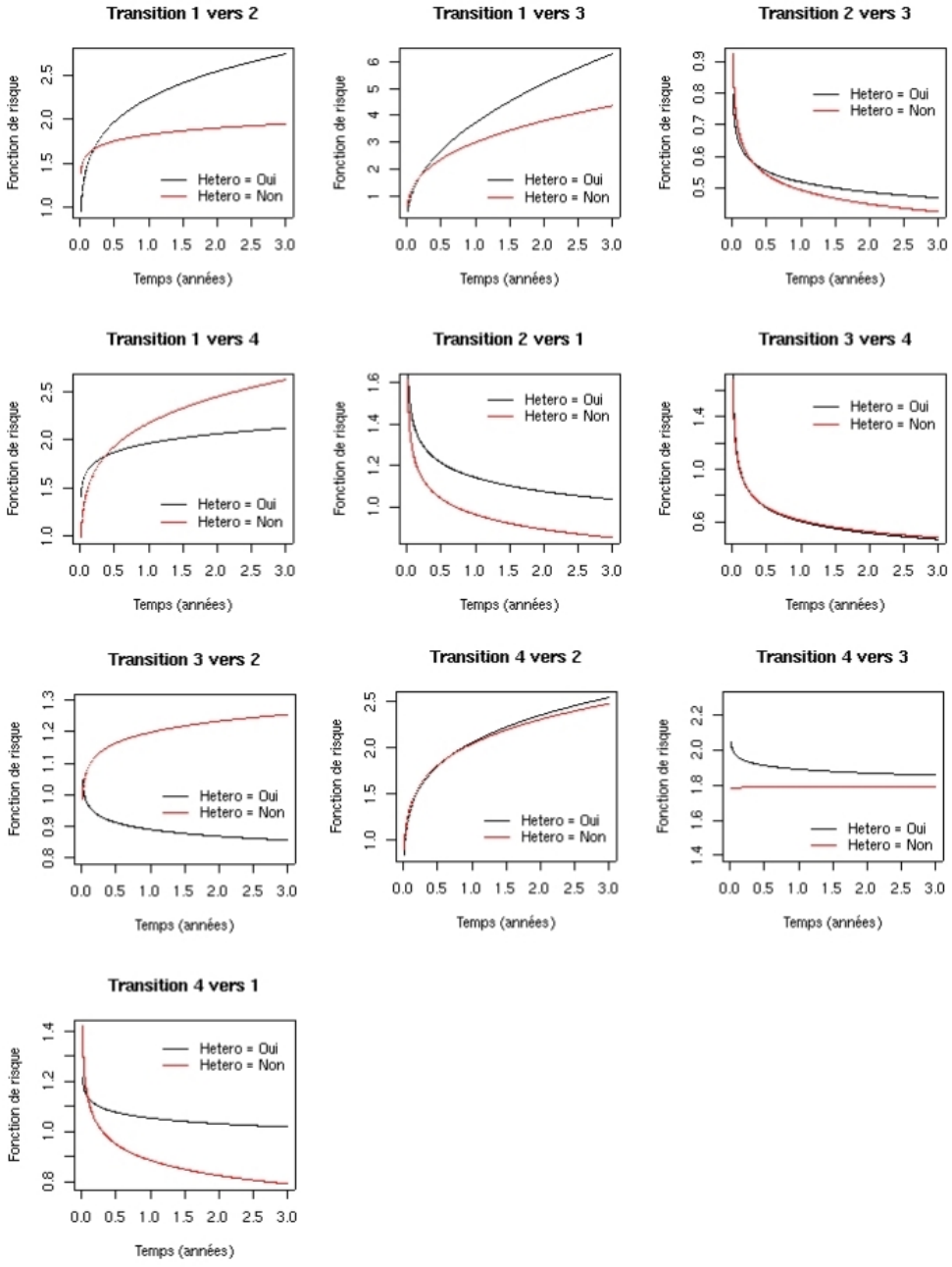

FIG. 3.6 – Fonctions de risque de type Weibull par transition et selon un mode de contamination hétérosexuelle

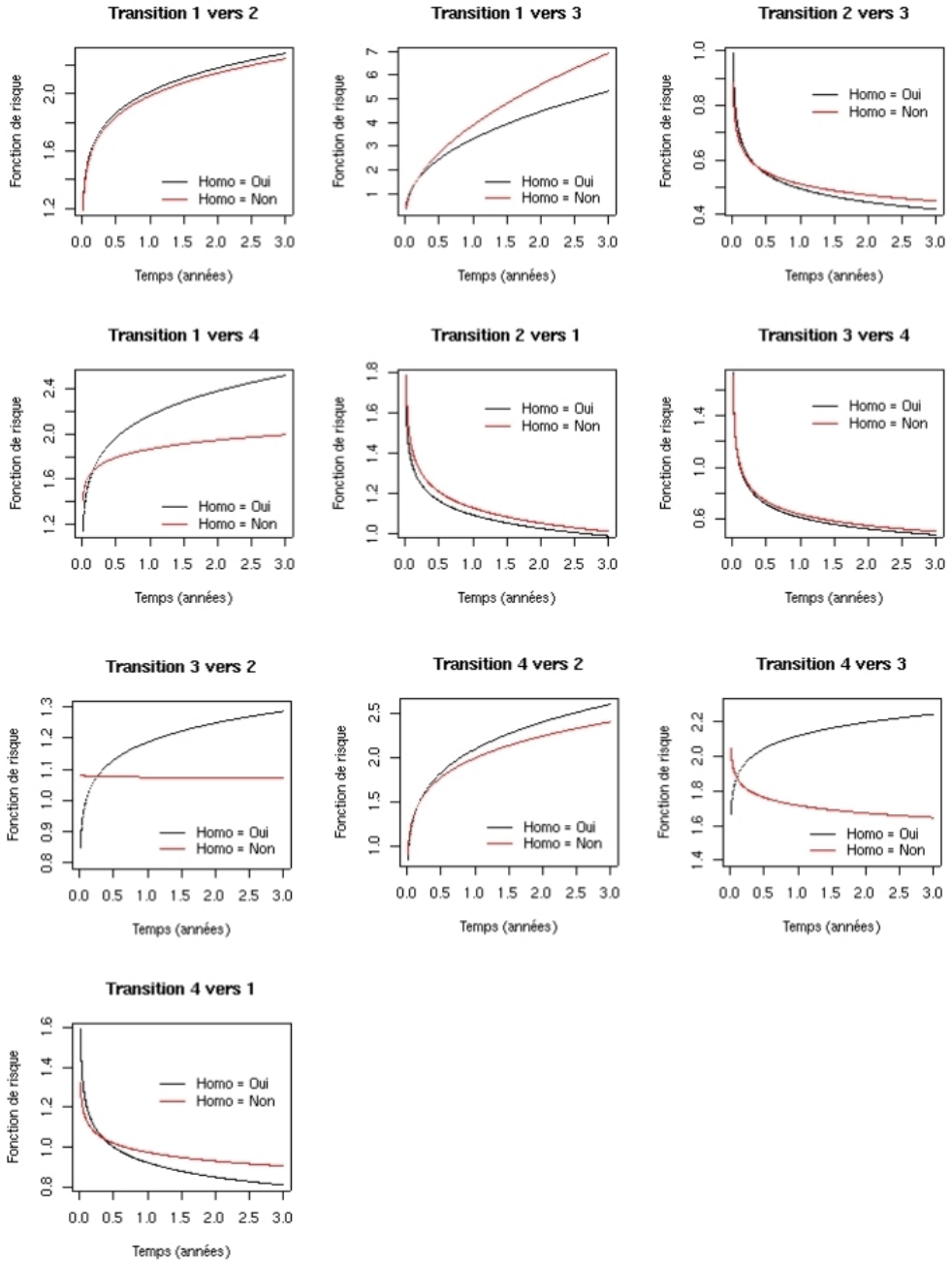

FIG. 3.7 – Fonctions de risque de type Weibull par transition et selon un mode de contamination homosexuelle

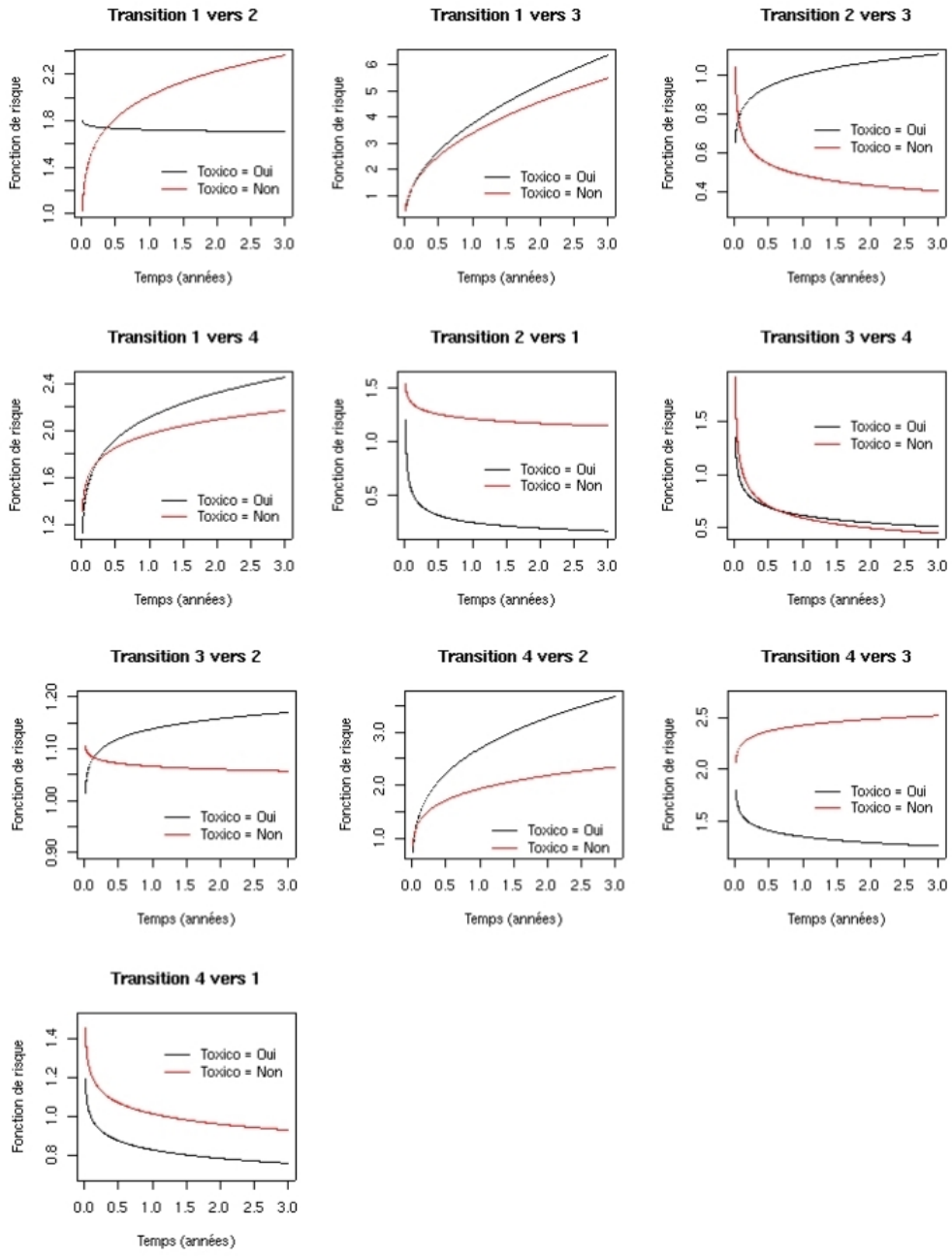

FIG. 3.8 – Fonctions de risque de type Weibull par transition et selon un mode de contamination par toxicomanie

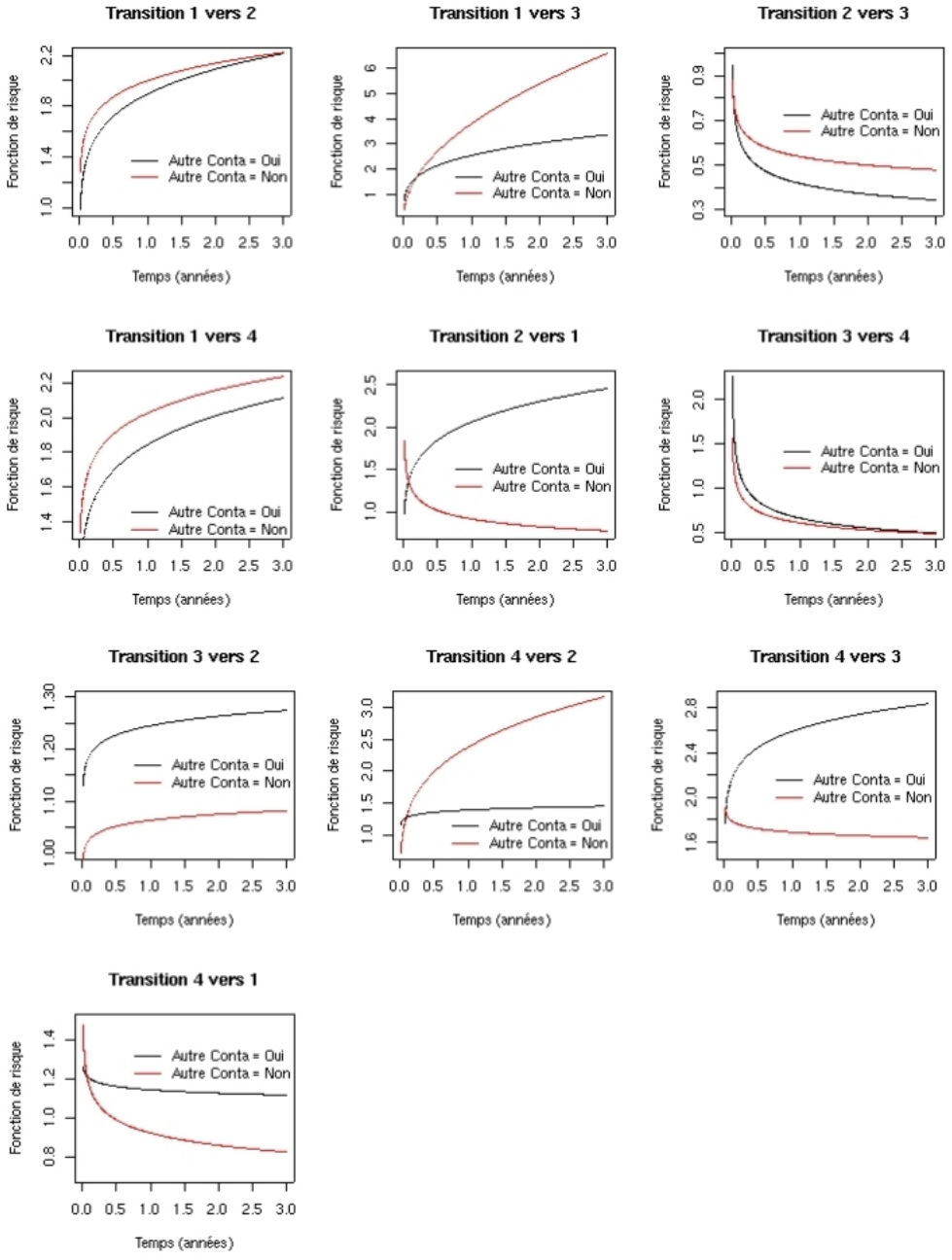

FIG. 3.9 – Fonctions de risque de type Weibull par transition et selon un mode de contamination autre

Annexe 4 : Modèle semi–Markovien stratifié de type Weibull généralisé

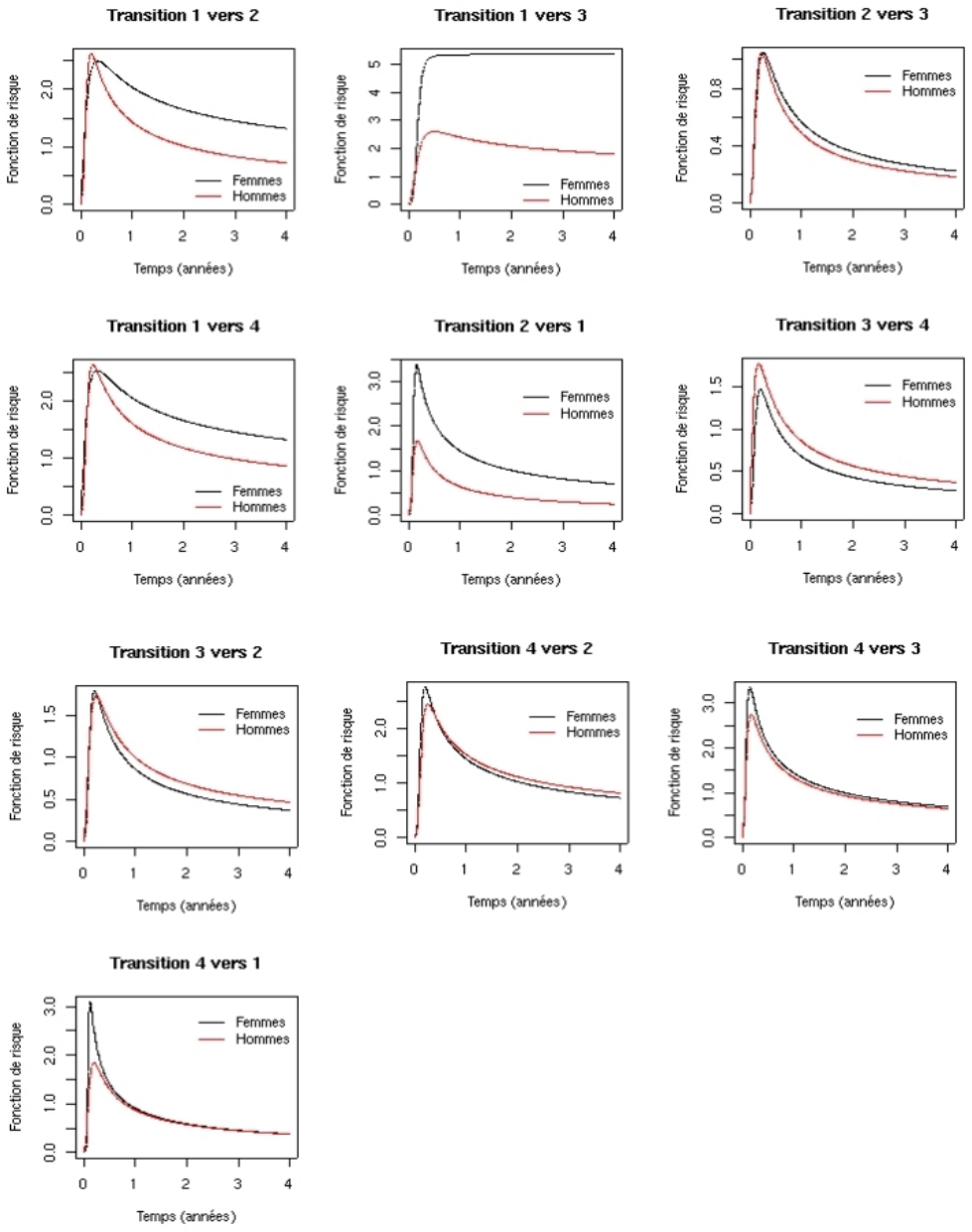

FIG. 3.10 – Fonctions de risque de type Weibull généralisé par transition et selon le sexe

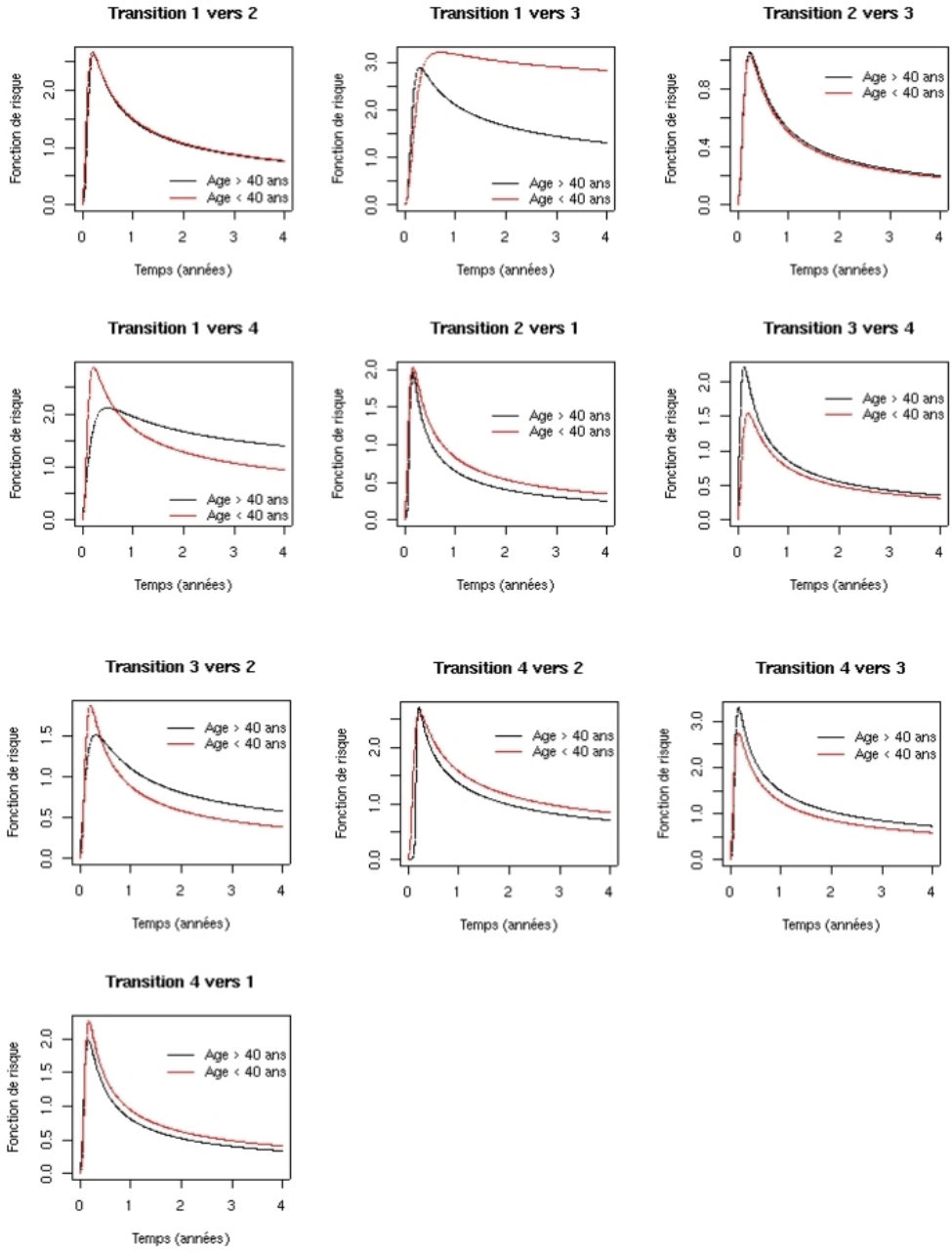

FIG. 3.11 – Fonctions de risque de type Weibull généralisé par transition et selon l'âge

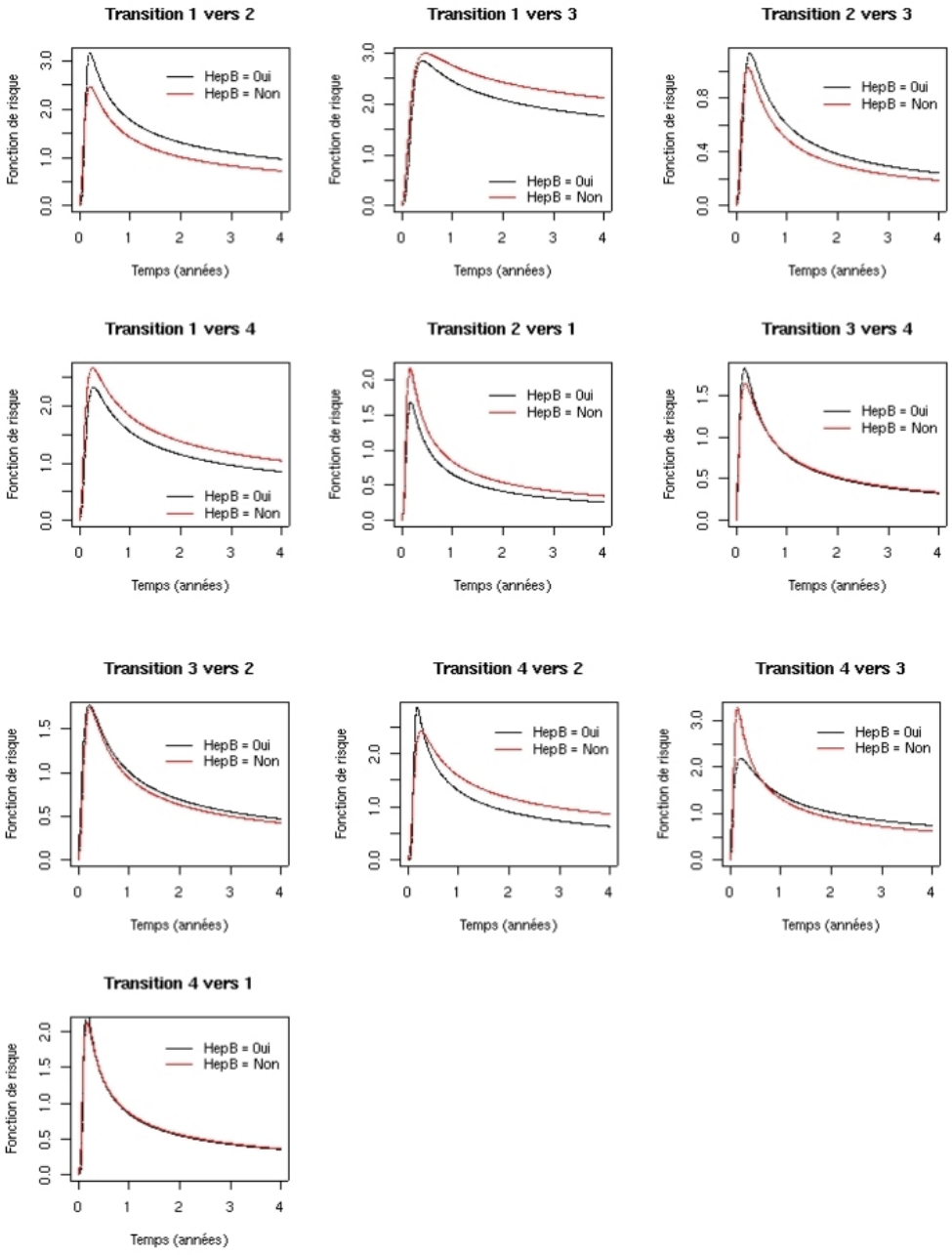

FIG. 3.12 – Fonctions de risque de type Weibull généralisé par transition et selon la coinfection par hépatite B

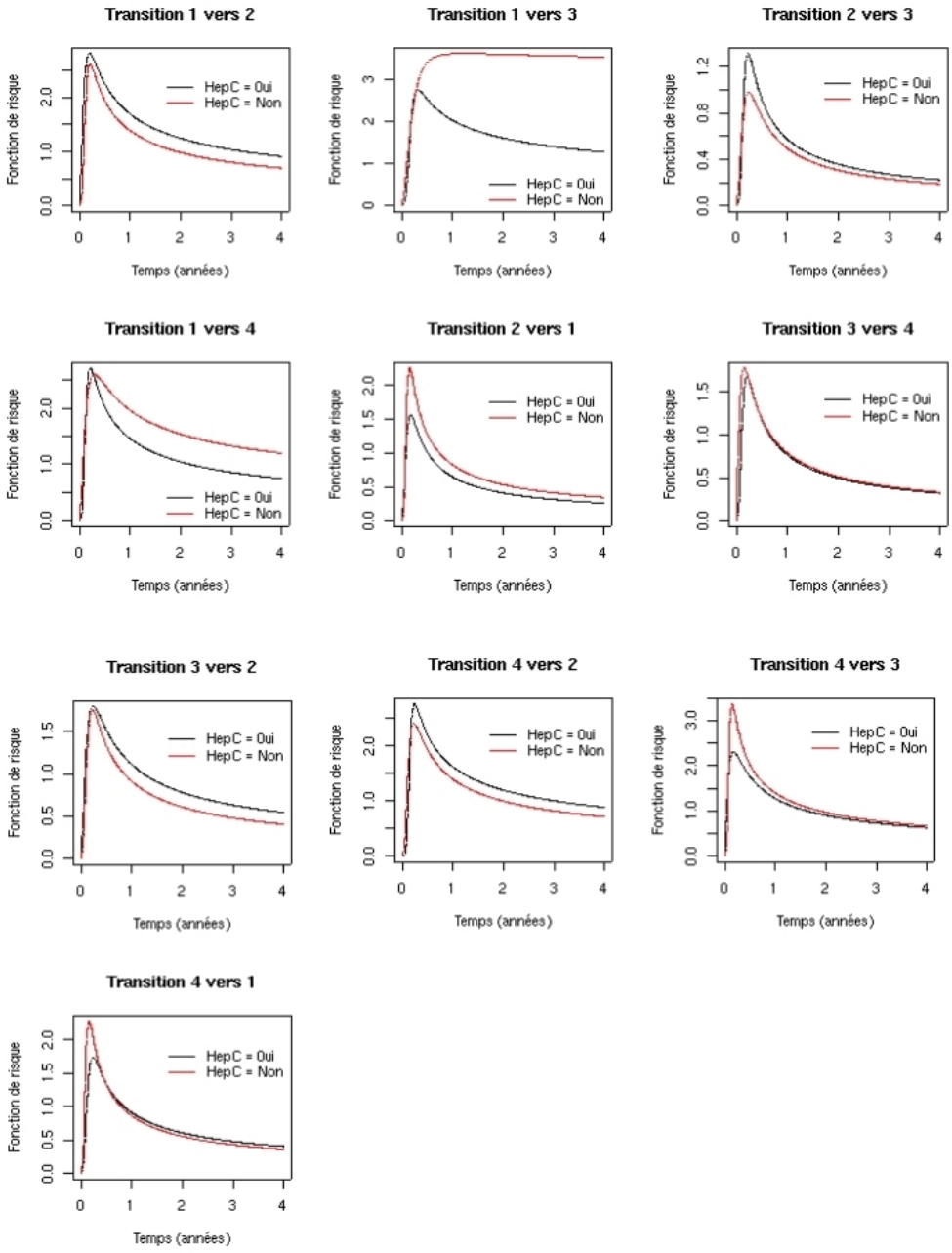

FIG. 3.13 – Fonctions de risque de type Weibull généralisé par transition et selon la coinfection par hépatite C

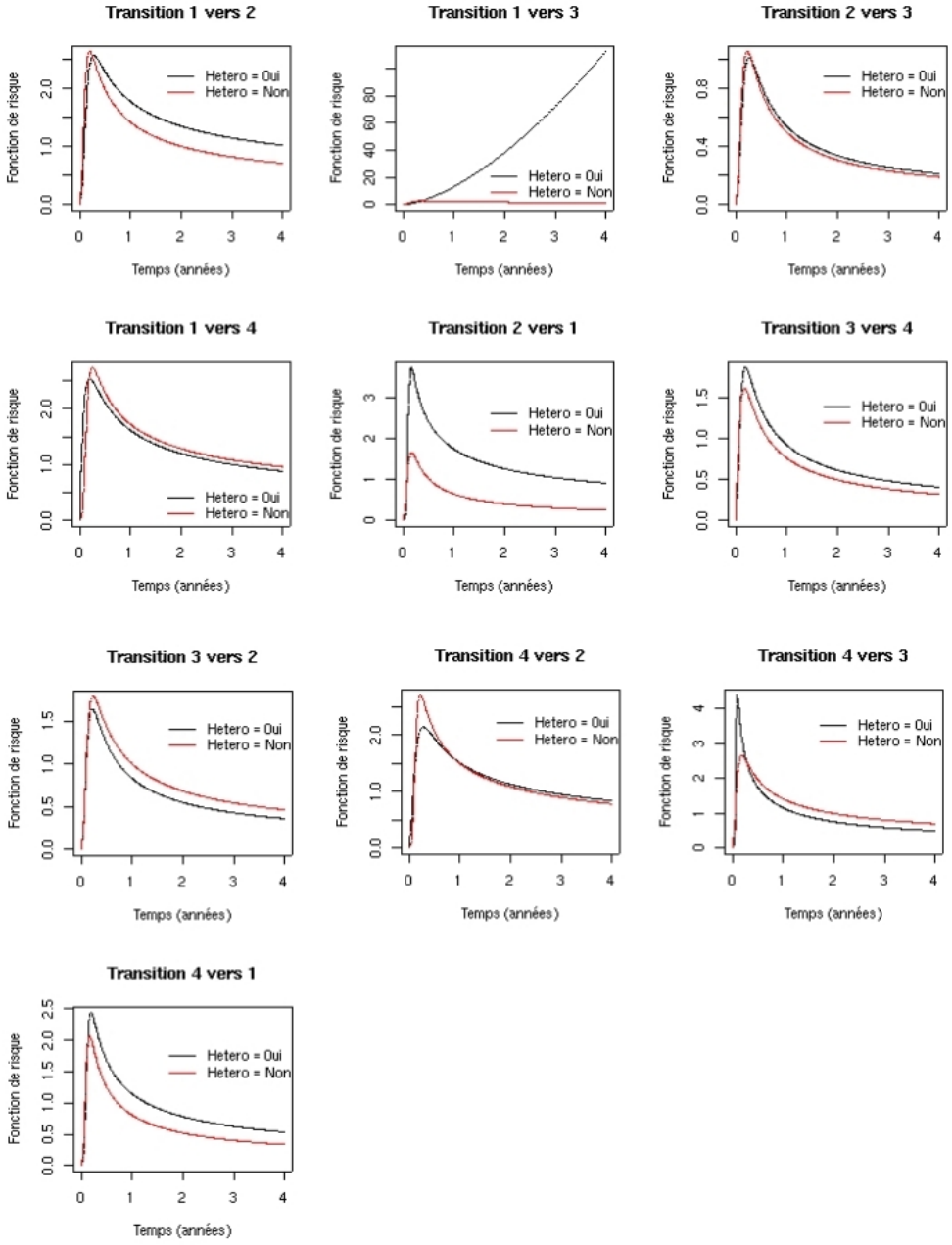

FIG. 3.14 – Fonctions de risque de type Weibull généralisé par transition et selon le mode de contamination hétérosexuelle

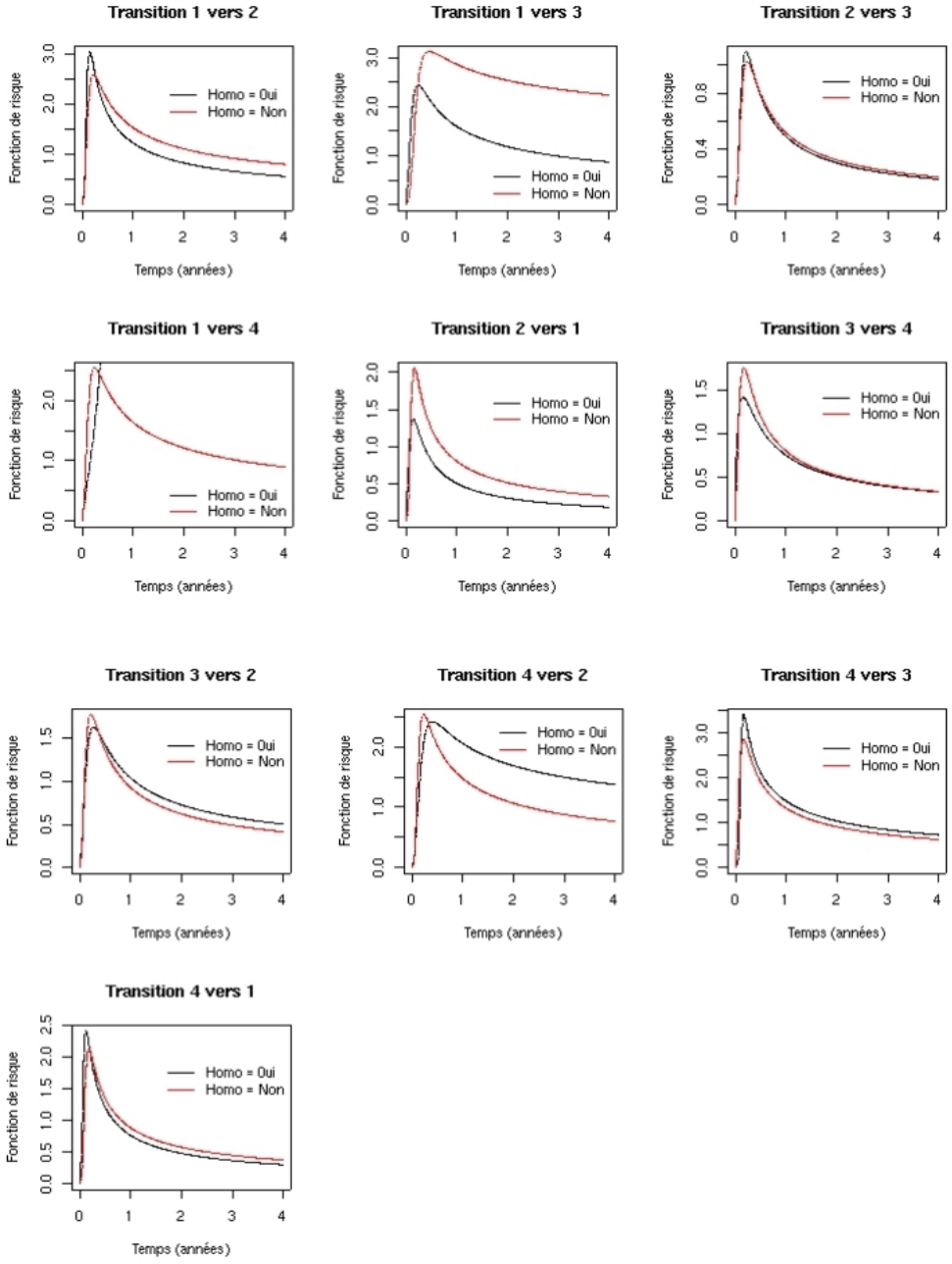

FIG. 3.15 – Fonctions de risque de type Weibull généralisé par transition et selon le mode de contamination homosexuelle

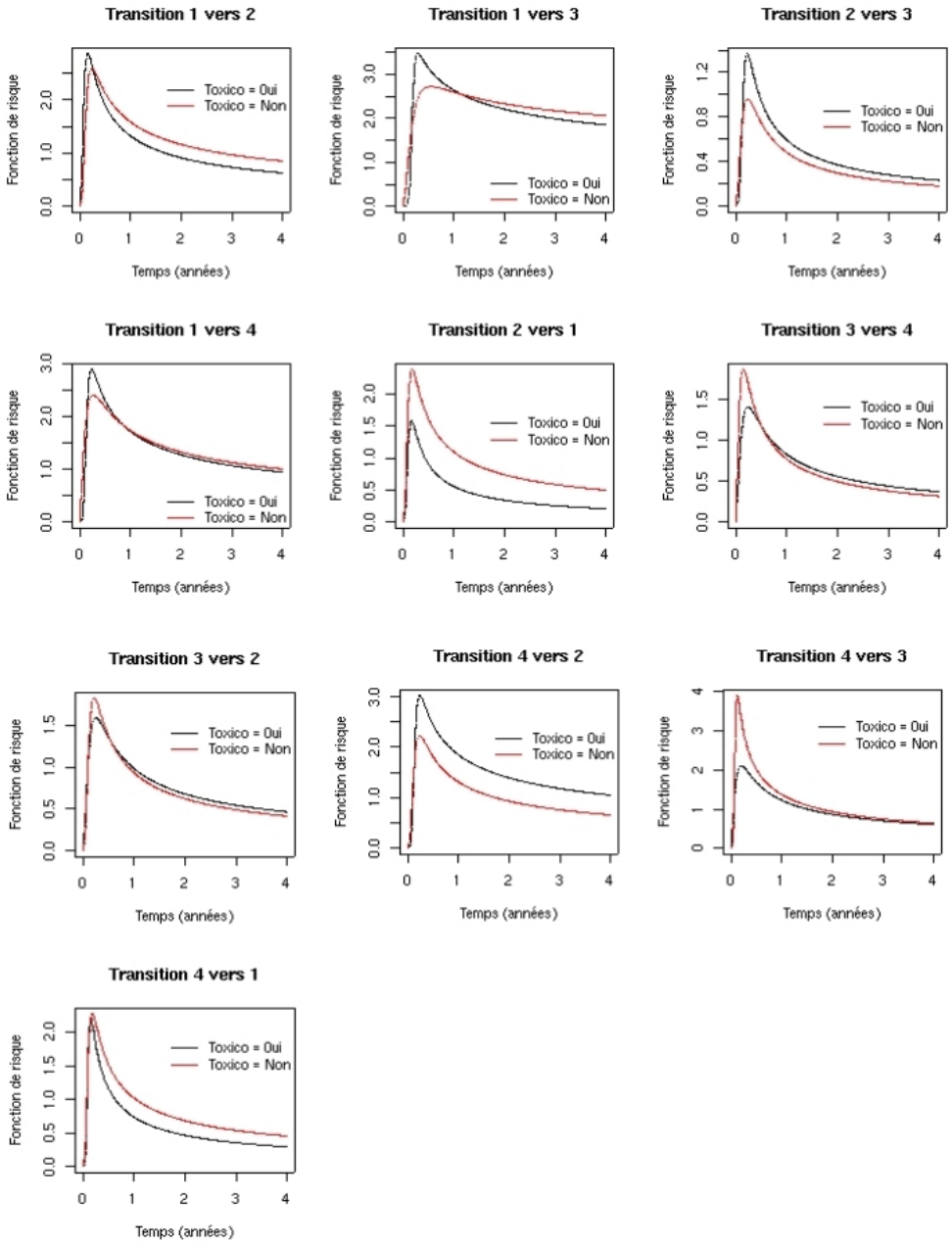

FIG. 3.16 – Fonctions de risque de type Weibull généralisé par transition et selon le mode de contamination par toxicomanie

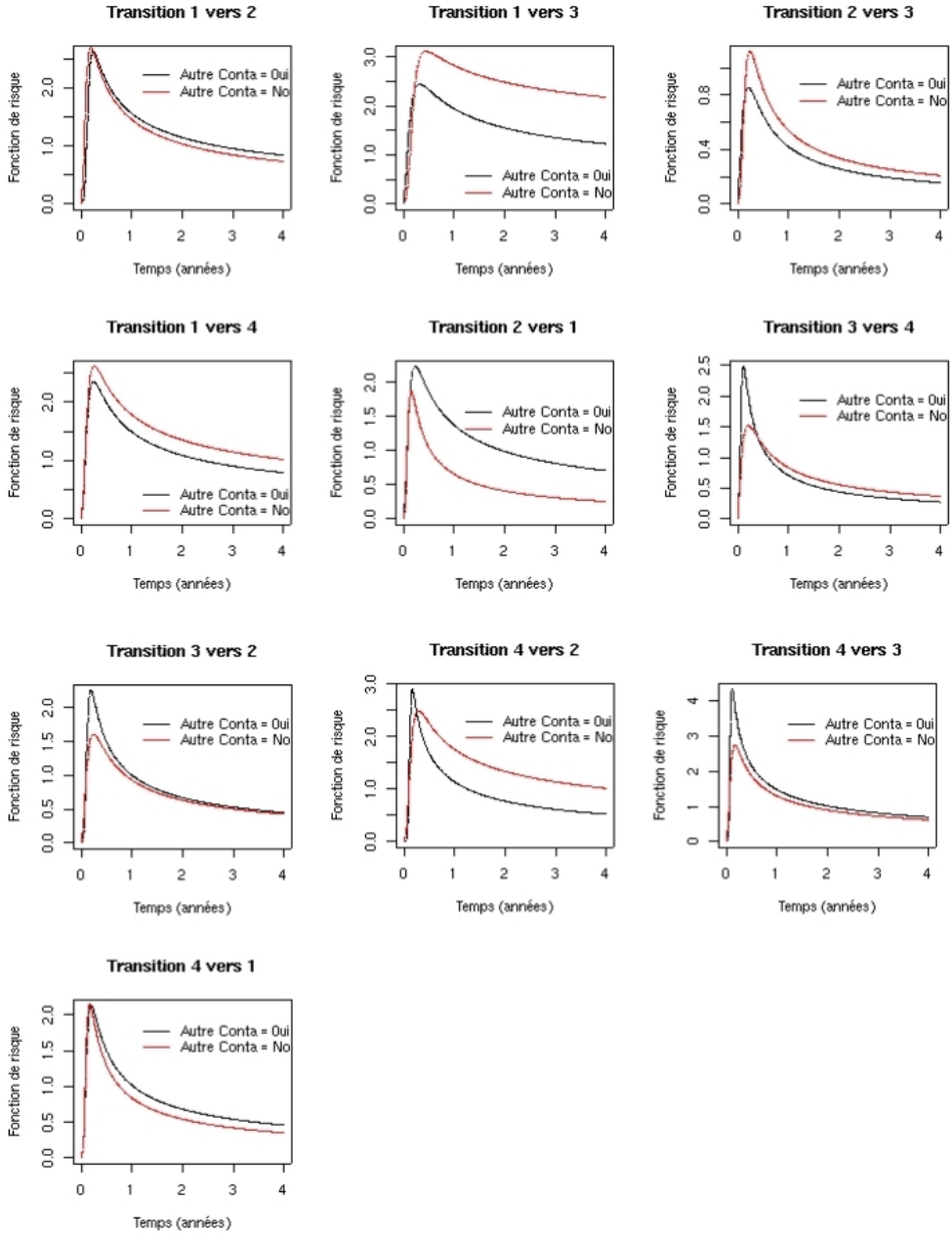

FIG. 3.17 – Fonctions de risque de type Weibull généralisé par transition et selon un mode de contamination autre

Annexe 5 : Modèle semi–Markovien multivarié de type Weibull

Paramètres	Coefficients	Écart–types
ν_{12}	1,11	0,05
ν_{13}	1,45	0,14
ν_{14}	1,10	0,06
ν_{21}	0,91	0,06
ν_{23}	0,87	0,02
ν_{32}	1,03	0,02
ν_{34}	0,78	0,04
ν_{41}	0,91	0,04
ν_{42}	1,19	0,09
ν_{43}	0,99	0,05
σ_{12}	0,59	0,04
σ_{13}	0,55	0,05
σ_{14}	0,58	0,05
σ_{21}	0,79	0,10
σ_{23}	1,88	0,07
σ_{32}	0,98	0,04
σ_{34}	1,84	0,26
σ_{41}	0,95	0,07
σ_{42}	0,63	0,06
σ_{43}	0,57	0,04
P_{12}	0,55	0,02
P_{13}	0,11	0,01
P_{21}	0,10	0,01
P_{32}	0,85	0,01
P_{41}	0,45	0,02
P_{42}	0,16	0,01
β_{32}^{VHC}	0,18	0,07
$\beta_{32}^{co.autre}$	0,18	0,07
β_{34}^{age}	0,51	0,17
β_{34}^{VHB}	0,34	0,18
$\beta_{43}^{co.autre}$	0,28	0,16

Tab. 3.6 – Résultats du modèle semi–Markovien multivarié de type Weibull

Annexe 6 : Modélisation semi–Markovienne de type Weibull généralisé

Transition	Sexe	Age	VHB	VHC	Co.Hétéro.	Co.Homo.	Co.Toxico.	Co.autre
$1 \rightarrow 2$	×		×	×	×			
$1 \rightarrow 3$	× O	×	×			×		×
$1 \rightarrow 4$	×		×	×				×
$2 \rightarrow 1$	×	× O	× O	× O	×	× O	× O	×
$2 \rightarrow 3$			×	×			×	× O
$3 \rightarrow 2$				×	× O			
$3 \rightarrow 4$	× O	× O			×			
$4 \rightarrow 1$		×			×		× O	×
$4 \rightarrow 2$				×			×	
$4 \rightarrow 3$		×						

Tab. 3.7 – Covariables retenues pour l'analyse multivariée après les stratégies stratifiées (×) et univariées (O)

Paramètres	Coefficients	Écart–types
ν_{12}	2,85	0,43
ν_{13}	2,86	0,65
ν_{14}	2,67	0,39
ν_{21}	3,04	0,49
ν_{23}	2,61	0,20
ν_{32}	2,75	0,21
ν_{34}	2,19	0,36
ν_{41}	3,50	0,60
ν_{42}	3,38	0,85
ν_{43}	2,90	0,41
σ_{12}	0,13	0,01
σ_{13}	0,23	0,05
σ_{14}	0,15	0,02
σ_{21}	0,12	0,01
σ_{23}	0,18	0,01
σ_{32}	0,15	0,01
σ_{34}	0,13	0,02
σ_{41}	0,11	0,01
σ_{42}	0,15	0,02
σ_{43}	0,10	0,01
θ_{12}	5,46	1,09
θ_{13}	3,61	1,25
θ_{14}	4,59	0,89
θ_{21}	26,23	6,62
θ_{23}	7,20	0,82
θ_{32}	6,28	0,63
θ_{34}	5,58	1,25
θ_{41}	8,49	1,72
θ_{42}	6,32	2,11
θ_{43}	6,23	1,13
P_{12}	0,55	0,02
P_{13}	0,11	0,01
P_{21}	0,20	0,02
P_{32}	0,86	0,01
P_{41}	0,44	0,02
P_{42}	0,16	0,01
β_{13}^{sexe}	0,58	0,33
β_{21}^{age}	0,65	0,19
β_{21}^{VHB}	0,85	0,22
$\beta_{21}^{co.homo}$	-0,55	0,28
$\beta_{21}^{co.toxico}$	0,44	0,22
$\beta_{23}^{co.autre}$	-0,19	0,09
$\beta_{32}^{co.hetero}$	-0,13	0,06
β_{34}^{sexe}	-0,43	0,19
β_{41}^{Toxico}	-0,28	0,13

Tab. 3.8 – Modèle semi–Markovien multivarié final de type Weibull généralisé

Bibliographie

[1] Kay R. A markov model for analysing cancer markers and disease states and survival studies. *Biometrics*, 42 :855–865, 1986.

[2] Kousignian I, Abgrall S, Duval X, Descamps D, Matheron S, Costagliola D. Modeling the time course of cd4 t–lymphocyte counts according to the level of virologic rebound in hiv–1–infected patients on highly active antiretroviral therapy. *J Acquir Immune Defic Syndr*, 34 :50–57, Sep 2003.

[3] Alioum A, Leroy V, Commenges D, Dabis F, Salamon R. Effect of gender, age, transmission category, and antiretroviral therapy on the progression of human immunodeficiency virus infection using multistate markov models. groupe d'épidemiologie clinique du sida en aquitaine. *Epidemiology*, 9 :605–612, 1998.

[4] Odd O. Aalen, Vernon T. Farewell, Daniela de Angelis, Nicholas E. Day, O. Nöel Gill. A markov model for hiv disease progression including the effect of hiv diagnosis and treatment : Application to aids prediction in england and wales. *Statistics in Medicine*, 16 :2191–2210, Oct 1997.

[5] Jackson C, Sharples L, Thompson S, Duffy S, Couto E. Multistate markov models for disease progression with classification error. *J Royal Statistical Soc D*, 52 :193–193, Jul 2003.

[6] Mauskopf J. Meeting the nice requirements : A markov model approach. *Value Health*, 3 :287–287, Jul 2000.

[7] Boudemaghe T, Daures JP. Modeling asthma evolution by a multi-state model. *Revue d'épidémiologie et de Santé Publique*, 48 :249–255, 2000.

[8] Perez–Ocon R, Ruiz–Castro JE. *Semi–Markov Models and Applications*, chapter 14, pages 229–238. Kluwer Academic Publishers, 1999.

[9] Dabrowska DM, Sun G, Horowitz MM. Cox regression in a markov renewal model : an application to the analysis of bone transplan data. *Journal of the American Statistical Association*, 89 :867–877, 1994.

[10] Karlin S, Taylor HM. *A first course in stochastic processes*, chapter 4. Academic Press, second edition, 1975.

[11] Grimmett GR, Stirzaker DR. *Probability and Random Processes*, chapter 6. Oxford Science Publications, second edition, 1992.

[12] Gill RD. Nonparametric estimation based on censored observations of a markov renewal process. *Zeitschrift Für Wahrscheinlichkeits Theorie Verwandte Gebiete*, 53 :97–116, 1980.

[13] Cox DR. Regression models and life–tables. *J. Roy. Stat. Soc.*, 34 :187–220, 1972.

[14] Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer–Verlag, 1993.

[15] Hill C, Com–Nougué c, Kramar A, Moreau T, O'Quiegley J, Senoussi R, Chastang C. *Analyse statistique des données de survie*. Médecine–Science Flammarion, seconde edition, 1996.

[16] Mudholkar GS, Srivastava DK, Freimer M. The exponetial weibull family : A reanalysis of the bus-motor–failure data. *Technometrics*, 37 :436–445, 1995.

[17] Kaplan EL, Meier P. Non–parametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53 :457–481, 1958.

Résumé

Dans l'étude de l'évolution des patients atteints d'une pathologie chronique, les modèles multi–états de type Markovien connaissent un essor singulier. Cependant, les modèles classiques homogènes ne correspondent pas forcément à la réalité clinique, où les forces de transition entre états ne sont pas forcément constantes. La théorie semi–Markovienne permet de définir explicitement les lois de temps de séjour dans les états et offre ainsi une alternative intéressante. Ce mémoire a pour objectifs d'expliciter cette approche et de l'appliquer à la dynamique des patients séropositifs. Il proposera aussi une stratégie d'analyse des données et apportera des généralisations à la formulation rencontrée dans la littérature [8, 9]. Nous introduirons en particulier la loi de Weibull généralisée. L'application est fondée sur un échantillon de 1244 patients suivis au CHU de Nice, inclus dans la cohorte NADIS.

Mots–clés : Modèle semi–Markovien, temps de séjour, semi–proportionnalité des risques, loi de Weibull généralisée, VIH.

Abstract

Markov multi–states models are inceasing in studies of patients evolution infected by chronic diseases. However, homogeneous classical models don't fit to the clinical reality, where transition forces between states are not constants. The semi–Markov theory defines explicitly the distribution of backward reccurence times and offers an intersting alternative. The objective of this work is to explain this approch and apply it to the dynamic of seropositiv patients. It will also propose a data analysis strategy and will bring generalisations to the literature formulation. We will introduce the generalized Weibull law. The application is based on a sample of 1244 patients from Nice Hospital, included in NADIS cohort.

Key–words : Semi–Markov model, backward reccurence times, semi–hazard proportionnality, generalized Weibull, HIV.