Thèse de doctorat de l'Université Montpellier 1

Formation Doctorale : Biostatistique

Ecole Doctorale : Information, Structures, Systèmes

présentée et soutenue publiquement par

Yohann Foucher

Pour obtenir le grade de

Docteur de l'Université Montpellier 1

Sujet de thèse:

Modèles semi-markoviens : Application à l'analyse de l'évolution de pathologies chroniques

soutenue le 04 Octobre 2007

devant le jury composé de:

M. Ahmadou ALIOUM	Professeur à l'Université de Bordeaux 2	Rapporteur
M. Jean-Pierre DAURES	Professeur à l'Université de Montpellier 1	Directeur
M. Gilles DUCHARME	Professeur à l'Université de Montpellier 2	Président du jury
Mme. Magali GIRAL	Praticien Hospitalier au CHU de Nantes	Examinateur
M. Paul LANDAIS	Professeur à l'Université de Paris 5	Rapporteur
M. Pierre LOMBRAIL	Professeur à l'Université de Nantes	Examinateur
M. Jean-Paul SOULILLOU	Professeur à l'Université de Nantes	Co-directeur
M. Norbert VICTOR	Professeur à l'Université d'Heidelberg	Rapporteur

Je dédie cette thèse à la mémoire de mon grand-père,
Edouard Foucher (1921-2005)

Remerciements

En premier lieu, je tiens à remercier Jean-Pierre Daurès, Professeur à l'Université de Montpellier 1. Il s'agit d'un homme dont le dévouement professionnel reste hors norme. J'ai beaucoup appris à ses côtés durant ces trois années de thèse, tant par ses conseils que par son dynamisme général. Jean-Pierre Daurès m'a aussi offert une opportunité d'évolution après une période difficile. Il m'a accueilli dans son équipe en me proposant d'excellentes conditions de travail. Pour toutes ces raisons, cette thèse est un peu la sienne. J'espère que ce travail ne reste qu'une introduction à une collaboration plus longue.

Je tiens aussi à témoigner ma reconnaissance à Magali Giral, Docteur au CHU de Nantes. Son implication dans ce projet a été déterminante. Avec Jean-Paul Soulillou, Professeur à l'Université de Nantes, ils ont soutenu ces développements méthodologiques, conscients de l'intérêt que ces derniers peuvent apporter à la recherche clinique.

Merci aux Professeurs Ahmadou Alioum, Gilles Ducharme, Paul Landais, Pierre Lombrail et Norbert Victor pour me faire l'honneur de constituer le jury.

Plusieurs collègues et amis m'ont offert leur aide. Je pense en particulier à Christel Castelli, Christophe Combescure, Christophe Demattei et Vanessa Rousseau. La qualité de ce travail a aussi été améliorée grâce aux corrections minutieuses de Nadège Dossat, Carole Gorichon, Laurent Molinier, Gaëlle Pédrono et Sandrine Soulier. Merci à tous.

Je ne serais oublier tous ceux que j'ai côtoyés à l'IURC. Je désire vous témoigner mes sentiments les plus chaleureux.

Toute ma gratitude revient aussi à mes proches. Cette thèse n'aurait pas eu lieu sans le soutien de mes parents, ils ont toujours été présents. Chantal et Paul, je sais vos sacrifices et je vous en serais éternellement reconnaissant.

Enfin, mes ultimes remerciements vont à Céline. Tu as toujours été un soutien important et tu m'as toujours suivi dans mes choix. Avec tout mon amour, merci.

Résumé

L'étude de l'évolution du pronostic de santé d'un patient constitue un domaine important en recherche clinique. Récemment, le développement des modèles multi-états a permis d'étudier cette dynamique en prenant en compte plusieurs états de santé. Dans ce manuscrit, nous utilisons plus particulièrement les modèles semi-markoviens. Ce type de processus distingue les temps de séjour dans les états et les trajectoires des transitions, contrairement à l'approche markovienne classique. Nous avons proposé plusieurs adaptations pour pouvoir appliquer ce type de modèle : la censure par intervalle, le choix des distributions des temps d'attente et l'introduction des covariables. Un test d'adéquation est aussi proposé pour vérifier l'hypothèse de stationnarité. Enfin, une méthode originale, incluant la théorie des courbes ROC, est présentée pour définir des états de santé pertinents au regard du pronostic. Ces développements sont principalement appliqués à une cohorte de patients greffés rénaux (base de données DIVAT).

Abstract

The study of the evolution of a patient constitutes an important field in clinical research. Recently, the development of the multi-state models allows to study this dynamics by taking into account several health states. In this manuscript, we use the semi-markovian models. This type of process distinguishes the durations in the states and the trajectories of the transitions, contrary to the traditional markovian approach. We proposed several adaptations to apply this type of model: the interval-censoring, the choice of the distributions of the durations and the introduction of the covariates. A goodness-of-fit statistic is also proposed to check the stationnarity assumption. Lastly, an original method, including the theory of the ROC curves, is presented to define relevant health states. These developments are mainly applied to kidney transplant recipient follow-up (DIVAT database).

Table des matières

Liste des tableaux

Table des figures

Introduction

Contexte

L'analyse de survie est un domaine important de la statistique appliquée à la santé. De nombreuses études cliniques s'intéressent en effet à l'analyse des temps d'événements, celles-ci ayant la particularité de faire appel à des données incomplètes, censurées ou tronquées. La méthode multivariée la plus utilisée pour ce type de problématique est le modèle de Cox [1], cette dernière ne fait aucune hypothèse sur la distribution du temps de survie. Une des limites de ce type d'approche est l'étude d'un événement unique. Il peut s'agir du décès de l'individu ou de la première déclaration d'une maladie. Certaines extensions plus récentes peuvent néanmoins prendre en compte la répétition ou le renouvellement de l'événement étudié par l'ajout de termes aléatoires (modèles de fragilité) [2, 3].

Dans ce contexte, les modèles multi-états connaissent un intérêt grandissant. Ils permettent d'étendre les modèles de survie classiques en dissociant l'évolution clinique de la maladie en plusieurs états de santé. En cancérologie par exemple [4], une fois le malade pris en charge, il peut rester en rémission ou bien rechuter. A partir de ces deux états cliniques, le patient peut décéder. Ce type de modèle stochastique correspond ainsi à une réalité clinique et permet donc une approche détaillée de la progression de la pathologie. Pour étudier une telle problématique à l'aide des modèles de survie, une solution classique est de réaliser un modèle pour chaque transition. En reprenant l'exemple précédent, deux modélisations peuvent être réalisées à partir de l'état de rémission afin d'identifier les covariables liées à une rechute et celles liées à un décès. Or, pour le modèle prédictif de la rechute du patient, le décès doit être considéré comme une censure à droite. Il est alors difficile de supposer l'indépendance entre événement et censure. De plus, ce type d'approche stratifiée par événement, ne permet pas de modéliser la trajectoire globale du processus.

Les modèles multi-états répondent à ces difficultés, puisqu'ils permettent l'analyse de processus stochastiques qui à tout temps peuvent occuper un état défini. En médecine, ces états peuvent être par exemple la bonne santé, un niveau de gravité d'une maladie, une rémission ou le décès. Un changement d'état est appelé transition ou événement. Lorsque le processus peut sortir d'un état, ce dernier est dit transitoire. A l'inverse, un état à

partir duquel un processus ne peut pas sortir est dit absorbant. Un modèle de survie peut donc être vu comme un modèle multi-états composé d'un état transitoire et d'un état absorbant. Beck et Pauker [5] ou Aalen et Johansen [6] font partie des premiers à avoir introduit les modèles de Markov dans l'analyse clinique multi-états. On peut aussi citer des applications plus récentes en particulier dans l'étude de pathologies chroniques : le Virus de l'Immunodéficience Humaine [7, 8, 9, 10], l'asthme [11, 12] ou le cancer [13, 14]. Il est probable que les modèles multi-états s'imposent à l'avenir comme une approche plus riche que les modèles de survie [15, 16].

Même si la structure générale états/transitions des modèles multi-états permet une grande diversité de schémas, on retrouve dans la littérature un certain nombre de structures standards [17], illustrées par la figure (1).

Le premier modèle permet d'étudier l'évolution d'un patient atteint d'une maladie irréversible, particulièrement lorsque cette maladie augmente le risque de décès. Il comporte un état absorbant et deux états transitoires. On peut alors comparer le taux de mortalité chez des sujets sains à celui des sujets malades. Ce modèle, une fois ses paramètres estimés, peut être utilisé pour la prévalence d'une maladie à l'échelle d'une population. Certaines hypothèses, en particulier sur le flux entrant dans l'état en bonne santé (natalité par exemple), doivent pour cela être formulées. Ce modèle peut être complété un ajoutant une possibilité de guérison.

La survie bivariée permet l'analyse de deux temps de survie non-indépendants. Cette approche, largement développée par Hougaard [3], peut être étendue au cas des données de survie multivariées. Une telle dépendance peut être due à différents facteurs. Par exemple, si le temps d'évènement est étudié au sein de mêmes familles, certains facteurs non-observés (environnementaux, génétiques, etc.) peuvent créer une dépendance au sein de ces groupes.

Les évènements récurrents peuvent aussi être vus comme un processus multi-états. Pour chaque individu, plusieurs événements identiques sont récoltés, formant ainsi un processus stochastique. Ce modèle comporte uniquement des états transitoires. C'est par exemple le cas pour l'étude de la fertilité chez la femme. Les événements sont alors les naissances successives. A chaque naissance, la femme passe à l'état suivant. Ce schéma ne permet pas de prendre en compte des naissances multiples, mais ce cas particulier pourrait être envisagé en considérant certains sauts d'états. Ce modèle permet ainsi de modéliser une dépendance des fonctions de risque d'avoir un enfant supplémentaire avec le nombre d'enfants déjà nés.

Enfin, le modèle à risques compétitifs peut-être appliqué lorsque plusieurs événements terminaux peuvent se produire, par exemple le décès du patient par différentes causes. Ce modèle possède un état transitoire et plusieurs états absorbants. Ces derniers sont considérés exclusifs, ils sont donc en compétition ou en concurrence [18]. Les intérêts de cette structure sont multiples. Un des avantages majeurs est de pouvoir évaluer simultanément

les intensités liées à chaque événement et l'effet propre de certaines covariables.

Ces structures, même si elles recouvrent beaucoup de problématiques cliniques rencontrées en pratique, peuvent être adaptées à des situations particulières. Par la suite, nous verrons les adaptations choisies pour répondre aux attentes cliniques.

Dans la définition du modèle utilisé, le choix de la structure représente donc un point majeur. L'autre difficulté est le choix de l'échelle de temps. En épidémiologie, où le niveau d'inférence est la population, le temps calendaire est souvent la principale unité. L'utilité de ce type de modèle dynamique est en effet d'estimer certains indicateurs dans un futur proche, par exemple la prévalence d'une maladie en 2010. A l'inverse, on s'intéresse à un niveau individuel, l'objectif étant la modélisation de l'évolution d'un patient atteint d'une pathologie chronique. Bien entendu, les deux niveaux sont étroitement liés. Cependant, l'échelle du temps de base reste le plus souvent différente. Dans notre contexte, ce choix peut varier selon les situations. Beaucoup de travaux choisissent l'âge du patient ou le temps écoulé depuis une date d'origine (inclusion dans l'étude, date de greffe, date d'apparition de la maladie, etc.).

L'hypothèse de modélisation de la fonction de risque est alors fondamentale. De nombreuses études utilisent les outils markoviens homogènes ayant l'inconvénient d'être sans mémoire, l'évolution du processus étant indépendante de la durée dans l'état actuel [4, 19]. Dans le domaine du vivant, cette contrainte est souvent trop forte. Les modèles markoviens non-homogènes permettent de se soustraire à cette hypothèse en modélisant une matrice de probabilités dépendantes du temps.

Dans cette thèse, nous nous intéressons à une approche semi-markovienne complémentaire. Le temps influençant les transitions est alors le temps depuis l'entrée dans l'état en cours. Ce temps est aussi appelé temps d'attente ou temps de séjour. Il est en effet souvent cohérent de supposer que l'évolution d'un patient dépend de la durée depuis laquelle il occupe son état de santé. Par exemple, la probabilité de guérison d'une maladie dépend le plus souvent du temps depuis l'infection. Dans ce type de modèles, la loi du temps de séjour dans l'état est alors explicite [20, 21]. Deux pathologies chroniques nous ont conduit à l'application de cette théorie semi-markovienne, elles sont exposées ci-après.

Problématiques cliniques

Evolution des patients atteints du VIH

Le VIH (Virus de l'Immuno-déficience Humaine) est une pathologie chronique, c'est à dire une maladie sans guérison possible. Ce virus appartient à la famille des rétrovirus. Il infecte une cellule en transcrivant sa molécule d'ARN en une molécule d'ADN viral,

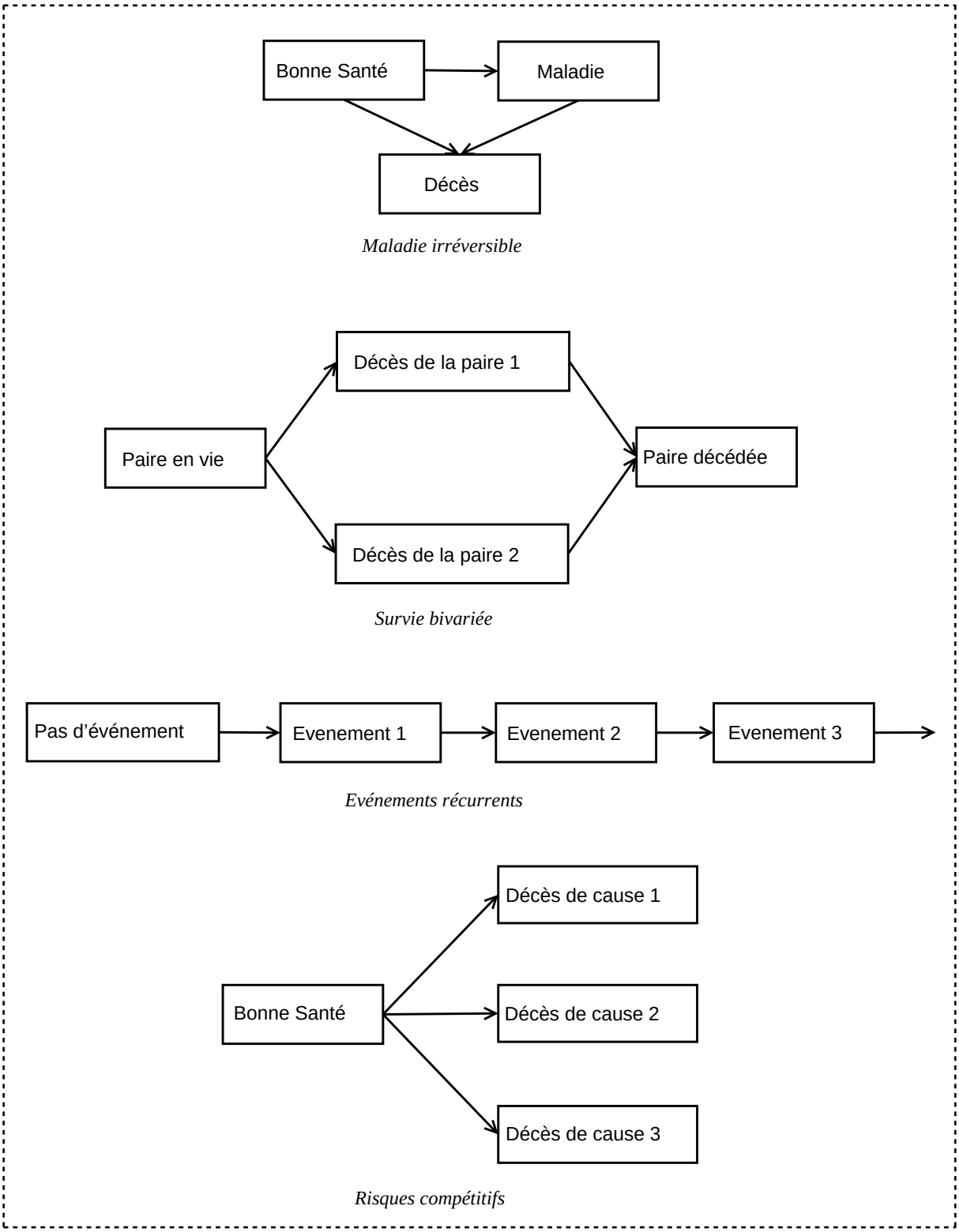

FIG. 1 – Structures standards de modèles multi-états

à partir de l'ADN de la cellule infectée. L'enzyme permettant ce mécanisme est appelée reverse transcriptase. Trois autres caractéristiques sont propres au VIH. D'une part, on peut observer une période de latence où le virus reste silencieux. L'ADN viral s'intègre alors à une extrémité de l'ADN cellulaire et est transmis aux cellules descendantes à chaque mitose. D'autre part, il possède une forte variabilité génétique due à des erreurs fréquentes de réplication. Ce pouvoir mutagène explique l'apparition de résistances aux traitements. Enfin, les lymphocytes T (CD4 et CD8) et les cellules macrophages-monocytes sont les principales cibles du VIH. Ce 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 (maladies dues à des germes habituellement peu agressifs mais qui sont susceptibles de provoquer de graves complications en affectant des personnes ayant un système immunitaire très affaibli).

Deux marqueurs de l'avancement de la maladie sont importants : la charge virale (CV) et la concentration de lymphocytes T CD4. La charge virale représente la quantité de virus dans le sang. Plus sa valeur est importante, moins le pronostic est bon. A l'inverse, le nombre de lymphocytes T CD4 représente la capacité immunologique de l'individu. Plus sa valeur est faible, moins le pronostic de la maladie est bon. De nombreuses études se sont donc attachées à étudier l'évolution quantitative de ces marqueurs de substitution au cours du temps en fonction de différentes covariables [22, 23, 24]. On sait ainsi que ces marqueurs sont liés à certains facteurs comme le mode de contamination, une éventuelle co-infection, le sexe du patient ou son âge.

La problématique posée par les cliniciens est de considérer conjointement et qualitativement les deux marqueurs, identifiant ainsi des états de gravité de l'infection. Ils considèrent quatre états en définissant un seuil pour chacune des deux variables (figure 2). Les transitions possibles entre états y sont aussi représentées. Elles ont été validées par les spécialistes et confirmées par leur représentativité dans la base de données utilisée (NADIS). Cette cohorte observationnelle sera décrite par la suite.

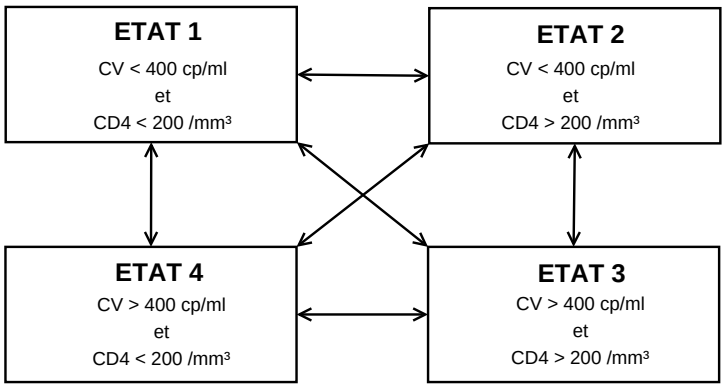

Fig. 2 – Modèle à 4 états transitoires caractérisant la gravité de l'infection VIH

Les états 2 et 4 constituent respectivement le stade le moins avancé et le plus avancé de la maladie. Les états 1 et 3 sont des stades intermédiaires de la maladie entre les deux états précédents. Tous ces états sont transitoires. La question posée est l'étude de l'évolution du patient à travers ces états et la mesure de l'effet de certaines covariables sur ces forces de transition. Dans la base de données, nous disposons de 8 covariables (même si de nombreuses autres sont relevées dans la littérature) : le sexe, l'âge (supérieur ou non à 40 ans), la co-infection par une hépatite C, la co-infection par une hépatite B et le mode de contamination (hétérosexuelle, homosexuelle, par toxicomanie, ou autres).

Ce type de modèle est sensiblement différent de ceux déjà utilisés dans le domaine du VIH. Longini et al. [25] définissent un modèle composé de quatre états transitoires (anti-corps négatif, anti-corps positif mais asymptomatique, symptômes pré-SIDA, SIDA) et un état absorbant (décès dû au SIDA). Ce modèle est devenu moins pertinent avec l'amélioration de la prise en charge des patients infectés par le VIH. Le stade final SIDA est devenu peu fréquent, ainsi que les décès dus à cette infection. Ceci justifie donc de remplacer la symptomatologie par un indicateur de la baisse des défenses immunitaires (CD4). Gentleman et al. [26] proposent ce type de structure pour laquelle les états transitoires sont définis à partir du nombre de CD4 $/mm^3$ (≥ 500, 200-499, < 200) et l'état absorbant est le stade SIDA. Joly et Commenges [27] utilisent un modèle progressif à trois états (non infecté, infecté, stade SIDA) pour décrire l'épidémiologie du virus.

La cohorte utilisée est extraite de la base de données NADIS [28, 29] qui regroupe tous les patients VIH positifs suivis au moins une fois en consultation au CHU de Nice. Le CISIH (Centre d'Information et de Soins de l'Immunodéficience Humaine) a développé le logiciel ADDIS permettant la saisie en temps réel des données par les médecins. Ce logiciel est fonctionnel depuis juin 1994. Ce dossier médical informatisé a fait l'objet d'une totale réécriture, dans un environnement compatible avec les nouvelles technologies de l'information. Ce travail, réalisé avec la participation de cinq autres CHU français (AP-Marseille, Toulouse, Paris-Pitié-Salpétrière, Nantes et Lille) a abouti à la fin de l'année 2000 au logiciel NADIS 2000.

Evolution des patients greffés rénaux

L'insuffisance rénale touche actuellement 2,8 millions de Français. Mais avec l'augmentation du diabète sucré, de l'hypertension et de l'allongement de l'espérance de vie, cette maladie est en forte progression, avec une hausse des cas de 5 à 7% chaque année. L'évolution peut conduire à une insuffisance rénale terminale qui nécessitera un traitement par dialyse ou une greffe rénale.

L'incidence et la sévérité des épisodes de rejets aigus ont été considérablement réduites durant la dernière décennie grâce à l'introduction de nouvelles thérapeutiques immunosuppressives. Une meilleure compréhension et prise en charge des facteurs de risque associés

à la perte des greffons ont par ailleurs permis une augmentation significative de la survie des greffes de rein à 1 an. En revanche, le taux de retour en dialyse sur le long terme n'a pas suivi la même tendance et est resté semblable au cours de ces dernières années. Cependant, l'introduction récente de molécules immunosuppressives comme le mycophenolate mofétyl et le tracrolimus dans le traitement d'entretien des greffes pourrait être à l'origine d'un infléchissement récent et sensible de la perte des greffons à long terme. Aucune étude spécifique ne l'a encore réellement démontré.

La dysfonction chronique, principale cause de retour en dialyse, est caractérisée par une dégradation progressive de la fonction rénale généralement associée à une augmentation de la protéinurie et une diminution de la clairance de la créatinine. Cette dysfonction est appelée rejet chronique dans les années 80-90 puis désignée par le terme néphropathie chronique du transplant, au début des années 2000. Comme pour le nombre de CD4 et la charge virale dans le domaine du VIH, de nombreux travaux tentent ainsi de prédire la clairance de la créatinine et la protéinurie au cours du temps et en fonction de différentes covariables [30, 31]. Parmi les marqueurs d'une aggravation du greffon, la clairance de la créatinine mesurée un an après la greffe est en effet beaucoup utilisée. Elle est calculée en ml/min à partir de la formule MDRD (Modification of Diet in Renal Disease) [32] :

$$CL = 32788 \times creatinine^{-1.154} \times age^{-0.203} \times constante \quad (1)$$

où $constante = 1$ pour les hommes et $constante = 0.742$ pour les femmes.

Si la néphropathie chronique du transplant représente encore la principale cause de perte des greffons, il existe cependant de nombreuses autres raisons liées à l'échec d'une greffe incluant : le décès avec greffon fonctionnel, la thrombose vasculaire en post-greffe immédiat, le rejet aigu (principalement vasculaire) et les autres types d'échec (complications urologiques, infections, récidives de maladie initiale, cancers et lymphomes). Gjertson et al. [33] proposent un modèle à risques compétitifs pour analyser plusieurs types d'échecs de la greffe : le rejet aigu, le rejet chronique, la mort avec un rein fonctionnel, et les autres types d'échec.

L'enjeu actuel de la transplantation est de définir des marqueurs précoces d'un retour en dialyse des patients sur le long terme conduisant à l'échec de la greffe. En reprenant notre problématique, il s'agit de définir des états transitoires d'aggravation du greffon, associés à un excès de risque d'échec de la greffe. Il s'agit aussi d'identifier les facteurs associés à ces forces de transition. Les échecs ont été séparés en deux catégories distinctes : le retour en dialyse et le décès du patient avec un greffon fonctionnel. Les états d'aggravation seront caractérisés par les mesures de protéinurie (PR) et de clairance de la créatinine (CL). La figure (3) présente la structure aggravation/échec ainsi choisie.

Il ne s'agit plus ici de tester un seul paramètre sur un seul événement, mais bien un ensemble de paramètres sur plusieurs événements caractérisant l'évolution du patient greffé rénal. Ces paramètres, qu'ils soient propres aux donneurs ou aux receveurs, sont nombreux. Les principaux sont : l'âge et le sexe du receveur et du donneur, l'ischémie froide

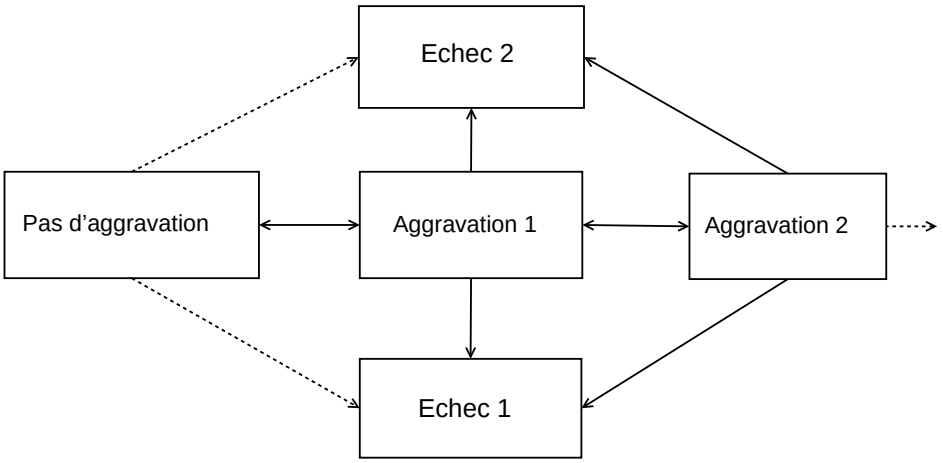

FIG. 3 – Structure aggravation/échec pour l'analyse des patients greffés rénaux

(temps écoulé entre le prélèvement et la transplantation), le nombre d'incompatibilités HLA, le taux d'immunisation anti HLA, le délai de reprise de fonction du greffon, l'année de la greffe et la maladie initiale.

Les données sont extraites de la cohorte DIVAT (Données Informatisées et Validées en Transplantation), coordonnée par l'Institut de Transplantation et de Recherche en Transplantation (ITERT) du CHU de Nantes. Cet institut est le premier centre français et européen de greffes de reins (171 greffes en 2004). Le recueil des données est effectué en temps réel, regroupant les paramètres du suivi biologique et médical des patients greffés d'un rein. Les CHU de Paris Necker, Nancy, Montpellier et Toulouse se sont associés à Nantes pour mettre en commun cet outil de recueil et créer une base de données unique fonctionnant en réseau, en temps réel via Internet et selon la même méthodologie de saisie des données. La banque de données est validée par un audit annuel. Actuellement la base en réseau DIVAT porte sur plus de 7000 patients. Cependant, nous étudierons uniquement les patients greffés au CHU de Nantes, les données de suivi biologique étant mieux renseignées. Depuis 1990, plus de 1000 patients ont été greffés dans ce centre.

Afin de mieux comprendre et interpréter les développements qui vont suivre, nous proposons dans cette introduction de décrire les covariables disponibles dans DIVAT.

(i) Le retard au démarrage des greffons (DGF). La traduction histologique du syndrome clinique que représente le DGF est la nécrose tubulaire aiguë. Cette nécrose est présente dans 22.7% des biopsies faites avant l'implantation du greffon et est associée statistiquement avec la néphropathie chronique du transplant qui conduit à l'échec de la greffe [34]. Les facteurs explicatifs de la survenue d'un DGF sont maintenant assez bien connus comme l'utilisation d'inotrope chez le donneur (en dehors de la dopamine) [35], l'hypovolémie du receveur, l'immunisation anti HLA et la re-transplantation, les facteurs prothrombotiques [36] et la durée de l'ischémie froide [37]. De nombreuses études montrent que le DGF est associé à une diminution de survie des greffons indépendamment

des épisodes de rejets aigus [38].

(ii) Le sexe du donneur et du receveur. Les premières études qui rapportent l'effet du sexe sur la survie des greffons remontent aux années 1990 et montrent une moins bonne survie des greffons à court et à long terme quand un rein de donneur femme est transplanté chez un receveur homme [39]. Ces observations sont confirmées ultérieurement sur de plus larges séries qui confirment que la survie des premières transplantations est significativement moins bonne quand le rein provient d'un donneur féminin quelque soit le sexe du receveur [40]. L'explication la plus répandue de cet effet du sexe du donneur sur la survie des greffons concerne l'inadéquation entre la "dose de néphrons" et la masse du receveur [41]. En effet, les études anatomiques sur les reins de femmes montrent qu'ils sont moins gros et ont moins de néphrons que les reins d'hommes mais cette observation n'est plus valable lorsque la taille du rein est corrigée par la surface corporelle [42]. Cependant, l'inadéquation entre la taille du rein et la masse du receveur, ne semble pas être le seul facteur explicatif de l'effet du sexe du donneur sur la survie du greffon. En effet, quelques études montrent que l'incidence du rejet aigu est plus élevée chez les receveurs masculins qui ont reçu un rein de femme. Une hypothèse, issue de modèles animaux, suggère que les reins de femelles expriment plus d'antigènes HLA et seraient ainsi plus antigéniques que les reins de mâles. L'effet du sexe est aussi retrouvé chez le receveur indépendamment du donneur [43, 44]. Cet effet a été récemment confirmé sur une population de 512 premières greffes de reins dans un seul centre où il est montré que la perte des greffons, due à une néphropathie chronique du transplant, est corrélée à la créatinémie, à la présence d'anticorps anti HLA de classe II, au rejet aigu, et au sexe masculin du receveur [45]. Une hypothèse hormonale pourrait être à l'origine, au moins en partie, de cette observation. Dans un modèle animal, des rats traités par testostérone présentent une protéinurie et des lésions du greffon plus marquées que les rats qui reçoivent de l'oestradiol.

(iii) L'incompatibilité HLA. Il existe un bénéfice clinique de l'effet de la compatibilité HLA entre donneur et receveur sur la survie à long terme des greffons comme le montre les registres de l'UNOS Cecka JM [46, 47]. Une étude prospective américaine, portant sur l'effet de la compatibilité HLA entre donneur et receveur en greffe de rein de cadavre, montre que la survie des greffons à 1 an est de 88% pour les greffons matchés en HLA comparé à 79% pour les greffons non matchés. L'estimation de la demi-vie est de 17,3 ans pour les greffes compatibles et seulement 7,8 ans pour les greffes incompatibles. Cicciarelli évoque la possibilité d'un seuil de compatibilité HLA au-delà duquel la survie des greffons est significativement meilleure ($\leq$ 2 incompatibilités vs $>$ 2) [48].

(iv) L'immunisation anti HLA (PRA pour Panel Reactive Antibodies). Il est actuellement admis que les patients qui présentent une pré-immunisation anti HLA sur le panel avant la transplantation, en rapport avec des greffes antérieures, des transfusions ou des grossesses, possèdent un risque accru de rejet aigu humoral ou cellulaire et de perte chronique des greffons comparés aux patients non immunisés [49, 50]. Le groupe de G. Opelz a aussi récemment montré que la présence d'une immunisation pré-greffe non HLA spé-

cifique pouvait jouer un rôle majeur sur la survie des greffes [51]. Ces antigènes cibles des anticorps, responsables de rejets chroniques, sont aussi parfois appelés Ag mineurs d'histocompatibilité et ne sont pas codés par le système HLA. L'apparition d'Ac contre ces antigènes mineurs conduirait préférentiellement au rejet chronique. Deux hypothèses sont envisagées pour expliquer le rôle de ces Ac dans la perte chronique des greffons. L'une est que les anticorps non HLA contre les Ag mineurs d'histocompatibilité apparaissent fréquemment avec les anticorps anti HLA, l'autre que la cross réactivité des anticorps anti HLA avec les épitopes des Ag mineurs d'histocompatibilités induisent des rejets chroniques tardifs.

(v) L'âge du donneur. Chez l'homme, l'âge du donneur semble augmenter l'immunogénicité du greffon et être à l'origine de l'incidence accrue d'épisode de rejets aigus chez les receveurs de vieux reins. Il s'agit essentiellement de rejet interstitiel de bas grade histologique (grade 1) survenant dans des délais semblables à ceux survenant sur des greffons plus jeunes. Ces rejets seraient cependant plus délétères sur la survie du greffon, probablement en raison de la moins bonne capacité de ces vieux tissus à réparer les lésions dues au rejet [52, 53]. Une étude récente, sur un modèle de greffe allogénique chez le rat, montre que le ratio des modifications structurales est doublé dans les greffons âgés et que leur fonction rénale est 5 fois inférieure à celle des reins jeunes. Cette étude montre aussi que les vieux reins présentent des modifications immunologiques. En effet, le nombre de cellules T et B (splénocytes et cellules périphériques circulantes) et de cellules alloréactives (Elispot, IFN gamma) augmentent dans les 6 premiers mois suivant la greffe plus significativement chez les receveurs de vieux reins [54], suggérant que la greffe de reins âgés engendre une plus forte réponse immune dans les 6 premiers mois de la transplantation.

(vi) L'âge du receveur. Le nombre de patients présentant une insuffisance rénale chronique terminale augmente régulièrement, essentiellement en raison de l'afflux croissant de sujets de plus de 65 ans sur les listes d'attente de greffe rénale. En Europe, le pourcentage de sujets âgés de plus de 65 ans est passé de 1,5% durant la période 1985-1989 à 17% de 2000 à 2004. Malgré une mortalité initiale accrue qui entoure la période péri-opératoire, les receveurs de plus de 65 ans qui reçoivent des reins de donneurs dits limites, vivent en moyenne 3,8 ans de plus que les patients équivalents en attente sur les listes de transplantation. Environ 50% des pertes de greffons chez les receveurs âgés sont en rapport avec le décès du patient avec un rein fonctionnel comparé à 15% chez les receveurs plus jeunes. Cependant, l'âge du receveur est en soi (indépendamment de l'âge du donneur, du décès, de la récidive de la maladie initiale, de la thrombose, du rejet aigu, des infections ou des échecs techniques) un facteur de risque de perte de greffon et ce malgré une incidence du rejet aigu chez les patients de plus de 60 ans généralement diminuée par rapport aux populations plus jeunes de receveurs [55, 56, 57, 58, 59]. Une des explications possibles, de l'effet délétère de l'âge du receveur, indépendamment des autres facteurs de perte de greffon (dont l'âge du donneur), pourrait être liée à des lésions vasculaires avancées retrouvées chez les vieux receveurs. En effet, la maladie vasculaire du receveur serait être à l'origine d'hypoxie et de lésions ischémiques du greffon favorisant le retard au démar-

rage du greffon puis les lésions de néphropathie chronique du transplant. De plus, si l'on considère que la plupart des équipes de transplantation respectent autant que possible une concordance d'âge entre le donneur et le receveur et donc attribuent généralement des reins limites à des receveurs âgés et vasculaires, on observe une synergie d'effet de l'âge sur la dysfonction chronique des greffons. Enfin, de façon intéressante il a aussi été montré, sur des modèles animaux, que l'âge du receveur était associé à des altérations de la réponse immune et avait pour conséquence une diminution de la survie des greffons. L'augmentation de l'âge du receveur résulte dans une diminution de la fonction rénale associée à des lésions histologiques chroniques et un infiltrat cellulaire marqué dans le greffon.

(vii) L'année de greffe. Les stratégies thérapeutiques ont été considérablement modifiées aux alentours des années 1996-1998. En effet, l'introduction de nouvelles drogues immunosuppressives comme le FK 506 (Prograf®, Astellas), le mycophenolate mofétil (Cellcept®, Roche) et le simulect® (anticorps monoclonal anti recepteur de l'interleukine 2, Novartis) a considérablement diminué l'incidence des épisodes de rejets aigus de plus de 35% avant 1998 à moins de 10% actuellement, sachant que le rejet aigu est la principale cause de dysfonction chronique des greffons. De plus, durant la même période, l'amélioration de la logistique d'organisation des greffes a permis de réduire très significativement la durée de l'ischémie froide, principal facteur de risque de survenue d'un retard au démarrage des greffons, lui-même facteur de risque d'échec de la greffe. Enfin, les techniques de dépistage des anticorps anti HLA pré-transplantation et le typage des motifs antigéniques HLA des donneurs et des receveurs avant greffe utilisent depuis ces dix dernières années des techniques hautement performantes et résolutives permettant d'éviter les rejets humoraux hyper aigus, très délétères pour les greffons et d'améliorer la compatibilité tissulaire entre les donneurs et les receveurs. Cependant durant la même période, l'âge des donneurs et celui des receveurs se sont considérablement accrus, n'allant cette fois ci pas dans le sens d'une amélioration des résultats des greffes dans ces tranches d'âge élevé.

Objectifs et structure de la thèse

L'objectif principal de ce travail est de développer une méthodologie adéquate à l'analyse multi-états des données cliniques présentées. Nous avons fait le choix d'un plan se déroulant selon les différentes étapes d'amélioration du modèle en rapport avec les difficultés méthodologiques rencontrées.

Le chapitre 1 décrit les notions de base à la compréhension des modèles semi-markoviens. Nous abordons les définitions propres à l'analyse de survie, en nous attachant particulièrement au choix des distributions des temps de survie et à l'extension de ces modèles à l'analyse des événements répétés ou groupés. Nous présentons aussi les processus mar-

koviens homogènes, afin de mieux appréhender l'intérêt de l'approche semi-markovienne pour l'analyse des processus stochastiques.

Le chapitre 2 présente le modèle semi-markovien. Il permet de définir explicitement la distribution des temps d'attente dans les états, tandis que la trajectoire du processus reste conditionnée par une chaîne de Markov. A ce niveau, le seul type de données incomplètes considéré est la censure à droite.

Deux difficultés majeures sont inhérentes à ces modèles : le choix des distributions des temps de séjour et l'introduction des covariables pour modéliser l'hétérogénéité de la population d'étude. Pour résoudre la première difficulté, Sternberg et Satten [60, 61] proposent une estimation non-paramétrique des fonctions de risque des temps d'attente en utilisant l'algorithme EM. L'avantage est de ne formuler aucune hypothèse sur la forme des fonctions de risque, néanmoins le nombre de paramètres est important et l'estimation instable. Joly et Commenges [27] utilisent des estimateurs basés sur des fonctions splines et calculés par maximisation de la vraisemblance pénalisée. Cette dernière approche est plus stable et nécessite moins de paramètres, même si les auteurs proposent 12 noeuds pour l'ajustement aux données. Dans ce contexte, les méthodes paramétriques constituent une alternative intéressante. En effet, même si certaines hypothèses doivent être faites sur les distributions utilisées, certaines lois permettent de modéliser une grande variété de dynamiques. Nous avons, en particulier, choisi la loi Weibull généralisée pouvant approcher des fonctions de risques non-monotones, en forme de $\cup$ ou $\cap$. Elle généralise ainsi la loi de Weibull, déjà bien adaptée à l'analyse de survie. Cette distribution utilise trois paramètres dont il est possible de tester la parcimonie par rapport à des lois moins complexes, comme Weibull ou Exponentielle. Concernant la seconde difficulté, les covariables sont introduites proportionnellement aux fonctions de risque des temps d'attente [20, 62]. Ce modèle est appliqué à l'évolution des patients atteints du VIH [63], en respectant la structure définie par la figure (2).

Les temps de transition sont considérés indépendants dans le second chapitre. Cependant, plusieurs transitions peuvent avoir lieu pour un même individu. Cette dépendance peut être gérée par l'introduction d'effets aléatoires. Cette généralisation est présentée dans le chapitre 3. Les transformées de Laplace sont utilisées pour obtenir une vraisemblance marginale analytique, c'est-à-dire sans résolution numérique d'intégrales.

Deux principales limites rendent les approches définies dans les chapitres 2 et 3 peu appropriées à l'analyse des données relatives à la transplantation rénale. Le chapitre 4 offre ainsi deux extensions. Premièrement, le modèle est adapté pour la gestion de la censure par intervalle, lorsque les temps des transitions sont seulement connus comme ayant eu lieu dans un intervalle. En effet, l'état de santé du patient évolue de manière continue au cours du temps, alors qu'il n'est renseigné qu'à certaines visites. Deuxièmement, l'initialisation du processus est prise en compte dans le modèle semi-markovien, le patient pouvant commencer son évolution dans différents états. La théorie des modèles pour variables dépendantes multinomiales [64] est adaptée pour inclure certains facteurs explicatifs de

cette initialisation. Ceci permet d'identifier une population à risque d'échec dès l'origine.

Dans le chapitre 5, l'hypothèse de proportionnalité des risques est levée. En effet, certains travaux récents en analyse de survie [65, 66] montrent que l'hypothèse de proportionnalité est peu réaliste dans beaucoup d'applications et peut alors entraîner de sérieux biais dans l'estimation de l'effet des covariables. Deux approches complémentaires sont mises en oeuvre. D'une part, l'effet des covariables peut être fonction du temps de survie et d'autre part, les covariables peuvent agir directement sur le temps d'attente. Cette dernière méthode permet de modifier l'échelle du temps de survie, d'où leur appellation *modèles de vie accélérée*.

Le modèle le plus abouti est présenté dans le chapitre 6. Les temps de transitions et les séquences d'états sont censurés par intervalle. Les chapitres 4 et 5 considéraient uniquement la censure des temps, supposant une régularité suffisante des visites pour identifier tous les états de santé occupés par un patient. De plus, alors que les covariables agissaient uniquement sur les temps d'attente dans les états, le modèle est aussi adapté pour pouvoir inclure des facteurs explicatifs à travers la chaîne de Markov. Les covariables peuvent ainsi influencer la trajectoire ou la vitesse du processus.

Une statistique de test, permettant d'évaluer l'hypothèse de stationnarité du modèle semi-markovien, est proposée dans le chapitre 7. Les deux difficultés de cette statistique de type Pearson sont de calculer les effectifs attendus et de définir sa distribution sous l'hypothèse nulle d'adéquation du modèle. Pour résoudre la seconde difficulté, nous proposons une méthode d'estimation basée sur un bootstrap semi-paramétrique.

Pour prendre en compte le biais dû à la période de la transplantation, un second modèle, mélangeant effets fixes et aléatoires, est défini dans le chapitre 8. Les transformées de Laplace n'étant pas applicables, une méthode d'approximation numérique de la vraisemblance marginale est définie. La quadrature de Gauss-Legendre est utilisée. Cette méthode d'intégration est d'ailleurs rencontrée à plusieurs reprises dans cette thèse.

Le chapitre 9 est consacré à la définition des états de gravité à partir d'un marqueur pronostique dépendant du temps. La méthode des courbes ROC dépendantes du temps est adaptée à cette problématique. Cette dernière partie est complémentaire aux précédents développements. Son intérêt est d'offrir des perspectives intéressantes dans la définition de la structure multi-états, point de départ indispensable à l'estimation du modèle semi-markovien.

Le dernier chapitre synthétise les résultats obtenus et décrit nos perspectives de travail.

Chapitre 1

Principes de modélisation en analyse de survie

La théorie semi-markovienne utilise à la fois des outils propres à l'analyse de survie et propres aux processus markoviens. Pour cette raison, nous commençons par aborder les notions spécifiques à ces deux domaines qui seront utiles à la compréhension de l'approche semi-markovienne. Cette partie ne se veut donc pas exhaustive, elle aborde uniquement les notions nécessaires aux chapitres suivants.

1.1 Définitions

Les modèles de survie permettent d'étudier le délai jusqu'à la survenue d'un événement particulier : l'âge de déclaration d'une maladie, le délai de rémission ou la mort par exemple. L'analyse porte donc sur une variable aléatoire continue. Plusieurs raisons rendent l'utilisation d'un modèle linéaire inadéquate. Tout d'abord, la durée est positive, s'opposant ainsi à l'hypothèse de normalité d'une régression linéaire. Ensuite, cette variable aléatoire n'est pas forcément observée (censure à droite). Ce type de données incomplètes ne se prête pas à une régression basée sur un vecteur observé de la variable dépendante. Enfin, par soucis d'interprétation, il apparaît plus pertinent de se focaliser sur le risque d'un événement au cours du temps, plutôt que sur la durée elle-même.

Cinq fonctions équivalentes définissent la loi de cette durée. Soit X, la variable aléatoire continue positive caractérisant ce délai.

(i) La fonction de survie $S(t)$. C'est la probabilité que l'événement ne se produise pas avant un temps t :

$$S(t) = P(X \geq t) \tag{1.1}$$

(ii) La fonction de répartition $F(t)$. C'est la probabilité que l'événement se produise avant un temps t :

$$F(t) = P(X < t) = 1 - S(t) \tag{1.2}$$

(iii) La fonction de densité $f(t)$. Elle représente la probabilité que l'événement se produise juste après t :

$$\begin{aligned} f(t) &= \lim_{dt \to 0^+} P(t < X < t + dt)/dt \\ &= -\partial S(t)/\partial t \\ &= \partial F(t)/\partial t \end{aligned} \tag{1.3}$$

(iv) La fonction de risque $\lambda(t)$. Elle représente la probabilité que l'événement se produise juste après t, conditionnellement au fait que l'événement n'a pas eu lieu jusqu'à t :

$$\begin{aligned} \lambda(t) &= \lim_{dt \to 0^+} P(t < X < t + dt | X \geq t)/dt \\ &= f(t)/S(t) \end{aligned} \tag{1.4}$$

(v) La fonction de risque cumulé $\Lambda(t)$. C'est l'intégrale de la fonction de risque :

$$\Lambda(t) = \int_0^t \lambda(u) du \tag{1.5}$$

Ces durées possèdent la caractéristique d'être le plus souvent incomplètes. On définit ainsi la censure à droite qui correspond au cas où l'individu n'aurait pas subi l'événement à sa dernière observation. A l'inverse, la censure à gauche correspond au cas où l'individu aurait subi l'événement avant un certain temps. La censure par intervalle correspond au cas où l'événement se serait produit entre deux temps. Une observation est dite tronquée si elle est conditionnelle à un autre événement. Ce type de données incomplètes n'est pas présent dans la suite de nos développements.

1.2 Modèles de régression

Le modèle de survie peut être vu comme un modèle multi-états particulier, défini par deux états et une transition. C'est ce que montre la figure 1.1. L'objectif est alors de modéliser l'intensité de transition entre l'état 1 et l'état 2, notée précédemment $\lambda(t)$. Plusieurs approches peuvent être envisagées entre les modèles non-paramétriques, semi-paramétriques et paramétriques.

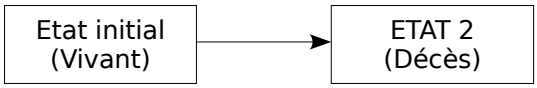

FIG. 1.1 – Représentation multi-états d'un modèle de survie

1.2.1 Modèles semi-paramétriques

Définitions

Le modèle de survie le plus rencontré est le modèle de Cox [1]. Il possède l'avantage de ne pas estimer la fonction de risque, et donc de ne faire aucune hypothèse sur cette dernière. Les paramètres d'intérêt sont les coefficients de régression. Ce modèle est basé sur l'hypothèse de proportionnalité des risques pour prendre en compte l'effet de certaines covariables sur le temps de survie. L'effet des covariables est alors multiplicatif de la fonction de risque et est indépendant du temps. De plus, l'utilisation de la fonction exponentielle permet de conserver la contrainte de positivité de la fonction de risque.

Soit un échantillon constitué de n individus indicés par i, $i = 1, ..., n$. Posons $z = (z_{i,1}, z_{i,2}, ..., z_{i,k})'$ le vecteur des k covariables pour l'individu i et $\beta = (\beta_1, \beta_2, ..., \beta_k)$ le vecteur des coefficients de régression associés. Alors, la fonction de risque est définie par

$$\lambda(t_i, z_i) = \lambda_0(t_i) exp(\beta z_i) \tag{1.6}$$

où $\lambda_0(t)$ est la fonction de risque de base (pour la population de référence) au temps t_i. Cette dernière n'est pas spécifiée. Cette méthode possède l'avantage principal d'interpréter l'exponentiel des paramètres de régression en terme de risques relatifs.

Considérons le seul cas d'une censure à droite. Soit $\mathcal{VP}$ la vraisemblance partielle d'un tel échantillon. La variable dépendante est caractérisée par le couple (T, δ). Si l'événement pour l'individu i est observé, alors δ_i vaut 1 et t_i est égal au temps de survenue de l'événement. En revanche si δ_i est égal à 0, alors l'individu est considéré censuré à droite et t_i est le temps de participation du sujet i. En considérant tous les sujets indépendants, ainsi que l'indépendance entre censure et événement [67], la vraisemblance partielle de cet échantillon s'écrit :

$$\mathcal{VP} = \prod_{i=1}^{n} \left\{ \frac{exp(\beta z_i)}{\sum_{k:t_k \geq t_i} exp(\beta z_k)} \right\}^{\delta_i} \tag{1.7}$$

où k indique les individus encore à risque au temps t_i.

Effets non-proportionnels des covariables

Cependant, l'hypothèse d'un effet constant et multiplicatif du risque peut ne pas correspondre aux données. Une extension directe est de considérer des effets dépendants du

temps en introduisant des interactions polynomiales avec le temps de survie [68]. Pour une covariable Z_1, on peut ainsi poser :

$$\lambda(t,z) = \lambda_0(t)exp(\beta_1 z_1 + \beta_2 z_1 t + \beta_3 z_1 t^2) \tag{1.8}$$

Le terme linéaire serait suffisant pour permettre une relation variant au cours du temps, cependant le terme quadratique permet de modéliser un risque non-monotone.

1.2.2 Modèles paramétriques

Choix des distributions

Un modèle paramétrique est construit en précisant la forme d'une des 5 fonctions précédentes. Cependant, on privilégie souvent la fonction de risque λ. Même s'il en existe beaucoup d'autres, nous définissons trois lois répandues pour l'analyse de survie dans le domaine du vivant :

(i) La loi Exponentielle, $\mathcal{E}(\sigma)$, où le risque est constant au cours du temps :

$$\lambda_{\mathcal{E}}(t) = \frac{1}{\sigma} \quad \forall\, \sigma > 0 \tag{1.9}$$

(ii) La loi de Weibull, $W(\sigma,\nu)$, où le risque est monotone au cours du temps :

$$\lambda_W(t) = \nu\Big(\frac{1}{\sigma}\Big)^{\nu} t^{\nu-1} \quad \forall\, \nu\ ,\ \sigma > 0 \tag{1.10}$$

(iii) La loi de Weibull Généralisée, $WG(\sigma,\nu,\theta)$, où le risque est non-monotone au cours du temps :

$$\lambda_{WG}(t) = \frac{1}{\theta}\Big(1 + (\frac{t}{\sigma})^{\nu}\Big)^{\frac{1}{\theta}-1} \frac{\nu}{\sigma}\Big(\frac{t}{\sigma}\Big)^{\nu-1} \quad \forall\, \nu\ ,\ \sigma\ ,\ \theta > 0 \tag{1.11}$$

Cette dernière distribution généralise les deux précédentes et possède de bonnes propriétés bien définies par Bagdonavicius et Nikulin [69]. Selon la valeur des paramètres, la fonction de risque peut-être constante, monotone (croissante ou décroissante), ou non-monotone ($\cup$ ou $\cap$). Pour ces fonctions non-monotones, il est intéressant de calculer le temps correspondant au maximum ou au minimum de la fonction de risque. Si $0 < \theta < \nu < 1$, alors la fonction (1.11) décroît de ∞ vers sa valeur minimale au temps

$$c = \sigma\Big(\frac{\theta - \nu\theta}{\nu - \theta}\Big)^{1/\nu} \tag{1.12}$$

puis croît jusqu'à ∞. La forme est en $\cup$. Si $\theta > \nu > 1$, alors le risque augmente de 0 à sa valeur maximum au temps c puis décroît vers 0. La forme est en $\cap$. Ces différentes formes sont représentées par la figure (1.2). L'intérêt de ces distributions est de pouvoir

facilement tester leur parcimonie aux données. En fixant θ égal à 1, on retrouve la loi de Weibull. De plus, si ν est égal à 1, la fonction de risque est exponentielle.

A partir, de la fonction de risque (1.11), on retrouve les autres fonctions de distribution. Ainsi, la fonction de densité est égale à :

$$\begin{aligned} f_{WG}(x) &= \lambda_{WG}(x) exp\Big(- \int_0^x \lambda_{WG}(u) du \Big) \\ &= \frac{1}{\theta}\Big(1 + (\frac{x}{\sigma})^{\nu}\Big)^{\frac{1}{\theta}-1} \frac{\nu}{\sigma}\Big(\frac{x}{\sigma}\Big)^{\nu-1} exp\Big(1 - \big(1 + (\frac{x}{\sigma})^{\nu}\big)^{\frac{1}{\theta}}\Big) \end{aligned} \tag{1.13}$$

En utilisant la propriété (1.4), la fonction de survie est égale à :

$$S_{WG}(x) = exp\Big(1 - \big(1 + (\frac{x}{\sigma})^{\nu}\big)^{\frac{1}{\theta}}\Big) \tag{1.14}$$

Estimation par maximum de vraisemblance

Considérons les mêmes notations que dans la section consacrée aux modèles semi-paramétriques, la vraisemblance de cet échantillon s'écrit :

$$\mathcal{V} = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{1.15}$$

La théorie classique du maximum de vraisemblance permet l'estimation des paramètres ainsi que leur variance.

Cette approche peut être assez facilement étendue pour la prise en compte de la censure à gauche et par intervalle en ajoutant deux composantes supplémentaires à la fonction (1.15). Comme proposé par Odell et al. [70], posons T_{0i} la valeur du temps pour laquelle si $T_{0i} > T_i$, alors T_{0i} est observé et T_i n'est pas observé. D'une manière similaire, posons T_{1i} la valeur du temps pour laquelle si $T_{1i} < T_i$, alors T_{1i} est observé et T_i n'est pas observé. Définissons δ_{Ri} l'indicateur de censure à droite $\{0 < T_{0i} \leq T_i < T_{1i} = \infty\}$, δ_{Li} l'indicateur de censure à gauche $\{0 = T_{0i} < T_i \leq T_{1i} < \infty\}$ et δ_{Ii} l'indicateur de censure par intervalle $\{0 < T_{0i} < T_i \leq T_{1i} < \infty\}$. Finalement, $\delta_{Ei} = 1 - \delta_{Ri} - \delta_{Li} - \delta_{Ii}$ est l'indicatrice des temps exactement observés. Toujours en supposant un processus de censure non-informatif, la vraisemblance s'écrit :

$$\mathcal{V} = \prod_{i=1}^{n} f(t_i)^{\delta_{Ei}} S(t_{0i})^{\delta_{Ri}} F(t_{1i})^{\delta_{Li}} (S(t_{0i}) - S(t_{1i}))^{\delta_{Ii}} \tag{1.16}$$

Introduction des covariables

La proportionnalité des risques est la méthode la plus utilisée. Une autre approche est de considérer un effet des covariables agissant directement sur l'échelle du temps. Le

principe de ce type de modèle est de considérer une action additive sur le logarithme du temps de survie :

$$log(T) = -\beta z + \sigma W$$

où W est une variable aléatoire indépendante de β. Il est équivalent d'écrire :

$$T' = Texp(\beta z)$$

où $T' = exp(\sigma W)$ possède une fonction de risque λ_0. Les fonctions de survie de T et T', respectivement $S(t)$ et $S_0(t)$ sont ainsi liées par la relation suivante :

$$\begin{aligned} S(t,z) &= P(T \geq t) = P(exp(-\beta z)T' \geq t) = P(T' \geq exp(\beta z)t) \\ &= S_0(exp(\beta z)t) \end{aligned} \tag{1.17}$$

La covariable ralentie ou accélère le temps de survie. La fonction de risque associée est égale à :

$$\lambda(t,z) = \lambda_0(texp(\beta z))exp(\beta z) \tag{1.18}$$

1.2.3 Modèles de fragilité paramétriques

Dans un certain nombre de situations, l'événement d'intérêt peut être répété pour un même individu. Dans ce cas, la dépendance des observations rend la vraisemblance (1.15) invalide. Pour un sujet donné, posons la fonction de risque égale à $\omega\lambda(t)$, où $\lambda(t)$ est la fonction de risque au temps t commune à tout l'échantillon et ω est le facteur aléatoire individuel (appelé fragilité). Conditionnellement à ω, les observations sont considérées indépendantes. Considérons un échantillon composé de n individus, $i = 1, ..., n$, chacun d'entre eux présentant n_i répétitions, $j = 1, ..., n_i$. Posons t_{ij} la i$ème$ observation de l'individu i. La vraisemblance conditionnelle est alors égale à :

$$\begin{aligned} \mathcal{V}_{cond} &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} f(t_{ij}|\omega_i)^{\delta_{ij}} S(t_{ij}|\omega_i)^{1-\delta_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} (\lambda(t_{ij}|\omega_i)S(t_{ij}|\omega_i))^{\delta_{ij}} S(t_{ij}|\omega_i)^{1-\delta_{ij}} \\ &= \prod_{i=1}^{n} exp(-\omega_i \sum_{j=1}^{n_i} \Lambda(t_{ij}))\omega_i^{\sum_{j=1}^{n_i}\delta_{ij}} \prod_{j=1}^{n_i} \lambda(t_{ij})^{\delta_{ij}} \end{aligned} \tag{1.19}$$

L'utilisation des transformées de Laplace permet une estimation de la vraisemblance marginale $\mathcal{V}$, avec $\mathcal{V} = E[\mathcal{V}_{cond}]$, sans intégration numérique [2] :

$$L(a) = E[exp(-a\omega)] \tag{1.20}$$

La r$ème$ transformée de Laplace est alors égale à :

$$L^{(r)}(a) = (-1)^{(r)}E[\omega^r exp(-a\omega)] \tag{1.21}$$

On obtient ainsi :

$$\mathcal{V} = \prod_{i=1}^{n} \left\{ (-1)^{n_{ij}} L^{(n_{ij})} (\sum_{j=1}^{n_i} \Lambda(t_{ij})) \prod_{j=1}^{n_i} \lambda(t_{ij})^{\delta_{ij}} \right\} \tag{1.22}$$

La distribution du terme de fragilité doit être choisie pour avoir une distribution explicite. Hougaard [3] décrit un certain nombre de possibilités. Cependant, seule la distribution Gamma offre des r*ièmes* dérivées de Laplace assez simples pour r grand. Pour les autres distributions, les dérivées utilisent des polynômes récursifs, déjà associés à des difficultés d'estimation dans les modèles de survie multivariée. Cette introduction ayant pour objectif de préparer à des développements dans le cadre semi-markovien, nous nous limitons à la loi Gamma. Sa densité est définie par :

$$g(\omega) = \left(\omega^{(\gamma^{-1}-1)} exp(-\omega\gamma^{-1}) \right) \Big/ \left(\Gamma(\gamma^{-1}) \gamma^{\gamma^{-1}} \right) \quad , \forall \gamma \geq 0 \tag{1.23}$$

où $\Gamma(\gamma^{-1}) = \int_0^\infty t^{(\gamma^{-1}-1)} exp(-t) dt$ est la fonction Gamma. L'espérance de ω est égale à 1 et respecte la contrainte d'identifiabilité. La variance de l'effet aléatoire est égale à γ. La transformée de Laplace, $L(a)$, est alors égale $(1+\gamma a)^{-\gamma^{-1}}$ et sa r*ième* dérivée à :

$$L^{(r)}(a) = -\gamma^{r-1}(1+\gamma a)^{-\gamma^{-1}-r} \prod_{k=1}^{r-1} (-\gamma^{-1} - k) \tag{1.24}$$

Bien entendu, d'autres auteurs ont proposé des méthodes d'estimation différentes. Klein [71] propose par exemple un algorithme EM pour estimer un modèle de Cox avec fragilité. Therneau [72] reprend un certain nombre de ces développements pour mettre en place l'algorithme d'estimation de référence, implémenté sous SAS ou S-plus.

1.3 Modèles markoviens à temps continu

Dans cette section, nous nous intéressons aux processus markoviens à temps continu et à espace d'états discret, tels que ceux développés par Karlin et Taylor [73]. L'objectif de cette première partie est d'introduire la notion de processus et de mettre en évidence l'apport des modèles semi-markoviens pour l'analyse des processus stochastiques.

1.3.1 Définitions

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 présent, autrement dit pour $t_0 < t_2 < t_1 < ... < t_n < t_{n+1}$, nous avons :

$$P(X(t_{n+1}) = j | X(t_n), X(t_{n-1}), ..., X(t_0)) = P(X(t_{n+1}) = j | X(t_n)) \tag{1.25}$$

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

$$P_{ii}(t, t+s) = 1 - \sum_{j \neq i} P_{ij}(t, t+s) \tag{1.26}$$

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 :

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

A partir de la propriété markovienne (1.25), 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 :

$$\mathbf{P}(0, t+s) = \mathbf{P}(0, t)\mathbf{P}(t, t+s) \tag{1.27}$$

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 :

$$\begin{aligned} \alpha_{ij}(t) &= \lim_{dt \to 0^+} P(X(t+dt) = j | X(t) = i)/dt \\ &= \lim_{dt \to 0^+} P_{ij}(t, t+dt)\ dt \end{aligned} \tag{1.28}$$

Notons que $\alpha_{ij}(t) \times dt$ représente la probabilité que le processus passe dans l'état j entre t et $t+dt$, 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 (1.26) :

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

d'où

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

En définissant :

$$\lim_{dt \to 0^+} (1 - P(X(t+dt) = i | X(t) = i))/dt = -\alpha_{ii}(t) \tag{1.29}$$

Nous obtenons alors :

$$\sum_{j \neq i} \alpha_{ij}(t) = -\alpha_{ii}(t) \quad \text{et} \quad \sum_{j} \alpha_{ij}(t) = 0$$

1.3.2 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 :

$$P_{ij}(t, t+s) = P_{ij}(0, s) = P_{ij}(s) \tag{1.30}$$

$P_{ij}(s)$ est indépendant de t, $\forall t \geq 0$. L'équation de Chapman-Kolmogorov (1.27) peut alors s'écrire :

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

donc

$$\begin{aligned} \frac{d\mathbf{P}(s)}{ds} &= \lim_{ds \to 0^+} (\mathbf{P}(s+ds) - \mathbf{P}(s))/ds \\ &= \lim_{ds \to 0^+} (\mathbf{P}(s)\mathbf{P}(ds) - \mathbf{P}(s))/ds \\ &= \lim_{ds \to 0^+} (\mathbf{P}(s)(\mathbf{P}(ds) - \mathbf{I}))/ds \\ &= \mathbf{P}(s) \lim_{ds \to 0^+} (\mathbf{P}(ds) - \mathbf{I})/ds \\ &= \mathbf{P}(s)\mathbf{Q} \end{aligned} \tag{1.31}$$

avec $\mathbf{I}$, la matrice identité, et où d'après les définitions (1.28) et (1.29), la matrice $\mathbf{Q}$ s'écrit :

$$\mathbf{Q} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1r} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{rr} \end{pmatrix}$$

Remarquons que les forces de transition ne dépendent pas du temps. L'équation différentielle (1.31) admet la solution :

$$\mathbf{P}(t) = exp\big(\mathbf{Q}t\big) \tag{1.32}$$

avec comme contrainte $\mathbf{P}(0) = \mathbf{I}$. Le calcul des termes diagonaux de la matrice $\mathbf{P}(t)$ est assez direct. En effet, d'après la définition (1.30) :

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

Développons cette propriété :

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

Or d'après la relation (1.29), nous avons :

$$\begin{aligned} & \lim_{du \to 0^+} (P_{ii}(u+du) - P_{ii}(u))/du = -\alpha_{ii} P_{ii}(u) \\ & \Longleftrightarrow dP_{ii}(u)/du = -\alpha_{ii} P_{ii}(u) \end{aligned} \tag{1.33}$$

où $dP_{ii}(u)/du$ 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 (1.33). 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 une loi Exponentielle :

$$\begin{aligned} P_{ii}(u) &= exp(-\alpha_{ii}u) \\ &= exp(-Q_{ii}u) \end{aligned} \tag{1.34}$$

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 transition. Comme nous l'avons abordé dans la section sur les modèles de survie paramétriques, nous pouvons aussi envisager des formes plus complexes non-monotones, en $\cup$ ou $\cap$ par exemple. L'intérêt des modèles semi-markoviens est de choisir explicitement la distribution du temps de séjour dans l'état.

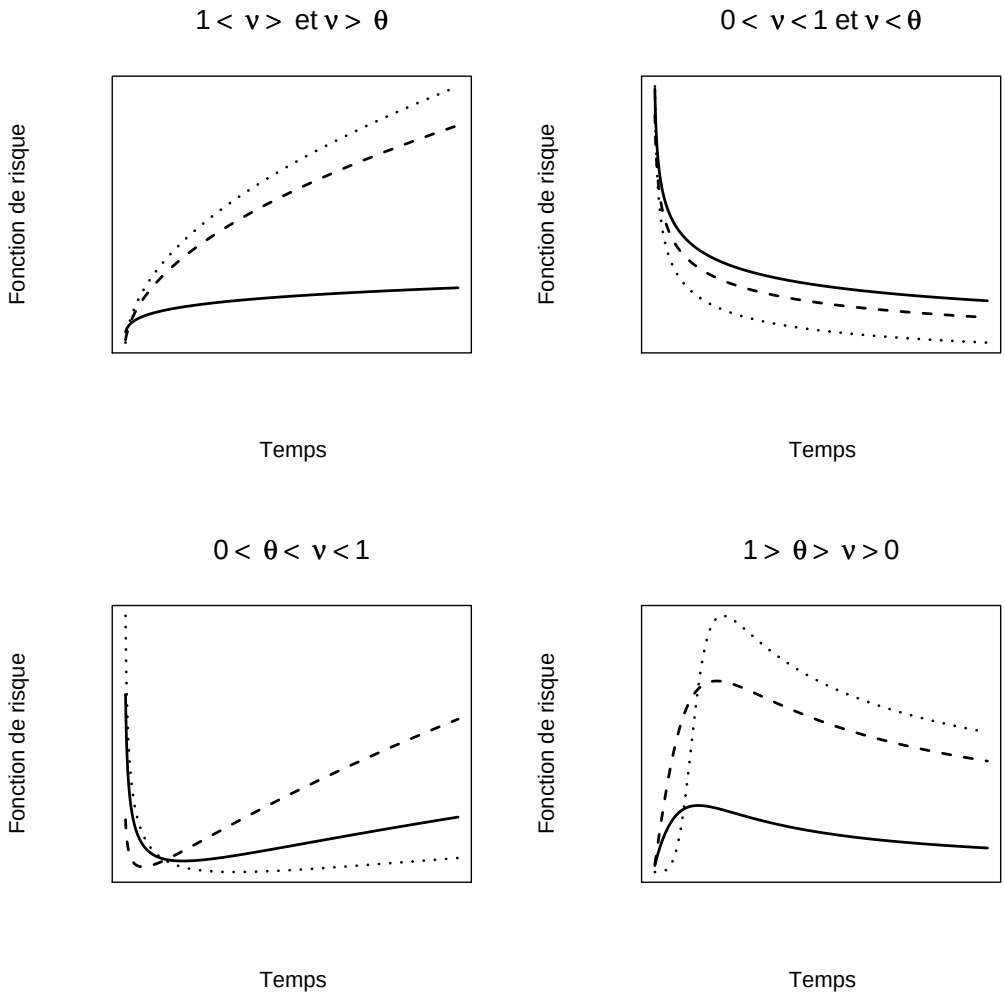

FIG. 1.2 – Les formes de fonctions de risque ajustables par une loi de Weibull généralisée selon les valeurs des paramètres σ, ν, θ

Chapitre 2

Modèle semi-markovien

L'intérêt de ce type de modèle est contenu dans le choix 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 type de modèle est présenté dans deux articles [21, 20] dont s'inspire ce document. Nous nous fonderons sur les notations issues des processus de comptage, définies par Gill [74]. Elles ont l'avantage d'être explicites et d'être transversales à de nombreuses problématiques markoviennes. Nous définirons dans une première partie le modèle. Ce dernier sera utilisé dans la seconde section pour l'analyse de l'évolution des patients atteints du VIH. La troisième section conclura ce chapitre.

2.1 Définition du modèle

2.1.1 Fonctions utiles

A des temps différents, le processus occupe des états définis. En l'absence de covariables, 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) :

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

où $\widetilde{N}_{ij}(t)$ représente le nombre de transitions $i \rightarrow j$ 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 en même temps). Ce processus

est dit *cadlag*. De plus, nous posons :

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

où $\widetilde{N}(t)$ est 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. Les probabilités de transition $i \to j$ associées à cette chaîne, notées P_{ij}, sont définies par :

$$P_{ij} = P(X_{n+1} = j | X_n = i) \tag{2.1}$$

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

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

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

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

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 ne gère pas le temps, mais les séquences des états indicés par le numéro de la transition. Les temps d'attente dans les états (ou temps de séjour) sont définis explicitement. Comme dans le cas markovien où l'histoire du processus est résumée dans l'état précédent (équation 1.25), le processus (T, X) est dit semi-markovien si la distribution des temps de séjour $(T_{n+1} - T_n)$ satisfait la condition suivante :

$$P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_0, T_0, .., X_n, T_n) = P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n) \tag{2.2}$$

Autrement dit, 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 stationnaire sur le temps chronologique. Parallèlement à l'analyse de survie et les équations (1.1) à (1.5), nous notons :

(i) la fonction de répartition :

$$F_{ij}(x) = P(T_{n+1} - T_n \leq x | X_{n+1} = j, X_n = i) \tag{2.3}$$

(ii) la fonction de survie :

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

(iii) la fonction de densité :

$$f_{ij}(x) = \lim_{dx \to 0^+} P(x < T_{n+1} - T_n < x + dx | X_{n+1} = j, X_n = i)/dx \tag{2.5}$$

(iv) la fonction de risque :

$$\lambda_{ij}(x) = \lim_{dx \to 0^+} P(x < T_{n+1} - T_n < x + dx | T_{n+1} - T_n \geq x, X_{n+1} = j, X_n = i)/dx \quad (2.6)$$

(v) la fonction de risque cumulé :

$$\Lambda_{ij}(x) = \int_0^x \lambda_{ij}(u)du \quad (2.7)$$

D'après le théorème de Bayes et les définitions (2.1) (2.3), nous pouvons préciser la condition (2.2) définissant les modèles semi-markoviens. Pour $i \neq j$:

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

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) :

$$\begin{aligned} F_{i.}(x) &= P(T_{n+1} - T_n \leq x | X_n = i) \\ &= \sum_j P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n = i) \\ &= \sum_j F_{ij}(x) P_{ij} \\ &= \sum_{j \neq i} F_{ij}(x) P_{ij} \ \text{ (puisque } P_{ii} = 0) \end{aligned} \quad (2.9)$$

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

$$\begin{aligned} S_{i.}(x) &= 1 - F_{i.}(x) \\ &= \sum_{j \neq i} S_{ij}(x) P_{ij} \end{aligned} \quad (2.10)$$

Et la fonction de densité marginale :

$$\begin{aligned} f_{i.}(x) &= \partial F_{i.}(x)/\partial x \\ &= \partial\big(\sum_{j \neq i} F_{ij}(x)\ P_{ij}\big)/\partial x \\ &= \sum_{j \neq i} P_{ij}\{\partial F_{ij}(x)/\partial x\} \\ &= \sum_{j \neq i} P_{ij} f_{ij}(x) \end{aligned} \quad (2.11)$$

Par définition, la fonction de risque instantané de i vers j du processus semi-markovien, correspond à la probabilité du processus à transiter juste après le temps x vers l'état j, sachant qu'il est dans l'état i depuis une durée x :

$$\begin{aligned}
\alpha_{ij}(x) &= \lim_{dx\to 0^+} P(x \leq T_{n+1} - T_n < x + dx, X_{n+1} = j | T_{n+1} - T_n \geq x, X_n = i)/dx \\
&= (P(T_{n+1} - T_n \geq x | X_n = i))^{-1} \\
&\times \lim_{dx\to 0^+} P(x \leq T_{n+1} - T_n < x + dx, X_{n+1} = j | X_n = i)/dx \\
&= (P(T_{n+1} - T_n \geq x | X_n = i))^{-1} \\
&\times \lim_{dx\to 0^+} (P(x \leq T_{n+1} - T_n < x + dx | X_{n+1} = j, X_n = i) P(X_{n+1} = j | X_n = i))/dx \\
&= P(X_{n+1} = j | X_n = i)/P(T_{n+1} - T_n \geq x | X_n = i) \\
&\times \lim_{dx\to 0^+} P(x \leq T_{n+1} - T_n < x + dx | X_{n+1} = j, X_n = i)/dx
\end{aligned} \tag{2.12}$$

d'où

$$\alpha_{ij}(x) = P_{ij} f_{ij}(x)/S_{i.}(x) \text{ avec } \begin{cases} i \neq j \\ i, j \in E \\ \alpha_{ii}(x) = -\sum_{j\neq i} \alpha_{ij}(x) \end{cases} \tag{2.13}$$

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, 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 des temps de séjour, $\lambda_{ij}(x)$, définie en (2.6). A partir de (2.13), nous pouvons écrire :

$$\begin{aligned}
\sum_{j\neq i} \alpha_{ij}(t) &= \sum_{j\neq i} P_{ij} f_{ij}(t)/S_{i.}(t) \\
&= (S_{i.}(t))^{-1} \sum_{j\neq i} P_{ij} f_{ij}(t) \\
&= (S_{i.}(t))^{-1} f_{i.}(t) \\
&= \alpha_{i.}(t)
\end{aligned} \tag{2.14}$$

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

2.1.2 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 de la séquence d'états markovienne X. X 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 de stationnarité précédemment abordée :

$$\begin{aligned} p_{ij}(l, l+t) &= P(Z(l+t) = j | Z(l) = i) \\ &= P(X_{\widetilde{N}(l+t)} = j | X_{\widetilde{N}(l)} = i) \\ &= P(X_{\widetilde{N}(t)} = j | X_{\widetilde{N}(0)} = i) \\ &= P(Z(t) = j | Z(0) = i) \\ &= p_{ij}(t) \qquad i, j \in E \text{ et } x \geq 0 \end{aligned} \tag{2.15}$$

La propriété de stationnarité contenue dans le processus markovien est donc transmise pour le processus semi-markovien sous la forme définie ci-dessus.

Pour calculer ces probabilités, considérons tout d'abord qu'au moins une transition se produit dans l'intervalle $[0, t]$. Ce conditionnement sur le premier événement est particulièrement utilisé dans la théorie du renouvellement [73]. 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 (2.13) :

$$\alpha_{ik}(x) S_{i.}(x) = P_{ik} f_{ik}(x) \tag{2.16}$$

En respectant la stationnarité sur le temps chronologique (2.15) 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 :

$$\begin{aligned} P(Z(t) = j | Z(x) = k) &= P(Z(t-x) = j | Z(0) = k) \\ &= p_{kj}(t-x) \end{aligned} \tag{2.17}$$

Le produit de convolution des deux probabilités (2.16) et (2.17) 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 :

$$P(Z(t) = j | X_0 = i, (T_1 - T_0) \leq t) = \sum_{k=1}^{r} \int_0^t P_{ik} f_{ik}(x) p_{kj}(t-x) dx \tag{2.18}$$

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 (2.18) la possibilité qu'aucun événement ne se produise dans $[0, t]$. En reprenant le résultat (2.10), cette probabilité est égale à :

$$\begin{aligned} P((T_1 - T_0) > t|X_0 = i) &= S_{i.}(t) \\ &= \sum_{l \neq i}^{r} P_{il} S_{il}(t) \end{aligned} \tag{2.19}$$

En posant $\delta_{ij} = 0$ si $i \neq j$ et 1 sinon, et à partir des résultats (2.18) et (2.19), nous obtenons finalement :

$$p_{ij}(t) = \sum_{k=1}^{r} \int_0^t P_{ik} f_{ik}(x) p_{kj}(t-x) dx \; + \; \delta_{ij} \sum_{l \neq i}^{r} P_{il} S_{il}(t) \tag{2.20}$$

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

2.1.3 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'états 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 $X_p^h \neq X_{p+1}^h$, $p = 1, ..., m_{h-1}$. 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 dans l'état i. La contribution de cette observation à la vraisemblance est :

$$\lim_{dx \to 0^+} P(x < T_{n+1} - T_n < x + dx, X_{n+1} = j|X_n = i)/dx = \alpha_{ij}(x) S_{i.}(x) = P_{ij} f_{ij}(x)$$

(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 :

$$P(T_{n+1} - T_n > x|X_n = i) = S_{i.}(x)$$

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

$$\begin{aligned} \mathcal{V} &= \prod_{h \in nc} \left[\prod_{r=1}^{m_h} \left\{ P_{X_{r-1}^h, X_r^h} f_{X_{r-1}^h, X_r^h}(T_{h,r} - T_{h,r-1}) \right\} \right] \\ &\times \prod_{h \in c} \left[\prod_{r=1}^{m_h - 1} P_{X_{r-1}^h, X_r^h} f_{X_{r-1}^h, X_r^h}(T_{h,r} - T_{h,r-1}) S_{X_{m_h-1}^h .}(T_{h,m_h} - T_{h,m_h-1}) \right] \end{aligned} \tag{2.21}$$

2.1.4 Introduction de covariables

Pour prendre en compte d'éventuelles covariables dans le modèle, nous reprendrons le même principe que celui utilisé par Cox [1]. Ainsi, parallèlement à l'égalité (1.6), on pose $z_{ij} = (z_{ij}^1, ..., z_{ij}^{n_{ij}})$, le vecteur des n_{ij} covariables propres à la transition de i vers j. La fonction de risque des temps d'attente dans l'état s'écrit alors comme :

$$\lambda_{ij}(x, z_{ij}) = \lambda_{0,ij}(x)\eta(z_{ij}) \tag{2.22}$$

où $\lambda_{0,ij}(x)$ est la fonction de risque de base propre à la transition i vers j au temps d'attente x et $\eta(z_{ij})$ est une fonction des covariables, multiplicative de la fonction de risque de la population de référence. Pour garantir une fonction de risque strictement positive et pour obtenir une interprétation des coefficients de régression en terme de risques relatifs, on définit :

$$\eta(z_{ij}) = exp(\beta_{ij}^T z_{ij}) \tag{2.23}$$

Comme pour le traitement des processus markoviens par Andersen [62], ce modèle est dit semi-proportionnel, car la proportionnalité des risques n'est supposée qu'au sein d'une même transition, aucune contrainte n'est imposée entre classes. De plus, les individus sont comparés à temps de séjour dans l'état fixé, cette proportionnalité des risques s'applique donc à des individus passant la même période de temps dans l'état. Enfin, ce modèle est plus *parcimonieux* que celui défini par Perez [20]. On attend par parcimonie la maximisation de la qualité d'ajustement à partir d'un nombre minimum de paramètres. En effet, dans ce dernier chaque covariable agit spécifiquement sur chaque transition, alors que les définitions précédentes permettent un vecteur de covariables de taille différente par transition. Les autres fonctions, comme la survie, découlent directement de la fonction de risque (2.22) :

$$\begin{aligned} S_{ij}(x, z_{ij}) &= exp(-\int_0^x \lambda_{ij}(u, z_{ij})du) \\ &= exp(-\eta(z_{ij}) \int_0^x \lambda_{0,ij}(u)du) \\ &= S_{0,ij}(x)^{\eta(z_{ij})} \end{aligned} \tag{2.24}$$

Pour estimer les paramètres, la vraisemblance (2.21) est reformulée en substituant les termes $f_{ij}(x)$, $S_{ij}(x)$ et $\lambda_{ij}(x)$, par leur forme avec covariables : $f_{ij}(x, z)$, $S_{ij}(x, z)$ et $\lambda_{ij}(x, z)$.

2.1.5 Choix des distributions

Précédemment, $f_{ij}(x)$ a été définie comme une fonction de densité dépendante de x et d'un vecteur de paramètres φ_{ij} propre à chaque type de transition de i vers j. Par

exemple pour une distribution de Weibull (1.10), φ_{ij} est égal au vecteur (σ_{ij}, ν_{ij}). Dans le but de généraliser le modèle précédent, notons explicitement :

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

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 le chapitre précédent, nous avons abordé trois distributions intéressantes en analyse de survie : Exponentielle (1.9), Weibull (1.10), et Weibull généralisée (1.11). Pour permettre un modèle plus parcimonieux, adoptons plutôt la notation suivante, qui ne change en rien le reste de la théorie abordée dans ce chapitre :

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

La forme et les paramètres des lois de distribution peuvent alors être spécifiques à chaque transition. La flexibilité du modèle est accrue. Il est en effet évident que la multiplication des paramètres à estimer constitue une difficulté majeure. Partant du modèle le plus complet, WG à 3 paramètres, nous testerons l'égalité des paramètres σ_{ij} et ω_{ij} à 1, permettant ainsi de se rapprocher d'une certaine parcimonie.

2.2 Application au VIH

2.2.1 Description du modèle et des données

Le modèle est défini par la figure (2). Cependant, comme le montre l'histogramme (2.1), les visites sont assez régulières au cours du suivi du patient. Il en découle que les transitions $3 \to 1$ et $2 \to 4$ sont peu représentées dans la base de données, ces transitions correspondant à un saut de deux niveaux dans les états de gravité. Le structure étudiée sera ainsi définie par quatre états transitoires et huit transitions (voir figure 2.2). Cette régularité des visites justifie aussi l'absence de la prise en compte de la censure par intervalle. Les transitions sont supposées se produire au milieu de l'intervalle entre deux visites.

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 sûr, selon leur date d'entrée dans la cohorte, ce nombre est plus ou moins important.

Les transitions possibles sont décrites dans le tableau (2.1). On peut voir que les allers et retours entre l'état 2 et l'état 3 sont les plus représentés, avec environ la moitié des transitions observées. Aux seuils choisis, il semble donc que la charge virale soit plus fluctuante

que le nombre de CD4. 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[a]
$1 \rightarrow 1$[b]	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
$2 \rightarrow 2$[b]	605	12,6 %	1,75
$2 \rightarrow 3$	994	20,7 %	0,81
$3 \rightarrow 2$	1340	27,9 %	0,78
$3 \rightarrow 3$[b]	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
$4 \rightarrow 4$[b]	65	1,4 %	1,18

[a] Temps de séjour médian en mois

[b] Censures à droite

TAB. 2.1 – Représentativité des transitions

Comme l'indique le tableau (2.2), 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 co-infectés par une hépatite B et C. Ces données sont comparables à la population cible des patients VIH positifs.

Covariable	Effectif	Pourcentage
Femmes	381	30,6 %
Age $>$ 40 ans	395	31,8 %
Co-infection VHB	121	9,7 %
Co-infection 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. 2.2 – Descriptif de la population d'étude

2.2.2 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 généralisée), 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 **-** Un modèle différent par modalité des covariables sera estimé (analyse en sous-groupes). 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 semi-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 (8×10), nous avons choisi de retenir, pour l'analyse multivariée, les facteurs dont la p-value est inférieure à 0,05 (test de Wald). Par souci de clarté, ces résultats ne seront pas présentés exhaustivement dans ce document. Seuls les principaux points seront explicités.

(iii) Analyse multivariée **-** Dans un troisième temps, toujours en supposant une distribution de type Weibull, toutes les variables précédemment retenues seront inclues 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 logvraisemblance et estimer la valeur des paramètres ainsi que leur matrice Hessienne. Cette fonction utilise l'algorithme de *quasi-Newton* [75]. 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 [76]. 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.

2.2.3 Résultats

Modèle semi-markovien de type Weibull

Analyse stratifiée **-** Compte tenu des graphiques, l'hypothèse de semi-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. A titre d'exemple, la figure (2.3) présente ces graphiques pour la variable sexe. Concernant 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 celui-ci ne respecte pas la proportionnalité.

Analyse univariée **-** Le tableau (2.3) représente ainsi les covariables sélectionnées par l'analyse stratifiée (signalées par le symbole $\times$). Cette première sélection, même si elle est subjective, présente 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 2.3, symbole O). Sur les 80 facteurs possibles, 11 ont été sélectionnés pour l'analyse multivariée.

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

TAB. 2.3 – Covariables retenues après les stratégies stratifiées (×) et univariées (O)

Analyse multivariée - Après sélection, le modèle final repose sur 5 covariables, soit 31 paramètres au total. Leurs estimations et leurs variances sont présentées dans les tableaux (2.4) et (2.5). 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. 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 supposer que certaines transitions obéissent à une loi Exponentielle. La prochaine étape consiste donc à identifier ces transitions sans mémoire.

Analyse multivariée 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 forces de transition, semblent suivre des lois Exponentielles, sans perte d'information ($ln\mathcal{V} = -6126,96$ et $AIC = 12307,92$). Il s'agit des transitions $4 \rightarrow 3$, $3 \rightarrow 2$, $2 \rightarrow 1$ et $1 \rightarrow 4$ (tableau 2.6). Les paramètres du modèle final, expliquant le maximum d'information à partir d'un minimum de paramètres, sont ainsi représentés dans le tableau (2.7). 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 sexuels et par toxicomanie.

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

	ν_{ij}		σ_{ij}	
Transition	Coeff.	Ecart-type	Coeff.	Ecart-type
$1 \rightarrow 2$	1,11	0,05	0,59	0,04
$1 \rightarrow 3$	1,45	0,14	0,55	0,05
$1 \rightarrow 4$	1,10	0,06	0,58	0,05
$2 \rightarrow 1$	0,91	0,06	0,79	0,10
$2 \rightarrow 3$	0,87	0,02	1,88	0,07
$3 \rightarrow 2$	1,03	0,02	0,98	0,04
$3 \rightarrow 4$	0,78	0,04	1,84	0,26
$4 \rightarrow 1$	0,91	0,04	0,95	0,07
$4 \rightarrow 2$	1,19	0,09	0,63	0,06
$4 \rightarrow 3$	0,99	0,05	0,57	0,04

TAB. 2.4 – Paramètres des distributions des temps d'attente pour le modèle de Weibull multi-états

Modèle semi-markovien de type Weibull généralisé

En utilisant les analyses stratifiées et univariées similaires à la partie précédente, 11 covariables ont été sélectionnées (voir tableau 2.8).

Après la stratégie multivariée descendante, 9 covariables sont définies comme influentes (tableaux 2.9 et 2.10). 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 co-infecté par une hépatite B ou avoir été contaminé par toxicomanie, semble 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, les modes de contamination par accident, hétérosexuel, 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, le critère AIC vaut 11498,16. Ce dernier est largement inférieur à celui obtenu pour le modèle final fondé sur des distributions Weibull et Exponentielle (12307,92). Toutes les transitions semblent suivre des fonctions de risque du type $\cap$. En effet, comme nous l'avons abordé dans le chapitre concernant les notions de base à l'analyse de survie (expression 1.12), les paramètres estimés permettent de vérifier les inégalités $\theta_{ij} > \nu_{ij} > 1$, $\forall\ ij$. 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, refletant une stabilisation. Plus la personne passe de temps dans un état, moins elle a de chance d'en sortir.

Paramètre	Coefficient	Ecart-type
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. 2.5 – Probabilités de transition et coefficients de régression du modèle semi-markovien multivarié de type Weibull

Modèle	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. 2.6 – Sélection du modèle le plus adéquat à partir des lois de Weibull et Exponentielle

2.3 Discussion

Il est évident, au vu de ces 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 permet la loi de Weibull, utilisée par Perez [20] pour la modélisation du cancer.

L'évolution des patients séropositifs nécessite une loi plus complexe en forme de $\cap$, comme la loi de Weibull généralisée. Outre la 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.

Cependant, ces premiers développements mettent en évidence un certain nombre de limites au modèle précédemment défini. L'indépendance des observations, permettant la construction de la vraisemblance (2.21), constitue une hypothèse forte. En effet, plusieurs transitions identiques sont observées pour un même patient, ce qui peut créer une corrélation de certains temps de transition. La section suivante permet l'introduction d'une telle dépendance avec la prise en compte de l'individu comme effet aléatoire.

Paramètre	Coefficient	Ecart-type
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. 2.7 – Modèle semi-markovien multivarié final de type Weibull et Exponentiel

Transit.	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. 2.8 – Covariables retenues pour l'analyse multivariée après les stratégies stratifiées (×) et univariées (O)

	ν_{ij}		σ_{ij}		θ_{ij}	
Transit.	Coeff.	Ecart-type	Coeff.	Ecart-type	Coeff.	Ecart-type
$1 \rightarrow 2$	2,85	0,43	0,13	0,01	5,46	1,09
$1 \rightarrow 3$	2,86	0,65	0,23	0,05	3,61	1,25
$1 \rightarrow 4$	2,67	0,39	0,15	0,02	4,59	0,89
$2 \rightarrow 1$	3,04	0,49	0,12	0,01	26,23	6,62
$2 \rightarrow 3$	2,61	0,20	0,18	0,01	7,20	0,82
$3 \rightarrow 2$	2,75	0,21	0,15	0,01	6,28	0,63
$3 \rightarrow 4$	2,19	0,36	0,13	0,02	5,58	1,25
$4 \rightarrow 1$	3,50	0,60	0,11	0,01	8,49	1,72
$4 \rightarrow 2$	3,38	0,85	0,15	0,02	6,32	2,11
$4 \rightarrow 3$	2,90	0,41	0,10	0,01	6,23	1,13

TAB. 2.9 – Paramètres des distributions des temps d'attente pour le modèle de Weibull généralisé multi-états

Paramètre	Coefficient	Ecart-type
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. 2.10 – Probabilités de transition et coefficients de régression du modèle semi-markovien multivarié de type Weibull généralisé

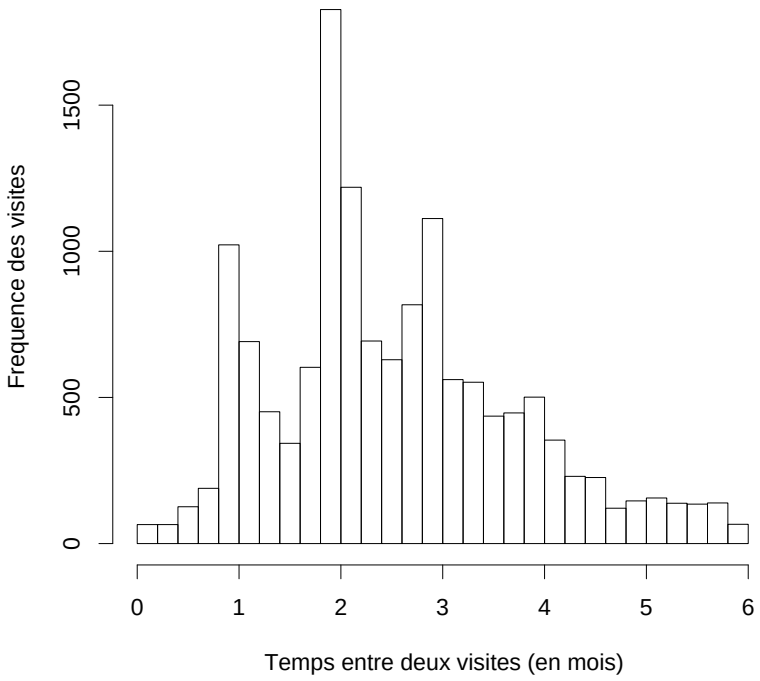

FIG. 2.1 – Répartition des délais entre visites

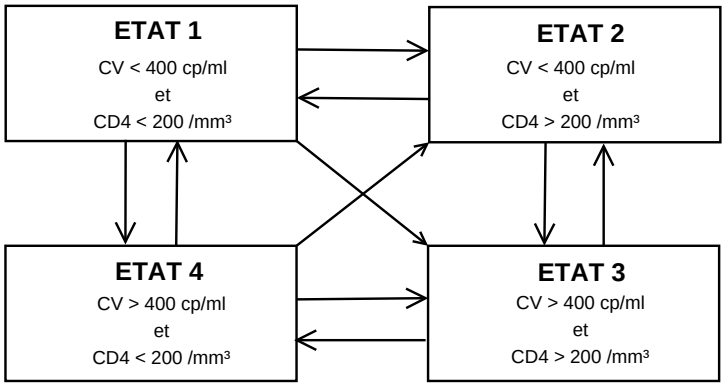

FIG. 2.2 – Graphique des transitions possibles pour l'étude du VIH

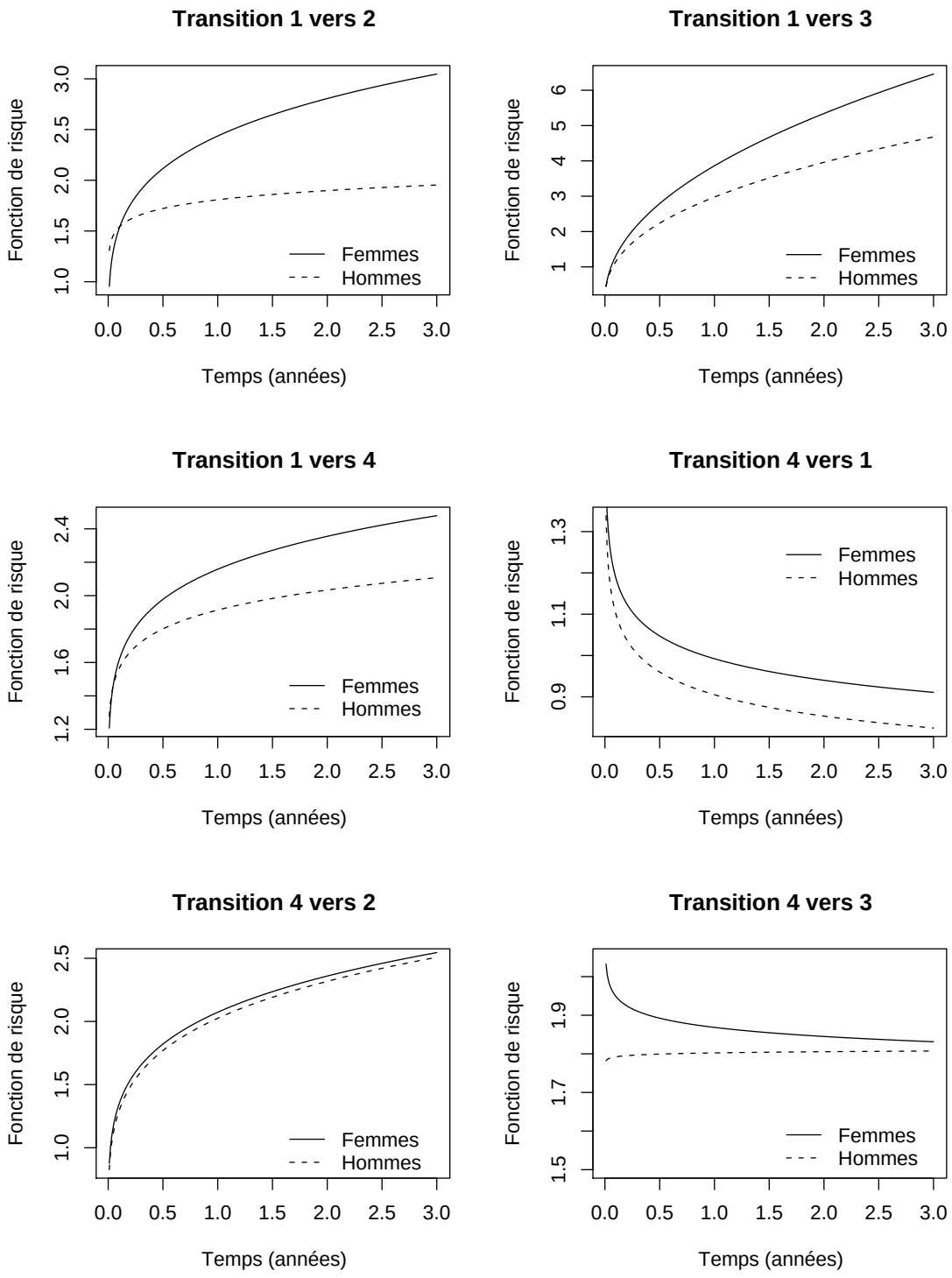

FIG. 2.3 – Fonctions de risque de type Weibull par transition et selon le sexe

Chapitre 3

Dépendance individuelle des observations

Comme évoqué dans l'introduction, les modèles de fragilité connaissent un succès grandissant dans le domaine médical. Ils permettent, en effet, l'analyse de données de survie non-indépendantes. C'est le cas en particulier pour les cohortes où certains événements peuvent être répétés chez un même individu.

Dans l'analyse multi-états, les données sont la plupart du temps considérées indépendantes [77, 78, 63]. Cependant, comme en analyse de survie, certaines corrélations peuvent apparaître dans des sous-groupes, dans le cas où les membres de ces sous-groupes posséderaient des caractéristiques communes non-observeés. Ce problème est soulevé dans l'application précédente où certaines transitions peuvent être répétées pour un même individu. Ce chapitre propose un modèle multi-états d'analyse des données de survie avec fragilité. Seul l'article de Ripatti et al. [79] propose un modèle à trois états (avec un état absorbant) permettant l'incorporation de fragilité. Cependant, ces derniers utilisent des fonctions de risque constantes par morceaux sans covariable.

3.1 Définition du modèle

Nous reprenons les mêmes notations et définitions que celles du chapitre 2. Le seul changement est dans la définition de l'échantillon observé. Pour écrire la vraisemblance du modèle, supposons que seule la dernière transition puisse être censurée à un temps d'attente égal à y. Les covariables, au moment du dernier passage dans l'état k, sont notées z_y. De plus, définissons δ l'indice de censure avec $\delta = 1$ si aucune censure n'est présente, et $\delta = 0$ sinon. A partir de la fonction (2.21) et en supposant qu'un sujet subisse n_{ij} fois la transition de l'état i vers l'état j après des durées d'attente $\{d_{ij1}, d_{ij2}, ..., d_{ijn_{ij}}\}$,

alors sa contribution individuelle à la vraisemblance peut s'écrire :

$$\mathcal{V} = \prod_{ij}\prod_{r=1}^{n_{ij}} \Big\{ P_{ij}\lambda_{ij}(d_{ijr}, z_{ijr}) exp\big(- \int_0^{d_{ijr}} \lambda_{ij}(u, z_{ijr})du\big)\Big\} \times \Big\{ \sum_{j\neq k} P_{kj} exp\big(- \int_0^{y} \lambda_{kj}(u, z_y)du\big)\Big\}^{\delta} \quad (3.1)$$

En posant $\Lambda_{ij}(x) = \int_0^x \lambda_{ij}(u)du$, l'équation (3.1) se simplifie :

$$\mathcal{V} = \prod_{ij} \Big\{ P_{ij}^{n_{ij}} \lambda_{ij}(d_{ij1}, z_{ij1})....\lambda_{ij}(d_{ijn_{ij}}, z_{ijn_{ij}}) exp\big(- [\Lambda_{ij}(d_{ij1}, z_{ij1}) + ... + \Lambda_{ij}(d_{ijn_{ij}}, z_{ijn_{ij}})]\big)\Big\} \times \Big\{ \sum_{j\neq k} P_{kj} exp\big(- \Lambda_{kj}(y, z_y)\big)\Big\}^{\delta} \quad (3.2)$$

Nous supposons que les fonctions de risque de base sont variables entre les individus et sont proportionnelles à un effet aléatoire (fragilité), ω. Les distributions de ces fragilités sont propres à chaque transition, on note donc la fonction de risque de base, au temps d'attente x et conditionnelle à la fragilité, égale à $\omega_{ij}\lambda_{ij}(x)$. Ainsi en reprenant l'équation (3.2), on peut écrire la contribution individuelle à la vraisemblance, conditionnellement à l'observation des termes de fragilité :

$$\mathcal{V}_{cond} = \prod_{ij} \Big\{ P_{ij}^{n_{ij}} \omega_{ij}^{n_{ij}} \lambda_{ij}(d_{ij1}, z_{ij1})....\lambda_{ij}(d_{ijn_{ij}}, z_{ijn_{ij}}) exp\big(- \omega_{ij}[\Lambda_{ij}(d_{ij1}, z_{ij1}) + ... + \Lambda_{ij}(d_{ijn_{ij}}, z_{ijn_{ij}})]\big)\Big\} \times \Big\{ \sum_{j\neq k} P_{kj} exp\big(- \omega_{kj}\Lambda_{kj}(y, z_y)\big)\Big\}^{\delta} \quad (3.3)$$

Pour simplifier, posons $V_{ij} = \Lambda_{ij}(d_{ij1}, z_{ij1}) + ... + \Lambda_{ij}(d_{ijn_{ij}}, z_{ijn_{ij}})$, l'intensité cumulée totale sur la période d'observation et $U_{kj} = \Lambda_{kj}(y, z_y)$. La contribution individuelle à la vraisemblance est alors égale à l'espérance sur Z de la forme conditionnelle (3.3) :

$$\begin{aligned} \mathcal{V} &= E\Big[\prod_{ij} \Big\{ P_{ij}^{n_{ij}} \omega_{ij}^{n_{ij}} \lambda_{ij}(d_{ij1}, z_{ij1})....\lambda_{ij}(d_{ijn_{ij}}, z_{ijn_{ij}}) exp\big(- \omega_{ij} v_{ij}\big)\Big\} \\ &\times \Big\{ \sum_{j\neq k} P_{kj} exp\big(- \omega_{kj} u_{kj}\big)\Big\}^{\delta}\Big] \end{aligned} \quad (3.4)$$

Les transformées de Laplace permettent une reformulation de (3.4). Reprenons la définition (1.20) en posant $L(a) = E[exp(aZ)]$, la r*ième* dérivée, $L^{(r)}(a)$, est alors égale à $(-1)^r E[Z^r exp(-aZ)]$. En faisant les hypothèses d'indépendance des fragilités entre les différentes transitions et de censure non-informative, l'équation (3.4) peut être reformulée :

$$\begin{aligned} \mathcal{V} &= \prod_{ij} \Big\{ (-1)^{n_{ij}} P_{ij}^{n_{ij}} \lambda_{ij}(d_{ij1}, z_{ij1})....\lambda_{ij}(d_{ijn_{ij}}, z_{ijn_{ij}}) L^{(n_{ij})}(v_{ij})\Big\} \\ &\times \Big\{ \sum_{j\neq k} P_{kj} L(u_{kj})\Big\}^{\delta} \end{aligned} \quad (3.5)$$

Finalement, la contribution individuelle à la logvraisemblance est donnée par :

$$\begin{aligned} ln\mathcal{V} &= \sum_{ij}\Big\{n_{ij}ln(P_{ij}) + \sum_{p=1}^{n_{ij}} ln\Big(\lambda_{ij}(d_{ijp},z_{ijp})\Big) + ln\Big((-1)^{n_{ij}}L^{(n_{ij})}(v_{ij})\Big)\Big\} \\ &+ \delta\, ln\Big(\sum_{j\neq k} P_{kj}L(u_{kj})\Big) \end{aligned} \tag{3.6}$$

Pour chaque transition, nous supposerons les ω_{ij} indépendants et identiquement distribués selon une loi Gamma, $G(\gamma_{ij}^{-1},\gamma_{ij}^{-1})$, dont la densité est donnée par (1.23). Cette distribution est l'une des plus utilisées, car elle possède de bonnes propriétés. Elle permet d'obtenir une espérance égale à 1 et une variance égale à γ_{ij}. Une grande valeur du paramètre γ_{ij} reflète ainsi une grande hétérogénéité. La transformée de Laplace, $L(a)$, pour une telle loi est égale à $(1+\gamma_{ij}a)^{-\gamma_{ij}^{-1}}$. La r*ième* transformée de Laplace sous l'hypothèse d'une distribution de la fragilité de type Gamma est alors égale à :

$$L^{(r)}(a) = -\gamma_{ij}^{r-1}(1+\gamma_{ij}a)^{(-1/\gamma_{ij}-r)}\prod_{p=1}^{r-1}(-1/\gamma_{ij}-p)$$

L'équation (3.6) peut alors être développée comme suit :

$$\begin{aligned} ln\mathcal{V} &= \sum_{ij}\Big\{n_{ij}ln(P_{ij}) + \sum_{p=1}^{n_{ij}} ln(\lambda_{ij}(d_{ijp},z_{ijp})) + ln\Big((-\gamma_{ij})^{n_{ij}-1}(1+\gamma_{ij}v_{ij})^{(-1/\gamma_{ij}-n_{ij})} \\ &\quad \prod_{p=1}^{n_{ij}-1}(-1/\gamma_{ij}-p)\Big)\Big\} + \delta\, ln\Big(\sum_{j\neq k} P_{kj}(1+\gamma_{kj}u_{kj})^{-1/\gamma_{kj}}\Big) \\ &= \sum_{ij}\Big\{n_{ij}ln(P_{ij}) + \sum_{p=1}^{n_{ij}} ln(\lambda_{ij}(d_{ijp},z_{ijp})) + (n_{ij}-1)ln(-\gamma_{ij}) + (-1/\gamma_{ij}-n_{ij}) \\ &\quad \times ln(1+\gamma_{ij}v_{ij}) + \sum_{p=1}^{n_{ij}-1} ln(-1/\gamma_{ij}-p)\Big\} + \delta\, ln\Big(\sum_{j\neq k} P_{kj}(1+\gamma_{kj}u_{kj})^{-1/\gamma_{kj}}\Big) \\ &= \sum_{ij}\Big\{n_{ij}ln(P_{ij}) + \sum_{p=1}^{n_{ij}} ln(\lambda_{ij}(d_{ijp},z_{ijp})) + (-1/\gamma_{ij}-n_{ij})ln(1+\gamma_{ij}v_{ij}) \\ &\quad + \sum_{p=1}^{n_{ij}-1}\Big(ln(-\gamma_{ij}) + ln(-1/\gamma_{ij}-p)\Big)\Big\} + \delta\, ln\Big(\sum_{j\neq k} P_{kj}(1+\gamma_{kj}u_{kj})^{-1/\gamma_{kj}}\Big) \end{aligned}$$

Or on a :

$$\begin{aligned} \sum_{p=1}^{n_{ij}-1}\Big(ln(-\gamma_{ij}) + ln(-1/\gamma_{ij}-p)\Big) &= \sum_{p=1}^{n_{ij}-1}\Big(ln(-\gamma_{ij}(-1/\gamma_{ij}-p))\Big) \\ &= \sum_{p=1}^{n_{ij}-1}\Big(ln(1+\gamma_{ij}p)\Big) \\ &= \sum_{p=1}^{n_{ij}}\Big(ln(1+(p-1)\gamma_{ij})\Big) \end{aligned}$$

On obtient alors :

$$\begin{aligned} ln\mathcal{V} &= \sum_{ij}\Big\{n_{ij}ln(P_{ij}) + \sum_{p=1}^{n_{ij}} ln(\lambda_{ij}(d_{ijp}, z_{ijp})) + (-1/\gamma_{ij} - n_{ij})ln(1+\gamma_{ij}v_{ij}) \\ &\quad + \sum_{p=1}^{n_{ij}}\Big(ln(1+(p-1)\gamma_{ij})\Big)\Big\} + \delta\, ln\Big(\sum_{j\neq k} P_{kj}(1+\gamma_{kj}u_{kj})^{-1/\gamma_{kj}}\Big) \\ &= \sum_{ij}\Big\{n_{ij}ln(P_{ij}) + \sum_{p=1}^{n_{ij}}\Big(ln(\lambda_{ij}(d_{ijp}, z_{ijp})) + ln(1+(p-1)\gamma_{ij})\Big) \\ &\quad -(1/\gamma_{ij} + n_{ij})ln(1+\gamma_{ij}v_{ij})\Big\} + \delta\, ln\Big(\sum_{j\neq k} P_{kj}(1+\gamma_{kj}u_{kj})^{-1/\gamma_{kj}}\Big) \end{aligned}$$

Pour clôturer la définition du modèle, il nous faut définir la distribution de la fonction de risque de base. La fonction de risque de type Weibull généralisée est donnée par (1.11). Par intégration, il en découle la fonction de risque cumulée suivante :

$$\Lambda_{0,ij}(x) = \Big(1 + \Big(\frac{x}{\sigma_{ij}}\Big)^{\nu_{ij}}\Big)^{1/\theta_{ij}} - 1$$

La logvraisemblance individuelle peut alors s'écrire :

$$\begin{aligned} ln\mathcal{V} &= \sum_{ij}\Big\{n_{ij}ln(P_{ij}) + \sum_{p=1}^{n_{ij}}\Big(ln\Big(\frac{1}{\theta_{ij}}\Big(1+(\frac{d_{ijp}}{\sigma_{ij}})^{\nu_{ij}}\Big)^{\frac{1}{\theta_{ij}}-1}\frac{\nu_{ij}}{\sigma_{ij}}\Big(\frac{d_{ijp}}{\sigma_{ij}}\Big)^{\nu_{ij}-1}exp(\beta_{ij}^T z_{ijp})\Big) \\ &\quad + ln\Big(1+(p-1)\gamma_{ij}\Big)\Big) - (1/\gamma_{ij}+n_{ij})ln\Big(1+\gamma_{ij}\sum_{p=1}^{n_{ij}}\Big(\Big(1+\Big(\frac{d_{ijp}}{\sigma_{ij}}\Big)^{\nu_{ij}}\Big)^{(1/\theta_{ij})}-1\Big) \\ &\times\ exp(\beta_{ij}^T z_{ijp})\Big)\Big\} + \delta\, ln\Big(\sum_{j\neq k} P_{kj}\Big(1+\gamma_{kj}\Big(\Big(1+\Big(\frac{y}{\sigma_{kj}}\Big)^{\nu_{kj}}\Big)^{1/\theta_{kj}}-1\Big) \\ &\times\ exp(\beta_{kj}^T z_y)\Big)^{-1/\gamma_{kj}}\Big) \end{aligned}$$

On retrouve finalement la logvraisemblance individuelle, dont la somme sur tous les sujets nous renvoie la logvraisemblance de l'échantillon.

$$\begin{aligned} ln\mathcal{V} &= \sum_{ij}\Big\{n_{ij}\Big[ln(P_{ij}) - ln(\theta_{ij}) + ln(\nu_{ij}) - \nu_{ij}ln(\sigma_{ij})\Big] + \sum_{p=1}^{n_{ij}}\Big((\frac{1}{\theta_{ij}}-1) \\ &\quad \times ln\Big(1+(\frac{d_{ijp}}{\sigma_{ij}})^{\nu_{ij}}\Big) + (\nu_{ij}-1)ln(d_{ijp}) + \beta_{ij}^T z_{ijp} + ln(1+(p-1)\gamma_{ij})\Big) \\ &\quad -(1/\gamma_{ij}+n_{ij})ln\Big(1+\gamma_{ij}\sum_{p=1}^{n_{ij}}\Big(\Big(1+\Big(\frac{d_{ijp}}{\sigma_{ij}}\Big)^{\nu_{ij}}\Big)^{(1/\theta_{ij})}-1\Big)exp(\beta_{ij}^T z_{ijp})\Big)\Big\} \\ &\quad +\delta\, ln\Big(\sum_{j\neq k} P_{kj}\Big(1+\gamma_{kj}\Big(\Big(1+\Big(\frac{y}{\sigma_{kj}}\Big)^{\nu_{kj}}\Big)^{1/\theta_{kj}}-1\Big)exp(\beta_{kj}^T z_y)\Big)^{-1/\gamma_{kj}}\Big) \end{aligned} \quad (3.7)$$

Il est intéressant de tester l'hypothèse nulle selon laquelle $\gamma_{ij} = 0$, correspondant à l'indépendance entre les observations d'un même sous-groupe. Ce test est à la limite de l'intervalle de définition du paramètre, $\gamma_{ij} \geq 0$. Néanmoins, Aalen et Husbey [2] ont montré que les valeurs de γ_{ij} peuvent être étendues à des valeurs légèrement négatives, rendant l'hypothèse nulle comme un point intérieur de l'espace de définition.

3.2 Application au VIH

Les mêmes données que celles utilisées dans le précédent chapitre ont été utilisées. Nous conservons aussi la structure multi-états (figure 2.2) et la stratégie de modélisation. L'estimation des coefficients de régression du modèle, obtenus par la maximisation de la fonction (3.7) est présentée dans le tableau (3.1). Cinq facteurs semblent être liés aux forces de transition, en particulier une co-infection par Hépatite B, associée aux transitions $1 \rightarrow 4$, $2 \rightarrow 1$ et $2 \rightarrow 3$. Par exemple, les patients co-infectés par Hépatite B ont 1,9 fois plus de chance de quitter l'état 2, sachant que l'état 1 suit. De la même manière, les patients qui n'ont pas été contaminés par une relation homosexuelle, semblent être associés à une transition $2 \rightarrow 1$ plus rapide.

Covariable	Transition	Estim.	ET	Risque Relatif	p-value
Hépatite B	$1 \rightarrow 4$	-0,67	0,30	0,51	0,0243
Hépatite B	$2 \rightarrow 1$	0,66	0,35	1,93	0,0581
Homosexualité	$2 \rightarrow 1$	-1,35	0,65	0,26	0,0384
Hépatite B	$2 \rightarrow 3$	-0,52	0,18	0,59	0,0044
Hépatite C	$3 \rightarrow 2$	0,31	0,13	1,36	0,0129

TAB. 3.1 – Paramètres de régression β_{ij} du modèle final avec fragilité

La figure (3.1) présente certaines fonctions de risque du processus semi-markovien, $\alpha_{ij}()$, comme défini par l'équation (2.13). Même si le rapport des risques associés aux temps d'attente est constant au cours du temps, l'interprétation du rapport de deux risques du processus semi-markovien n'est pas si simple. Ce dernier dépend du temps d'attente. Cependant, cette représentation graphique permet d'évaluer l'effet des covariables sans avoir à conditionner sur l'état qui suit (définition 2.6). Par exemple, la co-infection par Hépatite B influence la transition $1 \rightarrow 2$, même si cette covariable n'a pas été retenue comme associée directement à cette dernière transition.

Les paramètres des distributions des temps d'attente et des distributions des effets aléatoires sont présentés dans le tableau (3.2). Il est intéressant de noter que les variances des effets aléatoires, γ_{ij}, sont proches de la nullité, démontrant l'homogénéité des fonctions de risque entre individus. Une transformation logarithmique a été nécessaire pour mieux estimer ces valeurs limites. En utilisant un test de rapport de vraisemblance pour tester

l'hypothèse nulle selon laquelle γ_{ij}, $\forall i \neq j$, nous concluons que l'introduction des effets aléatoires n'est pas informative ($\chi^2 = 1,89$, $ddl = 8$ and $p = 0,9841$).

Comme dans l'analyse précédente supposant l'indépendance des observations, toujours à partir du tableau (3.2), remarquons que les valeurs des paramètres θ_{ij} et leurs écart-types justifient notre choix de fonction de risque des temps d'attente de type Weibull généralisé. Par exemple, l'intervalle de confiance à 95 % du paramètre θ_{12} est compris entre 2,72 et 7,14, celui-ci n'incluant pas la valeur 1. Cette distribution semble donc plus informative que celle de type Weibull. Ceci est aussi justifié par la figure (3.1).

	ν_{ij}		σ_{ij}		θ_{ij}		γ_{ij}	
Transit.	Estim	ET	Estim	ET	Estim	ET	Estim	exp(Estim)
$1 \rightarrow 2$	0,12	0,02	3,04	0,54	4,93	1,13	-8,18	0,0003
$1 \rightarrow 4$	0,13	0,02	3,29	0,76	4,52	1,38	-7,13	0,0008
$2 \rightarrow 1$	0,13	0,02	5,08	1,26	10,82	3,75	-1,11	0,3296
$2 \rightarrow 3$	0,18	0,03	2,64	0,34	4,24	0,94	-2,63	0,0721
$3 \rightarrow 2$	0,15	0,02	2,62	0,34	4,64	0,85	-9,14	0,0001
$3 \rightarrow 4$	0,11	0,02	2,30	0,44	4,50	1,17	-7,22	0,0007
$4 \rightarrow 1$	0,13	0,02	3,58	1,07	7,44	2,66	-8,29	0,0003
$4 \rightarrow 3$	0,11	0,02	2,81	0,59	5,15	1,39	-7,94	0,0004

TAB. 3.2 – Paramètres des distributions des temps de séjour du modèle final avec fragilité

3.3 Discussion

Dans ce chapitre, nous avons défini un modèle semi-markovien général avec effets aléatoires et des distributions des temps d'attente de type Weibull généralisé. Les effets aléatoires sont supposés suivre une loi Gamma, ainsi qu'être propre à chaque type de transition. En utilisant les transformées de Laplace pour obtenir une vraisemblance marginale, l'avantage de la méthode est d'utiliser la théorie classique du maximum de vraisemblance, ainsi que le test du rapport de vraisemblance.

Comme l'illustre notre application sur le VIH, ce modèle fournit un cadre utile pour les données multi-états répétées ou plus généralement groupées. Nous avons ici supposé que chaque individu forme un groupe de corrélation, une transition $i \rightarrow j$ donnée pouvant se produire plusieurs fois pour un même individu. Certaines covariables non-observées peuvent alors créer une variabilité entre sujets. Même si l'introduction de cette fragilité n'augmente pas significativement la vraisemblance du modèle pour cette application, cet apport pourrait se révéler pertinent dans d'autres cas. Il peut s'agir d'autres applications, ou d'un autre choix de modélisation de la fragilité. On pourrait par exemple envisager une distribution différente des effets aléatoires ou bien considérer une fragilité commune

à différentes transitions. La méthode proposée possède au moins l'avantage de pouvoir tester l'indépendance des temps de transition, hypothèse trop forte à formuler a priori.

La formulation du modèle et la vraisemblance correspondante (3.7) peuvent être utilisées pour modéliser différentes structures de corrélation. Par exemple pour prendre en compte un effet aléatoire du centre, d'une région géographique ou d'une famille. Ces effets peuvent être dus à des facteurs environnementaux ou génétiques qui ne sont pas pris en compte dans le modèle.

Les résultats obtenus concernant les facteurs de risque sont en accord avec la littérature médicale, excepté pour la co-infection par Hépatite. Cependant, à part pour les covariables associées à la transition $2 \rightarrow 1$, les résultats sont sensiblement différents de ceux obtenus par le même modèle supposant l'indépendance des données (tableau 2.10). Il serait peut être intéressant de réaliser un modèle différent pour les CD4 et la CV pour obtenir une analyse plus précise de cette covariable.

La principale limite de ce modèle, ainsi que de celui présenté dans le chapitre 2, est de ne pas prendre en compte la censure par intervalle des temps de transition observés. En effet, même si nous avons justifié l'absence de prise en compte de ce type de données incomplètes (figure 2.1), nous sommes bien confrontés au fait que la dynamique des deux marqueurs est continue, alors que le processus d'observation est discontinu. Dans les chapitres suivants, consacrés à l'analyse des greffes rénales, nous sommes contraints de prendre en compte ces censures par intervalle, la fréquence des observations pouvant être annuelle.

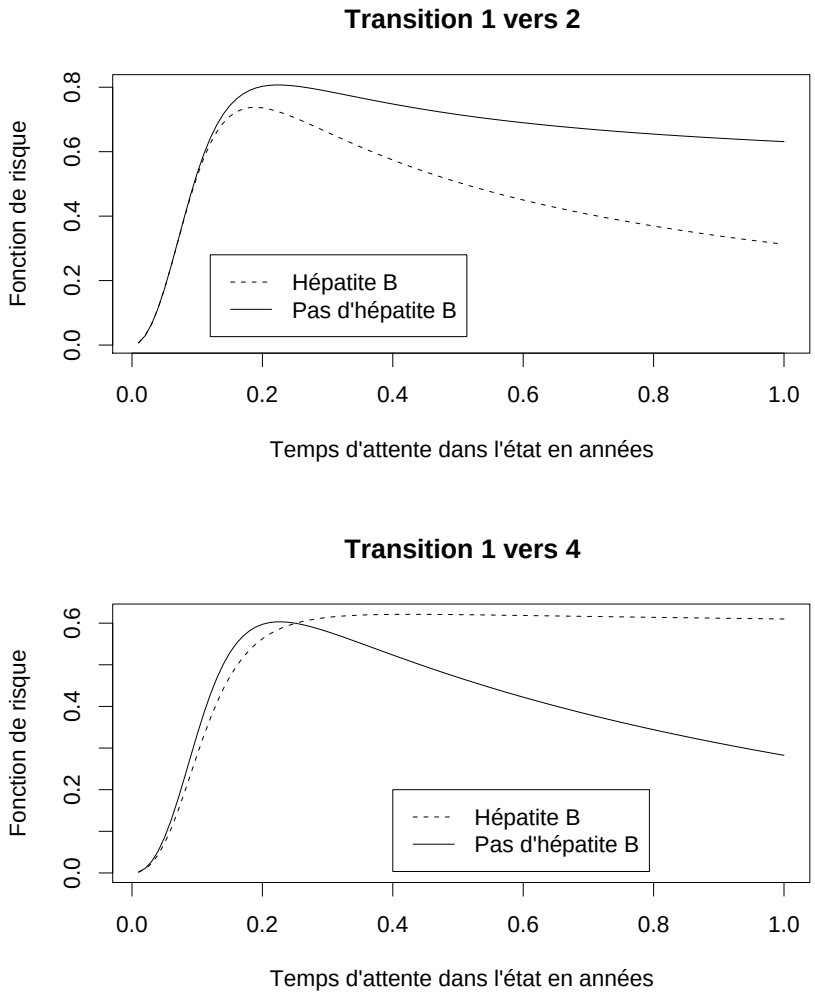

FIG. 3.1 – Fonction de risque du processus semi-markovien de l'état, $\alpha_{12}()$ et $\alpha_{14}()$, en fonction de la co-infection par hépatite B.

Chapitre 4

Censure par intervalle des durées

Fréquemment dans des études longitudinales, l'état d'une personne ne peut être connu qu'entre deux dates (visites médicales, mesures biologiques, etc.). La connaissance du temps de transition et des covariables temps-dépendantes correspondantes est alors censurée à droite et à gauche. Ce type de données incomplètes a été largement étudié en analyse de survie [70, 80, 81]. Concernant l'approche multi-états pour l'analyse de ce type de données incomplètes, Commenges [82] propose une modélisation markovienne non-homogène basée sur la vraisemblance pénalisée. Satten et Sternberg [61] montrent qu'un modèle semi-markovien non-paramétrique peut être adapté à la censure par intervalle.

Dans ce chapitre, nous définissons un modèle semi-markovien paramétrique basé sur des fonctions de risque Weibull généralisé et dont la vraisemblance permet la prise en compte de la censure par intervalle. Cette généralisation est motivée par l'analyse de l'évolution des patients transplantés d'un rein. Les individus sont régulièrement rappelés en consultation après avoir été transplantés. Deux marqueurs continus sont mesurés pour définir les états de gravité : la protéinurie (PR) et la clairance de la créatinine (CL). Les temps de transition entre ces états ne sont donc pas exactement connus. De plus, un patient peut commencer son évolution dans chacun de ces états de gravité. Nous introduisons des probabilités d'initialisation du processus dépendantes de covariables, permettant ainsi de gérer le processus dès son origine.

4.1 Modélisation statistique

Nous reprendrons les définitions du modèle semi-markovien simple, défini dans le chapitre 2. Deux composantes seront cependant ajoutées : la définition des probabilités d'initialisation et l'incorporation de la censure par intervalle dans la vraisemblance.

4.1.1 Probabilités initiales

Cette paramétrisation supplémentaire est pertinente lorsque le processus est observé dès son origine dans différents états. Ce type d'initialisation peut être modélisé en utilisant le principe des régressions multinomiales, bien défini par McCullagh [64, 83]. Posons $G = \{1, 2, ..., c\}$ l'espace fini d'états initiaux, avec $G \in E$. Comme nous l'avons défini précédemment, E est l'espace fini d'états total. Nous définissons :

$$\pi_{0j} = P(X_{h,1} = j) \quad \forall j \in G \tag{4.1}$$

comme la probabilité que le sujet h commence son suivi dans l'état j. En définissant $z_{h,0j}$, le vecteur de covariables associé à l'état j pour cet individu, la régression logistique multinomiale peut alors être définie par :

$$\pi_{0j} = exp(\gamma_{0j} + \beta_{0j} z_{h,0j}) / \sum_{k=1}^{c} exp(\gamma_{0k} + \beta_{0k} z_{h,0k}) \quad \text{for } j = 1, .., c \tag{4.2}$$

où γ_{0j} et β_{0j} sont respectivement l'intercept et le vecteur des coefficients de régression associés à $z_{h,0j}$. En supposant que les états initiaux sont mutuellement exclusifs et exhaustifs, nous devons vérifier la contrainte $\sum_{j=1}^{c} \pi_{0j} = 1$. En choisissant le c*ième* état initial comme état de référence, nous adoptons par convention la nullité de γ_{0c} et β_{0c}.

4.1.2 Censure par intervalle

Dans le chapitre 3, l'indicatrice d identifie une observation censurée à droite. Pour prendre en compte les censures à gauche et par intervalle, nous définissons plus de composantes, comme proposé par Odell et al. [70]. Soit $d_{h,r}$ le temps passé dans le r*ième* état pour le sujet h. Posons $d^0_{h,r}$ comme étant une durée telle que si $d^0_{h,r} < d_{h,r}$, alors $d^0_{h,r}$ est observée et $d_{h,r}$ ne l'est pas. D'une manière similaire, posons $d^1_{h,r}$ une durée telle que si $d^1_{h,r} > d_{h,r}$, alors $d^1_{h,r}$ est observée et $d_{h,r}$ ne l'est pas.

Dans notre contexte de l'analyse des données longitudinales, quatre types d'observation doivent être distingués. Nous considérons ici comme individu statistique la transition entre deux états. Chaque transition est considérée indépendante des éventuelles transitions supplémentaires d'un même sujet. Nous supposerons toujours le processus de censure de ces transitions non-informatif [67].

(i) Considérons le temps d'attente $d_{h,r}$ censuré à droite ($d^0_{h,r} \leq d_{h,r} < d^1_{h,r} = \infty$), indicé par $\delta^R_{h,r}$ pour construire la vraisemblance. La contribution d'une telle observation est égale à :

$$\begin{aligned} P(x > d^0_{h,r} | X_{h,r} = i) &= \sum_{j \neq i} P(X_{h,r+1} = j | X_{h,r} = i) \int_{d^0_{h,r}}^{\infty} f_{ij}(u) du \\ &= \sum_{j \neq i} P_{ij} S_{ij}(d^0_{h,r}) \end{aligned} \tag{4.3}$$

(ii) De la même manière, $\delta^I_{h,r}$ indique que la durée $d_{h,r}$ est censurée par intervalle $(0 < d^0_{h,r} < d_{h,r} \leq d^1_{h,r} < \infty)$. Ainsi, la contribution peut s'écrire :

$$\begin{aligned} P(d^0_{h,r} < x < d^1_{h,r}, X_{h,r+1} = j | X_{h,r} = i) &= P(X_{h,r+1} = i | X_{h,r} = i) \int_{d^0_{h,r}}^{d^1_{h,r}} f_{ij}(u)du \\ &= P_{ij}\Big(\int_0^{d^1_{h,r}} f_{ij}(u)du - \int_0^{d^0_{h,r}} f_{ij}(u)du \Big) \\ &= P_{ij}\Big(F_{ij}(d^1_{h,r}) - F_{ij}(d^0_{h,r}) \Big) \\ &= P_{ij}\Big(S_{ij}(d^0_{h,r}) - S_{ij}(d^1_{h,r}) \Big) \end{aligned} \tag{4.4}$$

(iii) La censure à gauche est un cas particulier de la censure par intervalle où $d^0_{h,r} = 0$. L'équation (4.4) devient alors :

$$\begin{aligned} P(x < d^1_{h,r}, X_{h,r+1} = j | X_{h,r} = i) &= P_{ij}(1 - S_{ij}(d^1_{h,r})) \\ &= P_{ij} F_{ij}(d^1_{h,r}) \end{aligned} \tag{4.5}$$

(iv) Enfin, considérons le cas d'une observation exactement observée, indiquée par $\delta^E_{h,r} = 1 - \delta^R_{h,r} - \delta^I_{h,r}$. Sa contribution à la vraisemblance est égale :

$$\lim_{\Delta d \to 0^+} P(d_{h,r} < x < d_{h,r} + \Delta d, X_{h,r+1} = j | X_{h,r} = i)/\Delta d = P_{ij} f_{ij}(d_{h,r})$$

Le produit de ces contributions permet d'obtenir la logvraisemblance suivante :

$$\begin{aligned} ln\mathcal{V} &= \sum_h \Big\{ \gamma_{0X_{h,1}} + \beta_{0X_{h,1}} z_{h,0X_{h,1}} - ln\Big(\sum_{i=1}^{c} exp(\gamma_{0i} + \beta_{0i} z_{h,0X_{h,1}}) \Big) \\ &+ \sum_{ij} \sum_{X_{h,r}=i, X_{h,r+1}=j} \Big\{ \delta^E_{h,r} \big[ln(P_{ij}) + ln\big(f_{ij}(d_{h,r}) \big) \big] \\ &+ \delta^I_{h,r} \big[ln(P_{ij}) + ln\big(S_{ij}(d^0_{h,r}, z_{h,ij}) - S_{ij}(d^1_{h,r}, z_{h,ij}) \big) \big] \Big\} \\ &+ \sum_{ij} \sum_{X_{h,r}=i} \Big\{ \delta^R_{h,r} \big[ln\Big(\sum_{j \neq i} P_{ij} S_{ij}(d^0_{h,r}, z_{h,ij}) \Big) \big] \Big\} \Big\} \end{aligned} \tag{4.6}$$

où par convention $\gamma_{03} = \beta_{03} = 0$. La logvraisemblance, correspondant à l'équation (4.6), est calculée explicitement dans l'annexe A, pour des fonctions de risque Weibull généralisé (1.11) et des covariables dont l'effet est proportionnel (2.22).

4.2 Application à la transplantation rénale

4.2.1 Présentation des données et de la structure multi-états

De tels développements, pour la prise en compte de la censure par intervalle, ont été motivés par l'étude de l'évolution des patients transplantés d'un rein à partir de la

cohorte DIVAT de Nantes. Nous considérons les greffes depuis l'année 1990. La structure du modèle unidirectionnel, représentée par la figure (4.1) a entièrement été définie par les cliniciens.

L'origine du suivi est définie trois mois après la date de transplantation. Le processus est alors considéré comme stabilisé. Les deux marqueurs d'aggravation du greffon sont pris en compte différemment. Considérons tout d'abord la protéinurie. Le rein est capable de retenir la plupart des protéines qui passent à son niveau. Par contre, en cas de maladies rénales et notamment dans la dysfonction chronique des greffons rénaux, la perméabilité du rein est augmentée et des protéines peuvent se retrouver dans les urines. La protéinurie rencontrée dans la dysfonction chronique des greffons serait due à une altération de la perm-séléctivité des glomérules. L'existence d'une protéinurie un an post-greffe serait corrélée avec la perte des greffons mais aussi avec le décès [84]. Dans une étude récente de Fernandez-Fresnedo et al., portant sur 3365 patients, la protéinurie à un an post-greffe a été analysée en fonction des seuils ($<$0,5, 0,5-1, $>$1 $g/jour$). La survie des greffons est corrélée à la présence mais aussi au degré de la protéinurie [85]. En reprenant ces classes de PR, trois états de gravité sont définis.

Il est complémentaire d'étudier la clairance de la créatinine. Ce dosage établit un rapport entre la créatinine sanguine et la créatinine urinaire. La clairance de la créatinine permet d'estimer le débit de filtration glomérulaire, c'est à dire la fonction des reins. Plus la valeur de CL diminue, moins bon est le pronostic. CL est calculée à partir de la formule MDRD (1). La forte variation de cette variable entre les individus, nous a amenée plutôt à la considérer en terme d'évolution relative. Si en moins d'un an ce marqueur diminue de plus de 20 %, on considère une aggravation de l'état 1 vers l'état 2. Une diminution de plus de 30 % est considérée comme plus grave en transitant vers l'état 3. Nous considérons que ces aggravations ne peuvent pas être récupérées. Cette progression irréversible explique le modèle unidirectionnel (4.1).

Excepté pour des cas particuliers, ces temps de transition ne sont pas exactement connus. En effet, ces marqueurs évoluent continuellement, alors que les mesures biologiques sont discontinues. Deux états absorbants sont aussi définis dans la structure (4.1), leur date d'occurrence est exactement renseignée. Il s'agit du retour en dialyse et du décès du patient avec un greffon fonctionnel. Une analyse préliminaire a montré que toutes les transitions possibles entre états transitoires et absorbants ne sont pas représentées. Ainsi, un patient ne peut transiter vers un état absorbant qu'à partir de l'état 3, c'est-à-dire le stade d'aggravation du greffon le plus élevé. Cette observation justifie la structure du modèle choisi.

Un des objectifs de ce modèle est d'analyser l'évolution des patients à travers cette structure en prenant en compte les éventuels effets de 5 covariables : le sexe (1 si homme et 0 sinon), l'ischémie froide (1 si supérieure ou égale à 16 heures et 0 sinon), l'année de la transplantation (1 si inférieure à 1998 et 0 sinon), l'âge du receveur à la date de greffe (1 si supérieur ou égal à 55 ans et 0 sinon) et le délai de reprise de l'activité rénale après

la greffe (1 si supérieur ou égal à 6 jours et 0 sinon).

L'échantillon est constitué de 997 patients, représentant 22131 mesures biologiques et 1980 transitions (qu'elles soient exactement observées ou censurées). Le tableau (4.1) décrit les fréquences de ces transitions. On peut voir que seulement 15,8 % et 12,2 % du total des transitions $3 \rightarrow 4$ et $3 \rightarrow 5$ correspondent à des temps exactement observés. En effet, même si la date d'entrée dans un des états absorbants est connue, la date d'entrée dans le dernier état transitoire est le plus souvent censurée par intervalle. Ces deux pourcentages concernent donc des patients qui commencent directement leur évolution dans l'état 3. Excepté pour ces cas particuliers, toutes les autres transitions sont soit censurées par intervalle, soit à gauche, soit à droite.

Transition	Effectif	Pourcentage	Observation exacte[1]
$0 \rightarrow 1$	807	80,9 %	.
$0 \rightarrow 2$	112	11,2 %	.
$0 \rightarrow 3$	78	7,8 %	.
total	997	100,0 %	.
$1 \rightarrow 2$	396	20,0 %	0,0 %
$1 \rightarrow 3$	299	15,1 %	0,0 %
$1 \rightarrow$ censure	103	5,2 %	.
$2 \rightarrow 3$	330	16,7 %	0,0 %
$2 \rightarrow$ censure	159	8,0 %	.
$3 \rightarrow 4$	139	7,0 %	15,8 %
$3 \rightarrow 5$	41	2,1 %	12,2 %
$3 \rightarrow$ censure	513	25,9 %	.
Total	1980	100,0 %	.

[1] Pourcentage de temps exactement observés pour une transition donnée

TAB. 4.1 – Répartition des transitions selon leur contribution à la vraisemblance

4.2.2 Stratégie de modélisation

La même stratégie d'analyse des données que celle utilisée pour les applications sur le VIH a été mise en place ici. Les covariables agissant sur les forces de transition sont sélectionnées en univarié avec un seuil de signification égal à 0,20. Le vecteur final de covariables, spécifique à chaque transition, est obtenu par une procédure de sélection descendante ($p \leq 0,05$). La dernière étape consiste à évaluer la pertinence des distributions de type Weibull généralisé pour la modélisation des forces de transition. Ainsi, toujours en utilisant le test du rapport de vraisemblance, on teste l'hypothèse nulle selon laquelle le paramètre θ_{ij} est égal à la valeur 1 ($\forall\, ij \in \{12, 13, 23, 34, 35\}$). Comme précédemment, la maximisation de la logvraisemblance (4.6) a été réalisée sous R. L'algorithme de *quasi-Newton* [75] permet d'obtenir les estimations et les variances des paramètres du modèle.

Concernant le test de l'hypothèse de semi-proportionnalité des risques, un modèle est calculé pour chaque modalité des covariables. On peut alors identifier si elle est abusive. On utilise une approche graphique simple, en représentant le logarithme de la fonction de risque cumulé en fonction du temps d'attente dans l'état, pour chaque groupe et chaque transition [86, 87]. En effet, pour une covariable z à deux modalités (1 ou 0) et d'après l'équation (2.24) :

$$S_{ij}(x,z) = S_{0,ij}(x)^{exp(\beta_{ij}z)} \tag{4.7}$$

d'où

$$\begin{aligned} ln(-ln(S_{ij}(x,z))) &= ln(-ln(S_{0,ij}(x))) + \beta_{ij}z \\ &= ln\Big(\Lambda_{0,ij}(x)\Big) + \beta_{ij}z \end{aligned} \tag{4.8}$$

Sous l'hypothèse de semi-proportionnalité, les deux fonctions ainsi tracées dans chacun des groupes doivent être séparées par une constante β_{ij}. Autrement dit, elles doivent être parallèles. Cette méthode est plus aisée à évaluer visuellement que la simple représentation des fonctions de risque dans chaque groupe.

4.2.3 Résultats

Vérification de l'hypothèse de semi-proportionnalité

Nous présentons ici uniquement les résultats concernant l'âge du receveur pour ne pas surcharger le document en graphiques. La figure (4.2) présente ainsi, pour différentes transitions, le logarithme de la fonction de risque cumulé pour les patients de plus ou moins 55 ans à la date de greffe. On peut voir que l'hypothèse semble être respectée uniquement pour la transition $3 \rightarrow 4$. Pour les autres transitions et covariables, de nombreuses associations ne semblent pas proportionnelles. Seules celles respectant l'hypothèse sont incluses dans l'analyse.

Estimation des paramètres

L'analyse univariée a permis de réduire les covariables sélectionnées au nombre de 14. Après la procédure de sélection multivariée et descendante, 9 facteurs sont finalement retenus ($ln\mathcal{V} = -4864,62$). Les résultats de l'estimation des coefficients de régression sont présentés dans le tableau (4.2). Pour interpréter ces quelques facteurs influençant l'évolution du patient, nous sommes contraints de conditionner sur l'état qui suit (voir définitions 2.6 et 2.22). Les patients recevant un rein avant 1998 ont 2,2 fois plus de risque de subir une transition $1 \rightarrow 2$ par rapport à ceux greffés à partir de 1998. En respectant la même formulation, l'année de transplantation constitue un facteur de risque des transitions $1 \rightarrow 3$ et $2 \rightarrow 3$. Finalement, être une femme ou être âgé de plus de 55 ans

à la date de greffe semblent constituer des facteurs augmentant respectivement les forces de transition $1 \to 3$ et $3 \to 5$.

Transition	Covariable	Estim.	ET	RR	p-value
$0 \to 1$	Intercept	2,85	0,19	.	0,0001
$0 \to 1$	Sexe du receveur	-0,39	0,17	.	0,0226
$0 \to 1$	Délai de reprise	-0,53	0,17	.	0,0014
$0 \to 2$	Intercept	-0,67	0,44	.	0,1258
$0 \to 2$	Ischémie froide	1,13	0,44	.	0,0092
$1 \to 2$	Année de la greffe	-0,80	0,12	0,45	0,0001
$1 \to 3$	Sexe du receveur	0,29	0,15	1,34	0,0484
$1 \to 3$	Année de la greffe	-1,20	0,21	0,30	0,0001
$2 \to 3$	Année de la greffe	-0,54	0,12	0,59	0,0001
$3 \to 5$	Age du receveur	1,48	0,39	4,41	0,0001

TAB. 4.2 – Coefficients de régression du modèle multivarié final

D'autres facteurs, présentés dans la partie supérieure du tableau (4.2), semblent influencer les probabilités d'initialisation du processus. Ces probabilités sont calculées à partir des estimations précédentes et sont fournies dans le tableau (4.3). La majorité des patients commence dans l'état 1, cependant les hommes ont une probabilité plus faible de commencer dans cet état par rapport aux femmes (0,82 contre 0,87 pour une ischémie froide supérieure ou égale à 16 heures et un délai de reprise inférieur à 6 jours). Le délai de reprise et l'ischémie froide constituent aussi des covariables associées à ces probabilités initiales. Une ischémie supérieure à 16 heures augmente le risque de commencer dans l'état 2, alors qu'un délai de reprise inférieur à 6 jours augmente les chances de commencer dans l'état 1.

		Femmes			Hommes		
		π_{01}	π_{02}	π_{03}	π_{01}	π_{02}	π_{03}
Ischémie froide	DGF $<$ 6 jours	0,92	0,03	0,05	0,88	0,04	0,08
$<$ 16 heures	DGF $\geq$ 6 jours	0,87	0,04	0,09	0,82	0,06	0,12
Ischémie froide	DGF $<$ 6 jours	0,87	0,08	0,05	0,82	0,11	0,07
$\geq$ 16 heures	DGF $\geq$ 6 jours	0,80	0,13	0,08	0,72	0,17	0,11

TAB. 4.3 – Probabilités d'un patient à commencer dans l'état j, π_{0j} $(j = 1, 2, 3)$

La figure (4.3) présente certains exemples de fonctions de risque du processus semi-markovien, $\alpha_{34}()$ et $\alpha_{35}()$, selon les modalités de l'âge du receveur. Il est intéressant de remarquer qu'une covariable peut influencer indirectement une transition $i \to j$, même si une seule association significative est retenue pour une autre transition $i \to k$, avec le même état de départ. De plus, la figure (4.4) montre le rapport entre ces deux fonctions

de risque. Les valeurs sont supérieures à 1, démontrant le rôle protecteur chez les moins de 55 ans à la date de greffe. Ce risque n'est pas constant au cours du temps d'attente dans l'état 3.

Les résultats concernant les paramètres des distributions des temps d'attente sont présentés dans le tableau (4.4). Excepté pour la transition de l'état 3 à l'état 4, les valeurs de θ_{ij} montrent que l'utilisation de lois de type Weibull généralisé est informative par rapport à des fonctions de type Weibull. En effet, la statistique LRS pour tester l'hypothèse nulle $\{\theta_{ij} = 1 \ ; \ \forall ij = 12, 13, 23, 35\}$ est égale à 34,37, correspondant à une p-value inférieure à 0,0001. Néanmoins, une loi de Weibull apparaît suffisante pour la transition de l'état 3 vers l'état 4 ; la statistique LRS pour l'hypothèse nulle $\{\theta_{34} = 1\}$ est égale à 0,69 (p=0,9520). La fonction de risque est alors monotone croissante.

	σ_{ij}		ν_{ij}		θ_{ij}	
Transition	Estim.	ET	Estim.	ET	Estim.	ET
$1 \rightarrow 2$	36,14	31,97	0,53	0,03	0,24	0,09
$1 \rightarrow 3$	34,11	65,20	0,52	0,05	0,19	0,15
$2 \rightarrow 3$	33,40	31,34	0,56	0,03	0,30	0,13
$3 \rightarrow 4$	10,16	1,56	1,49	0,11	.	.
$3 \rightarrow 5$	18,48	47,62	1,14	0,23	1,46	3,75

TAB. 4.4 – Paramètres des distributions des temps d'attente pour le modèle multivarié final ($P_{12} = 0,59$ $ET = 0,02$, $P_{34} = 0,74$ $ET = 0,10$)

La figure (4.5) montre les fonctions de survie, $S_{ij}()$, et de risque, $\lambda_{ij}()$, associées aux paramètres du tableau (4.4). Les fonctions de risque des transitions $1 \rightarrow 2$, $1 \rightarrow 3$ et $2 \rightarrow 3$ sont en forme de $\cup$. Leurs minima (voir équation 1.12) correspondent respectivement à des temps d'attente égaux à 6,10, 2,99 et 9,95 années. Après être entré dans un état, le risque de transiter à nouveau est grand mais décroît ensuite. Cette dynamique correspond à une réalité clinique, une transition récente indiquant une instabilité du patient. Si ce patient reste quelques temps dans ce nouvel état, sa stabilité est traduite par une diminution de la fonction de risque. Cependant, après un temps significatif passé dans l'état, le vieillissement du processus induit une augmentation de la force de transition. A l'inverse de ces formes en $\cup$, la fonction de risque propre à la transition $3 \rightarrow 5$ est de type $\cap$, avec un maximum pour 12,86 années passées dans l'état 3.

En supplément à l'ensemble de ces résultats, il peut être pertinent d'étudier l'effet marginal d'une covariable sur l'occurrence d'un événement final, et ceci en fonction du temps depuis la greffe (et non pas en fonction du temps passé dans l'état). Ce type de calcul est particulièrement intéressant lorsqu'un facteur est associé à plusieurs transitions. Par exemple, l'année de la greffe est associée aux transitions $1 \rightarrow 2$, $1 \rightarrow 3$ et $2 \rightarrow 3$, montrant ainsi que cette covariable est associée à l'aggravation du greffon. La question est alors d'identifier l'impact global de ce facteur sur un événement terminal, que ce soit

le retour en dialyse ou le décès du patient. Cette fonction de probabilité marginale d'un événement final k ($k = 4, 5$) est déduite des équations (2.1), (2.5) et (4.2), en utilisant le produit de convolution :

$$\begin{aligned} p_k(t) &= \pi_{03} P_{3k} f_{3k}(t) + \pi_{02} \int_0^t P_{23} f_{23}(x) P_{3k} f_{3k}(t-x) dx + \pi_{01} \Big[\int_0^t P_{13} f_{13}(x) \\ \times & \quad P_{3k} f_{3k}(t-x) dx + \int_0^t P_{12} f_{12}(x) \int_0^{t-x} P_{23} f_{23}(y) P_{3k} f_{3k}(t-x-y) dy dx \Big] \end{aligned} \quad (4.9)$$

Dans notre application, nous avons décidé de représenter l'effet marginal de l'année de greffe et du délai de reprise sur le pronostic final du greffon. La figure (4.6) montre les probabilités de retour en dialyse et de décès. Pour les patients transplantés avant 1998, un meilleur pronostic du rein peut clairement être identifié. L'effet du délai de reprise ne semble pas avoir un impact aussi important, même si son unique association avec les probabilités initiales modifie légèrement les probabilités d'occurrence d'un événement terminal.

4.3 Discussion

Dans ce chapitre, nous avons défini un modèle pour l'analyse des données longitudinales multi-états et censurées par intervalle. Cette approche est appliquée à l'évolution des patients greffés d'un rein, pour laquelle l'utilisation de distributions non-monotones se révèle une nouvelle fois significativement plus informative qu'une loi Weibull. De plus, l'adaptation de la régression logistique multinomiale et multivariée, pour l'analyse de l'initialisation du processus, apparaît aussi comme un apport dans la modélisation du processus stochastique.

La limite principale de cette modélisation est de considérer chaque transition comme indépendante, c'est à dire sans considérer l'information apportée par les autres transitions ayant lieu chez un même sujet. Une solution pour s'affranchir de ce problème est de ne plus considérer la transition comme individu statistique, mais le sujet lui-même. Dans ce cas, même si une transition est censurée par intervalle, le fait que le saut se produise entre deux dates apporte de l'information pour les temps des transitions qui vont suivre pour ce même individu. Il faut alors prendre en compte le temps chronologique depuis la greffe dans l'écriture de la vraisemblance du modèle, en plus des temps de séjour. Cette prise en compte plus précise et plus adéquate de la censure par intervalle sera explicitée plus en détails dans le chapitre 6.

Comme pour les applications concernant le VIH, une des limites de l'approche présentée est l'hypothèse de semi-proportionnalité. L'analyse stratifiée, permettant d'examiner graphiquement la validité de cette hypothèse, nous contraint à ne pas modéliser l'effet de

covariables sur certaines transitions. Le chapitre suivant est consacré à la modélisation de risques non-proportionnels, généralisant ainsi le modèle présenté ici.

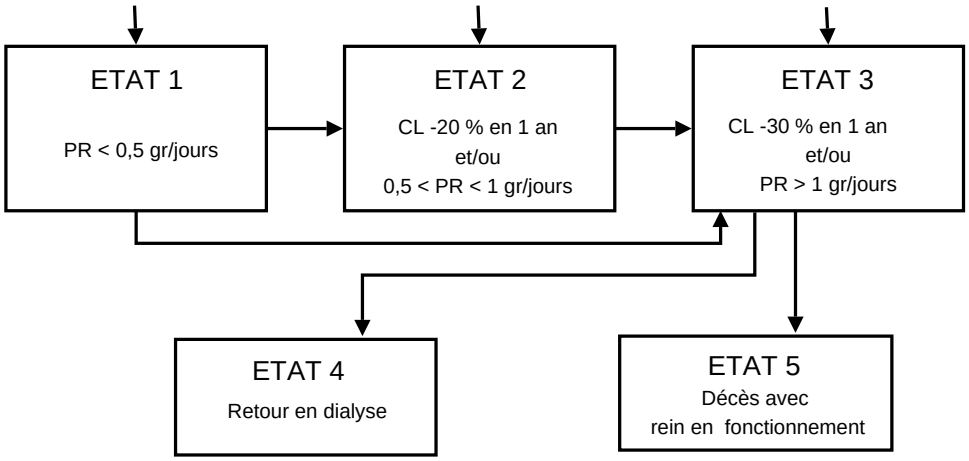

FIG. 4.1 – Structure du modèle multi-états pour l'analyse de la progression des patients transplantés d'un rein

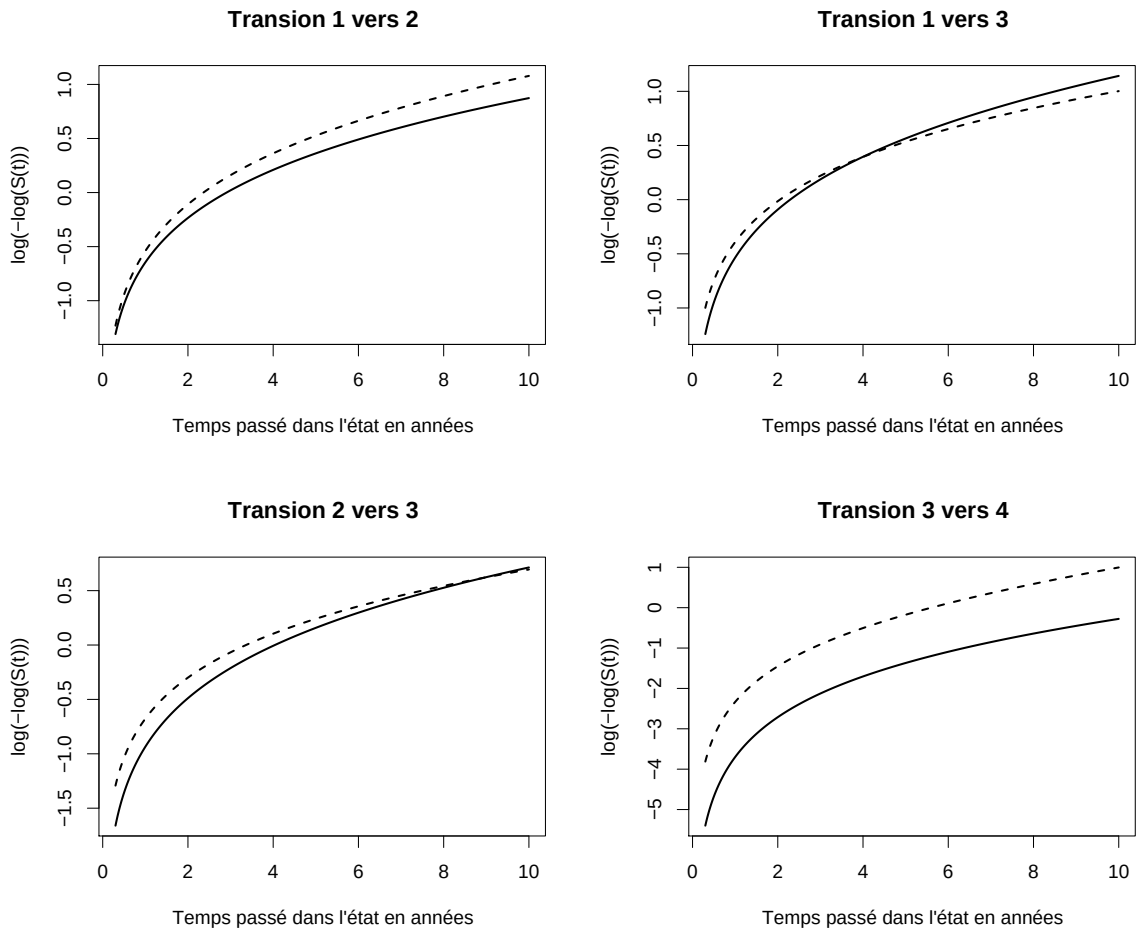

FIG. 4.2 – Test graphique de l'hypothèse de semi-proportionnalité des fonctions de risque pour l'âge du receveur (- - - si supérieur ou égal à 55 ans et —– sinon)

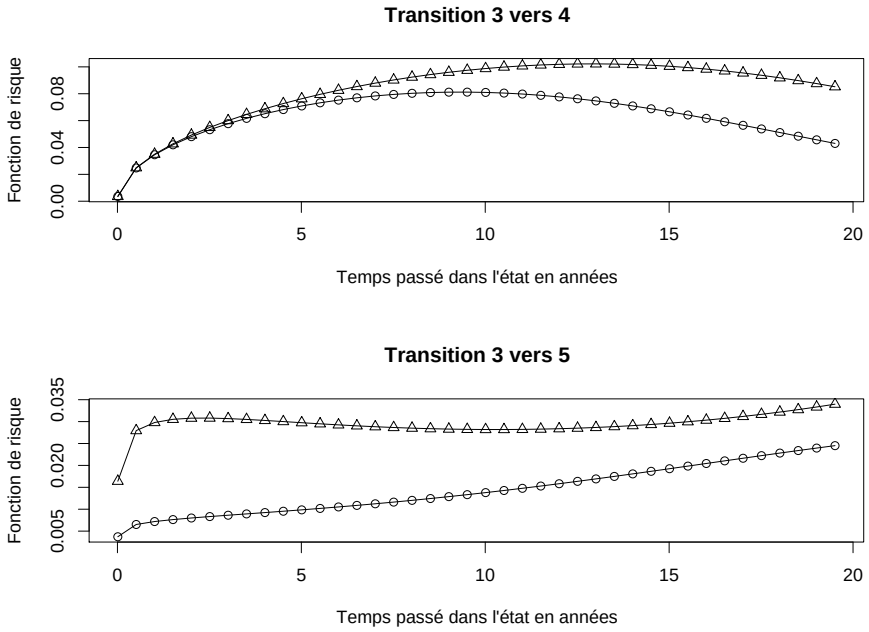

FIG. 4.3 – Fonctions de risque du processus semi-markovien $3 \rightarrow 4$ et $3 \rightarrow 5$, $\alpha_{34}()$ et $\alpha_{35}()$ respectivement. $\triangle\ \triangle\ \triangle$ âge du receveur ≥ 55 ans ; o o o âge du receveur < 55 ans.

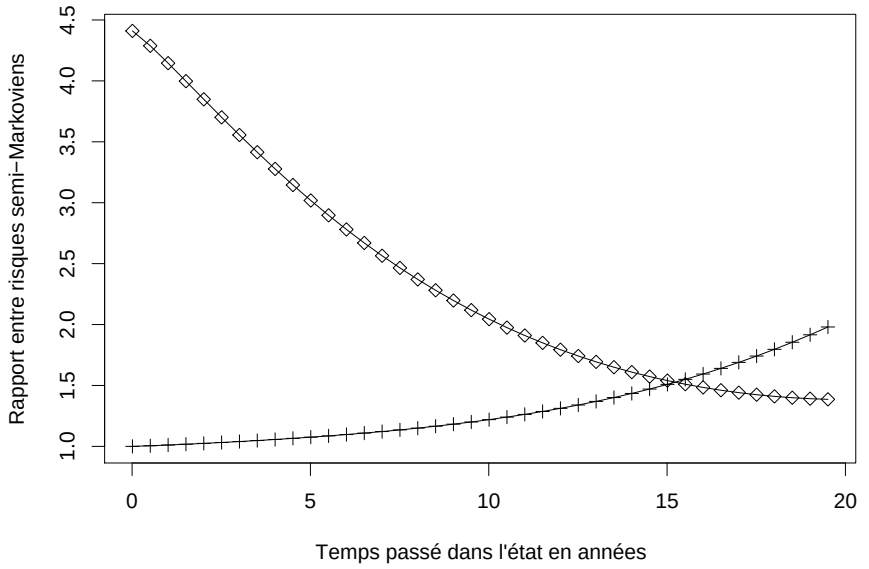

FIG. 4.4 – Effets de l'âge du receveur sur les transitions $3 \rightarrow 4$ et $3 \rightarrow 5$. $\diamond\ \diamond\ \diamond$ âge ≥ 55 ans vs âge < 55 ans (transition $3 \rightarrow 4$) ; $+\ +\ +$ âge ≥ 55 ans vs âge < 55 ans (transition $3 \rightarrow 5$).

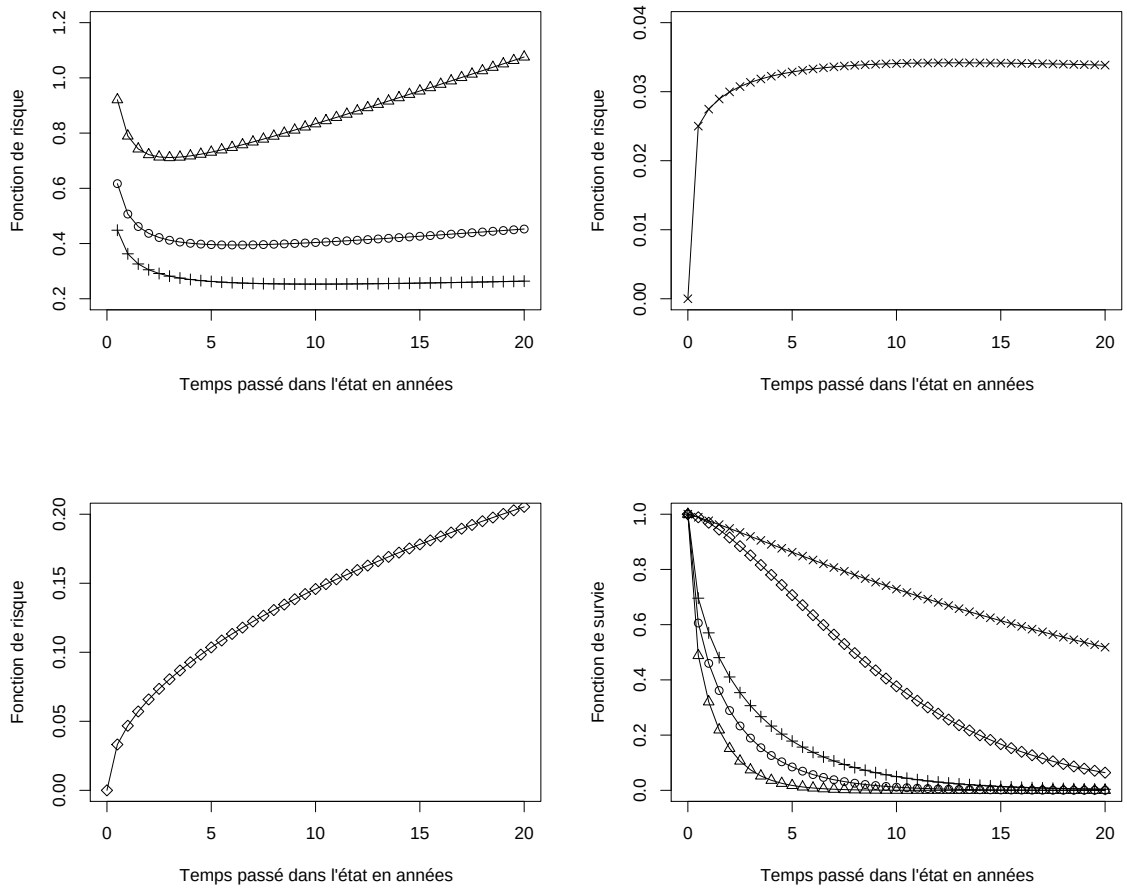

FIG. 4.5 – Fonctions de risque et de survie de base des temps d'attente, $\lambda_{0,ij}()$ et $S_{0,ij}()$ respectivement. $\triangle$ $\triangle$ $\triangle$ transition $1 \rightarrow 2$; o o o transition $1 \rightarrow 3$; $+$ $+$ $+$ transition $2 \rightarrow 3$; $\diamond$ $\diamond$ $\diamond$ transition $3 \rightarrow 4$; $\times$ $\times$ $\times$ transition $3 \rightarrow 5$.

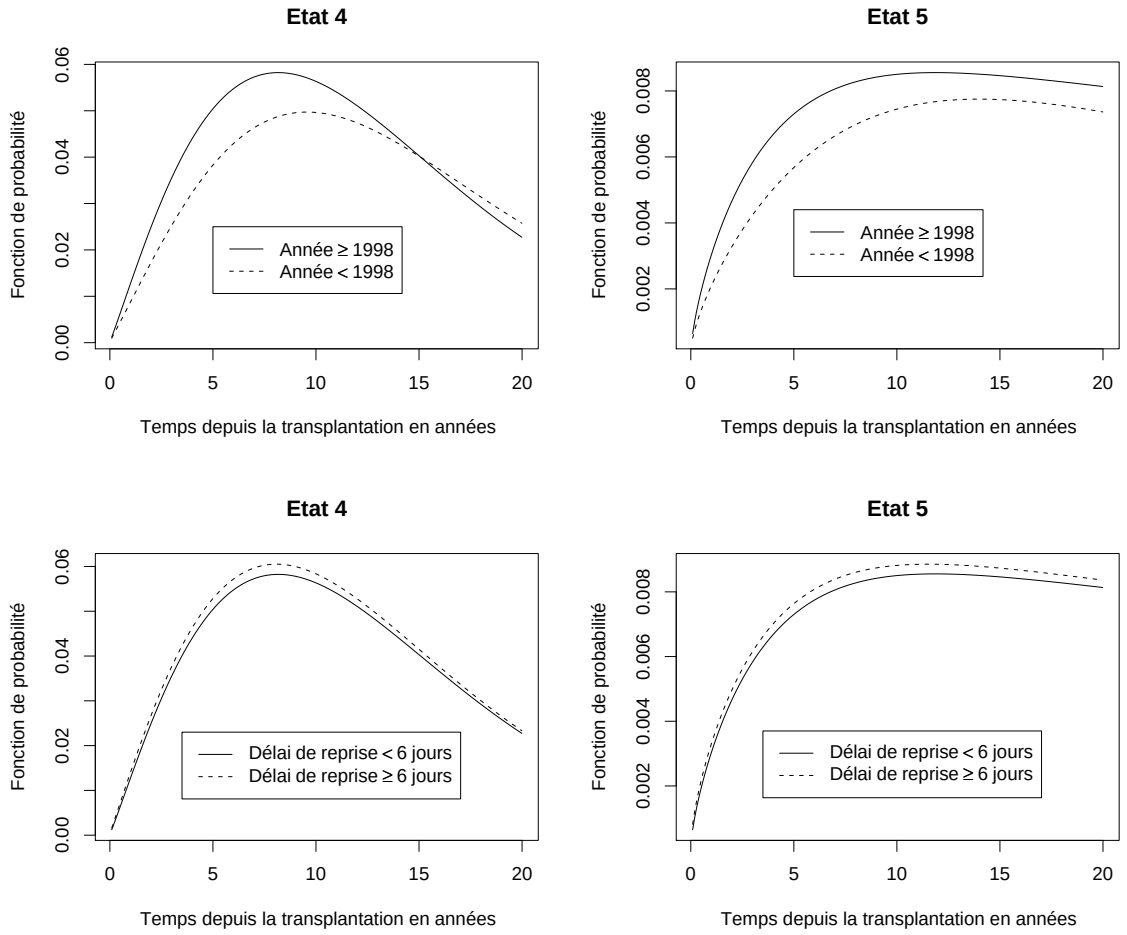

FIG. 4.6 – Probabilité de retour en dialyse (état 4) ou de décès (état 5) en fonction du temps depuis la transplantation (toutes les autres covariables sont fixées à 0).

Chapitre 5

Proportionnalité des risques

En analyse de données de survie, le modèle le plus utilisé est celui de Cox [1], ce dernier ne faisant aucune hypothèse sur la forme de la fonction de risque de base. Cependant, quelques travaux plus récents montrent que l'hypothèse de proportionnalité des risques ne tient pas dans certaines applications et peut entraîner des biais substantiels [65, 66].

Le même problème peut être rencontré dans l'analyse de processus multi-états plus complexes. Dans la majorité des travaux basés sur l'approche semi-markovienne, l'introduction des covariables reprend l'hypothèse de semi-proportionnalité introduite par Andersen [62]. On parle de semi-proportionnalité puisque la proportionnalité des risques est supposée pour chaque transition, mais ne tient pas pour des transitions différentes. De nombreux exemples respectant ce principe peuvent être cités [21, 20, 79, 63].

Dans notre précédente application sur l'évolution des patients greffés d'un rein, nous avons pu montrer que cette contrainte était trop forte, plusieurs covariables n'étant pas proportionnelles pour certaines transitions. Afin d'évaluer le respect de cette contrainte, nous avons choisi l'approche graphique consistant à représenter le logarithme de la fonction de risque cumulé selon le temps de survie. L'objectif principal de ce chapitre est de proposer un modèle semi-markovien assez flexible pour prendre en compte des risques non-proportionnels. Deux stratégies sont exposées : l'introduction d'effets dépendants du temps passé dans l'état, ainsi que des covariables pouvant modifier directement l'échelle de temps (modèle de vie accélérée). Nous reprendrons les bases du modèle défini dans le chapitre 4, permettant ainsi la prise en compte de forces de transition non-monotones (en forme de $\cup$ ou $\cap$), de temps de transition censurés par intervalle et d'une hétérogénéité à l'initialisation du processus. La construction de la vraisemblance reste aussi inchangée.

5.1 Incorporation des covariables

L'hypothèse de semi-proportionnalité a été définie dans le chapitre 2. Ainsi, parallèlement à l'équation (2.24), la fonction de survie, spécifique à la transition de l'état i vers l'état j, s'écrit :

$$S_{ij}(x, \eta_{h,ij}^{PH}) = S_{0,ij}(x)^{exp(\eta_{h,ij}^{PH})} \tag{5.1}$$

où $S_{0,ij}(x)$ est la fonction de survie de base au temps d'attente x et $\eta_{h,ij}^{PH}$ est le prédicteur linéaire, proportionnel à la fonction de risque, pour le h*ème* sujet. Formellement, $\eta_{h,ij}^{PH} = \beta_{ij}^T z_{h,ij}$ où $\beta_{ij} = (\beta_{ij}^1, \beta_{ij}^2, ..., \beta_{ij}^{n_{ij}})$ est le vecteur des coefficients de régression associé à $z_{h,ij}$, le vecteur des n_{ij} covariables $(z_{h,ij}^1, z_{h,ij}^2, ..., z_{h,ij}^{n_{ij}})$. Il est équivalent de définir la fonction de risque

$$\lambda_{ij}(x, \eta_{h,ij}^{PH}) = \lambda_{0,ij}(x) exp(\eta_{h,ij}^{PH}) \tag{5.2}$$

où $\lambda_{0,ij}(x)$ est la fonction de risque de base propre à la transition $i \rightarrow j$.

Pour permettre une plus grande flexibilité dans la modélisation de l'effet des covariables, définissons la fonction de survie suivante :

$$S_{ij}\big(x, \eta_{h,ij}^{AFT}(x), \eta_{h,ij}^{NPH}(x)\big) = S_{0,ij}\big(x exp(\eta_{h,ij}^{AFT}(x))\big)^{exp(\eta_{h,ij}^{NPH}(x))} \tag{5.3}$$

En comparaison avec la définition précédente (5.1), l'effet des covariables peut varier au cours du temps et peut directement influencer le temps de survie, par accélération ou ralentissement. Cette dernière propriété, reprenant le principe des modèles de vie accélérée (AFT pour *accelerated failure time*), peut être vue comme un changement d'échelle du temps. Ainsi, $\eta_{h,ij}^{AFT}(x)$ est le prédicteur linéaire de covariables directement associé au temps d'attente avant la transition de l'état i vers l'état j. Formellement,

$$\eta_{h,ij}^{AFT}(x) = \gamma_{ij}(x)^T y_{h,ij}$$

où $\gamma_{ij}(x) = (\gamma_{ij}^1(x), \gamma_{ij}^2(x), ..., \gamma_{ij}^{m_{ij}}(x))$ est le vecteur des coefficients de régression associé à $y_{h,ij}$, le vecteur des m_{ij} covariables agissant sur l'échelle de temps. D'une manière similaire, $\eta_{h,ij}^{NPH}(x)$ est le prédicteur linéaire agissant multiplicativement sur la fonction de risque. Contrairement à précédemment, l'effet de ce prédicteur n'est pas proportionnel puisqu'il dépend du temps d'attente dans l'état x. Plus précisément,

$$\eta_{h,ij}^{NPH}(x) = \beta_{ij}(x)^T z_{h,ij}$$

où $\beta_{ij}(x) = (\beta_{ij}^1(x), \beta_{ij}^2(x), ..., \beta_{ij}^{n_{ij}}(x))$ est le vecteur des coefficients de régression associé à $z_{h,ij}$, le vecteur des n_{ij} covariables agissant de manière multiplicative à la fonction de risque de base.

Pour obtenir un modèle interprétable et identifiable, nous supposons qu'une même covariable ne peut pas se trouver à la fois dans les deux composantes, $\eta_{h,ij}^{AFT}(x)$ et $\eta_{h,ij}^{NPH}(x)$,

pour une transition donnée. A partir de la définition de la survie (5.3), la fonction de survie pour une distribution de Weibull généralisé est égale à :

$$S_{ij}(x) = exp\big(1 - \big(1 + \big(xexp(\eta_{h,ij}^{AFT}(x))/\sigma_{ij}\big)^{\nu_{ij}}\big)^{\theta_{ij}^{-1}}\big)^{exp(\eta_{h,ij}^{NPH}(x))} \tag{5.4}$$

La fonction de risque correspondante est égale à :

$$\begin{aligned}\lambda_{ij}\big(x, \eta_{h,ij}^{AFT}(x), \eta_{h,ij}^{NPH}(x)\big) &= exp\big(\eta_{h,ij}^{NPH}(x)\big)\big\{\lambda_{0,ij}\big(xexp(\eta_{h,ij}^{AFT}(x))\big)exp(\eta_{h,ij}^{AFT}(x)) \\ &\times \ (1 + x\partial\eta_{h,ij}^{AFT}(x)/\partial x) - (\partial\eta_{h,ij}^{NPH}(x)/\partial x) \\ &\times \ \big(1 - \big(1 + \big(xexp(\eta_{h,ij}^{AFT}(x))/\sigma_{ij}\big)^{\nu_{ij}}\big)^{\theta_{ij}^{-1}}\big)\big\}\end{aligned} \tag{5.5}$$

où $\lambda_{0,ij}()$ est la fonction de risque de type Weibull généralisé (1.11). En supprimant le terme multiplicatif de la fonction de risque, $\eta_{h,ij}^{NPH}(x) = 0$, et en supposant constant l'effet des covariables associées à l'échelle de temps, $\eta_{h,ij}^{AFT}(x) = \eta_{h,ij}^{AFT}$, on obtient à partir de l'équation (5.5) :

$$\lambda_{ij}\big(x, \eta_{h,ij}^{AFT}, 0\big) = \lambda_{0,ij}\big(xexp(\eta_{h,ij}^{AFT})\big)exp(\eta_{h,ij}^{AFT}) \tag{5.6}$$

Cette fonction de risque correspond alors à la définition traditionnelle des modèles de vie accélérée (1.18). D'une manière similaire, en supprimant le prédicteur influençant l'échelle de temps, $\eta_{h,ij}^{AFT}(x) = 0$, et en considérant les autres covariables proportionnelles, $\eta_{h,ij}^{NPH}(x) = \eta_{h,ij}^{NPH}$, on obtient :

$$\lambda_{ij}\Big(x, 0, \eta_{h,ij}^{NPH}\Big) = \lambda_{0,ij}(x)exp(\eta_{h,ij}^{NPH}) \tag{5.7}$$

On retrouve le modèle initial respectant l'hypothèse de semi-proportionnalité (5.2).

Pour prendre en compte des effets variants au cours du temps, nous adoptons la méthode définie par Stablein et al. [68], en introduisant des interactions avec le temps. Par exemple, pour une simple covariable z_{ij} et pour un individu h donné :

$$\eta_{h,ij}^{NPH}(x) = \beta_{ij}^{(1)} z_{h,ij} + \beta_{ij}^{(2)} z_{h,ij} x + \beta_{ij}^{(3)} z_{h,ij} x^2$$

Même si le terme d'interaction linéaire serait suffisant pour permettre une relation variant au cours du temps, le terme quadratique permet une relation non-monotone. Ce type de fonction polynomiale possède l'avantage d'être facilement dérivable pour calculer les composantes $\partial\eta_{h,ij}^{AFT}(x)/\partial x$ et $\partial\eta_{h,ij}^{PH}(x)/\partial x$ de l'équation (5.5). Le nombre de paramètres à estimer reste de plus raisonnable.

5.2 Stratégie de modélisation

La première étape consiste à tester le respect de l'hypothèse de semi-proportionnalité, en traçant le logarithme de la fonction de risque cumulé. Cette étape a montré dans le chapitre précédent que pour de nombreux couples transition-covariable, cette hypothèse ne pouvait pas être retenue.

Dans un second temps, pour chacun de ces couples, 6 modèles univariés sont calculés :

- $\eta_{h,ij}^{PH}(x) = \beta_{ij}^{(1)} z_{h,ij}$ et $\eta_{h,ij}^{AFT}(x) = 0$
- $\eta_{h,ij}^{NPH}(x) = \beta_{ij}^{(1)} z_{h,ij} + \beta_{ij}^{(2)} z_{h,ij} x$ et $\eta_{h,ij}^{AFT}(x) = 0$
- $\eta_{h,ij}^{NPH}(x) = \beta_{ij}^{(1)} z_{h,ij} + \beta_{ij}^{(2)} z_{h,ij} x + \beta_{ij}^{(3)} z_{h,ij} x^2$ et $\eta_{h,ij}^{AFT}(x) = 0$
- $\eta_{h,ij}^{AFT}(x) = \gamma_{ij}^{(1)} y_{h,ij}$ et $\eta_{h,ij}^{PH}(x) = 0$
- $\eta_{h,ij}^{AFT}(x) = \gamma_{ij}^{(1)} y_{h,ij} + \gamma_{ij}^{(2)} y_{h,ij} x$ et $\eta_{h,ij}^{PH}(x) = 0$
- $\eta_{h,ij}^{AFT}(x) = \gamma_{ij}^{(1)} y_{h,ij} + \gamma_{ij}^{(2)} y_{h,ij} x + \gamma_{ij}^{(3)} y_{h,ij} x^2$ et $\eta_{h,ij}^{PH}(x) = 0$

La modélisation retenue est celle qui minimise le critère AIC. La covariable ainsi définie pour une transition donnée est inclue dans le modèle multivarié si le rapport des vraisemblances entre le modèle univarié et le modèle sans covariable conclut à une p-value inférieure ou égale à 0,15.

Dans un troisième et dernier temps, les coefficients de régression et les paramètres des distributions des temps d'attente seront supprimés selon une procédure descendante avec un seuil de signification défini à 0,05.

5.3 Application à la transplantation rénale

Nous considérons la même structure multi-états (figure 4.1) et la même base de données (tableau 4.1). Le codage des covariables reste aussi inchangé. Comme le montre le tableau (5.1), plus de covariables sont retenues que dans le modèle supposant la semi-proportionnalité. La logvraisemblance de ce modèle multi-états est égale à -4846,70. En utilisant le test du rapport de vraisemblance, il semble que cette nouvelle flexibilité soit plus adéquate que l'approche semi-proportionnelle pour l'étude de l'évolution des patients greffés d'un rein ($p<0{,}0001$).

Il semble qu'être greffé avant 1998 soit un facteur de ralentissement de la transition $1 \rightarrow 2$. Cette covariable agit directement sur l'échelle de temps (AFT). Cependant, en considérant l'interaction significative d'ordre 1, cet effet diminue avec le temps passé dans l'état. Un effet similaire de l'année de la greffe est retenu pour le temps de transition de l'état 2 vers l'état 3, néanmoins l'effet est ici supposé multiplicatif de la fonction de risque (NPH). Enfin, toujours pour ce facteur période, l'hypothèse de semi-proportionnalité est retenue pour les transitions $1 \rightarrow 3$ et $3 \rightarrow 4$. Cette dernière association n'est pas significative au seuil 0,05, cependant le facteur est conservé dans le modèle pour ajuster l'effet de l'ischémie froide sur la force de transition $3 \rightarrow 4$. Globalement, être transplanté avant 1998 constitue un facteur protecteur d'aggravation du greffon.

Dans le chapitre précédent, lorsque l'hypothèse de semi-proportionnalité est supposée, l'effet de l'âge du receveur n'est pas retenu concernant la transition $2 \rightarrow 3$. Cependant,

la figure (4.2) montre des courbes divergentes. Pour obtenir le modèle le plus parcimonieux, cette covariable est considérée ici multiplicative de la fonction de risque avec une interaction avec le temps d'ordre 2. Ainsi, une association significative est déduite. Etre âgé de plus de 55 ans au moment de la greffe semble être un facteur de risque de passer de l'état 2 à l'état 3, cet effet diminuant linéairement au cours du temps et augmentant quadratiquement.

Toujours en opposition avec les résultats du chapitre 4, un effet variant au cours du temps de l'ischémie froide est retenu comme influençant la vitesse de la transition $3 \rightarrow 4$. Ce facteur est aussi considéré de manière multiplicative. Une ischémie froide supérieure à 16 heures augmente le risque de subir ce passage de l'état le plus grave vers un retour en dialyse, cet effet diminuant linéairement avec le temps passé dans l'état et augmentant quadratiquement.

L'âge du receveur est retenu comme un facteur influençant de manière constante la vitesse de la transition $3 \rightarrow 5$. Cependant, l'effet est ici associé à l'échelle de temps. Les patients âgés de plus de 55 ans à la date de greffe survivent environ 2,3 fois moins longtemps dans l'état le plus grave, sachant que l'état suivant est le décès.

Transition	Mode	Covariable	Estim.	ET	p-value
$0 \rightarrow 1$	.	Intercept	2,83	0,19	0,0001
$0 \rightarrow 1$	.	Sexe du receveur	-0,38	0,17	0,0270
$0 \rightarrow 1$	.	Délai de reprise	-0,52	0,17	0,0016
$0 \rightarrow 2$	.	Intercept	-0,66	0,43	0,1275
$0 \rightarrow 2$	.	Ischémie froide	1,12	0,43	0,0096
$1 \rightarrow 2$	AFT	Année de la greffe	-1,84	0,27	0,0001
$1 \rightarrow 2$	AFT	Année de la greffe $\times x$	0,14	0,03	0,0001
$1 \rightarrow 3$	PH	Année de la greffe	-1,31	0,17	0,0001
$2 \rightarrow 3$	NPH	Année de la greffe	-0,75	0,14	0,0001
$2 \rightarrow 3$	NPH	Année de la greffe $\times x$	0,06	0,02	0,0009
$2 \rightarrow 3$	NPH	Age du receveur	0,34	0,18	0,0551
$2 \rightarrow 3$	NPH	Age du receveur $\times x$	-0,14	0,06	0,0199
$2 \rightarrow 3$	NPH	Age du receveur $\times x^2$	0,01	0,01	0,0219
$3 \rightarrow 4$	NPH	Ischémie froide	1,57	0,60	0,0093
$3 \rightarrow 4$	NPH	Ischémie froide $\times x$	-0,42	0,13	0,0014
$3 \rightarrow 4$	NPH	Ischémie froide $\times x^2$	0,02	0,01	0,0037
$3 \rightarrow 4$	PH	Année de la greffe	0,41	0,23	0,0792
$3 \rightarrow 5$	AFT	Age du receveur	0,86	0,29	0,0033

TAB. 5.1 – Coefficients de régression du modèle multivarié final sans hypothèse de semi-proportionnalité

Excepté pour la transition $1 \rightarrow 3$, la distribution de type Weibull apparaît suffisante

par rapport à sa généralisation (voir tableau 5.2). Seul le paramètre θ_{13} semble être significativement différent de la valeur 1 ($LRS = 10,59$, $p = 0,0011$). La figure (5.1) montre que la fonction de risque de cette transition de l'état le moins grave au plus grave suit une forme en $\cup$. Son minimum correspond à temps d'attente de 4,5 années dans l'état 1 pour les patients transplantés à partir de 1998 (équation 1.12).

L'intérêt de l'utilisation de la loi Weibull généralisé apparaît donc plus limité lorsque des interactions avec le temps sont inclues dans les prédicteurs linéaires. En effet, comme tend à le montrer la figure (5.2) concernant la transition $3 \rightarrow 4$, des formes de risque complexes peuvent être obtenues à partir d'une simple distribution de type Weibull.

	σ_{ij}		ν_{ij}		θ_{ij}	
Transition	Estim.	Et	Estim.	ET	Estim.	ET
$1 \rightarrow 2$	1,43	0,20	0,56	0,04	1,00	.
$1 \rightarrow 3$	5,86	7,92	0,55	0,06	0,36	0,20
$2 \rightarrow 3$	2,54	0,42	0,59	0,05	1,00	.
$3 \rightarrow 4$	9,65	1,62	2,21	0,30	1,00	.
$3 \rightarrow 5$	10,09	2,96	1,31	0,17	1,00	.

TAB. 5.2 – Paramètres des lois des temps d'attente du modèle multivarié final sans hypothèse de semi-proportionnalité ($P_{12} = 0,60$ $ET = 0,02$, $P_{34} = 0,85$ $ET = 0,03$)

L'interprétation des coefficients se révèle plus complexe que dans le chapitre 4. Calculer l'impact marginal d'une covariable sur la probabilité de subir un événement terminal en fonction du temps depuis la transplantation constitue donc un bon moyen d'interpréter la dynamique globale du processus. A partir de l'équation (4.9), ces fonctions sont représentées dans la figure (5.3). Il semble ainsi qu'être greffé avant 1998, avoir un délai de reprise de l'activité rénale supérieur ou égal à 6 jours, ou bien une ischémie supérieure à 16 heures constituent des facteurs augmentant la probabilité de retour en dialyse. Concernant la probabilité de décès, être greffé à partir de 1998 et être âgé de plus de 55 ans sont autant de facteurs de risques. Cependant, l'effet de l'âge a tendance à s'inverser avec l'augmentation du temps depuis la greffe.

5.4 Discussion

Pour étudier l'évolution des patients transplantés d'un rein, nous avons pu montrer, par une représentation stratifiée du logarithme de la fonction de risque cumulé, que l'utilisation d'un modèle semi-markovien basé sur l'hypothèse de semi-proportionnalité est trop restrictive. Le modèle plus flexible proposé semble être mieux adapté aux données. Nous avons ainsi pu nous rendre compte que sous l'hypothèse de semi-proportionnalité, certaines covariables semblant influencer les forces de transition ne sont pas retenues et

l'interprétation de l'effet des covariables peut être biaisée.

Les fonctions de risque associées aux temps d'attente dans les états sont aussi sensiblement différentes. Il semble que la prise en compte de covariables, dont l'effet peut varier au cours du temps et peut directement être associé à l'échelle de temps, permet une flexibilité accrue dans la modélisation des forces de transition.

L'approche multinomiale d'initialisation du processus reste inchangée. L'effet de ces covariables reste donc logiquement identique.

Quelques limites à ce modèle sont à noter. Bien que plus flexible que celui supposant la proportionnalité, nous avons restreint la dépendance temporelle de l'effet des covariables à un polynôme d'ordre 2. D'autres approches moins contraignantes, comme l'utilisation de fonctions splines, auraient pu être envisagées. Cependant, notre choix d'une régression paramétrique et polynomiale permet une interprétation plus simple de l'effet des covariables et aboutit à une logvraisemblance moins complexe.

Notons aussi les problèmes liés à la définition de la fonction de risque 5.5. En effet, cette dernière n'est pas strictement définie positive. Ceci rend plus difficile l'estimation des paramètres (en particulier les coefficients de régression) devant respecter cette contrainte. La solution a été de choisir un codage de covariables permettant d'obtenir une classe de référence protectrice. De cette manière, la fonction de risque est strictement positive dans la population de référence (toutes les covariables sont nulles) et elle est encore supérieure pour les autres sous-groupes.

Une autre difficulté des modèles utilisés jusqu'ici est de considérer des états d'aggravation d'une pathologie par des marqueurs continus. Ces variables, contrairement à des événements cliniques, sont soumises à certaines fluctuations à court terme : erreur de la mesure, variation des conditions de la mesure, etc. Par exemple, la protéinurie dépend de l'alimentation, de la pratique d'une activité physique ou de l'horaire du prélèvement. Cette fluctuation peut être à l'origine d'erreurs dans le diagnostic de l'état. Cette problématique a déjà fait l'objet de travaux dans le cadre markovien [88, 89]. Le choix des seuils de ces marqueurs pour la définition des états de gravité a, jusqu'ici, été réalisé sur avis d'experts. Il serait intéressant de pouvoir définir ces seuils par des méthodes statistiques.

La dernière limite importante de notre approche est la définition donnée à la censure par intervalle. Toutes les transitions observées, comme l'indique le schéma 4.1, se produisent entre deux états de gravité contigus. Seul le passage de l'état 1 à l'état 3 échappe à cette règle. L'hypothèse implicite dans la construction de la vraisemblance est qu'aucune autre transition n'a eu lieu entre les observations de ces deux états. La double transition $1 \rightarrow 2$ et $2 \rightarrow 3$ est en effet possible entre deux observations. En effet, la transition $1 \rightarrow 3$ est directement possible en cas de chute de plus de 30 % de la clairance de la créatinine en moins d'un an, mais il est possible que cette transition soit le fait d'une évolution brutale de la protéinurie d'une valeur inférieure à 0,5 g/jour à une valeur supérieure à 1 g/jour.

Dans ce dernier cas, le processus est passé par l'état intermédiaire 2. Nous supposons donc dans ce chapitre que le processus d'observation est assez régulier et rapproché pour ne pas omettre de tels événements. De plus, comme nous l'avons abordé dans la discussion du chapitre 4, les transitions sont considérées indépendantes les unes des autres, alors que le temps de transition, même censuré par intervalle, apporte une information aux autres transitions du même individu. Le chapitre suivant tente de répondre à ces différentes difficultés, en considérant dans le calcul de la vraisemblance les trajectoires individuelles, et non plus les transitions.

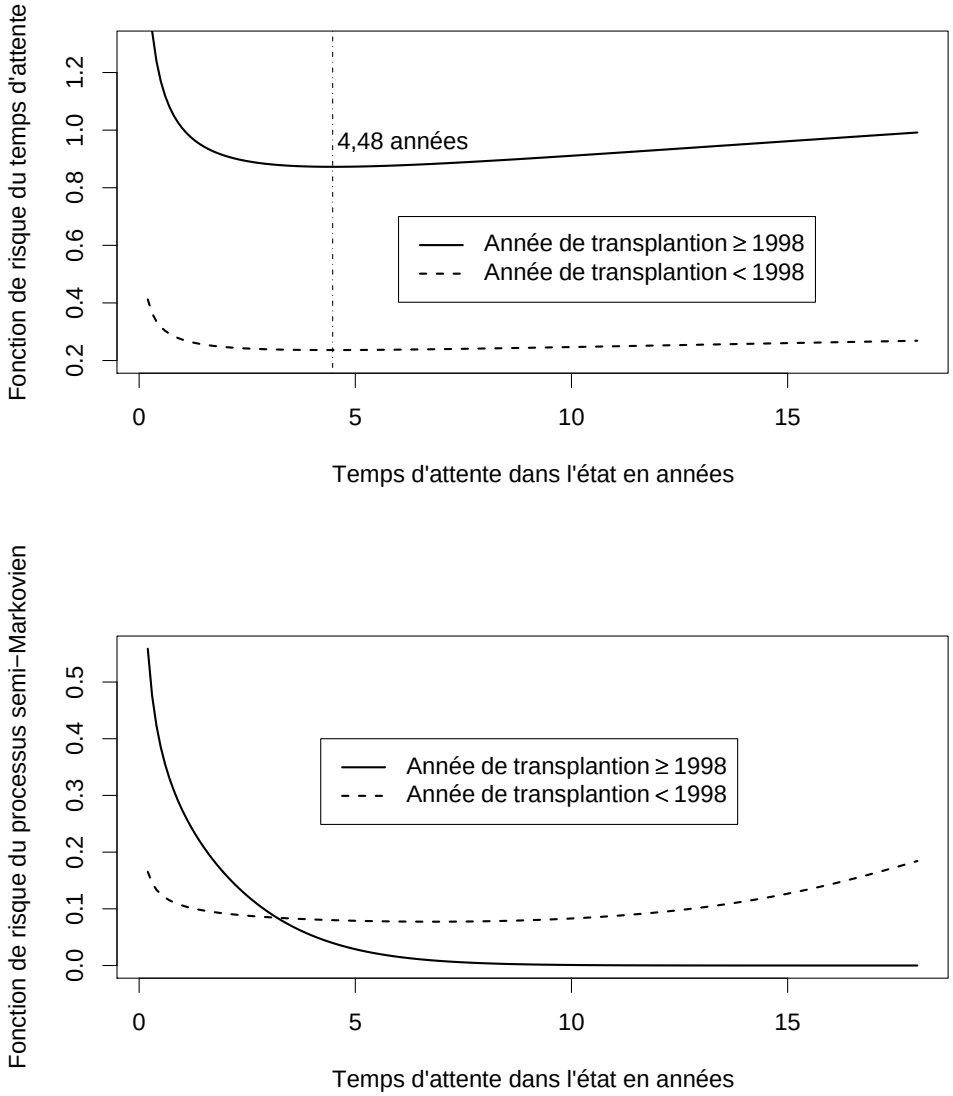

FIG. 5.1 – Fonction de risque du temps d'attente dans l'état 1 avant de transiter vers l'état 3, $\lambda_{13}()$, et fonction de risque du processus semi-markovien associée, $\alpha_{13}()$

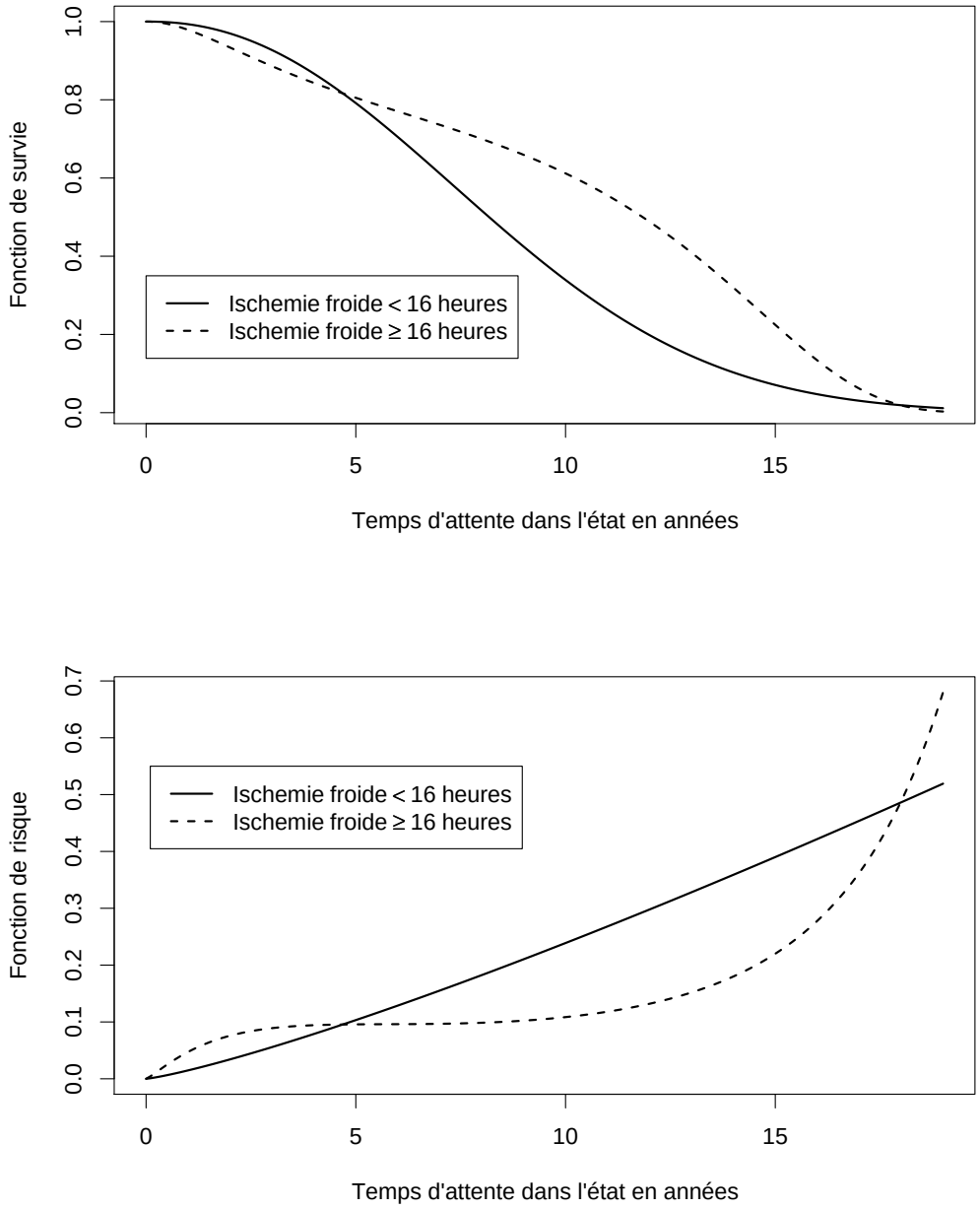

FIG. 5.2 – Fonction de survie et de risque du temps d'attente dans l'état 3 avant de transiter vers l'état 4, $S_{34}()$ et $\lambda_{34}()$, en fonction de l'ischémie froide

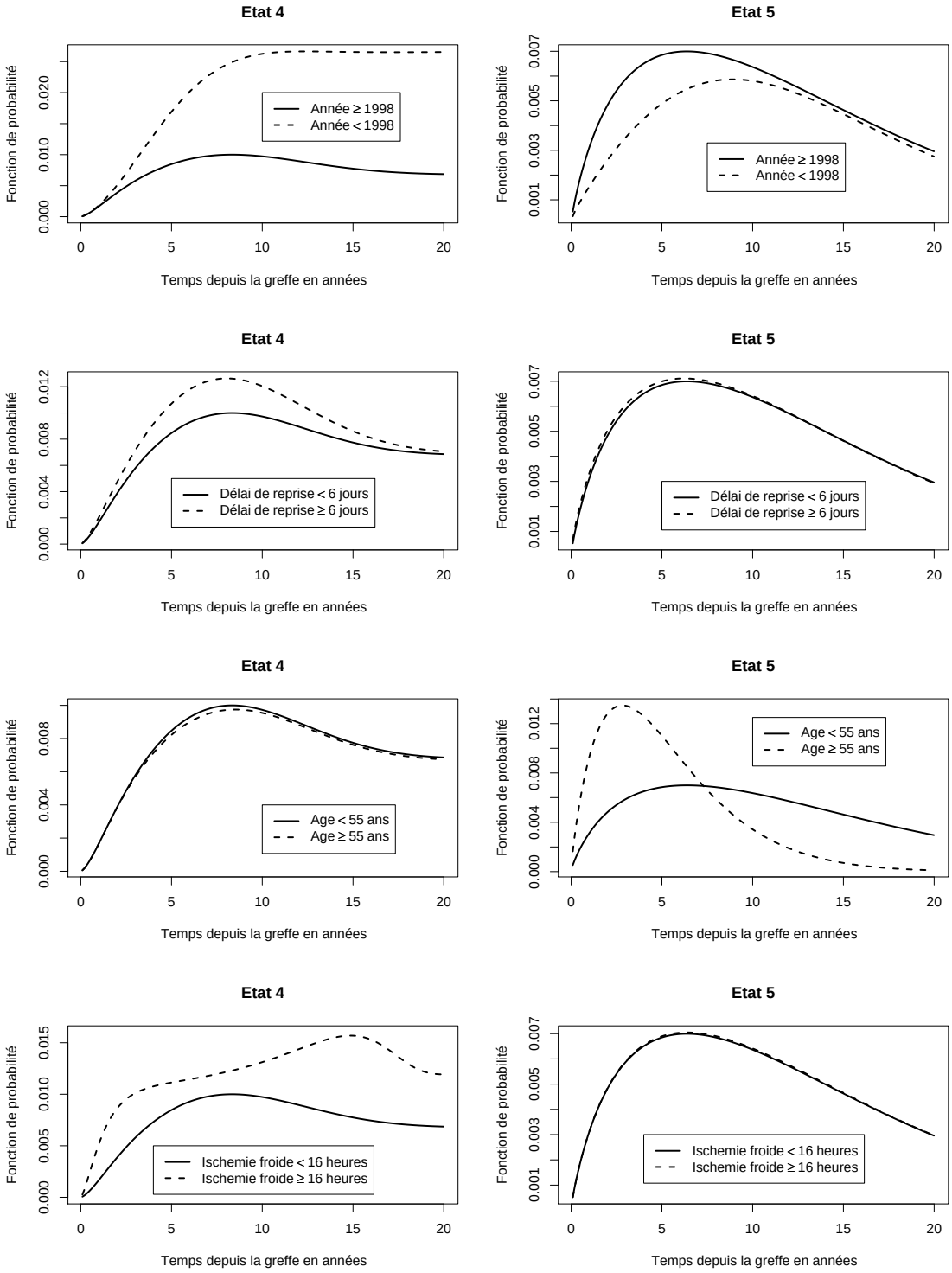

FIG. 5.3 – Fonction de probabilité de retour en dialyse (état 4) et de décès (état 5) en fonction du temps écoulé depuis la transplantation (pour des hommes receveurs)

Chapitre 6

Censure par intervalle des séquences d'états

Plusieurs limites ont été soulevées précédemment. La principale d'entre elle est la prise en compte de la censure par intervalle. Dans nos développements précédents, la censure n'est considérée que pour le temps de transition [90]. Autrement dit, la seule information disponible est que la transition $i \rightarrow j$ a lieu entre deux temps. Or, il apparaît pertinent de considérer aussi la censure d'une transition. Dans ce cas, le patient est passé de l'état i à l'état j entre deux temps, avec la possibilité d'être passé par un autre état k dans cet intervalle. Ce double type de censure peut être pris en compte assez directement grâce à notre choix paramétrique. La loi du délai nécessaire à une double transition correspond à la somme des deux variables aléatoires des temps d'attente dans les états consécutifs. Cette loi peut alors être calculée à partir de la convolution de ces deux variables [91].

Bien entendu, il est toujours préférable que le processus d'observation soit le plus rapproché possible pour minimiser ce type de censure. Dans la base de données DIVAT, la mesure de la clairance de la créatinine est plus régulière que celle de la protéinurie. Cette constatation justifie notre choix de considérer uniquement la clairance comme marqueur de gravité, sans prendre en compte la protéinurie. La question posée est alors la règle de classification des états selon ce marqueur. Peu de travaux discutent cette difficulté. Dans la grande majorité des études de survie, la valeur de la clairance à un an est choisie comme marqueur d'une éventuelle dysfonction chronique du greffon [92]. Le premier inconvénient de cette approche est d'exclure les rejets précoces. Le second est de ne pas considérer l'information apportée par la dynamique de cette variable.

Pour rendre compte du lien entre la dynamique de la clairance et l'échec de la greffe, nous avons représenté les valeurs mesurées en fonction du temps depuis la greffe pour des patients retournant en dialyse (figure 6.1). Pour certains individus, on peut observer une première phase d'augmentation de la clairance. En revanche, tous ces patients terminent leur suivi par une phase de diminution, cette dernière étant souvent assez longue.

A partir de ces observations, l'idée est de distinguer trois stades de gravité d'un patient transplanté rénal. Cette classification est caractérisée par le schéma (6.2). L'état 1, représentant un stade de faible risque de retour en dialyse, est caractérisé par une clairance en augmentation depuis la greffe. L'état de gravité intermédiaire est identifié par une diminution de ce marqueur. Autrement dit, un patient peut débuter son évolution dans l'état 2 dès l'origine, si aucune phase d'augmentation n'a été observée après la greffe. Enfin, l'état 3, où le risque d'échec est le plus important, est toujours caractérisé par une chute de la clairance, mais la valeur de cette dernière est alors inférieure à un certain pourcentage de la valeur à l'entrée dans l'état 2. Le premier problème est donc de définir la valeur de ce pourcentage.

L'autre difficulté, au regard de la figure (6.1), est la fluctuation de ce marqueur. A court terme, on peut observer de multiples diminutions et augmentations de la clairance. Ces variations sont cependant plus faibles d'un jour à l'autre que la tendance à long terme : soit en rapport avec l'élévation du taux de créatinémie (traduisant une dégradation vraie de la fonction rénale), soit du fait du vieillissement de l'individu, soit du fait de certains facteurs de confusion (variation pondérale, etc.). Une élévation isolée de la créatinine ne correspond pas obligatoirement à une dégradation irréversible de la fonction rénale. Par exemple, un épisode de déshydratation du patient ou une toxicité médicamenteuse, peuvent faire monter la créatinine sanguine (donc diminuer la clairance de la créatinine) de façon provisoire. Pour apprécier la réelle dégradation de la fonction rénale, il est donc souhaitable de lisser son évolution, en particulier pour les mesures rapprochées.

Certains auteurs se sont déjà intéressés à cette problématique, qui peut être vue comme une erreur de classement de l'état de gravité due à la fluctuation à court terme [88, 89]. Dans leurs applications, les auteurs distinguent deux processus : le véritable état du patient et l'état observé à travers le marqueur continu. Le véritable état est défini par d'autres critères, comme l'avis d'experts. Il est dit "caché" au sens où il n'est pas forcément observable en pratique. L'objectif est alors de modéliser la dynamique de ce processus caché, tout en mesurant l'erreur commise si l'on se base sur le processus observé. Notre problématique est différente, puisque le véritable état sous-jacent n'est pas connu. Le seul critère de classification de l'état du rein disponible est la clairance de la créatinine. Nous proposons ici de retrouver une valeur sous-jacente de la clairance par lissage. Cette approche possède le double avantage d'éliminer la fluctuation et d'identifier plus clairement les différentes phases précédentes.

En considérant les trois difficultés soulevées dans cette introduction, ce chapitre se divise en autant de sections. La première concerne le lissage de l'évolution de la clairance de la créatinine. La seconde partie présente la méthode d'estimation du pourcentage de diminution de la clairance associée à l'entrée dans l'état 3. La dernière section définit le modèle multi-états utilisé. Nous reprendrons les différents points théoriques abordés dans les chapitres précédents, mais en modélisant la non-proportionnalité des covariables d'une manière légèrement différente, moins contraignante concernant l'estimation des coefficients

de régression. De plus, les covariables seront introduites à travers la chaîne de Markov sous-jacente. Le modèle semi-markovien ainsi obtenu permet à la fois d'identifier les facteurs associés aux vitesses de transitions et aux trajectoires du processus.

6.1 Lissage du marqueur par B-splines

Le principe des régressions splines est de modéliser une variable dépendante continue, Y, en fonction d'une variable explicative, X. Supposons que le suivi d'un patient h ($h = 1, 2, ..., n$) soit composé de $n_h + 1$ visites aux temps depuis la greffe $\{v_{h,0}, v_{h,1}, ..., v_{h,n_h}\}$. A chaque visite, la clairance de la créatinine, notée $y_{h,r}$ ($r = 0, 1, ..., n_h$), est mesurée. Les données sont alors composées des couples (y_r, v_r), $r = 1, ..., n_h$. Un modèle de régression peut être écrit d'une manière générale :

$$y_r = f(v_r) + \epsilon_r \tag{6.1}$$

où ϵ_r est une erreur aléatoire, supposée le plus souvent indépendante et identiquement distribuée selon une loi normale de moyenne nulle et de variance σ^2. La problématique est ici de définir non-paramétriquement la fonction $f()$. L'utilisation de polynômes par morceaux constitue une méthode reconnue [93]. Ces morceaux ou intervalles sont définis par une séquence de noeuds. Plus le nombre de noeuds est important, plus la régression est flexible. De plus, il est commun de forcer la continuité des polynômes au niveau des noeuds. Même si beaucoup de configurations sont possibles, la plus utilisée consiste en des polynômes cubiques par morceaux avec des premières et secondes dérivées continues aux noeuds.

Deux difficultés sont propres à ce type de méthodes : le choix du vecteur de noeuds et le choix de la base de fonctions. Considérons K noeuds intérieurs, notés $\xi_1 < ... < \xi_K$, et deux noeuds extrêmes ξ_0 et ξ_{K+1}. Intuitivement, le choix de fonctions de puissance tronquées est assez simple. Par exemple, pour une spline cubique :

$$f(v) = \beta_0 + \beta_1 v + \beta_2 v^2 + \sum_{j=1}^{K} \theta_j (v - \xi_j)_+^3 \tag{6.2}$$

où a_+ correspond à la partie positive de a. Les notations sous forme de bases permettent de simplifier cette écriture pour une portée plus générale. Pour un polynôme de degré $m - 1$, posons

$$B_j(v) = v^{j-1},\ j = 1, ..., m \tag{6.3}$$

$$B_{m+j}(v) = (v - \xi_j)_+^{m-1},\ j = 1, ..., K \tag{6.4}$$

En définissant le vecteur des coefficients de régression $\gamma_j = (\beta_0, ..., \beta_m, \theta_1, ..., \theta_K)$, on obtient alors :

$$f(v) = \sum_{j=1}^{m+K} \gamma_j B_j(v) \tag{6.5}$$

Les splines cubiques (m = 4) sont habituellement favorisées, car elles permettent d'obtenir des courbes assez lisses tout en conservant un nombre raisonnable de paramètres. Toutefois, lorsque la méthode d'estimation est plus adaptative et lorsque le choix des noeuds est dirigé par les données, on doit généralement se restreindre à des polynômes de degré 1 ou 2 de façon à limiter le temps de calcul.

Le lissage par B-splines, initialement introduit par de Boor [94], constitue un type de régression splines populaire grâce à ses bonnes propriétés numériques. Pour calculer ces B-splines, supposons toujours K noeuds intérieurs, notés $\xi_1 < ... < \xi_K$, et deux noeuds extrêmes ξ_0 et ξ_{K+1}. De plus, pour une spline de degré $m-1$, considérons $2m$ noeuds extrêmes supplémentaires : $\xi_{-(m-1)} = ... = \xi_{-1} = \xi_0$ (noeud limite inférieur) et $\xi_{K+1} = ... = \xi_{K+m-1} = \xi_{K+m}$ (noeud limite supérieur). Dans ce cas, nous avons $K+m-1$ fonctions de base $B_{-(m-1),m}(v)$, ..., $B_{K,m}(v)$ de degré $m-1$, définies récursivement par :

$$B_{j,m}(v) = \frac{v - \xi_j}{\xi_{j+m-1} - \xi_j} B_{j,m-1}(v) + \frac{\xi_{j+m} - v}{\xi_{j+m} - \xi_{j+1}} B_{j+1,m-1}(v) \tag{6.6}$$

pour $j = -(m-1), ..., k$, en utilisant

$$B_{j,1}(v) = \begin{cases} 1, \ v \in [\xi_j, \xi_{j+1}[, \\ 0, \text{sinon} \end{cases}$$

La fonction de régression B-splines de degré $m-1$ avec K noeuds intérieurs est alors une combinaison linéaire de ces fonctions de base

$$f(v) = \sum_{j=-(m-1)}^{K} \alpha_j B_{j,m}(v) \tag{6.7}$$

Notons qu'une fois les bases de splines calculées, l'équation (6.7) définit alors un modèle linéaire avec comme paramètres de régression α_j $(j = -(m-1), ..., K)$. Ainsi, l'utilisation de fonctions B-splines implique l'estimation de $m + K$ paramètres. La méthode des moindres carrés est adaptée à ce type d'estimation. En reprenant les notations relatives à l'expression (6.1), les paramètres α_j sont ceux qui minimisent la quantité :

$$\sum_{r=1}^{n_h} \left\{ y_r - \sum_{j=-(m-1)}^{K} \alpha_j B_{j,m}(v_r) \right\}^2$$

Une fois les paramètres estimés, les valeurs prédites de y_r pour chacune des observations de v_r, notées $\tilde{y}_r$, sont supposées représenter la véritable valeur du processus sous-jacent épurée de la fluctuation ϵ_r $(\tilde{y}_r = y_r - \epsilon_r)$.

Bien entendu, plus le modèle sera considéré flexible, plus les erreurs ϵ_r seront faibles. La difficulté consiste donc à choisir le modèle qui correspond le mieux aux données observées, mais avec un lissage suffisant afin de supprimer les fluctuations. Nous aborderons dans la section suivante comment ce consensus est réalisé.

Définissons a priori une base de fonctions cubiques et 5 noeuds intérieurs. L'emplacement des noeuds est choisi aux quantiles correspondants. Ce lissage est représenté par la figure (6.1) pour les patients retournant en dialyse. Cette représentation nous montre ce principe d'élimination du bruit et d'estimation des valeurs sous-jacentes de la clairance soumises à fluctuation.

6.2 Définition des états de gravité

6.2.1 Choix du modèle

A partir des valeurs sous-jacentes du marqueur, la question est alors de définir les états de gravité les plus annonciateurs d'un retour en dialyse ou d'un décès. La méthode choisie est l'utilisation d'un modèle de survie, l'événement étudié étant le retour en dialyse ou le décès du patient, l'état de gravité étant pris en compte comme facteur explicatif. Le contexte est ici un peu particulier. L'estimation de la fonction de risque de base ne nous intéresse pas, le seul objectif étant de définir le codage des états de gravité le plus prédictif de l'incidence d'un événement terminal. Comme nous l'avons défini en introduction de ce chapitre, trois phases sont distinguées : une phase d'augmentation de la clairance, puis une période de diminution pendant laquelle la clairance reste supérieure à un certain pourcentage de la valeur maximum entre les deux premières phases, et la dernière phase caractérisée par le passage sous ce seuil. Cette classification est caractérisée par le schéma (6.2). La covariable n'est donc pas fixe au cours du temps.

Pour répondre à ces trois contraintes, nous utilisons un modèle de Cox étendu pour la prise en compte de covariables temps-dépendantes. Parallèlement à l'expression (1.6), la fonction de risque est définie par

$$\lambda(v_{h,r}, z_h(v_{h,r}))) = \lambda_0(v_{h,r}) exp(\beta z_h(v_{h,r})) \tag{6.8}$$

où le vecteur $z_h(v_{h,r}) = (z_{h,1}(v_{h,r}), z_{h,2}(v_{h,r}))$ représente l'état de gravité occupé par l'individu h au temps $v_{h,r}$ depuis la transplantation. $z_{h,1}(v_{h,r})$ est égal à 1 pour le second état de gravité et 0 sinon. De même, $z_{h,2}(v_{h,r})$ est égal à 1 pour le troisième état de gravité et 0 sinon. $\beta = (\beta_1, \beta_2)$ est le vecteur des coefficients de régression associé à $z_h(v_{h,r})$.

Le temps étudié est le délai entre la transplantation et le retour en dialyse ou le décès du patient. Si l'individu h présente un de ces deux événements à la fin de son suivi au temps v_{h,n_h}, alors δ_h vaut 1. En revanche si δ_h est égale à 0, alors v_{h,n_h} représente son temps de participation. La vraisemblance partielle (1.7) est adaptée pour l'estimation de β dans ce contexte de variables temps-dépendantes [95].

$$\mathcal{VP} = \prod_{h=1}^{n} \left\{ exp(\beta z_h(v_{h,n_h})) / \sum_{R(v_{h,n_h})} exp(\beta z_q(v_{h,n_h})) \right\}^{\delta_h} \tag{6.9}$$

où $R(v_{h,n_h})$ dénote la somme sur tous les indices q tels que $v_{q,n_q} \geq v_{h,n_h}$, c'est à dire la somme sur tous les sujets à risque au temps v_{h,n_h}.

Certains paramètres, implicites dans l'équation (6.8), doivent être définis : le nombre de noeuds, l'ordre du polynôme de lissage, ainsi que le seuil de diminution de la clairance permettant de distinguer les états 2 et 3. L'emplacement des noeuds est choisi aux quantiles correspondants. Comme de nombreux auteurs, nous supposons les polynômes cubiques pour leurs bonnes propriétés de lissage. Le couple formé par le nombre de noeuds intérieurs et le seuil de diminution, (k, s), est celui maximisant la fonction (6.9) pour certaines combinaisons :

$$(\hat{k}, \hat{s}) = Argmax_{(k,s)}\{\mathcal{VP}\};\ \ k = 1, ..., 15\ ;\ s = 10\%, ..., 90\% \tag{6.10}$$

Aucune pénalisation à la vraisemblance partielle n'est définie. En effet, le seul objectif est d'obtenir la trajectoire du patient la plus informative pour prédire un échec. Si l'objectif était de la modélisation, une pénalisation en fonction du nombre de noeuds utilisés aurait pu être prise en compte.

6.2.2 Application aux données de transplantation

Comme il a été démontré dans les chapitres 4 et 5, l'année de la greffe est un facteur prédictif important de l'évolution d'un patient transplanté rénal. Pour réduire ce biais période, nous avons limité l'inclusion des individus greffés à partir de 1996. Cette année correspond à un double changement : meilleur renseignement du suivi de la clairance et changement des traitements d'induction. Les autres critères d'inclusion sont conservés : patients majeurs au moment de la greffe, aucune transplantation de pancréas associée et uniquement le centre hospitalier nantais. Au total, 819 patients sont inclus dans l'analyse. A la date de greffe, les patients sont âgés en moyenne de 48,5 ans, tandis que l'âge moyen des donneurs est de 46,9 ans. 62 % des receveurs sont des hommes, contre 63 % chez les donneurs. La répartition selon l'année de greffe est équilibrée. La figure (6.5) présente les distributions de ces principales variables continues.

Même si l'échantillon est restreint à partir de 1996, la figure (6.6) permet d'évaluer l'éventuel biais dû à l'évolution des profils des patients en fonction de l'année de greffe. On voit clairement que les âges moyens du donneur et du receveur augmentent. Les autres marqueurs ne semblent pas suivre une telle tendance, que se soit l'ischémie froide, le délai de reprise au démarrage ou le sexe du donneur ou du receveur.

La figure (6.7) illustre l'estimation du nombre de noeuds et du pourcentage de diminution de la clairance caractérisant l'état 3. Dans notre application, on retient ainsi un lissage par B-splines cubiques de degré 8, c'est-à-dire 4 noeuds intérieurs ($\hat{k} = 4$). A partir des observations de clairance ainsi lissées, il apparaît assez nettement un maximum de vraisemblance pour une valeur du seuil $\hat{s}$ égale à 45 %. La clairance ainsi lissée permet d'obtenir le modèle de survie le plus prédictif d'un des deux événements terminaux. Le modèle

	Coef	ET	RR	$IC_{95\%}$(RR)	p-value
Etat de gravité 1	1				
Etat de gravité 2	0,21	0,34	1,23	[0,62 ; 2,45]	0,5500
Etat de gravité 3	22,64	0,30	22,64	[12,59 ; 40,69]	0,0001

TAB. 6.1 – Relation entre les 3 états de gravité et le retour en dialyse ou le décès.

	Coef	ET	RR	$IC_{95\%}$(RR)	p-value
Etat de gravité 1	1				
Etat de gravité 2	3,10	0,22	22,20	[14,30 ; 34,40]	0,0001

TAB. 6.2 – Relation entre les 2 états de gravité et le retour en dialyse ou le décès.

correspondant est présenté dans le tableau (6.1). Les patients dans l'état 2 ont 1,2 fois plus de risque de décéder ou de retourner en dialyse que dans l'état 1 ($IC_{95\%} = [0,62; 2,45]$). Ce risque est égal à 22,6 pour les patients dans l'état 3 ($IC_{95\%} = [12,59; 40,69]$). En terme de marqueur de risque, l'état intermédiaire ne semble donc pas significativement différent du premier l'état (p=0,5500). Ce résultat remet en question la structure multi-états admettant 3 états de gravité. Un modèle basé sur deux états de gravité apparaît plus pertinent. Le schéma (6.3) représente cette nouvelle classification.

A partir de cette nouvelle hypothèse, la même stratégie de modélisation est appliquée. Les résultats sont présentés par la figure (6.8). Une base de splines cubiques avec 2 noeuds internes est retenue pour le lissage de la clairance de la créatinine. L'adéquation du modèle (6.8) est maximale pour un passage à l'état 2 caractérisé par une chute de plus de 30 % du pic de la clairance post-transplantation. Le modèle de survie correspondant est présenté dans le tableau (6.2). Le risque de décès ou de retour en dialyse est multiplié par plus de 22,2 lors du passage de l'état 1 à l'état 2 ($IC_{95\%}(RR) = [14,30; 34,40]$). 73,3 % des 116 événements terminaux observés ont lieu dans l'état 2.

6.3 Modèle semi-markovien et censure par intervalle

6.3.1 Description de la structure multi-états

La structure multi-états de type "aggravation/échec" est définie par le schéma (6.4). Elle est constituée de 2 états de gravité (transitoires), 2 états absorbants et 5 transitions. Contrairement au modèle utilisé dans les chapitres 4 et 5, tous les patients commencent leur évolution dans l'état 1. Ils peuvent soit transiter vers l'état 2, soit subir un événement terminal. L'état 3 représente le retour en dialyse et l'état 4 caractérise le décès du patient. A partir de l'état transitoire le plus grave, seul le passage vers un événement terminal est

possible. Les patients peuvent être censurés à droite dans les deux états transitoires.

Le tableau (6.3) présente la répartition des patients selon leur trajectoire observée. Ces premiers résultats montrent que la structure de gravité semble plus associée au retour en dialyse qu'au décès. En effet, parmi les 77 retours observés, 61 individus (79,2 %) étaient dans l'état 2 au moins à la dernière visite. En revanche, la répartition des décès est équilibrée selon l'état à la dernière visite. Concernant les temps moyens d'apparition d'un événement terminal, les délais sont en moyenne plus courts lorsqu'un patient n'a jamais été observé dans l'état 2. Ceci justifie la conservation des transitions directes de l'état 1 aux états k ($k = 3, 4$).

Etats observés	Effectif	Pourcentage	Moyenne[1]	Médiane[2]
1[3]	537	64,0 %	44,9	40,7
1 ; 2[3]	190	22,7 %	48,9	45,3
1 ; 2 ; 3	61	7,3 %	44,6	43,5
1 ; 2 ; 4	18	2,1 %	36,3	31,8
1 ; 3	16	1,9 %	13,5	7,0
1 ; 4	17	2,0 %	31,4	14,3
total	839	100,0 %		

TAB. 6.3 – Répartition des patients selon leur trajectoire observée.

6.3.2 Procédure d'estimation

La structure du modèle est sensiblement différente de celle abordée dans les chapitres précédents : seulement 2 états de gravité, à partir desquels il est possible de transiter vers un des deux états absorbants. De plus, tous les patients commencent leur évolution dans l'état 1. La vraisemblance s'en trouve considérablement modifiée. En effet, une des hypothèses était que la régularité du processus d'observation permettait de supposer que tous les états occupés par un patient étaient renseignés (même si les temps de transition n'étaient pas exactement connus). Or, comme le montre le tableau (6.3), de nombreuses transitions de l'état 1 vers un état terminal sont observées. L'éventuel passage par l'état 2 doit donc être pris en compte. A partir de ces observations, la censure rend à la fois incomplets les temps de transition et les séquences d'états.

Les notations précédentes sont adaptées à ce nouveau contexte. Soit un échantillon constitué de n sujets, $h = 1, ..., n$. A ce niveau de l'analyse, à chacune des visites de l'individu h aux temps $v_{h,r}$ ($r = 0, 1, ..., n_h$), correspond une valeur lissée de clairance de la créatinine $\tilde{y}_{h,r}$ et un état observé $W_{h,r}$. Par définition, tous les patients commencent

[1] Temps moyen du suivi avant la censure ou l'événement terminal en mois.
[2] Temps médian du suivi avant la censure ou l'événement terminal en mois.
[3] Trajectoires censurées à droite.

leur évolution dans l'état 1, $W_{h,0} = 1$. Si l'événement final du suivi est le retour en dialyse, $W_{h,n_h} = 3$. S'il s'agit du décès du patient, alors $W_{h,n_h} = 4$. Sinon, pour toutes les autres visites $\{W_{h,r}, r = 1, ..., n_h\} \in \{1, 2\}$.

A partir de cette séquence d'états, il est directement déduit la séquence d'états observés et distincts, notée $\{X_{h,r}, r = 0, ..., m_h\}$. m_h indique donc le nombre de transitions observées chez le sujet h. Cette séquence forme la chaîne de Markov sous-jacente correspondant à l'expression (2.1). $d_{h,r}$ représente le temps passé dans l'état $X_{h,r}$, c'est-à-dire l'état suivant la r*ième* transition. Rappelons que v_{h,n_h} correspond au temps de fin de suivi depuis la date de greffe pour l'individu h. Il est associé soit à un événement terminal, soit aux dernières nouvelles. Distinguons les différentes possibilités de trajectoire d'un patient, comme le montre la figure (6.9).

(i) L'individu h est observé dans les deux états de gravité avant de déclarer un événement terminal k $(k = 3, 4)$. Notons $C_{h,1}$ la contribution à la vraisemblance d'un tel individu, avec $\delta_{h,1}$ l'indicatrice égale à 1 si le sujet h respecte cette trajectoire.

$$\begin{aligned}
C_{h,1} &= \lim_{\Delta d \to 0^+} \Big\{ P(v_{h,n_h} - d_{h,0} < d_{h,1} < v_{h,n_h} - d_{h,0} + \Delta d, X_{h,2} = k, \\
&\quad d^0_{h,0} < d_{h,0} < d^1_{h,0}, X_{h,1} = 2 | X_{h,0} = 1)/\Delta d \Big\} \\
&= \lim_{\Delta d \to 0^+} \Big\{ P(v_{h,n_h} - d_{h,0} < d_{h,1} < v_{h,n_h} - d_{h,0} + \Delta d, X_{h,2} = k | \\
&\quad d^0_{h,0} < d_{h,0} < d^1_{h,0}, X_{h,1} = 2, X_{h,0} = 1) \\
&\quad \times P(d^0_{h,0} < d_{h,0} < d^1_{h,0}, X_{h,1} = 2 | X_{h,0} = 1)/\Delta d \Big\} \\
&= \lim_{\Delta d \to 0^+} \Big\{ P(X_{h,2} = k | X_{h,1} = 2) P(v_{h,n_h} - d_{h,0} < d_{h,1} < v_{h,n_h} - d_{h,0} + \Delta d | \\
&\quad X_{h,2} = k, X_{h,1} = 2)/\Delta d \Big\} P(X_{h,1} = 2 | X_{h,0} = 1) \\
&\quad \times P(d^0_{0,h} < d_{h,0} < d^1_{h,0} | X_{h,1} = 2, X_{h,0} = 1) \\
&= P_{12} P_{2k} \int_{d^0_{h,0}}^{d^1_{h,0}} f_{12}(u) f_{2k}(v_{h,n_h} - u) du
\end{aligned} \tag{6.11}$$

(ii) L'individu h n'est pas observé dans l'état 2 avant de déclarer un événement terminal k $(k = 3, 4)$. Notons $C_{h,2}$ la contribution à la vraisemblance d'un tel individu, avec $\delta_{h,2}$ l'indicatrice égale à 1 si les observations du sujet h respectent ce schéma. La structure (6.4) montre que dans ce cas, le patient peut être directement passé de l'état 1 à l'état k, mais il peut aussi être passé par l'état 2 entre $d^0_{h,0}$ et v_{h,n_h}. Distinguons les deux possibilités. Premièrement, supposons que le patient subit la transition $1 \to k$ directement :

$$\begin{aligned}
&\lim_{\Delta d \to 0^+} \Big\{ P(v_{h,n_h} < d_{h,0} < v_{h,n_h} + \Delta d, X_{h,1} = k | X_{h,0} = 1)/\Delta d \Big\} \\
&\quad = \lim_{\Delta d \to 0^+} \Big\{ P(X_{h,1} = k | X_{h,0} = 1) \\
&\quad \times P(v_{h,n_h} < d_{h,0} < v_{h,n_h} + \Delta d | X_{h,1} = k, X_{h,0} = 1)/\Delta d \Big\} \\
&\quad = P_{1k} f_{1k}(v_{h,n_h})
\end{aligned} \tag{6.12}$$

Deuxièmement, considérons le cas où le passage à l'état 2 a été censuré entre $d^0_{h,0}$ et v_{h,n_h}. En reprenant les mêmes développements aboutissant à l'expression (6.11), on obtient :

$$\begin{aligned}&\lim_{\Delta d\to 0^+}\Big\{P(v_{h,n_h}-d_{h,0}<d_{h,1}<v_{h,n_h}-d_{h,0}+\Delta d, X_{h,2}=k,\\ &\quad d^0_{h,0}<d_{h,0}<v_{h,n_h}, X_{h,1}=2|X_{h,0}=1)/\Delta d\Big\}\\ &=P_{12}P_{2k}\int_{d^0_{h,0}}^{v_{h,n_h}} f_{12}(u)f_{2k}(v_{h,n_h}-u)du \qquad (6.13)\end{aligned}$$

A partir des expressions (6.12) et (6.13), on en déduit :

$$C_{h,2}=P_{1k}f_{1k}(v_{h,n_h})+P_{12}P_{2k}\int_{d^0_{h,0}}^{v_{h,n_h}} f_{12}(u)f_{2k}(v_{h,n_h}-u)du \qquad (6.14)$$

(iii) L'individu h est censuré dans l'état 1, la date de dernières nouvelles correspondant au temps v_{h,n_h} depuis la greffe. Sa contribution $C_{h,3}$ est indiquée par $\delta_{h,3}$.

$$\begin{aligned}C_{h,3} &= P(d_{h,0}>v_{h,n_h}|X_{h,0}=1)\\ &= \sum_{j\neq 1}P(X_{h,1}=j|X_{h,0}=1)P(d_{h,0}>v_{h,n_h}|X_{h,1}=j,X_{h,0}=1)\\ &= \sum_{j\neq 1}P_{1j}S_{1j}(v_{h,n_h})=S_{1.}(v_{h,n_h}) \qquad (6.15)\end{aligned}$$

(iv) L'individu h est censuré dans l'état 2, le temps de censure correspondant à la dernière visite v_{h,n_h}. Sa contribution $C_{h,4}$ est indiquée par $\delta_{h,4}$.

$$\begin{aligned}C_{h,4} &= P(d_{h,1}>v_{h,n_h}-d_{h,0}, X_{h,1}=2, d^0_{h,0}<d_{h,0}<d^1_{h,0}|X_{h,0}=1)\\ &= P(d_{h,1}>v_{h,n_h}-d_{h,0}|X_{h,1}=2, d^0_{h,0}<d_{h,0}<d^1_{h,0}, X_{h,0}=1)\\ &\quad\times P(d^0_{h,0}<d_{h,0}<d^1_{h,0}, X_{h,1}=2|X_{h,0}=1)\\ &= \Big\{\sum_{j=3}^{4}P(X_{h,2}=j|X_{h,1}=2)P(d_{h,1}>v_{h,n_h}-d_{h,0}|X_{h,2}=j,X_{h,1}=2)\Big\}\\ &\quad\times P(X_{h,1}=2|X_{h,0}=1)P(d^0_{h,0}<d_{h,0}<d^1_{h,0}|X_{h,1}=2,X_{h,0}=1)\\ &= \int_{d^0_{h,0}}^{d^1_{h,0}}P_{12}f_{12}(u)\Big\{\sum_{j=3}^{4}P_{2j}S_{2j}(v_{h,n_h}-u)\Big\}du\\ &= P_{12}\int_{d^0_{h,0}}^{d^1_{h,0}}f_{12}(u)S_{2.}(v_{h,n_h}-u)du \qquad (6.16)\end{aligned}$$

A partir de l'ensemble de ces contributions, la vraisemblance d'un échantillon composé de n sujets s'écrit alors :

$$\mathcal{V}=\prod_{h=1}^{n}\prod_{i=1}^{4}(C_{h,i})^{\delta_{h,i}} \qquad (6.17)$$

La maximisation de la fonction (6.17) sous-entend la résolution de certaines fonctions intégrales. En choisissant les distributions des temps d'attente exponentielles, ces intégrales peuvent trouver une solution analytique. L'utilisation de lois Weibull généralisées ne permet pas l'obtention de formes explicites. Une approximation numérique est alors nécessaire.

La méthode des trapèzes constitue l'approche la plus simple et la plus intuitive pour le calcul numérique d'une intégrale, cette dernière étant approchée par l'aire de polynômes interpolateurs, en l'occurrence des trapèzes. Soit $[a, b]$, l'intervalle d'intégration d'une fonction f, divisé en n sous-intervalles de longueur h. La valeur de la primitive est alors approchée par la somme des aires des trapèzes sur chaque sous-intervalle :

$$\int_a^b f(u)du \approx n^{-1}(b-a)\big(0,5(f(a)+f(b)) + \sum_{k=1}^{n-1} f(a+kh)\big) \tag{6.18}$$

L'inconvénient majeur de cette méthode est le temps de calcul, n doit être suffisamment grand pour minimiser l'erreur commise dans l'approximation. Ce problème est d'autant plus important dans notre contexte, puisque les intégrales interviennent dans un algorithme de maximisation. Les méthodes de quadrature constituent des approximations plus efficaces en terme de ressources. La quadrature de Gauss-Legendre est l'une des plus utilisées, l'expression (6.18) devient alors :

$$\int_a^b f(u)du \approx 0,5(b-a)\sum_{q=1}^{Q} \tilde{w}_q f\big(0,5(b-a)\tilde{u}_q + 0,5(a+b)\big) \tag{6.19}$$

où $\tilde{u}_q$ sont les racines du polynôme de Legendre de degré Q et où $\tilde{w}_q$ sont les poids associés à ces racines [96]. Des tables permettent d'obtenir les valeurs des points et des poids. Pour l'intégration de fonctions simples, une bonne précision est obtenue à partir de 3 noeuds. Nous utiliserons $Q = 10$, dont les racines et les poids sont présentés dans le tableau (B.1) de l'annexe B.

6.3.3 Introduction des covariables

Dans le chapitre précédent, les covariables associées aux vitesses de transition ont été modélisées de manière non-proportionnelle à la fonction de risque des temps d'attente dans les états. L'application aux données de transplantation a montré l'intérêt de cette approche. Cependant, certaines contraintes doivent être posées lors du calcul des paramètres pour respecter la positivité des fonctions de risque. Ce dernier aspect rend difficile l'estimation du modèle. Pour résoudre cette difficulté, plutôt que de définir en premier lieu la fonction de survie, nous posons la fonction de risque du temps d'attente dans l'état i avant de passer à l'état j égale à :

$$\lambda_{ij}\big(x, \eta_{h,ij}(x)\big) = \lambda_{0,ij}(x) exp\big(\eta_{h,ij}(x)\big) \tag{6.20}$$

où $\eta_{h,ij}(x)$ est le prédicteur linéaire des covariables dont les coefficients de régression peuvent dépendre du temps passé dans l'état, x, grâce à des interactions polynomiales dont le degré est limité à 2. Par exemple, pour une covariable $z_{h,ij}$:

$$\eta_{h,ij}(x) = \beta_{ij}^{(1)} z_{h,ij} + \beta_{ij}^{(2)} z_{h,ij} x + \beta_{ij}^{(3)} z_{h,ij} x^2 \tag{6.21}$$

La fonction de survie correspondante à (6.20) n'a pas de solution formelle. En reprenant la méthode de quadrature de Gauss-Legendre (6.19), on a :

$$\begin{aligned} S_{ij}\big(x, \eta_{h,ij}(x)\big) &= exp\Big(- \int_0^x \lambda_{ij}\big(u, \eta_{h,ij}(u)\big) du \Big) \\ &\approx exp\Big(- 0.5x \sum_{q=1}^{Q} \tilde{w}_q \lambda_{ij}\big(0.5x(\tilde{u}_q + 1), \eta_{h,ij}(0.5x(\tilde{u}_q + 1))\big)\Big) \end{aligned} \tag{6.22}$$

L'approximation (6.22) pose un problème lorsque la fonction de risque tend vers l'infini lorsque x tend vers 0. En effet, il est très difficile d'approcher une singularité par un polynôme. Plusieurs centaines de noeuds sont alors nécessaires. La relation de Chasles permet de résoudre ce problème, en répartissant les noeuds d'une manière plus adéquate. Soit c une valeur positive proche de 0 :

$$\begin{aligned} S_{ij}\big(x, \eta_{h,ij}(x)\big) &= exp\Big(- \int_0^c \lambda_{ij}\big(u, \eta_{h,ij}(u)\big) du - \int_c^x \lambda_{ij}\big(u, \eta_{h,ij}(u)\big) du \Big) \\ &\approx exp\Big(- 0.5c \sum_{q=1}^{Q_1} \Big\{ \tilde{w}_q \lambda_{ij}\big(0.5c(\tilde{u}_q + 1), \eta_{h,ij}(0.5c(\tilde{u}_q + 1))\big)\Big\} \\ &- 0.5(x-c) \sum_{q=1}^{Q_2} \Big\{ \tilde{w}_q \lambda_{ij}\big(0.5(x-c)\tilde{u}_q + 0.5(x+c), \\ &\quad \eta_{h,ij}(0.5(x-c)\tilde{u}_q + 0.5(x+c))\big)\Big\}\Big) \end{aligned} \tag{6.23}$$

Ainsi, le nombre de noeuds au début du domaine d'intégration, Q_1, peut être choisi assez important, sans être contraint d'utiliser un grand nombre de noeuds au total. Comme précédemment, en supposant $Q_2 = 10$, quelques calculs préliminaires nous montrent que la combinaison $c = 0,02$ et $Q_1 = 30$ offre une bonne approximation de ce type d'intégrales singulières en 0. Les racines et les poids sont présentés dans le tableau (B.2) de l'annexe B.

La définition des états de gravité impose que tous les individus commencent dans l'état 1. L'introduction de covariables associées aux probabilités des états initiaux est donc inutile. Cependant, en reprenant le même principe de régression logistique multinomiale, il est possible de se soustraire à l'homogénéité de la chaîne de Markov sur tous les individus. Ainsi, à partir des expressions (2.1) et (4.2), la probabilité que l'état consécutif à l'état i soit l'état j s'écrit :

$$P_{ij}(y_{h,ik}, k \neq i) = exp(\gamma_{ij} + \beta_{ij} y_{h,ij}) / \sum_{k \neq i} exp(\gamma_{ik} + \beta_{ik} y_{h,ik}) \tag{6.24}$$

Transition	Variable	Coef.	ET	p-value
$1 \rightarrow 2$	Intercept	1,88	0,68	0,0063
$1 \rightarrow 2$	Délai de reprise	0,73	0,34	0,0306
$1 \rightarrow 2$	Age du donneur	-2,04	0,70	0,0034
$1 \rightarrow 3$	Intercept	-2,81	0,53	0,0001
$2 \rightarrow 3$	Intercept	1,14	0,34	0,0007
$2 \rightarrow 3$	Incompatibilités A+B+DR	0,93	0,47	0,0483

TAB. 6.4 – Coefficients de régression associés à la chaîne de Markov

où γ_{ij} et β_{ij} sont respectivement l'intercept et le vecteur des coefficients de régression associé à $y_{h,ij}$, le vecteur des covariables associées à la trajectoire $i \rightarrow j$. La contrainte markovienne, $\sum_{j \neq i} P_{ij}(y_{h,ik}, k \neq i) = 1$ impose la définition d'un état de référence. Par convention, nous adoptons la nullité de γ_{ij} et de β_{ij} pour l'état j ayant le plus grand numéro. Par exemple, à partir de l'état 1, il est possible de transiter vers les états 2, 3 ou 4. La transition $1 \rightarrow 4$ est alors la transition de référence ($\gamma_{14} = \beta_{14} = 0$). De la même manière, $\gamma_{24} = \beta_{24} = 0$.

Dans le modèle ainsi défini, les covariables peuvent agir soit sur les vitesses de transition, soit sur les trajectoires du processus.

6.3.4 Application aux données de transplantation

Parallèlement aux chapitres précédents, les covariables et les paramètres des lois des temps d'attente ont d'abord été sélectionnés en univarié ($p \leq 0.20$) et l'hypothèse de proportionnalité des risques a été examinée graphiquement. Si cette hypothèse n'est pas respectée pour certains couples transition/covariable, l'expression (6.21) est utilisée pour modéliser cette relation. Après la stratégie de sélection multivariée descendante ($p \leq 0.05$), 11 facteurs explicatifs sont retenus ($log(\mathcal{V}) = -1550.70$). Les estimations des coefficients de régression,γ_{ij} et β_{ij} ($ij \in \{12, 13, 14, 23, 24\}$), sont présentées dans le tableau (6.4) pour les variables associées aux séquences d'états et dans le tableau (6.5) pour les variables agissant sur les temps de séjour dans les états.

Concernant les trajectoires, la séquence $X_{h,0} = 1$ et $X_{h,1} = 2$ est plus probable chez les patients dont le délai de reprise de l'activité du greffon est supérieur ou égal à 6 jours. A l'inverse, recevoir un rein d'un donneur âgé de plus de 55 ans augmente le risque d'une transition directe de l'état initial vers le décès. Enfin, sachant que l'état occupé est le second stade de gravité, le retour en dialyse est plus probable que le décès pour les greffes présentant plus de 4 incompatibilités HLA (A+B+DR).

Certaines covariables agissent parallèlement sur les durées de séjour. Sachant que l'état 2 est le suivant, le temps passé dans l'état initial semble plus court pour les femmes traités

Transition	Variable	Coef.	ET	p-value
$1 \rightarrow 2$	Traitement d'induction	0,36	0,14	0,0063
$1 \rightarrow 2$	Sexe du receveur	-0,26	0,13	0,0541
$1 \rightarrow 2$	Age du donneur	0,96	0,23	0,0001
$1 \rightarrow 3$	Ischémie froide	5,02	1,20	0,0001
$2 \rightarrow 3$	Incompatibilité A+B+DR	0,90	0,29	0,0017
$2 \rightarrow 3$	PRA	1,10	0,35	0,0016
$2 \rightarrow 3$	PRA $\times d$ *	-0,47	0,22	0,0309
$2 \rightarrow 4$	Délai de reprise	2,01	0,60	0,0008
$2 \rightarrow 4$	Sexe du receveur	1,52	0,64	0,0174
$2 \rightarrow 4$	Sexe du receveur $\times d$ *	-4,31	1,19	0,0003
$2 \rightarrow 4$	Sexe du receveur $\times d^2$ *	1,30	0,32	0,0001

* Interaction avec la durée dans l'état, d.

TAB. 6.5 – Coefficients de régression associés aux temps de séjours

par Simulect et ayant reçu un greffon d'un donneur de plus de 55 ans. De plus, sachant que le patient retourne directement en dialyse à partir de l'état initial, cette transition semble accélérée lorsque l'ischémie froide est d'au moins 24 heures. De la même manière, si le patient occupe l'état 2 avant de retourner en dialyse, cette transition est plus rapide pour les greffes à forte incompatibilité. Cette dernière relation est aussi observée pour les patients avec un taux d'immunisation supérieur à 0%, mais cet effet semble diminuer avec le temps passé dans l'état 2. Enfin, concernant le décès à partir de cet état de santé aggravé, la durée dans l'état 2 apparaît plus courte lorsque le délai de reprise de l'activité rénale a été long. Encore une fois, cet effet n'est pas constant.

Les résultats relatifs aux lois des temps d'attente dans les états sont présentés dans le tableau (6.6). La fonction de risque de la transition de l'état de gravité 1 à l'état de gravité 2 est en forme de $\cup$ avec un minimum environ 4 ans après la greffe (équation 1.12). Toujours en utilisant le test LRS et une stratégie de sélection descendante, il semble que les autres transitions soient distribuées suivant une loi Exponentielle sans mémoire (risque constant au cours du temps). On peut aussi voir que la fonction de risque de la transition $2 \rightarrow 3$ est égale à plus de 3 fois celle de la transition $1 \rightarrow 3$, montrant ainsi le rôle de marqueur précoce de l'état 2 pour le retour en dialyse.

6.4 Discussion

Dans ce chapitre, nous avons défini un modèle semi-markovien adapté pour la prise en compte de censures par intervalle. Les données incomplètes concernent à la fois la séquence des états et les temps de transition. Cette approche se révèle plus adéquate que

	σ_{ij}		ν_{ij}		θ_{ij}	
Transition	Estim.	ET	Estim.	ET	Estim.	ET
$1 \rightarrow 2$	68,84	82,00	0,77	0,05	0,25	0,21
$1 \rightarrow 3$	42,91	47,44	1	.	1	.
$1 \rightarrow 4$	109,80	73,31	1	.	1	.
$2 \rightarrow 3$	12,96	3,34	1	.	1	.
$2 \rightarrow 4$	6,30	3,93	1	.	1	.

TAB. 6.6 – Paramètres associés aux lois d'attente dans les états

celle définie dans les chapitres 4 et 5, où les transitions étaient supposées indépendantes, alors que ces dernières sont dépendantes de la trajectoire globale du sujet au cours de son suivi.

L'autre amélioration concerne la modélisation des facteurs influençant la dynamique du processus. D'une part, les variables agissant sur les vitesses de transition sont introduites sans supposer la proportionnalité avec la fonction de risque et sans imposer de contraintes afin de respecter la positivité de la fonction de risque. Pour cela, l'exponentiel du prédicteur linéaire des covariables est multiplié à la fonction de risque de base, certaines interactions avec le temps pouvant être introduites. La difficulté de cette méthode, par rapport à celle définie dans le chapitre 5, est que la fonction de survie associée n'est pas explicite et nécessite une approximation numérique plus lourde en terme de temps de calcul. Que se soit pour l'estimation de ces fonctions de survie ou pour le calcul des contributions à la vraisemblance, la quadrature de Gauss-Legendre est utilisée car elle nécessite un faible nombre de noeuds. D'autre part, le modèle permet l'inclusion de facteurs explicatifs dans la chaîne de Markov sous-jacente, permettant ainsi de modéliser l'hétérogénéité des séquences d'états parallèlement à la modélisation des vitesses précédentes.

L'avantage de cette approche paramétrique est l'estimation par maximum de vraisemblance. Le test LRS est alors à la base de la stratégie de sélection des paramètres du modèle semi-markovien. Il permet en effet l'évaluation de la parcimonie des distributions associées aux temps de séjours dans les états et l'identification des coefficients de régression significatifs. Par exemple, l'application aux données a montré que seul le temps d'attente, dans l'état initial avant de transiter vers l'état d'aggravation, est distribué suivant une loi de Weibull généralisée, alors que les autres transitions semblent associées à une fonction de risque constante. Cependant, l'hypothèse de stationnarité (2.15) ne peut pas être examinée par ce test. Les développements du chapitre suivant, basés sur les travaux de Aguirre-Hernandez et Farewell [97] dans le cadre markovien, se penchent sur cette difficulté en proposant un test d'adéquation.

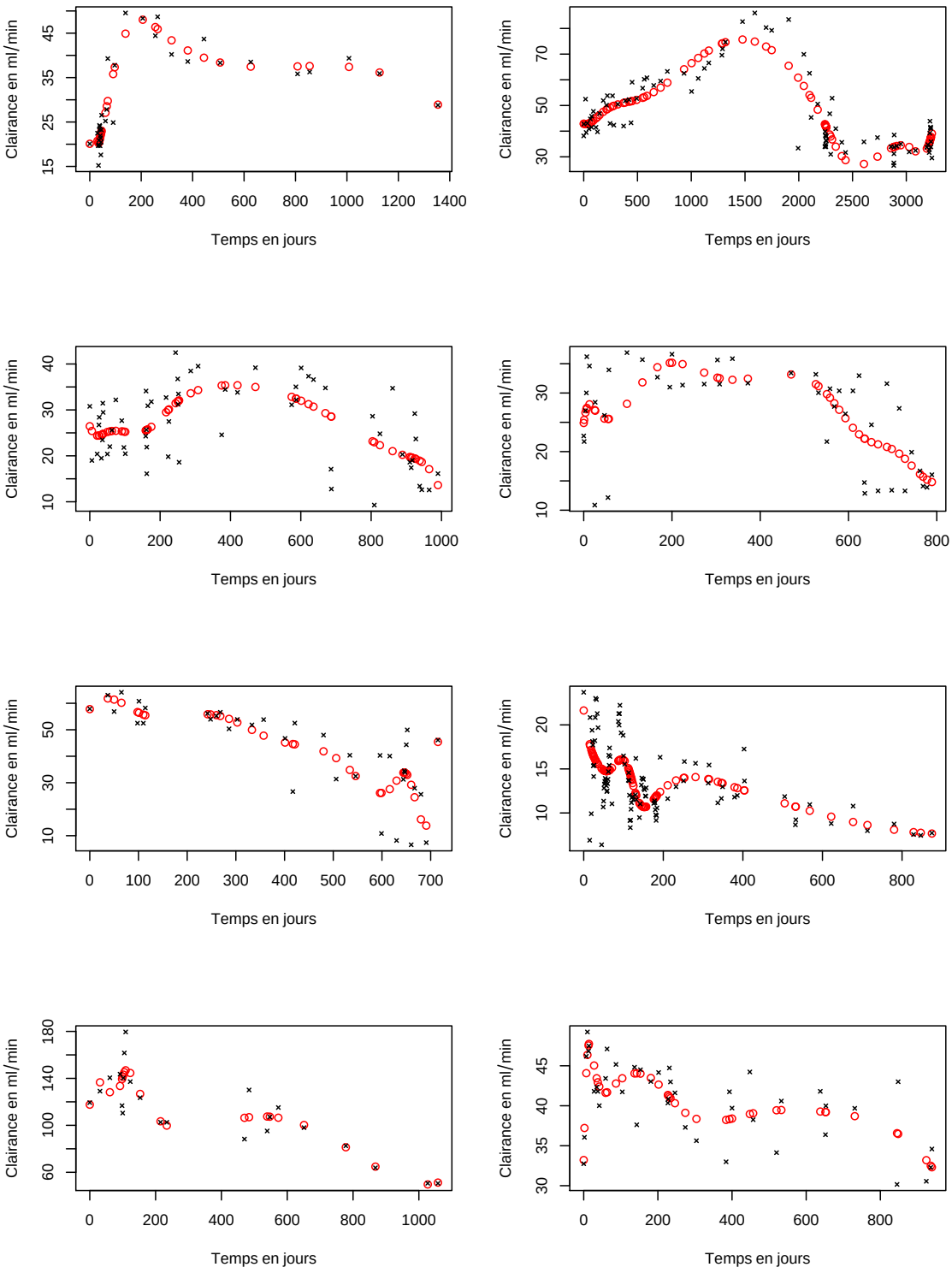

FIG. 6.1 – Diminution de la variabilité à court terme de la clairance de la créatinine chez des patients retournés en dialyse. × × × valeurs observées ; o o o valeurs sous-jacentes.

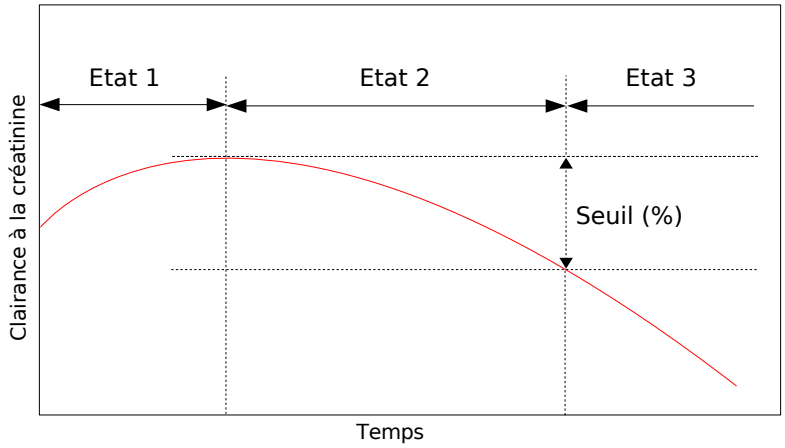

FIG. 6.2 – Classification en 3 états de gravité selon la clairance de la créatinine.

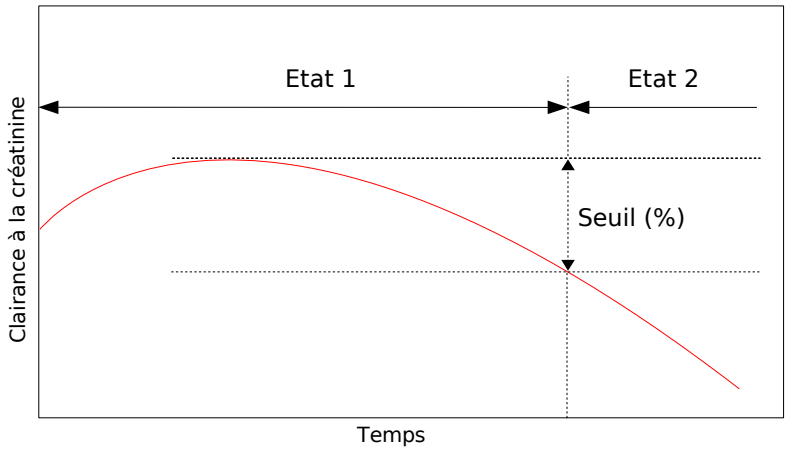

FIG. 6.3 – Classification en 2 états de gravité selon la clairance de la créatinine.

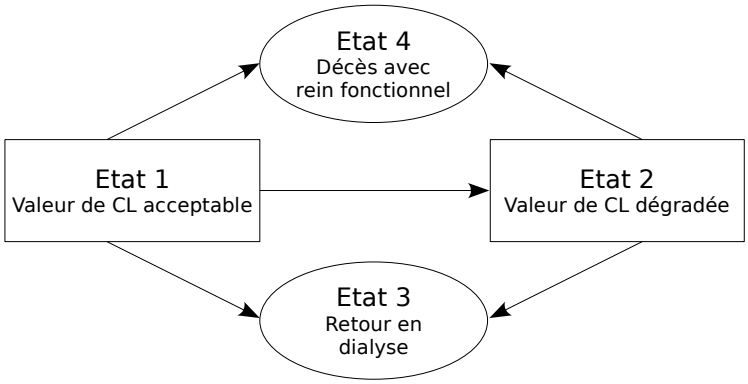

FIG. 6.4 – Structure du modèle multi-états avec deux états de gravité transitoires.

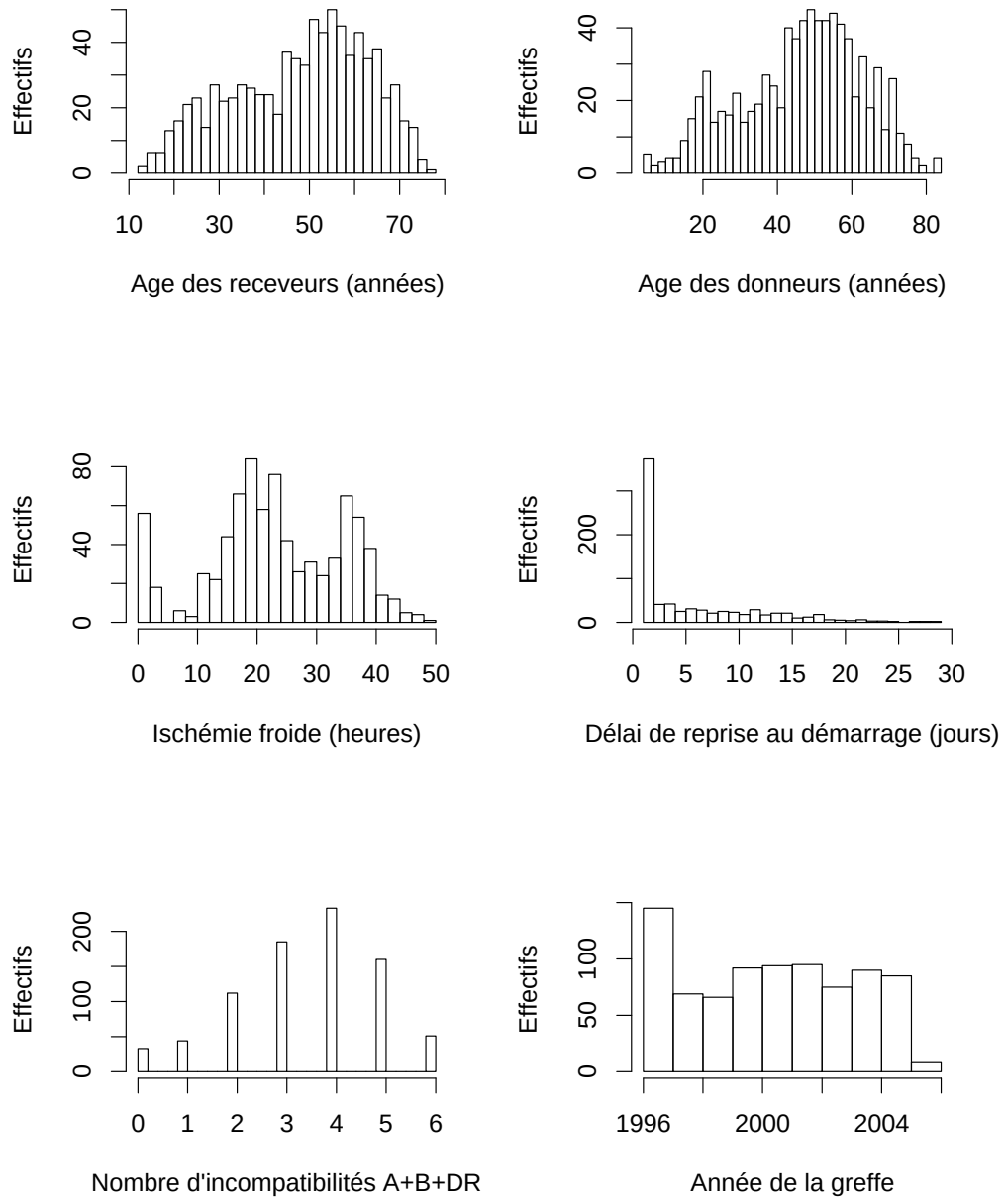

FIG. 6.5 – Répartitions des patients au moment de la greffe en fonction de caractéristiques continues.

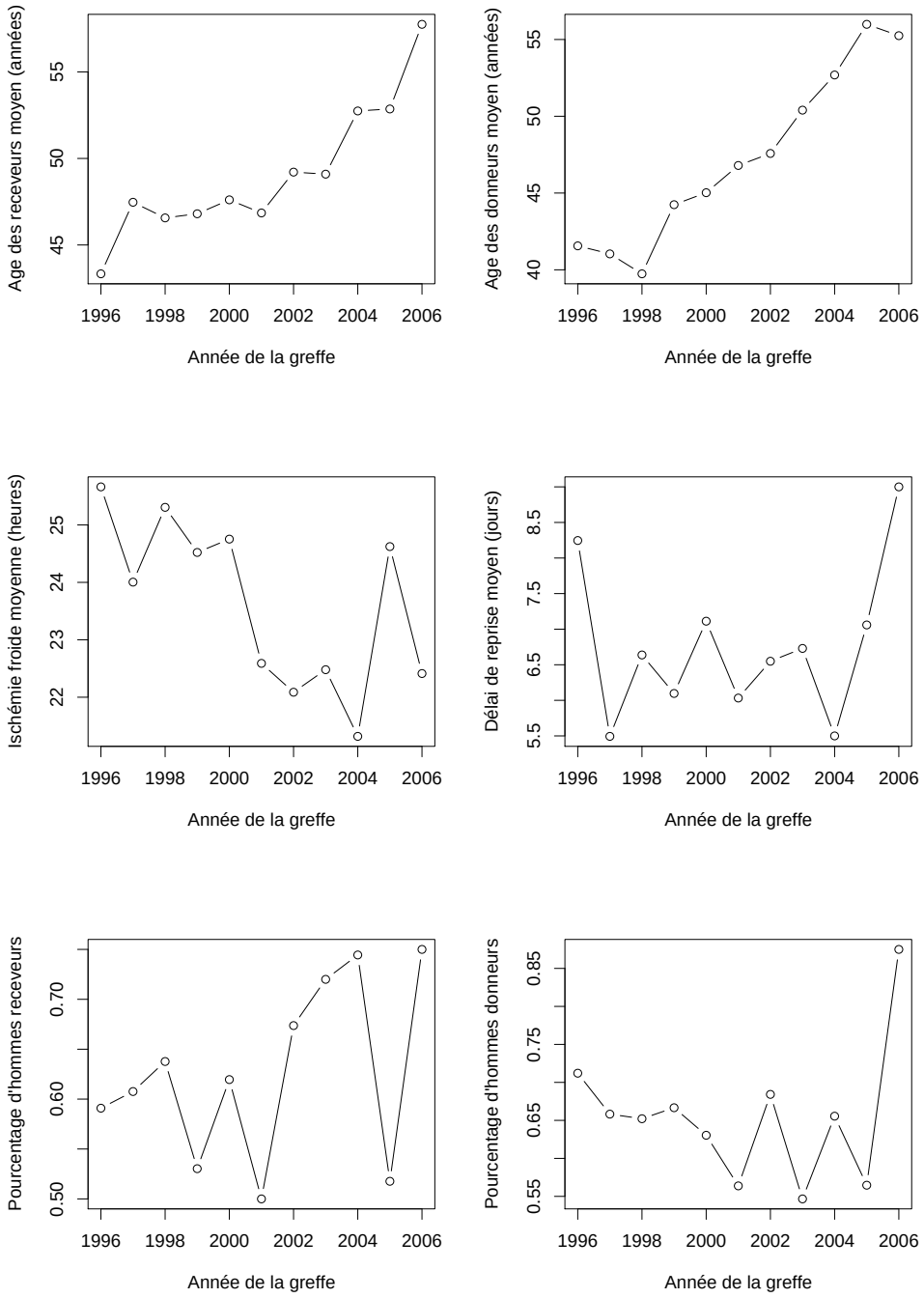

FIG. 6.6 – Evolutions du profil des patients en fonction de l'année de greffe.

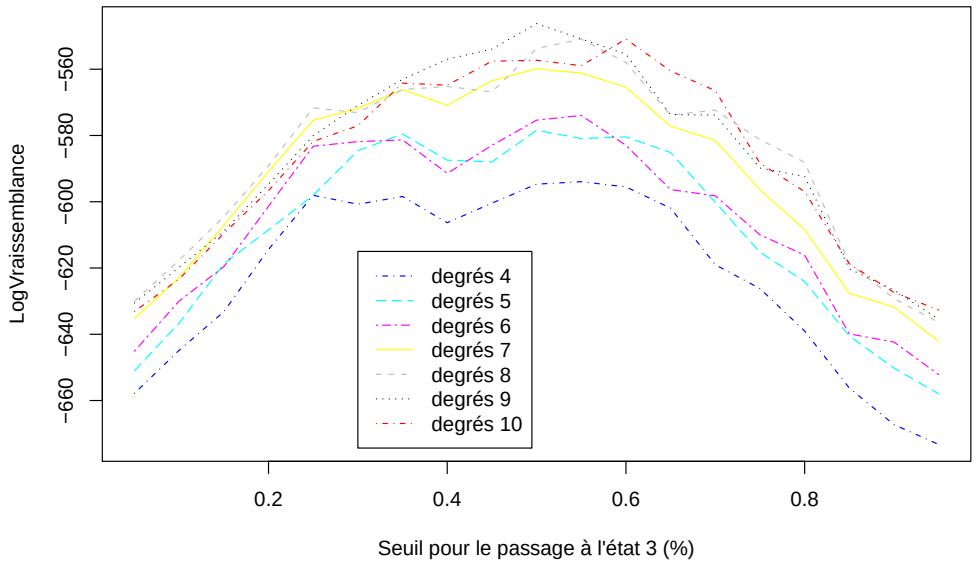

FIG. 6.7 – Sélection du nombre de noeuds, k, et du seuil de diminution de la clairance de la créatinine, s. Structure à 5 états (3 états de gravité transitoires)

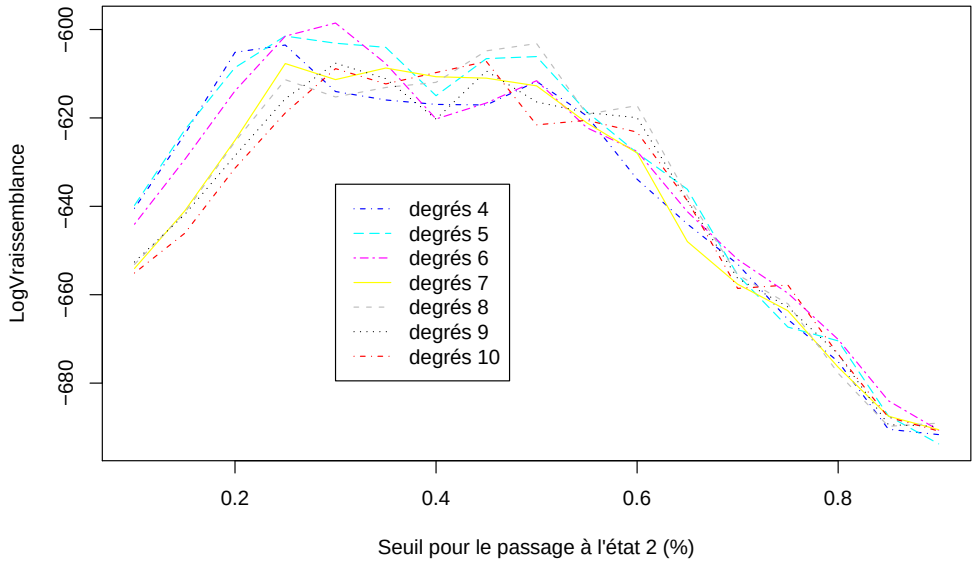

FIG. 6.8 – Sélection du nombre de noeuds, k, et du seuil de pourcentage de diminution de la clairance de la créatinine, s. Structure à 4 états (2 états de gravité transitoires)

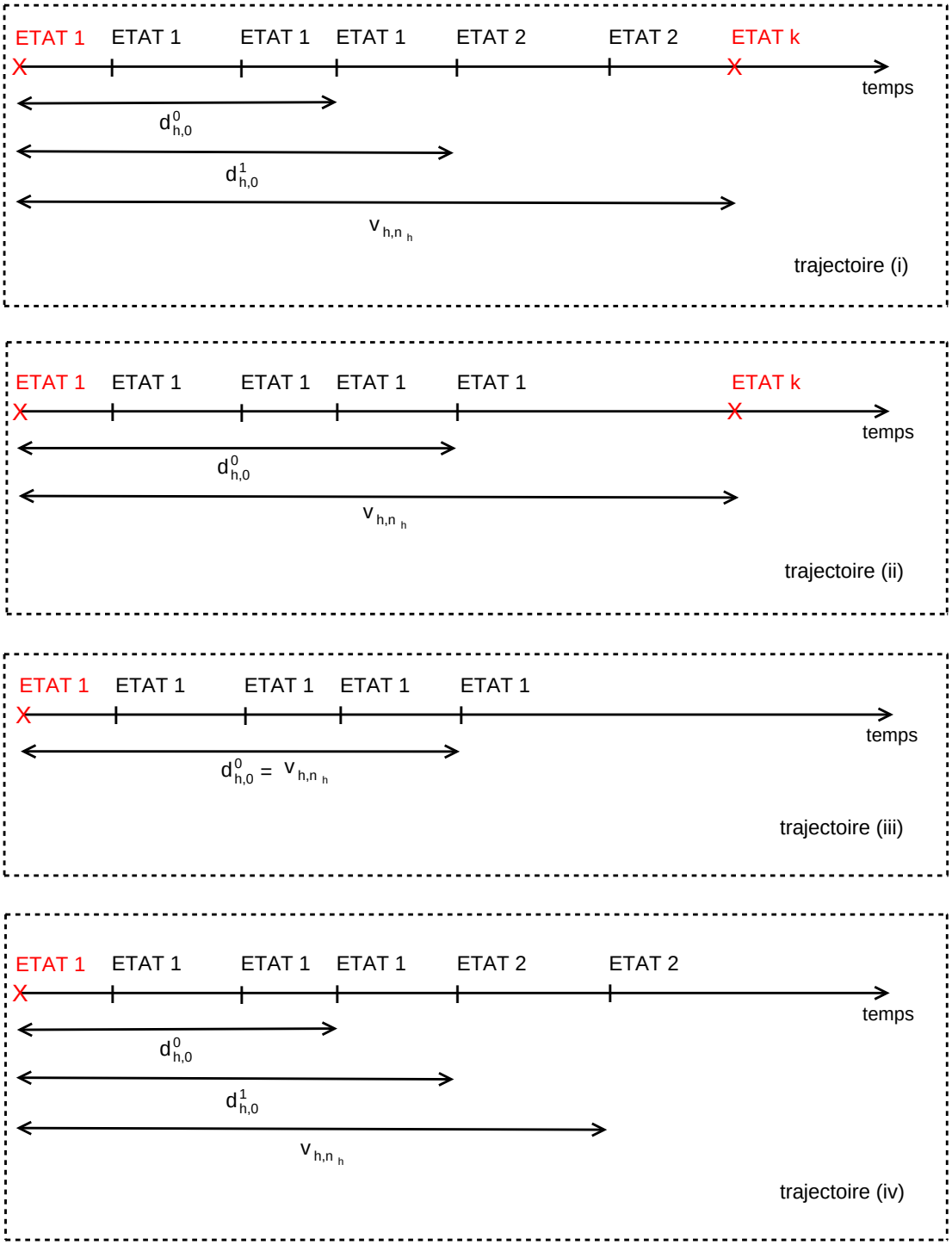

FIG. 6.9 – Trajectoires possibles d'un patient transplanté selon la structure à 4 états. | | | états observés aux visites ; × × × événements dont la date est exactement connue.

Chapitre 7

Test d'adéquation

Même si l'utilisation de modèles multi-états est de plus en plus répandue dans l'analyse de données longitudinales, peu d'attention est accordée à l'évaluation de la qualité d'ajustement aux données. La plupart des tests d'adéquation concerne les modèles markoviens. Pour chaque sujet, le nombre de mesures de la variable d'intérêt, c'est à dire l'état de santé, est considéré fixe. De plus, aucune covariable n'est prise en compte. Comme l'expliquent Bishop et al. [98], ce type de données longitudinales peut être représenté dans des tableaux de contingence. Les auteurs montrent que l'analyse de ces tables est équivalente à l'analyse de tableaux de contingence basés sur des modèles log-linéaires.

Lorsque les mesures longitudinales sont également espacées, Kalbfleisch et Lawless [99] ont aussi proposé la construction de ce type de tableaux de contingence répertoriant les effectifs de transitions observées et attendues pour chaque intervalle de temps. La matrice des probabilités de transition est utilisée pour calculer ce nombre attendu. La statistique du Chi-deux est calculée pour chaque table. Ces tableaux peuvent être considérés indépendants si le modèle markovien est d'ordre 1, la statistique de test est alors obtenue par la somme des statistiques calculées pour chaque intervalle. Les auteurs supposent que cette dernière possède asymptotiquement une distribution du Chi-deux. Ce test d'adéquation a aussi été proposé par Stavola [100].

Gentleman et al. [26] proposent une adaptation de cette statistique lorsque les intervalles de temps ne sont pas également espacés et lorsque aucune covariable n'est présente. Une partition de l'échelle de temps est définie, mais certaines mesures peuvent être manquantes. Une approximation du nombre observé est alors obtenue en supposant qu'un individu non-observé à un temps t soit resté dans le même état que le précédent.

Dans le cadre d'essais thérapeutiques, la plupart des analyses à mesures répétées sont définies avec une certaine périodicité. Cependant, en santé publique ou en médecine, la variabilité entre les visites peut être telle que l'échelle de temps des visites doit être considérée continue. Les méthodes d'adéquation classiques sont alors inadaptées, d'autant

plus qu'elles ne prennent pas en compte l'effet d'éventuelles covariables.

Aguirre-Hernandez et Farewell [97] ont défini plus récemment une statistique de type Pearson pour examiner la qualité d'ajustement d'un modèle markovien stationnaire à temps continu d'ordre 1. Cette statistique est construite pour les modèles dont les forces de transition dépendent de covariables. Elle est aussi adaptée aux données dont l'espacement entre les mesures et le nombre total d'observations peut varier d'un individu à l'autre. De plus, ils ne supposent pas de distribution asymptotique et proposent une méthode de bootstrap pour calculer la distribution de la statistique de test.

Notre problématique est équivalente puisque le processus d'observation d'un individu est complètement aléatoire, chaque individu possédant des intervalles différents entre visites. De plus, le modèle dépend de covariables. Cependant quelques différences persistent : le modèle est semi-markovien et les temps d'apparition des événements absorbants sont exactement renseignés. Nous proposons une statistique de test de type Pearson. L'hypothèse d'une distribution asymptotique de cette statistique étant trop forte, nous définissons une méthode de boostrap semi-paramétrique en adaptant les travaux de Aguirre-Hernendez et Farewell [97] et de Lawless et Babineau [101]. Nous l'appliquerons au dernier modèle estimé dans le chapitre 6.

7.1 La statistique de test

Dans le modèle de régression semi-markovien défini au chapitre 6, les taux de transitions dépendent des covariables, des séquences d'états observés et du temps passé dans l'état. Tous ces facteurs doivent être considérés dans la statistique de test.

(i) Regroupement en fonction des covariables. Comme toutes les covariables utilisées sont qualitatives ou discrètes, ce classement est assez direct. Les catégories ainsi définies sont dénotées par l'indice c ($c = 1, ..., C$).

(ii) Regroupement en fonction des événements étudiés. Si le nombre d'états possible est K, alors en considérant que dans le cas semi-markovien le passage d'un état vers lui-même est impossible, il existe un maximum de $K^2 - K$ transitions observables. En fonction de la structure multi-états étudiée, cet effectif peut-être restreint. C'est par exemple le cas si le modèle est uni-directionnel ou si certains états sont absorbants. Les événements seront notés par l'indice k ($k = 1, ..., K$).

(iii) Regroupement en fonction des temps d'incidence des événements. Pour obtenir des effectifs équilibrés [102], les quantiles des temps d'apparition des événements sont utilisés pour définir les intervalles de classement. On définit L le nombre d'intervalles avec comme bornes $(t_0, t_1, ..., t_l, ..., t_L)$. Ce découpage en fonction du temps chronologique permet de tester la stationnarité du processus. En effet, il est supposé que les forces de

transitions dépendent uniquement du temps d'attente dans l'état, indépendamment du temps chronologique (voir expression 2.15).

Soit $e_{l,k,c}$ le nombre de transitions attendues dans la cellule relative aux catégories (l,k,c). De la même manière, posons $n_{l,k,c}$ le nombre de transitions observées pour cette cellule. La statistique d'adéquation de type Pearson, permettant de tester l'adéquation du modèle semi-markovien est égale à :

$$G = \sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{c=1}^{C}\left\{(n_{l,k,c} - e_{l,k,c})^2/e_{l,k,c}\right\} \tag{7.1}$$

La distribution de la statistique (7.1) dépend à la fois de la taille de l'échantillon n, du nombre de cellules indépendantes du tableau de contingence R et du nombre de paramètres estimés dans le modèle semi-markovien η. Si R est petit $(R > \eta)$ et n grand, la statistique est supposée approximativement suivre une loi du Chi-deux à $(R-\eta)$ degrés de liberté. Cependant, ces hypothèses ne sont pas valides dans notre cas où R et η deviennent vite assez grands par rapport à la taille de l'échantillon. L'approximation du Chi-deux n'est donc pas adéquate a priori.

La définition de la distribution exacte est d'autant plus insurmontable que le nombre total d'observations n'est pas fixe. En effet, l'entrée dans un état absorbant termine l'observation d'un sujet. L'approximation du Chi-deux serait adéquate en considérant le nombre total fixe d'observations, alors la probabilité $P(n_{1,1,1} = N_{1,1,1}, ..., n_{L,K,C} = N_{L,K,C})$ est la somme des distributions multinomiales indépendantes et non-identiques.

Des simulations peuvent être utilisées pour fournir des intervalles d'estimation ou des p-values de certains tests lorsque la distribution asymptotique n'est pas disponible. Une méthode d'estimation de la distribution de (7.1) est la génération de B échantillons indépendants de bootstrap, à partir du modèle à tester sous l'hypothèse nulle, selon laquelle la régression est adéquate. On calcule alors une statistique pour chacun de ces échantillons. Comme B tend vers l'infini, la distribution de bootstrap de (7.1) approchera la vraie distribution de la statistique de test sous l'hypothèse nulle [103]. A partir de la statistique de test calculée sur l'échantillon initial, la p-value peut alors être calculée à partir de la distribution de la statistique. Cette méthode de bootstrap est appelée semi-paramétrique.

7.2 Application aux données de transplantation

7.2.1 Définition du tableau de contingence

Concernant le regroupement des événements, nous avons choisi de nous intéresser aux états absorbants, leur temps d'apparition étant exactement connu et l'entrée dans l'état

de gravité 2 étant censurée par intervalle. Ceci permet de facilement les classer selon le temps depuis la greffe. Nous considérons ainsi deux types d'observations ($K = 2$) : passage de l'état 1 ou 2 vers l'état 3 ($e \rightarrow 3$, avec $e = 1, 2$) et le passage de l'état 1 ou 2 vers l'état 4 ($e \rightarrow 4$, avec $e = 1, 2$). K représente donc le nombre d'états absorbants et terminaux. Pour simplifier les développements en respectant les notations précédentes, on définit $k = 3, 4$.

Nous avons choisi 5 intervalles de temps ($L = 5$) pour obtenir des effectifs consistants dans chaque cellule. Les bornes sont définies de telle sorte que les effectifs d'événements absorbants soient répartis de façon égale : t_l $(l = 1, ..., 5) = \{0{,}011\,;\ 0{,}689\,;\ 2{,}168\,;\ 3{,}826\,;\ 5{,}213\,;\ 9{,}158\}$.

Aucun regroupement selon les modalités des covariables n'est cliniquement intuitif. La stratégie consistant à grouper les individus selon certains facteurs de risque aboutirait à des cellules de très faibles effectifs pour les groupes plutôt protecteurs des événements terminaux. De plus, l'effet des covariables est testé graphiquement et par la statistique du rapport de vraisemblance. Les observations seront ainsi étudiées quelque soit le profil des patients ($C = 1$). L'objectif central de la statistique de test (7.1) est donc d'examiner la validité de l'hypothèse de stationnarité du modèle semi-markovien.

7.2.2 Calcul des effectifs

En reprenant l'expression (6.14), la probabilité qu'un individu h transite vers l'état k ($k = 3, 4$) au temps t depuis la greffe est égale à :

$$P_{1k}(y_{h,12}, y_{h,13}) f_{1k}(t, \eta_{h,1k}(t)) + P_{12}(y_{h,12}, y_{h,13}) P_{2k}(y_{h,23}) \int_0^t f_{12}(u, \eta_{h,12}(u)) f_{2k}(t-u, \eta_{h,2k}(t-u)) du$$

Ainsi, par intégration, la probabilité que cet événement k se produise entre t_{l-1} et t_l ($l = 1, ..., L$) est égale à :

$$P_{1k}(y_{h,12}, y_{h,13}) \int_{t_{l-1}}^{t_l} f_{1k}(t, \eta_{h,1k}(t)) dt + P_{12}(y_{h,12}, y_{h,13}) P_{2k}(y_{h,23}) \int_{t_{l-1}}^{t_l} \int_0^t f_{12}(u, \eta_{h,12}(u)) f_{2k}(t-u, \eta_{h,2k}(t-u)) du dt$$

En prenant en compte la censure à droite, l'effectif attendu de transitions dans la cellule (l, k) s'écrit :

$$e_{l,k} = \sum_{R(t_{l-1})} P_{1k}(y_{h,12}, y_{h,13}) \int_{t_{l-1}}^{min(v_{h,n_h}, t_l)} f_{1k}(t, \eta_{h,1k}(t)) dt + P_{12}(y_{h,12}, y_{h,13}) P_{2k}(y_{h,23}) \times \int_{t_{l-1}}^{min(v_{h,n_h}, t_l)} \int_0^t f_{12}(u, \eta_{h,12}(u)) f_{2k}(t-u, \eta_{h,2k}(t-u)) du dt \quad (7.2)$$

où $min(v_{h,n_h}, t_l)$ est le minimum entre l'éventuel temps de censure à droite pour le sujet h (v_{h,n_h}) et la borne de l'intervalle (t_l). $R(t_{l-1})$ dénote la somme sur tous les individus h qui n'ont pas été censurés avant le temps t_{l-1}.

L'effectif de transitions observées est simplement égal au nombre d'événements k se produisant dans l'intervalle :

$$n_{l,k} = \sum_{R(t_{l-1})} I_{\{v_{h,n_h} \leq t_l \text{ et } X_{h,n_h}=k\}}$$

où $I_{\{a\}} = 1$ si la condition a est respectée et 0 sinon.

7.2.3 Bootstrap semi-paramétrique

Comme nous l'avons vu dans l'introduction de ce chapitre, cette méthode permet d'estimer la distribution de la statistique (7.1) sans formuler d'hypothèses asymptotiques. Ce calcul se déroule en plusieurs étapes.

(i) Génération de B échantillons de bootstrap chacun de taille n. Chaque individu h^* ($h^* = 1, ..., n$) est observé aux temps de visites depuis la greffe $\{v^*_{h^*,0}, v^*_{h^*,1}, ..., v^*_{h^*,n_{h^*}}\}$. Ces temps de visites ne sont donc pas simulés, mais ils correspondent aux temps observés dans l'échantillon initial. Ceci justifie le terme semi-paramétrique.

(ii) Simulation de la trajectoire de chaque individu h^* sous l'hypothèse nulle selon laquelle son évolution obéit au modèle semi-markovien estimé initialement. Tous les individus entrent dans l'état 1 : $X^*_{h^*,0} = 1$. A partir des coefficients de régression associés à la chaîne de Markov, $\{\hat{\beta}_{12}, \hat{\beta}_{13}\}$, le second état, $X^*_{h^*,1}$, est simulé à l'aide d'une loi multinomiale de paramètres $\{P_{1j}(\psi_{h^*,12}, \psi_{h^*,13}), j = 2, 3, 4\}$. Si $X^*_{h^*,1} = 2$, alors la valeur de $X^*_{h^*,2}$ est simulée à partir d'une loi binomiale de paramètres $\{P_{2j}(\psi_{h^*,23}), j = 3, 4\}$.

(iii) Simulation des temps d'attente dans les états sous l'hypothèse nulle. A partir des paramètres estimés des lois d'attente dans les états, la durée dans l'état 1 pour l'individu h, $D^*_{h^*,0}$, suit une loi de densité $f_{1X^*_{h^*,1}}$. La fonction de répartition correspondante est notée $F_{1X^*_{h^*,1}}$. Rappelons que $F_{1X^*_{h^*,1}}$ est continue et strictement croissante sur $\mathbb{R}$ et qu'elle a pour limites 0 en $-\infty$ et 1 en $+\infty$. Dans ce cas particulier, $F_{1X^*_{h^*,1}}$ réalise une bijection de $\mathbb{R}$ sur $]0,1[$ et admet donc une fonction inverse $F^{-1}_{1X^*_{h^*,1}} :]0,1[\rightarrow \mathbb{R}$. Si U est une variable aléatoire de loi uniforme sur $]0,1[$, alors $Y^*_{h^*,0} = F^{-1}_{1X^*_{h^*,1}}(U)$ a même loi que $D^*_{h^*,0}$. On le vérifie facilement en calculant la fonction de répartition de $Y^*_{h^*,0}$ [104] :

$$P(Y^*_{h^*,0} \leq d^*_{h^*,0}) = P(F^{-1}_{1X^*_{h^*,1}}(U) \leq d^*_{h^*,0})) = P(U \leq F_{1X^*_{h^*,1}}(d^*_{h^*,0})) = F_{1X^*_{h^*,1}}(d^*_{h^*,0}) \quad (7.3)$$

Cette propriété (7.3) permet donc, à partir d'un générateur aléatoire fournissant des réalisations d'une variable uniforme, de simuler une variable aléatoire de même loi que $D^*_{h^*,0}$, en admettant que l'on sache calculer $F^{-1}_{1X^*_{h^*,1}}$.

Si la transition de l'état 1 à l'état $X^*_{h^*,1}$ respecte l'hypothèse de proportionnalité des risques (voir équation 2.22), la fonction de répartition inverse d'une loi de Weibull généralisée pour l'individu h^* est égale à :

$$F^{-1}_{1X^*_{h^*,1}}(u) = \sigma_{1X^*_{h^*,1}}\Big(\big(1-(exp(\eta_{h^*,1X^*_{h^*,1}})log(1-u)\big)^{\theta_{1X^*_{h^*,1}}}\big)-1\Big)^{(1/\nu_{1X^*_{h^*,1}})} \tag{7.4}$$

Comme U est une variable aléatoire de loi uniforme sur $]0,1[$, on peut exploiter le fait que $1-U$ a même loi que U, alors la fonction (7.4) s'écrit :

$$F^{-1}_{1X^*_{h^*,1}}(u) = \sigma_{1X^*_{h^*,1}}\Big(\big(1-(exp(\eta_{h^*,1X^*_{h^*,1}})log(u)\big)^{\theta_{1X^*_{h^*,1}}}\big)-1\Big)^{(1/\nu_{1X^*_{h^*,1}})} \tag{7.5}$$

Néanmoins, si la transition de l'état 1 à l'état $X^*_{h^*,1}$ ne respecte pas l'hypothèse de proportionnalité des risques, le prédicteur linéaire $\eta_{h^*,1X^*_{h^*,1}}$ devient dépendant du temps. En reprenant la formulation (6.20), la fonction de répartition s'écrit sous la forme de l'intégrale :

$$\begin{aligned} F_{1X^*_{h^*,1}}(x) &= 1-exp\Big(-\int_0^x \theta^{-1}_{1X^*_{h^*,1}}\Big(1+(t\sigma^{-1}_{1X^*_{h^*,1}})^{\nu_{1X^*_{h^*,1}}}\Big)^{1/\theta_{1X^*_{h^*,1}}-1} \\ &\times \nu_{1X^*_{h^*,1}}\sigma^{-1}_{1X^*_{h^*,1}}\Big(t\sigma^{-1}_{1X^*_{h^*,1}}\Big)^{\nu_{1X^*_{h^*,1}}-1} exp(\eta_{h^*,1X^*_{h^*,1}}(t))dt\Big) \end{aligned} \tag{7.6}$$

La fonction inverse de répartition correspondante à (7.6) n'est pas calculable explicitement. Une solution est de l'estimer non-paramétriquement à partir des coordonnées $(x; F_{1X^*_{h^*,1}}(x))$, $\forall x \in \mathbb{R}$. Posons $\tilde{F}^{-1}_{1X^*_{h^*,1}}(u)$ cette fonction telle que :

$$x = \tilde{F}^{-1}_{1X^*_{h^*,1}}(F_{1X^*_{h^*,1}}(x)) + \epsilon \tag{7.7}$$

La méthode de régression par B-splines, définie dans le chapitre 6, est utilisée pour l'estimation de $\tilde{F}^{-1}_{1X^*_{h^*,1}}$. La base de fonction est choisie cubique et le degré est fixé à 10. La flexibilité de la régression, ainsi obtenue, permet une estimation très précise de la fonction.

De la même manière, pour les individus dont le second état observé de bootstrap est égal à 2 ($X^*_{h^*,1} = 2$), il est possible de simuler les temps d'attente à partir des fonctions de répartition $F_{2X^*_{h^*,2}}$ et d'un générateur aléatoire fournissant des réalisations d'une variable aléatoire uniforme.

(iv) Estimation du modèle à partir de l'échantillon de bootstrap. Pour chaque individu h^* de l'échantillon de bootstrap, on dispose, grâce aux simulations précédentes, de la séquence d'états $X^*_{h^*,r}$, $(r = 1, ..., m^*_{h^*})$, où $m^*_{h^*}$ indique le nombre de transitions simulées chez le sujet h^*, ainsi que des temps d'attente correspondant $d^*_{h^*,r}$. On dispose parallèlement des temps de visites $\{v^*_{h^*,0}, v^*_{h^*,1}, ..., v^*_{h^*,n_{h^*}}\}$. A partir de ces informations, les contributions sont obtenues comme suit.

Si $d^*_{h^*,0} > v^*_{h^*,n_{h^*}}$ alors l'individu h^* est censuré dans l'état 1, dont la date de dernières nouvelles correspond à $v^*_{h^*,n_{h^*}}$. La contribution correspond alors à l'expression (6.15) :

$$C^*_{h^*,3} = S_{1.}(v^*_{h^*,n_{h^*}}) \tag{7.8}$$

Sinon si $v^*_{h^*,r-1} < d^*_{h^*,0} \leq v^*_{h^*,r}$ $(r = 1, ..., n_{h^*})$ et $X^*_{h^*,1} = k$ $(k = 3, 4)$, alors l'individu h^* n'est pas observé dans l'état 2 avant de déclarer un événement terminal k, la date d'apparition de ce dernier étant exactement connue $(d^*_{h^*,0})$. La contribution est alors équivalente à (6.14) :

$$C^*_{h^*,2} = P_{1k} f_{1k}(d^*_{h^*,0}) + P_{12} P_{2k} \int_{v^*_{h^*,r-1}}^{d^*_{h^*,0}} f_{12}(u) f_{2k}(d^*_{h^*,0} - u) du \tag{7.9}$$

Sinon si $v^*_{h^*,r-1} < d^*_{h^*,0} \leq v^*_{h^*,r}$ $(r = 1, ..., n_{h^*})$, $X^*_{h^*,1} = 2$ et $d^*_{h^*,0} + d^*_{h^*,1} > v^*_{h^*,n_{h^*}}$, alors l'individu h^* est censuré dans l'état 2. Parallèlement à l'expression (6.16), on a :

$$C^*_{h^*,4} = P_{12} \int_{v^*_{h^*,r-1}}^{v^*_{h^*,r}} f_{12}(u) S_{2.}(v^*_{h^*,n_{h^*}} - u) du \tag{7.10}$$

Sinon si $v^*_{h^*,r-1} < d^*_{h^*,0} \leq v^*_{h^*,r}$ $(r = 1, ..., n_{h^*})$, $X^*_{h^*,1} = 2$ et $d^*_{h^*,0} + d^*_{h^*,1} \leq v^*_{h^*,n_{h^*}}$, alors l'individu h^* est observé dans l'état 2 avant de déclarer un événement terminal k au temps $(d^*_{h^*,0} + d^*_{h^*,1})$ depuis la transplantation. A partir de la contribution (6.11), on obtient pour cette trajectoire :

$$C^*_{h^*,1} = P_{12} P_{2k} \int_{v^*_{h^*,r-1}}^{v^*_{h^*,r}} f_{12}(u) f_{2k}(d^*_{h^*,0} + d^*_{h^*,1} - u) du \tag{7.11}$$

A partir de ces contributions individuelles et de la fonction de vraisemblance (6.17), le modèle correspondant à l'échantillon de bootstrap peut alors être estimé.

(v) Calcul de la p-value du test. Pour chaque couple composé de l'échantillon de bootstrap et du modèle semi-markovien estimé à partir de ce dernier, la statistique de test G^*_b $(b = 1, ..., B)$ est calculée à partir des expressions (7.1) et (7.2). La p-value, c'est à dire la probabilité d'erreur de rejeter l'hypothèse nulle alors que cette dernière est vraie, est alors facilement calculable :

$$p = B^{-1} \sum_{b=1}^{B} \mathbb{1}_{\{G^*_b \geq G\}} \tag{7.12}$$

où $\mathbb{1}_{\{G^*_b \geq G\}}$ vaut 1 si la statistique G^*_b est supérieure ou égale à G, la statistique calculée à partir de l'échantillon initial, et 0 sinon.

7.2.4 Résultats

Le tableau (7.1) présente les transitions observées et théoriques associées au modèle estimé dans le chapitre 6 et résumé dans les tableaux (6.4), (6.5) et (6.6). Les deux premières lignes correspondent aux nombres de retours en dialyse et de décès entre 0,011 et 0,689 années après la transplantation. Pour ces deux types d'événements, les cellules

Temps chronologique (en années)		Transition $e \to 3$	Transition $e \to 4$	Pourcentage $e \to 3$	Pourcentage $e \to 4$
$]0,011; 0,689]$	Observé	12	8	5,19%	10,12%
	Attendu	9,38	5,25		
$]0,689; 2,168]$	Observé	13	8	**21,21%**	2,89%
	Attendu	20,91	10,02		
$]2,168; 3,826]$	Observé	16	5	15,73%	11,73%
	Attendu	23,17	8,81		
$]3,826; 5,213]$	Observé	17	4	2,14%	4,45%
	Attendu	14,87	6,18		
$]5,213; 9,158]$	Observé	14	7	**25,31%**	0,24%
	Attendu	23,08	7,51		

TAB. 7.1 – Tableau de contingence des transitions attendues et observées vers un état final

contribuent respectivement à 5,19% et 10,12% au calcul de la statistique, avec $G = 14,12$. Les deux cellules en gras contribuent à hauteur de 46,52% et correspondent aux retours en dialyse. L'hypothèse de stationnarité semble donc plus adaptée à la prédiction des décès. Pour des raisons de temps de calcul, 400 échantillons de bootstrap ont été simulés. 159 statistiques de bootstrap sont supérieures ou égales à G, correspondant à une p-value de 0,3975. Globalement et en considérant un risque de première espèce de 5%, le modèle semi-markovien estimé semble adéquat aux données de transplantations rénales. Quelques quantiles de cette fonction de probabilité cumulée de la statistique sont présentés dans le tableau (7.2).

Probabilités	0,75	0,50	0,25	0,10	0,05	0,01
Quantiles	9,79	12,77	15,26	18,27	20,95	27,37

TAB. 7.2 – Quantiles de la distribution de bootstrap de la statistique de test

7.2.5 Discussion

Dans les chapitres précédents, seuls les tests de Wald et LRS étaient utilisés pour tester l'adéquation du modèle concernant les lois des temps d'attente et les coefficients de régression associés aux facteurs explicatifs. Nous avons considéré ici un test d'adéquation complémentaire permettant d'examiner la stationnarité du modèle semi-markovien.

La statistique de test est de type Pearson. Le tableau de contingence regroupe les événements observés selon des intervalles de temps chronologique, c'est à dire le temps écoulé depuis la greffe. Le modèle présenté dans le chapitre 6 est examiné et semble

adéquat aux données de transplantation.

La limite principale de cette statistique est qu'elle ne considère que les événements terminaux, la prise en compte de la transition de l'état 1 à l'état 2 n'est considérée que indirectement dans le produit de convolution entre densités (équation 7.2).

A ce niveau des développements, deux échelles de temps ont été considérées : les durées dans les états et le temps depuis la transplantation. Le temps calendaire, autrement dit l'année de la greffe dans notre application, avait été modélisé dans les chapitres 4 et 5 en tant que facteur explicatif des vitesses de transition. Son effet était principalement expliqué par des biais périodes (évolution de la prise en charge thérapeutique, changements de profils des receveurs, amélioration de la qualité de la base de données, etc.). Cette solution était donc peu satisfaisante puisque l'interprétation de l'effet associé à la période de greffe est difficile. Pour limiter ce biais, nous avons considéré le sous-groupe des patients transplantés à partir de 1996, date à partir de laquelle le Cellecept fait son apparition et la créatinine est mieux renseignée. Le problème majeur de cette approche, outre la perte de puissance, est de ne pas considérer l'effet continu de l'année de la greffe. En effet, même si 1996 constitue une année charnière, l'évolution de la prise en charge n'est pas discontinue.

Pour conserver l'année de la transplantation, sans avoir à modéliser l'année de greffe en effet fixe (nombre de paramètres à estimer trop important), nous proposons dans le chapitre suivant de prendre en compte l'hétérogénéité due au temps calendaire en introduisant des termes aléatoires multiplicatifs des fonctions de risque de base.

Chapitre 8

Modélisation de l'effet période

Pour prendre en compte le biais période dans le modèle, nous avons choisi l'introduction de l'année de la transplantation comme effet aléatoire. Ce choix se justifie par le nombre de modalités trop important pour prendre en compte ce facteur comme un effet fixe et par son interprétation difficile (changement du profil des donneurs et des receveurs, modification du recueil des données, prise en charge évolutive, etc.).

Ce type de modèle a déjà été abordé dans le troisième chapitre. Pour cette application sur le VIH, la vraisemblance était construite de telle manière que l'individu statistique était la transition entre deux états. Plusieurs transitions pouvant être observées par patient, ce dernier était défini comme le groupe de corrélation [105].

L'application qui nous intéresse maintenant est sensiblement différente. En effet, la prise en compte de la censure par intervalle ne permet pas d'obtenir une forme de vraisemblance rendant possible l'utilisation des transformées de Laplace.

8.1 Méthodes

8.1.1 Définition du modèle

Les notations définies dans le chapitre 6 sont adaptées à ce nouveau contexte. Soit un échantillon constitué de A années de greffe, indicées par a ($a = 1, ..., A$). Considérons n_a sujets greffés à l'année a, indicés par h ($h = 1, ..., n_a$). Ainsi, $\{X_{ah,r}, r = 0, ..., m_{ah}\}$ représente la séquence d'états observés et distincts chez le h*ième* sujet de l'année a. m_{ah} indique donc son nombre de transitions observées. De plus, notons $d_{ah,r}$ le temps qu'il a passé dans l'état $X_{ah,r}$, c'est-à-dire l'état suivant sa r*ième* transition. Posons aussi t_{ah}, le temps écoulé entre la transplantation et la fin du suivi. Rappelons qu'il peut correspondre à un événement terminal ou à la date de dernières nouvelles.

En reprenant le principe d'une fragilité qui soit d'une part multiplicative de la fonction de risque de base et d'autre part spécifique à chaque transition, notée $\omega_{a,ij}$ pour l'année a et pour le transition de l'état i vers l'état j, l'expression (6.20) devient :

$$\lambda_{ij}\big(x, \eta_{ah,ij}(x), \omega_{a,ij}\big) = \omega_{a,ij}\lambda_{0,ij}(x)exp\big(\eta_{ah,ij}(x)\big) \text{ avec } \omega_{a,ij} > 0 \text{ et } x \geq 0 \qquad (8.1)$$

où $\eta_{ah,ij}(x)$ est le prédicteur linéaire des covariables dont les coefficients de régression peuvent dépendre du temps passé dans l'état, x, pour le h*ième* sujet de l'année a. La fonction de survie correspondante à (8.1) s'écrit :

$$S_{ij}\big(x, \eta_{ah,ij}(x), \omega_{a,ij}\big) = exp\Big(-\omega_{a,ij}\int_0^x \lambda_{0,ij}(u)exp\big(\eta_{ah,ij}(u)\big)\big)du\Big) \qquad (8.2)$$

La fonction de densité, $f_{ij}\big(x, \eta_{ah,ij}(x), \omega_{a,ij}\big)$, est égale au produit des fonctions (8.1) et (8.2).

Les différentes contributions individuelles à la vraisemblance peuvent alors être définies directement en reprenant celles définies dans le chapitre 6. Pour ne pas alourdir les notations, les covariables ne sont pas considérées dans ces fonctions.

(i) Le h*ième* individu de l'année a transite de l'état 1 à l'état 2 entre les temps $d^0_{ah,0}$ et $d^1_{ah,0}$ depuis la date de transplantation. De plus, il entre dans un état terminal k au temps t_{ah} $(k = 3, 4)$. Notons $C_{ah,1}(\omega_{a,ij}, ij = (12, 2k))$ la contribution à la vraisemblance d'un tel individu, avec $\delta_{ah,1}$ l'indicatrice égale à 1 si ce sujet respecte cette trajectoire.

$$C_{ah,1}(\omega_{a,ij}, ij = (12, 2k)) = P_{12}P_{2k}\int_{d^0_{ah,0}}^{d^1_{ah,0}} f_{12}(u, \omega_{a,12})f_{2k}(t_{ah} - u, \omega_{a,2k})du$$

(ii) Le h*ième* individu de l'année a reste dans l'état 1 jusqu'au temps $d^0_{ah,0}$ et déclare un événement terminal k $(k = 3, 4)$ au temps t_{ah}. Notons $C_{ah,2}(\omega_{a,ij}, ij = (1k, 12, 2k))$ la contribution à la vraisemblance d'un tel individu, avec $\delta_{ah,2}$ l'indicatrice correspondante, alors :

$$\begin{aligned} C_{ah,2}(\omega_{a,ij}, ij = (1k, 12, 2k)) &= P_{1k}f_{1k,\omega_{a,1k}}(t_{ah}, \omega_{a,1k}) \\ &+ P_{12}P_{2k}\int_{d^0_{ah,0}}^{t_{ah}} f_{12}(u, \omega_{a,12})f_{2k}(t_{ah} - u, \omega_{a,2k})du \end{aligned}$$

(iii) Le h*ième* individu de l'année a est censuré dans l'état 1, la date de dernières nouvelles correspond au temps t_{ah} depuis la greffe. Sa contribution $C_{ah,3}(\omega_{a,1j}, j \neq 1)$ est indiquée par $\delta_{ah,3}$.

$$C_{ah,3}(\omega_{a,1j}, j \neq 1) = \sum_{j \neq 1} P_{1j}S_{1j}(t_{ah}, \omega_{a,1j})$$

(iv) Le h*ième* individu de l'année a est censuré dans l'état 2, le temps de censure correspondant à la dernière visite t_{ah}. Le temps de transition de l'état 1 à l'état 2 est censuré

dans l'intervalle $[d^0_{ah,0}; d^1_{ah,0}]$. Sa contribution $C_{ah,4}(\omega_{a,ij}, ij = (12, 23, 24))$ est indiquée par $\delta_{ah,4}$.

$$C_{ah,4}(\omega_{a,ij}, ij = (12, 23, 24)) = \int_{d^0_{ah,0}}^{d^1_{ah,0}} P_{12} f_{12}(u, \omega_{a,12}) \Big\{ \sum_{j=3}^{4} P_{2j} S_{2j}(t_{ah} - u, \omega_{a,2j}) \Big\}$$

Conditionnellement aux effets aléatoires, ces contributions sont supposées indépendantes. Ainsi, la logvraisemblance conditionnelle est donnée par :

$$\begin{aligned} \mathcal{V}_{cond}(\omega_{a,ij}, \forall ij) &= \sum_{a=1}^{A} \sum_{h=1}^{n_a} \Big\{ \delta_{ah,1} log\Big(C_{ah,1}(\omega_{a,ij}, ij = (12, 2k)) \Big) \\ &+ \delta_{ah,2} log\Big(C_{ah,2}(\omega_{a,ij}, ij = (1k, 12, 2k)) \Big) \\ &+ \delta_{ah,3} log\Big(C_{ah,3}(\omega_{a,1j}, j \neq 1) \Big) \\ &+ \delta_{ah,4} log\Big(C_{ah,4}(\omega_{a,ij}, ij = (12, 23, 24)) \Big) \Big\} \end{aligned}$$

La logvraisemblance marginale est alors définie par :

$$\mathcal{V} = \int_0^\infty ... \int_0^\infty \mathcal{V}_{cond}(\omega_{a,ij}, \forall ij) g(\omega_{a,12}, ..., \omega_{a,24}) d\omega_{a,12} ... d\omega_{a,24} \tag{8.3}$$

où $g(\omega_{a,12}, ..., \omega_{a,23})$ est la fonction de densité jointe des effets aléatoires $\omega_{a,ij}$. Si on suppose ces fragilités indépendantes, la fonction (8.3) devient :

$$\mathcal{V} = \int_0^\infty ... \int_0^\infty \mathcal{V}_{cond}(\omega_{a,ij}, \forall ij) g(\omega_{a,12}) ... g(\omega_{a,24}) d\omega_{a,12} ... d\omega_{a,24} \tag{8.4}$$

8.2 Application

8.2.1 Stratégie d'analyse

Nous considérons comme modèle initial celui retenu dans le chapitre 6. Les estimations des paramètres sont présentées dans les tableaux (6.4), (6.5), et (6.6). La construction ascendante suivante est adoptée. Un effet aléatoire pour une seule transition est estimé. 5 modèles univariés sont calculés. L'information apportée par chaque fragilité est évaluée par un test de rapport de vraisemblance. Si plusieurs termes aléatoires sont retenus significatifs, ils sont inclus dans le modèle du plus au moins significatif.

En univarié, la fonction (8.4) est approchée en utilisant la quadrature de Gauss-Legendre. Etant donné les faibles valeurs prises par les fonctions de risque (tableau 6.6), les variances attendues des fragilités $\omega_{a,ij}$ seront aussi faibles. Pour une meilleure précision, l'intégration sur $\mathbb{R}^+$ de l'expression (8.4) est décomposée en deux parties grâce à

Transition	Estim.	$log(\mathcal{V})$	LRS *	p-value
$1 \rightarrow 2$	0,126	-1548,584	4,232	**0,0397**
$1 \rightarrow 3$	0,037	-1550,716	0,032	0,8580
$1 \rightarrow 4$	0,356	-1550,642	0,116	0,7334
$2 \rightarrow 3$	0,033	-1550,649	0,102	0,7494
$2 \rightarrow 4$	0,846	-1549,087	3,226	0,0724

* Logvraisemblance du modèle sans fragilité : -1550,70

TAB. 8.1 – Paramètres des lois Gamma pour chaque transition

la relation de Chasles. Par exemple, pour le modèle prenant uniquement en compte la variabilité annuelle sur la transition de l'état 1 à l'état 2 :

$$\begin{aligned}
\mathcal{V} &= \int_0^{\infty} \mathcal{V}_{cond}(\omega_{a,12}) g(\omega_{a,12}) d\omega_{a,12} \\
&= \int_0^{c} \mathcal{V}_{cond}(\omega_{a,12}) g(\omega_{a,12}) d\omega_{a,12} + \int_c^{\infty} \mathcal{V}_{cond}(\omega_{a,12}) g(\omega_{a,12}) d\omega_{a,12} \\
&\approx 0.5\Big\{ c \sum_{q=1}^{Q} \tilde{w}_q \mathcal{V}_{cond}\big(0,5c(u_q+1)\big) g\big(0,5c(\tilde{u}_q+1)\big) + (d-c) \\
&\times \sum_{q=1}^{Q} \tilde{w}_q \mathcal{V}_{cond}\big(0,5(d-c)\tilde{u}_q + 0,5(d+c)\big) g\big(0,5(d-c)\tilde{u}_q + 0,5(d+c)\big) \Big\} \quad (8.5)
\end{aligned}$$

Rappelons que $\tilde{w}_q$ et $\tilde{u}_q$ représentent respectivement les Q poids et les Q racines du $Qième$ polynôme de Legendre. Quelques analyses préliminaires ont montré qu'une bonne approximation était obtenue avec $c = 1$, $d = 5$ et $Q = 10$. Pour les mêmes arguments que dans la chapitre 3, les effets aléatoires sont supposés suivre une distribution Gamma à un paramètre dont la densité est donnée par (1.23).

8.2.2 Résultats

Les estimations univariées des paramètres de la loi Gamma pour chaque transition sont présentées dans le tableau (8.1). Seule la fonction de risque du temps d'attente dans l'état 1 avant d'entrer dans l'état 2 semble varier significativement selon l'année de la greffe ($p = 0,0397$). Ce modèle est donc retenu pour prendre en compte l'effet période. Les fonctions de risque des autres transitions ne semblent pas subir une telle variation.

Les autres paramètres du modèle sont présentés en annexe C. Ils sont en effet très proches de ceux obtenus dans le chapitre 6. La seule différence notable est que la fonction de risque propre à la transition $1 \rightarrow 2$ semble distribuée selon une loi de Weibull (tableau C.2). Le modèle final décrit dans le chapitre 6 était basé sur une loi Weibull généralisée pour cette transition.

8.3 Discussion

Deux échelles de temps ont été abordées dans les chapitres précédents : le temps d'attente dans l'état et le temps chronologique depuis la date d'origine. Les compléments, présentés dans ce chapitre 8, permettent de prendre en compte la troisième échelle, c'est-à-dire le temps du calendrier. La solution, consistant à modéliser l'année de greffe comme effet fixe, est peu satisfaisante. Nous avons choisi une méthode basée sur l'introduction de termes aléatoires.

Concernant l'estimation du modèle, une approche alternative aux transformées de Laplace est proposée. En effet, la forme plus complexe de la vraisemblance conditionnelle ne permet pas d'utiliser l'estimation formelle des termes de fragilité définie dans le chapitre 3. Nous proposons une approximation de la vraisemblance marginale par la méthode de Gauss-Legendre.

La limite principale de cette dernière est le temps de calcul associé à la maximisation d'une fonction telle que l'expression (8.5). Dans notre application, cette solution reste cependant acceptable puisque seul le terme aléatoire associé à la transition de l'état 1 à l'état 2 semble informatif. Dans l'hypothèse où plusieurs effets aléatoires auraient été retenus, l'optimisation d'une fonction basée sur des intégrales multiples serait certainement plus difficile.

Notons enfin que les résultats obtenus sont en accord avec la réalité des données. Il est assez logique d'observer un effet période associé à la transition entre l'état 1 et l'état 2. Ces deux états sont basés sur la clairance de la créatinine, alors que ce marqueur est inégalement renseigné depuis 1996. Parallèlement, la mortalité ou le taux de retour en dialyse n'ont pas subi un tel changement.

Chapitre 9

Définition des états de gravité et courbe ROC

Dans la première partie du chapitre 6, l'objectif était de définir le passage d'un état à faible risque d'échec à un état de mauvais pronostique. A partir de la CL mesurée tout au long du suivi, la question était de définir le seuil de décision le plus adéquat pour identifier cette transition. Nous proposons, dans ce chapitre, une méthode alternative à celle exposée précédemment, basée sur la maximisation de la vraisemblance.

L'objectif fixé est proche de la théorie relative aux tests diagnostiques faisant appel à des notions étroitement liées à la décision médicale. Quelques rappels sont donnés ici. Posons X un marqueur diagnostique continu défini sur $\mathbb{R}$. Supposons que des valeurs importantes soient plutôt en faveur de l'événement (maladie, décès, etc). Le test est dit positif si $X > c$, c étant le seuil recherché. Les erreurs issues de ce test peuvent être précisées en deux catégories : diagnostiquer l'événement à tort (faux positifs) et diagnostiquer l'absence d'événement à tort (faux négatifs). Les courbes ROC (Receiver Operator Characteristic) constituent des méthodes reconnues pour résumer l'exactitude de ce type de tests diagnostiques. Soit D l'indicatrice de l'événement. Lorsqu'il n'est pas temps-dépendant, on définit la sensibilité comme la capacité du test à bien détecter l'événement :

$$se(c) = P(X > c|D = 1) \tag{9.1}$$

Parrallèlement, la spécificité représente la capacité du test à bien détecter l'absence d'événement :

$$sp(c) = P(X \leq c|D = 0) \tag{9.2}$$

La courbe ROC correspondante est monotone croissante sur $[0, 1]$. Elle représente $P(X > c|D = 1)$ en fonction de $P(X > c|D = 0)$, $\forall c \in]-\infty, \infty[$. En médecine, cette courbe est intéressante pour plusieurs raisons. Tout d'abord, elle permet de décrire la qualité d'un marqueur diagnostique sans définir de seuil. En effet, plus l'aire sous la courbe (AUC) est importante, plus le marqueur possède de bonnes propriétés de discrimination. Dans

Valeur de l'AUC	Interprétation
AUC = 0,5	Aucune discrimination
0,7 ≤ AUC < 0,8	Discrimination acceptable
0,8 ≤ AUC < 0,9	Discrimination excellente
AUC ≥ 0,9	Discrimination exceptionnelle

TAB. 9.1 – Interprétation des aires sous la courbe

le contexte du diagnostic, Hosmer et Lemeshow [102] définissent les conclusions résumées dans le tableau (9.1). Ensuite, différents marqueurs peuvent être comparés, même si leurs échelles sont complètement différentes. Enfin, ces courbes ROC ne dépendent pas de la prévalence de l'événement, ce qui permet leur estimation à partir d'études cas-témoins.

Afin de définir le seuil de décision c, posons C_{FN} et C_{FP} les coûts respectifs d'un faux négatif et d'un faux positif. Alors, le coût total dû aux individus mal-classés et associé au seuil c, noté $C(c)$, s'écrit :

$$C(c) = n_{FN}(c)C_{FN} + n_{FP}(c)C_{FP}$$

où $n_{FN}(c)$ et $n_{FP}(c)$ sont les effectifs de faux négatifs et de faux positifs issus du test $X > c$. Si n est la taille de l'échantillon, alors :

$$n_{FN}(c) = nP(X \leq c|D = 1)P(D = 1) \text{ et } n_{FP}(c) = nP(X > c|D = 0)P(D = 0)$$

D'où

$$\begin{aligned} C(c) &= n\{P(X \leq c|D = 1)P(D = 1)C_{FN} + P(X > c|D = 0)P(D = 0)C_{FP}\} \\ &= n\{(1 - se(c))P(D = 1)C_{FN} + (1 - sp(c))(1 - P(D = 1))C_{FP}\} \end{aligned}$$

Le seuil optimal, noté $\hat{c}$, est celui qui minimise ce coût total :

$$\partial C(c)/\partial c = n\{-(\partial se(c)/\partial c)P(D = 1)C_{FN} - (\partial sp(c)/\partial c)(1 - P(D = 1))C_{FP}\}$$

Or, si f_1 et f_0 représentent, respectivement, les densités du marqueur X chez les malades et les non-malades, alors :

$$se(c) = P(X > c|D = 1) = \int_c^{\infty} f_1(x)dx \text{ et } sp(c) = P(X \leq c|D = 0) = \int_{-\infty}^{c} f_0(x)dx$$

Les dérivées de $se(c)$ et de $sp(c)$ par rapport à c sont respectivement égales à $-f_1(c)$ et $f_0(c)$. L'annulation de la dérivée précédente permet d'obtenir l'égalité suivante :

$$f_1(\hat{c})P(D = 1)C_{FN} = f_0(\hat{c})(1 - P(D = 1))C_{FP}$$

Finalement, en définissant le poids d'un faux positif par rapport à un faux négatif, avec $C_{FP} = kC_{FN}$, le seuil optimal respecte l'égalité suivante :

$$f_1(\hat{c})/f_0(\hat{c}) = k(1 - P(D = 1))/P(D = 1) \tag{9.3}$$

Dans notre contexte, le problème de ces définitions est de considérer l'événement D comme fixe au cours du temps. L'inégalité $X > c$ représente un test diagnostique. La censure des données n'est pas considérée et le seuil optimal $\hat{c}$ nécessite d'être estimé temps de pronostic. En se basant sur les travaux de Heagerty et al. [106], l'objet des développements qui vont suivre est l'adaptation de cette méthodologie à des événements temps-dépendants, notés $D(t)$. Nous parlerons plutôt de tests pronostiques, la question étant de savoir si la mesure du marqueur continu à un temps t_0 permet de pronostiquer le statut du patient à un temps t ($t_0 < t$). L'événement peut ainsi être censuré et le seuil optimal est considérer comme une fonction du temps de pronostic.

Les développements de Heagerty et al. sont basés sur un estimateur non-paramétrique de la survie sans covariable. Pour permettre l'ajustement sur certains facteurs, nous adaptons la méthode au cas multivarié en considérant un modèle paramétrique. De plus, l'approche est étendue pour prendre en compte l'information apportée par la répétition de la mesure du marqueur.

Ce chapitre est constitué d'une première section présentant le travail de Heagerty et al. [106]. La seconde section est consacrée aux généralisations de cette méthode à l'aide de modèles paramétriques. La section 3 est consacrée à l'application de ces développements aux données de transplantation. Une discussion de la méthode et des résultats clôture cette partie.

9.1 Méthodes non-paramétriques

9.1.1 Estimateur de Kaplan-Meier

Considérons un échantillon constitué de n individus. Posons T_i le délai entre la transplantation et l'échec de la greffe (retour en dialyse ou décès du patient) et x_i la valeur du marqueur pour le sujet i à l'origine ($i = 1, ..., n$). Posons C_i le temps de censure et $Z_i = min(T_i, C_i)$ le temps de suivi, avec $\delta_i = 1$ si l'échec se produit ($T_i \leq C_i$) et $\delta_i = 0$ si le suivi est censuré ($T_i > C_i$). Parallèlement aux processus de comptage, on note $D_i(t) = 1$ lorsque l'événement s'est produit avant le temps t ($T_i < t$). De la même manière, $D_i(t) = 0$ si $T_i \geq t$.

Les définitions de la sensibilité (9.1) et de la spécificité (9.2) sont adaptées à ce contexte temps-dépendant :

$$se(c,t) = P(X > c|D(t) = 1) \tag{9.4}$$

$$sp(c,t) = P(X \leq c|D(t) = 0) \tag{9.5}$$

La courbe ROC dépend du temps de pronostic. Le problème central est l'estimation des probabilités conditionnelles (9.4) et (9.5). Elles peuvent être réécrites à l'aide du théorème

de Bayes :

$$\begin{aligned} se(c,t) &= P(X > c|D(t) = 1) \\ &= P(X > c|T \leq t) \\ &= P(T \leq t|X > c)P(X > c)/P(T \leq t) \\ &= \{1 - S(t|X > c)\}P(X > c)/\{1 - S(t)\} \end{aligned}$$

$$\begin{aligned} sp(c,t) &= P(X \leq c|D(t) = 0) \\ &= P(X \leq c|T > t) \\ &= P(T > t|X \leq c)P(X \leq c)/P(T > t) \\ &= S(t|X \leq c)P(X \leq c)/S(t) \end{aligned}$$

où $S(t)$ est la fonction de survie telle que $S(t) = P(T > t)$ et $S(t|X > c)$ est la fonction de survie dans le groupe où le test est positif. Une méthode d'estimation largement répandue de $S(t)$ est celle de Kaplan et Meier [76]. Posons τ_n les temps uniques des événements, alors l'estimateur de Kaplan-Meier (KM) s'écrit :

$$\hat{S}_{KM}(t) = \prod_{s \in \tau_n, s \leq t} \Big\{1 - \{\sum_j \mathbb{1}\{z_j = s\}\delta_j\}/\{\sum_j \mathbb{1}\{z_j \geq s\}\}\Big\} \tag{9.6}$$

Il utilise ainsi toute l'information disponible, en incluant les données censurées. Des estimations simples de la sensibilité et de la spécificité sont alors obtenues à partir de l'estimateur de KM et de la fonction de répartition empirique du marqueur X :

$$\hat{se}_{KM}(c,t) = \{1 - \hat{S}_{KM}(t|X > c)\}\{1 - \hat{F}_X(c)\}/\{1 - \hat{S}_{KM}(t)\} \tag{9.7}$$

$$\hat{sp}_{KM}(c,t) = \hat{S}_{KM}(t|X \leq c)\hat{F}_X(c)/\hat{S}_{KM}(t) \tag{9.8}$$

où $\hat{F}_X(c) = n^{-1}\sum_i \mathbb{1}\{x_i \leq c\}$. Le premier problème avec ces estimations est qu'elles ne garantissent pas une monotonie de la sensibilité et de la spécificité. Par exemple pour la sensibilité, l'inégalité $P(X > c|D(t) = 1) \geq P(X > c'|D(t) = 1)$ si $c' > c$, devrait être respectée. Le second problème est qu'elles supposent que le processus de censure ne dépend pas de X. Cette hypothèse n'est pas valide si l'intensité du suivi dépend du pronostic initial.

La fonction de coût permet de définir un seuil du marqueur X optimal pour discriminer deux sous-groupes selon le risque d'échec. Le seuil choisi est celui qui minimise cette fonction. Soit $C(c,t)$ le coût total de la décision basée sur le seuil c pour prédire les événements jusqu'au temps t. Conservons les notations précédentes avec C_{FP} le coût d'un faux positif et C_{FN} le coût d'un faux négatif. Le coût total est alors égal à :

$$C(c,t) = C_{FP} \times n_{FP}(c,t) + C_{FN} \times n_{FN}(c,t) \tag{9.9}$$

où $n_{FP}(c,t)$ et $n_{FN}(c,t)$ sont respectivement le nombre de faux positifs et de faux négatifs au temps t issus du test basé sur le seuil c. Les valeurs des coûts sont du domaine de l'hypothèse (experts, littérature, etc.), la difficulté est donc d'estimer $n_{FP}(c,t)$ et $n_{FN}(c,t)$:

$$\begin{aligned} n_{FP}(c,t) &= nP(X > c, D(t) = 0) \\ &= nP(T > t|X > c)P(X > c) \\ &= n\hat{S}_{KM}(t|X > c)\{1 - \hat{F}_X(c)\} \end{aligned}$$

$$\begin{aligned} n_{FN}(c,t) &= nP(X \leq c, D(t) = 1) \\ &= nP(T \leq t|X \leq c)P(X \leq c) \\ &= n\{1 - \hat{S}_{KM}(t|X \leq c)\}\hat{F}_X(c) \end{aligned}$$

A partir de ces effectifs, la fonction du coût (9.9) peut alors être estimée par :

$$\hat{C}_{KM}(c,t) = n[C_{FP}\hat{S}_{KM}(t|X > c)\{1 - \hat{F}_X(c)\} + C_{FN}\{1 - \hat{S}_{KM}(t|X \leq c)\}\hat{F}_X(c)]$$

En posant $C_{FP} = kC_{FN}$ et puisque la proportionnalité de $C(c,t)$ est suffisante pour l'estimation de c :

$$\hat{C}_{KM}(c,t) \propto k\hat{S}_{KM}(t|X > c)\{1 - \hat{F}_X(c)\} + \{1 - \hat{S}_{KM}(t|X \leq c)\}\hat{F}_X(c) \tag{9.10}$$

9.1.2 Estimateur d'Akritas

Une solution plus valide pour l'estimation de la courbe ROC est obtenue à partir de la distribution bivariée $S(c,t) = P(X > c, T > t)$ fournie par Akritas [107] :

$$S(c,t) = \int_c^{\infty} S(t|X = x)dF_X(x) \tag{9.11}$$

où $F_X(x)$ est la fonction de répartition de X. Akritas propose d'approcher cette expression par :

$$\hat{S}_{\lambda_n}(c,t) = n^{-1}\sum_i^n \hat{S}_{\lambda_n}(t|X = x_i)\mathbb{1}\{x_i \leq c\}$$

où $\hat{S}_{\lambda_n}(t|X = x_i)$ est un estimateur de la fonction de survie conditionnelle, basé sur un noyau des plus proches voisins, dépendant du paramètre λ_n. Il s'écrit simplement comme un estimateur de KM pondéré :

$$\hat{S}_{\lambda_n}(t|X = x_i) = \prod_{s \in \tau_n, s \leq t} \left\{ 1 - \frac{\sum_j K_{\lambda_n}(X_j, X_i)\mathbb{1}\{z_j = s\}\delta_j}{\sum_j K_{\lambda_n}(X_j, X_i)\mathbb{1}\{z_j \geq s\}} \right\} \tag{9.12}$$

où $K_{\lambda_n}(X_j, X_i)$ est le noyau des plus proches voisins, le principe étant de choisir les patients éligibles, c'est-à-dire les patients j tels que la valeur de leur marqueur X_j soit

proche de la valeur X_i d'intérêt. Cet estimateur (9.12) est donc équivalent à celui de KM appliqué à un sous-groupe de l'échantillon. Le noyau est dit binaire puisqu'il permet simplement de sélectionner ou d'exclure les individus éligibles :

$$K_{\lambda_n}(X_j, X_i) = \mathbb{1}\{-\lambda_n < \hat{F}_X(x_i) - \hat{F}_X(x_j) < \lambda_n\}$$

$2\lambda_n \in [0;1]$ représente le pourcentage d'individus de l'échantillon à inclure comme éligibles (exception faite aux limites de la distribution de X). Les estimateurs de la sensibilité et de la spécificité sont alors facilement calculables. Premièrement, pour la sensibilité, à partir de la définition (9.4) et du théorème des probabilités totales, nous avons :

$$\begin{aligned} se(c,t) &= P(X > c | D(t) = 1) \\ &= P(T \leq t, X > c)/P(T \leq t) \\ &= \{P(X > c) - P(T > t, X > c)\}/P(T \leq t) \end{aligned}$$

d'où

$$\hat{se}_{\lambda_n}(c,t) = \{(1 - \hat{F}_X(c)) - \hat{S}_{\lambda_n}(c,t)\}/\{1 - \hat{S}_{\lambda_n}(t)\} \tag{9.13}$$

avec $\hat{S}_{\lambda_n}(t) = \hat{S}_{\lambda_n}(-\infty, t)$. Remarquons que le numérateur du rapport (9.13) est égal à $n^{-1}\sum_i^n \{1 - \hat{S}_{\lambda_n}(t|X = x_i)\}\mathbb{1}\{x_i \leq c\}$ et est donc monotone et décroissant quand c augmente. Deuxièmement, pour la spécificité :

$$\begin{aligned} sp(c,t) &= P(X \leq c | D(t) = 0) \\ &= P(X \leq c, T > t)/P(T > t) \\ &= \{P(T > t) - P(X > c, T > t)\}/P(T > t) \\ &= 1 - \{P(X > c, T > t)/P(T > t)\} \end{aligned}$$

d'où

$$\hat{sp}_{\lambda_n}(c,t) = 1 - \{\hat{S}_{\lambda_n}(c,t)/\hat{S}_{\lambda_n}(t)\} \tag{9.14}$$

Le numérateur de cette fonction (9.14) est là aussi monotone, ce qui rend ces estimateurs plus adéquats que ceux basés sur la simple méthode de KM. De plus, l'autre avantage de l'approche bivariée est de considérer le processus de censure comme dépendant du marqueur.

Concernant l'estimation de la fonction de coût (9.9), les effectifs de faux positifs et de faux négatifs peuvent être calculés comme suit :

$$n_{FP}(c,t) = nP(T > t, X > c) = n\hat{S}_{\lambda_n}(t,c)$$

$$\begin{aligned} n_{FN}(c,t) &= nP(T \leq t, X \leq c) \\ &= n\{P(X \leq c) - P(T > t, X \leq c)\} \\ &= n\{P(X \leq c) - P(T > t) + P(T > t, X > c)\} \\ &= n\{\hat{F}_X(c) - \hat{S}_{\lambda_n}(t) + \hat{S}_{\lambda_n}(t,c)\} \end{aligned}$$

La fonction à minimiser s'écrit :

$$\begin{aligned}\hat{C}_{\lambda_n}(c,t) &= C_{FP} \times n\hat{S}_{\lambda_n}(t,c) + C_{FN} \times n\{\hat{F}_X(c) - \hat{S}_{\lambda_n}(t) + \hat{S}_{\lambda_n}(t,c)\} \\ &\propto k\hat{S}_{\lambda_n}(t,c) + \{\hat{F}_X(c) - \hat{S}_{\lambda_n}(t) + \hat{S}_{\lambda_n}(t,c)\} \\ &\propto (k+1)\hat{S}_{\lambda_n}(t,c) + \hat{F}_X(c) \end{aligned} \tag{9.15}$$

9.2 Méthodes paramétriques

9.2.1 Une mesure du marqueur avec ajustement

Les deux méthodes qui viennent d'être abordées sont non-paramétriques, que ce soit l'estimation des fonctions de survie ou l'estimation des fonctions de répartition du marqueur. Les résultats de l'article de Heagerty et al. [106] montrent de très faibles différences entre les deux méthodes, même si celle basée sur la distribution bivariée reste plus adéquate (monotonie et dépendance du marqueur avec la censure).

Le principal avantage des méthodes non-paramétriques est de ne pas formuler d'hypothèse quant aux distributions des variables aléatoires T et X. Cependant, elles ne permettent pas l'introduction de covariables. Soit Y le vecteur des facteurs d'ajustement et

$$\lambda(t|X=x, Y=y) = \lambda_0(t)exp(\beta x + \gamma y + \alpha xy) \tag{9.16}$$

le modèle de régression du temps de survie où $exp(\beta)$ est le risque relatif associé au marqueur X et $exp(\gamma)$ est le vecteur des risques relatifs associés aux autres facteurs d'ajustement Y. D'éventuelles interactions entre le marqueur X et certaines covariables Y peuvent aussi être introduites. α représente les coefficients de régression associés à ces interactions. La fonction de survie associée à (9.16) s'écrit :

$$S(t|X=x, Y=y) = exp(-\int_0^t \lambda_0(u)exp(\beta x + \gamma y + \alpha xy)du) = S_0(t)^{exp(\beta x + \gamma y + \alpha xy)}$$

Le risque de base, $\lambda_0(t)$, est choisi distribué selon une loi Weibull généralisée (1.11). Parallèlement à la distribution bivariée (9.11), on définit :

$$S(c,t|Y=y) = P(X>c, T>t|Y=y) = \int_c^\infty S(t|X=x, Y=y)dF_{X|y}(x)$$

où $F_{X|y}(x)$ est la fonction de répartition du marqueur X conditionnellement aux facteurs d'ajustement Y. La nécessité de prendre en compte certaines covariables, pour analyser la distribution de X, justifie l'utilisation des modèles linéaires généralisés adaptés à de nombreuses configurations [64]. Nous verrons dans les applications que X, une transformée de la CL, sera supposée suivre une loi normale :

$$X \rightsquigarrow \mathcal{N}(\mu, \sigma)$$

avec $E[X] = \mu = \alpha_0 + \alpha y$. α_0 représente l'intercept et α est le vecteur des coefficients de régression associés aux covariables Y. Qu'il s'agisse des paramètres du modèle de survie ou de ceux de la fonction de répartition, les estimations sont obtenues par maximum de vraisemblance. Pour simplifier les développements qui vont suivre, nous ne distinguerons pas les paramètres de leur estimation. Dans le cadre multivarié, la sensibilité inclut un double conditionnement :

$$\begin{aligned} se(c,t,y) &= P(X > c | D(t) = 1, Y = y) \\ &= P(T \leq t, X > c | Y = y) / P(T \leq t | Y = y) \\ &= \{P(X > c | Y = y) - P(T > t, X > c | Y = y)\} / P(T \leq t | Y = y) \end{aligned}$$

d'où

$$\begin{aligned} se(c,t,y) &= \Big\{ \int_c^\infty f_{X|y}(x)dx - \int_c^\infty S(t|X = x, Y = y) f_{X|y}(x)dx \Big\} \\ &\times \Big\{ 1 - \int_{\mathbb{R}} S(t|X = x, Y = y) f_{X|y}(x)dx \Big\}^{-1} \\ &= \big(1 - S(t|Y = y)\big)^{-1} \int_c^\infty \{1 - S(t|X = x, Y = y)\} f_{X|y}(x)dx \\ &= F(t|Y = y)^{-1} \int_c^\infty F(t|X = x, Y = y) f_{X|y}(x)dx \end{aligned} \tag{9.17}$$

De la même manière, on pose la spécificité égale à :

$$\begin{aligned} sp(c,t,y) &= P(X \leq c | D(t) = 0, Y = y) \\ &= P(X \leq c, T > t | Y = y) / P(T > t | Y = y) \\ &= \{P(T > t | Y = y) - P(X > c, T > t | Y = y)\} / P(T > t | Y = y) \\ &= 1 - \{P(X > c, T > t | Y = y) / P(T > t | Y = y)\} \end{aligned}$$

d'où

$$sp(c,t,y) = 1 - S(t|Y = y)^{-1} \int_c^\infty S(t|X = x, Y = y) f_{X|y}(x)dx \tag{9.18}$$

Pour le calcul du seuil optimal, la fonction de coût (9.9) doit ici tenir compte des facteurs d'ajustement Y. Elle s'écrit alors :

$$C(c,t,y) = C_{FP} \times n_{FP}(c,t,y) + C_{FN} \times n_{FN}(c,t,y)$$

où $n_{FP}(c,t,y)$ et $n_{FN}(c,t,y)$ sont, respectivement, le nombre de faux positifs et de faux négatifs au temps t sachant $Y = y$. Formellement, si n_y représente l'effectif de patients à l'inclusion respectant le profil y :

$$n_{FP}(c,t,y) = n_y P(X > c, D(t) = 0 | Y = y) = n_y \int_c^\infty S(t|X = x, Y = y) f_{X|y}(x)dx$$

et

$$n_{FN}(c,t,y) = n_y P(X \leq c, D(t) = 1 | Y = y) = n_y \int_{-\infty}^c F(t|X = x, Y = y) f_{X|y}(x)dx$$

La fonction de coût à minimiser est donc finalement définie par :

$$C(c,t,y) \propto k \int_c^{\infty} S(t|X=x,Y=y) f_{X|y}(x)dx + \int_{-\infty}^{c} F(t|X=x,Y=y) f_{X|y}(x)dx \tag{9.19}$$

A l'inverse des méthodes non-paramétriques, le seuil optimal est calculé analytiquement :

$$\partial C(c,t,y)/\partial c = -kS(t|X=c,Y=y) f_{X|y}(c) + F(t|X=c,Y=y) f_{X|y}(c)$$

En annulant cette fonction, $\tilde{c}$ satisfait l'égalité suivante :

$$S(t|X=\tilde{c},Y=y) = 1/(1+k) \tag{9.20}$$

Remarquons que l'utilisation d'un modèle à risques proportionnels (équation 9.16) permet de calculer directement les seuils $\tilde{c}$. Si Λ_0 est la fonction de risque cumulée de base, alors :

$$\tilde{c} = \{log(log(1+k)) - log(\Lambda_0(t)) - \gamma y\}/\{\beta + \alpha y\} \tag{9.21}$$

9.2.2 Deux mesures du marqueur sans ajustement - Méthode simplifiée

Supposons maintenant, pour simplifier les développements, qu'aucune variable d'ajustement n'est présente. L'objectif est de préciser le test pronostique précédant avec plusieurs mesures du marqueur X au cours du temps.

Considérons deux mesures x_0 et x_1 aux temps t_0 et t_1. On pose $t_0 = 0$, x_0 représentant ainsi la mesure initiale du marqueur X. L'objectif est alors de définir les seuils c_0 et c_1, tels que si $x_0 > c_0$ et $x_1 > c_1$, le pronostic est en faveur d'un événement avant le temps t : $D(t) = 1$ $(t > t_1 > t_0)$. Dans cette section, tous les sujets sont supposés survivre jusqu'au temps t_1. Ainsi, la sensibilité est définie par :

$$se(c_0,c_1,t) = P(X_0 > c_0, X_1 > c_1 | D(t) = 1)$$

Cette probabilité est difficile à estimer dans le cadre non-paramétrique nécessitant des estimations en sous-groupes, tout en conservant des effectifs consistants. L'approche paramétrique se révèle plus appropriée :

$$\begin{aligned}
se(c_0,c_1,t) &= P(X_0 > c_0, X_1 > c_1 | T \leq t) \\
&= P(X_0 > c_0, X_1 > c_1, T \leq t)/P(T \leq t) \\
&= F(t)^{-1} \int_0^t \int_{c_0}^{\infty} \int_{c_1}^{\infty} f(u,x_0,x_1) du dx_0 dx_1 \\
&= F(t)^{-1} \int_{c_0}^{\infty} \int_{c_1}^{\infty} \Big(\int_0^t f(u|x_0,x_1) du \Big) f_{X_0,X_1}(x_0,x_1) dx_0 dx_1 \\
&= F(t)^{-1} \int_{c_0}^{\infty} \int_{c_1}^{\infty} F(t|x_0,x_1) f_{X_0,X_1}(x_0,x_1) dx_0 dx_1
\end{aligned} \tag{9.22}$$

L'écriture de la spécificité est plus complexe. Un test est défini positif si $X_0 > c_0$ et $X_1 > c_1$. Un test négatif correspond donc au complémentaire, tel que :

$$\begin{aligned} sp(c_0, c_1, t) &= P(X_0 \leq c_0, X_1 > c_1 | T > t) + P(X_0 > c_0, X_1 \leq c_1 | T > t) \\ &\quad + P(X_0 \leq c_0, X_1 \leq c_1 | T > t) \\ &= P(T > t)^{-1}\{P(T > t) - P(X_0 > c_0, X_1 > c_1, T > t)\} \\ &= 1 - S(t)^{-1}\Big\{\int_{c_0}^{\infty}\int_{c_1}^{\infty} S(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1\Big\} \end{aligned} \quad (9.23)$$

Dans les équations (9.22) et (9.23), les fonctions $F(t|x_0, x_1)$ et $S(t|x_0, x_1)$ sont déduites du modèle de survie ajusté sur les deux covariables X_0 et X_1. Seuls les individus ayant un temps de suivi supérieur à t_1 sont pris en compte. La difficulté consiste en la gestion de la distribution jointe de X_0 et X_1. Une solution serait d'utiliser le conditionnement tel que la densité $f_{X_0,X_1}(x_0, x_1)$ soit modélisée par $f_{X_1|x_0}(x_1|x_0) f_{X_0}(x_0)$. Cependant, ce type d'écriture est difficilement généralisable dans le cas où plus de deux mesures du marqueur sont réalisées. Considérons plutôt la distribution jointe de X_0 et de X_1. Dans les applications suivantes, le marqueur X sera supposé suivre une loi Normale, ainsi :

$$(X_0, X_1) \sim \mathcal{N}\left(\left|\begin{array}{c} \mu_0 \\ \mu_1 \end{array}\right|, \left|\begin{array}{cc} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{array}\right|\right)$$

où ρ est le coefficient de corrélation, avec $\rho \in [-1, 1]$. La densité s'écrit alors :

$$f_{X_0,X_1}(x_0, x_1) = (2\pi\sigma_0\sigma_1\sqrt{1-\rho^2})^{-1} exp(-0,5Q(x_0, x_1)) \quad (9.24)$$

où

$$Q(u, v) = \frac{1}{1-\rho^2}\Big\{\Big(\frac{x_0 - \mu_0}{\sigma_0}\Big)^2 + \Big(\frac{x_1 - \mu_1}{\sigma_1}\Big)^2 - 2\rho\frac{(x_0 - \mu_0)(x_1 - \mu_1)}{\sigma_0\sigma_1}\Big\}$$

La fonction coût, correspondant aux expressions (9.22) et (9.23), dépend des deux seuils c_0 et c_1, ainsi que du temps t. Si $n_{FP}(c_0, c_1, t)$ et $n_{FN}(c_0, c_1, t)$ sont les effectifs attendus de faux positifs et de faux négatifs au temps t, alors :

$$\begin{aligned} C(c_0, c_1, t) &= C_{FP} n_{FP}(c_0, c_1, t) + C_{FN} n_{FN}(c_0, c_1, t) \\ &= nC_{FP} P(T > t, X_0 > c_0, X_1 > c_1) + nC_{FN}\{P(T \leq t, X_0 \leq c_0, X_1 > c_1) \\ &+ P(T \leq t, X_0 > c_0, X_1 \leq c_1) + P(T \leq t, X_0 \leq c_0, X_1 \leq c_1)\} \\ &\propto C_{FP} P(T > t, X_0 > c_0, X_1 > c_1) \\ &+ C_{FN}\{P(T \leq t) - P(T \leq t, X_0 > c_0, X_1 > c_1)\} \end{aligned}$$

Sous la forme d'intégrales, on obtient :

$$\begin{aligned} C(c_0, c_1, t) &\propto C_{FP}\int_{c_0}^{\infty}\int_{c_1}^{\infty}\int_{t}^{\infty} f(t, x_0, x_1) dt dx_0 dx_1 \\ &+ C_{FN}\Big\{F(t) - \int_{c_0}^{\infty}\int_{c_1}^{\infty}\int_{0}^{t} f(t, x_0, x_1) dt dx_0 dx_1\Big\} \\ &\propto C_{FP}\int_{c_0}^{\infty}\int_{c_1}^{\infty} S(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \\ &+ C_{FN}\Big\{F(t) - \int_{c_0}^{\infty}\int_{c_1}^{\infty} F(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1\Big\} \end{aligned}$$

Toujours en considérant $C_{FP} = kC_{FN}$, on a alors :

$$\begin{aligned} C(c_0, c_1, t) &\propto k \int_{c_0}^{\infty} \int_{c_1}^{\infty} S(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \\ &+ F(t) - \int_{c_0}^{\infty} \int_{c_1}^{\infty} F(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \end{aligned} \tag{9.25}$$

Pour être complet, il nous faut considérer le cas où le test est positif si $X_0 > c_0$ ou si $X_1 > c_1$. Alors la sensibilité est égale à :

$$\begin{aligned} se(c_0, c_1, t) &= P(X_0 > c_0, X_1 > c_1 | T \leq t) + P(X_0 \leq c_0, X_1 > c_1 | T \leq t) \\ &+ P(X_0 > c_0, X_1 \leq c_1 | T \leq t) \\ &= \{P(T \leq t) - P(X_0 \leq c_0, X_1 \leq c_1, T \leq t)\}/P(T \leq t) \\ &= 1 - \{P(X_0 \leq c_0, X_1 \leq c_1, T \leq t)\}/P(T \leq t) \\ &= 1 - F(t)^{-1} \Big\{ \int_{-\infty}^{c_0} \int_{-\infty}^{c_1} F(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \Big\} \end{aligned} \tag{9.26}$$

Et la spécificité devient :

$$\begin{aligned} sp(c_0, c_1, t) &= P(X_0 \leq c_0, X_1 \leq c_1 | T > t) \\ &= P(X_0 \leq c_0, X_1 \leq c_1, T > t)/P(T > t) \\ &= S(t)^{-1} \Big\{ \int_{-\infty}^{c_0} \int_{-\infty}^{c_1} S(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \Big\} \end{aligned} \tag{9.27}$$

Remarquons que la sensibilité (9.22), dans le cas de la première règle de décision ($x_0 > c_0$ et $x_1 > c_1$), est strictement inférieure à la sensibilité définie en (9.26). Parallèlement, la spécificité (9.23) est strictement inférieure à l'expression (9.27). La première règle de décision est donc logiquement plus conservatrice. Autrement dit, pour c_0 et c_1 fixés, le nombre de faux positifs pour la première règle sera plus faible que pour la seconde. De la même manière, le nombre de faux négatifs sera plus important en utilisant la première règle de décision. En considérant que le test est positif si $X_0 > c_0$ ou si $X_1 > c_1$, la fonction de coût devient :

$$\begin{aligned} C(c_0, c_1, t) &= C_{FP} n_{FP}(c_0, c_1, t) + C_{FN} n_{FN}(c_0, c_1, t) \\ &= nC_{FP} \Big\{ P(T > t, X_0 > c_0, X_1 > c_1) + P(T > t, X_0 \leq c_0, X_1 > c_1) \\ &+ P(T > t, X_0 > c_0, X_1 \leq c_1) \Big\} + nC_{FN} P(T \leq t, X_0 \leq c_0, X_1 \leq c_1) \\ &\propto k\{P(T > t) - P(T > t, X_0 \leq c_0, X_1 \leq c_1)\} + P(T \leq t, X_0 \leq c_0, X_1 \leq c_1) \end{aligned}$$

D'où

$$\begin{aligned} C(c_0, c_1, t) &\propto k \Big\{ S(t) - \int_{-\infty}^{c_0} \int_{-\infty}^{c_1} S(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \Big\} \\ &+ \int_{-\infty}^{c_0} \int_{-\infty}^{c_1} F(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \end{aligned} \tag{9.28}$$

9.2.3 Deux mesures du marqueur sans ajustement - Méthode complète

La méthode simplifiée précédente utilise l'information apportée par deux mesures répétées d'un même marqueur. Pour simplifier les développements, nous avons considéré uniquement les individus dont le suivi est supérieur au temps t_1. Il est logique de considérer que l'information apportée par x_1 en terme de pronostic est pertinente uniquement pour les individus n'ayant toujours pas subi d'échec avant t_1. En revanche, la valeur x_0 devrait être informative pour tous les événements depuis l'inclusion. En considérant l'ensemble de l'échantillon et en choisissant la première règle de décision ($X_0 > c_0$ et $X_1 > c_1$), la sensibilité peut alors s'écrire :

$$\begin{aligned} se(c_0, c_1, t) &= P(X_0 > c_0, X_1 > c_1, T \leq t)/P(T \leq t) \\ &= \{P(X_0 > c_0, T \leq t_1) + P(X_0 > c_0, X_1 > c_1, t_1 < T \leq t)\} \\ &\times \{P(T \leq t_1) + P(t_1 < T \leq t)\}^{-1} \\ &= \Big\{ \int_{c_0}^{\infty} F(t_1|x_0) f_{X_0}(x_0) dx_0 + \int_{c_0}^{\infty} \int_{c_1}^{\infty} F(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \Big\} \\ &\times \Big\{ \int_{\mathbb{R}} F(t_1|x_0) f_{X_0}(x_0) dx_0 + \int_{\mathbb{R}} \int_{\mathbb{R}} F(t|x_0, x_1) f_{X_0,X_1}(x_0, x_1) dx_0 dx_1 \Big\}^{-1} \end{aligned} \tag{9.29}$$

La fonction de répartition, $F(t|x_0)$, est estimée sur tout l'échantillon, avec $t_0 = 0$, et $F(t|x_0, x_1)$ est estimée dans le sous-groupe des sujets encore en vie à t_1. La spécificité reste inchangée à l'expression (9.23), puisque le conditionnement porte sur $T > t$ alors que par définition $t_1 < t$. En effet, pour $t < t_1$, on retombe dans le contexte de la section 9.2.1. où seul X_0 est pris en compte. La fonction de coût correspondante est définie par :

$$\begin{aligned} C(c_0, c_1, t) &= C_{FP} n_{FP}(c_0, c_1, t) + C_{FN} n_{FN}(c_0, c_1, t) \\ &= nC_{FP} P(T > t, X_0 > c_0, X_1 > c_1) + nC_{FN}\{P(T \leq t, X_0 \leq c_0, X_1 > c_1) \\ &+ P(T \leq t, X_0 > c_0, X_1 \leq c_1) + P(T \leq t, X_0 \leq c_0, X_1 \leq c_1)\} \\ &\propto kP(T > t, X_0 > c_0, X_1 > c_1) + P(T \leq t) - P(T \leq t, X_0 > c_0, X_1 > c_1) \\ &\propto kP(T > t, X_0 > c_0, X_1 > c_1) + P(T \leq t_1) + P(t_1 < T \leq t) \\ &- P(t_1 < T \leq t, X_0 > c_0, X_1 > c_1) - P(T \leq t_1, X_0 > c_0, X_1 > c_1) \end{aligned}$$

A ce niveau du développement, la probabilité

$$P(T \leq t_1, X_0 > c_0, X_1 > c_1) \tag{9.30}$$

pose un problème puisque le marqueur X_1 est mesuré au temps t_1, alors que $T \leq t_1$. Cette difficulté peut être résolue en remarquant que la probabilité (9.30) est égale à

$$P(X_1 > c_1 | T \leq t_1, X_0 > c_0) \times P(T \leq t_1, X_0 > c_0) \tag{9.31}$$

Puisque l'événement a lieu avant t_1, le premier terme du produit (9.31) ne dépend pas de X_1. Autrement dit, soit c_1 tend vers $-\infty$, soit c_1 tend vers ∞. Or, la probabilité (9.30) n'est pas obligatoirement nulle, donc c_1 tend vers $-\infty$ et $P(X_1 > c_1 | T \leq t_1, X_0 > c_0)$ est égale à 1. On obtient que

$$P(T \leq t_1, X_0 > c_0, X_1 > c_1) = P(T \leq t_1, X_0 > c_0) \tag{9.32}$$

On peut alors reprendre le développement de la fonction de coût :

$$\begin{aligned} C(c_0, c_1, t) &\propto kP(T > t, X_0 > c_0, X_1 > c_1) + P(T \leq t_1) + P(t_1 < T \leq t) \\ &- P(t_1 < T \leq t, X_0 > c_0, X_1 > c_1) - P(T \leq t_1, X_0 > c_0) \\ &\propto kP(T > t, X_0 > c_0, X_1 > c_1) + P(t_1 < T \leq t) \\ &- P(t_1 < T \leq t, X_0 > c_0, X_1 > c_1) + P(T \leq t_1, X_0 \leq c_0) \end{aligned}$$

Sous la forme d'intégrales, on trouve finalement que :

$$\begin{aligned} C(c_0, c_1, t) &\propto k \int_{c_0}^{\infty} \int_{c_1}^{\infty} S(t|x_0, x_1) f_{X_0, X_1}(x_0, x_1) dx_0 dx_1 + \int_{-\infty}^{c_0} F(t_1|x_0) f_{X_0}(x_0) dx_0 \\ &+ \int_{\mathbb{R}} \int_{\mathbb{R}} F(t|x_0, x_1) f_{X_0, X_1}(x_0, x_1) dx_0 dx_1 \\ &- \int_{c_0}^{\infty} \int_{c_1}^{\infty} F(t|x_0, x_1) f_{X_0, X_1}(x_0, x_1) dx_0 dx_1 \end{aligned} \tag{9.33}$$

Complétons les développements en considérant le test positif si $X_0 > c_0$ ou si $X_1 > c_1$. La sensibilité devient alors :

$$\begin{aligned} se(c_0, c_1, t) &= \{P(X_0 > c_0, X_1 > c_1, T \leq t) + P(X_0 \leq c_0, X_1 > c_1, T \leq t) \\ &+ P(X_0 > c_0, X_1 \leq c_1, T \leq t)\} / P(T \leq t) \\ &= \{P(T \leq t) - P(X_0 \leq c_0, X_1 \leq c_1, T \leq t)\} / P(T \leq t) \\ &= 1 - \{P(X_0 \leq c_0, X_1 \leq c_1, T \leq t)\} / P(T \leq t) \\ &= 1 - \{P(X_0 \leq c_0, X_1 \leq c_1, t_1 < T \leq t) + P(X_0 \leq c_0, T \leq t_1)\} \\ &\times \{P(t_1 < T \leq t) + P(T \leq t_1)\}^{-1} \end{aligned}$$

On obtient alors :

$$\begin{aligned} se(c_0, c_1, t) &= 1 - \Big\{ \int_{-\infty}^{c_0} F(t_1|x_0) f_{X_0}(x_0) dx_0 \\ &+ \int_{-\infty}^{c_0} \int_{-\infty}^{c_1} F(t|x_0, x_1) f_{X_0, X_1}(x_0, x_1) dx_0 dx_1 \Big\} \\ &\times \Big\{ \int_{\mathbb{R}} F(t_1|x_0) f_{X_0}(x_0) dx_0 + \int_{\mathbb{R}} \int_{\mathbb{R}} F(t|x_0, x_1) f_{X_0, X_1}(x_0, x_1) dx_0 dx_1 \Big\} \end{aligned} \tag{9.34}$$

Comme pour la première règle, la spécificité (9.27) reste inchangée puisque le conditionnement porte sur $T > t$ alors que $t_1 < t$. La fonction de coût correspondante est égale

à :

$$\begin{aligned} C(c_0,c_1,t) &= C_{FP}n_{FP}(c_0,c_1,t) + C_{FN}n_{FN}(c_0,c_1,t) \\ &\propto k\{P(T>t,X_0>c_0,X_1>c_1) + P(T>t,X_0\leq c_0,X_1>c_1) \\ &+ P(T>t,X_0>c_0,X_1\leq c_1)\} + P(T\leq t,X_0\leq c_0,X_1\leq c_1) \\ &\propto k\{P(T>t) - P(T>t,X_0\leq c_0,X_1\leq c_1)\} \\ &+ P(X_0\leq c_0,T\leq t_1) + P(X_0\leq c_0,X_1\leq c_1,t_1<T\leq t) \end{aligned}$$

Et sous forme d'intégrales, on a :

$$\begin{aligned} C(c_0,c_1,t) &\propto k\Big\{\int_{\mathbb{R}}\int_{\mathbb{R}} S(t|x_0,x_1)f_{X_0,X_1}(x_0,x_1)dx_0dx_1 \\ &- \int_{-\infty}^{c_0}\int_{-\infty}^{c_1} S(t|x_0,x_1)f_{X_0,X_1}(x_0,x_1)dx_0dx_1\Big\} \\ &+ \int_{-\infty}^{c_0} F(t_1|x_0)f_{X_0}(x_0)dx_0 + \int_{-\infty}^{c_0}\int_{-\infty}^{c_1} F(t|x_0,x_1)f_{X_0,X_1}(x_0,x_1)dx_0dx_1 \end{aligned}$$

9.2.4 Marqueur comme variable dépendante du temps

Les développements précédents ont permis d'étudier le pouvoir pronostique d'un marqueur X, ce dernier étant au maximum mesuré deux fois au cours du temps. Or, notre objectif initial est de détecter à quel moment du suivi un patient peut être considéré dans un état de mauvais pronostic. Il est alors nécessaire d'étudier le marqueur comme un processus variant continuellement au cours du temps. Pour l'estimation du modèle de survie défini à partir de ce type de covariables, nous nous baserons sur les développements de Sparling et al. [108].

En restant cohérent avec les notations du chapitre 6, notons $\{v_{h,0}, v_{h,1}, ..., v_{h,n_h}\}$ le vecteur des n_h temps de visite où les valeurs du marqueur sont mesurées pour le sujet h $(h = 1, 2, ..., n)$. Le plus souvent, $v_{h,0} = 0$ représente la date d'origine de l'étude. Soit $x_{h,v_{h,r}}$ la r*ième* mesure du marqueur au temps $v_{h,r}$ $(r = 0, ..., n_h)$ de l'individu h. Notons $X_{h,v_{h,n_h}}$ la variable aléatoire correspondante. Parallèlement, posons $x_{h[t]}$ la séquence des mesures du marqueur jusqu'au temps t. Ainsi, $x_{h[v_{h,n_h}]}$ représente l'ensemble des valeurs observées du marqueur pour le sujet h. Notons enfin t_h, son temps de fin de suivi. Il correspond soit à un événement $(\delta_h = 1)$, soit à une censure à droite $(\delta_h = 0)$. La vraisemblance d'un tel échantillon est alors définie par :

$$\mathcal{V} = \prod_{h=1}^{n} f(t_h|x_{h[t_h]})^{\delta_h} S(t_h|x_{h[t_h]})^{1-\delta_h} \tag{9.35}$$

où $f(t_h|x_{h[t_h]})$ est la densité du temps de survie conditionnelle aux mesures du marqueur jusqu'en t_h. $S(t_h|x_{h[t_h]})$ est la fonction de survie correspondante. Les auteurs définissent

une forme générale de la fonction de risque en ajoutant un paramètre supplémentaire aux distributions Weibull et log-logistique :

$$\lambda(v_{h,v_{h,r}}|x_{h,v_{h,r}}) = \{\alpha\beta_{h,v_{h,r}}v_{h,r}^{\alpha-1}\}/\{(1+\beta_{h,v_{h,r}}v_{h,r}^{\alpha})^{\kappa}\} \tag{9.36}$$

où $\beta_{h,v_{h,r}} = exp(\theta+\gamma x_{h,v_{h,r}})$ et $\alpha > 0$. Par construction, $\beta_{h,v_{h,r}}$ est constant dans l'intervalle $[v_{h,r}; v_{h,r+1}[$ $(r = 0, 1, ..., n_h - 1)$.

Des valeurs spécifiques de α et de κ aboutissent à des distributions spécifiques. Si $\alpha = 1$ et $\kappa = 0$, alors une distribution Exponentielle est obtenue. Plus généralement, si $\kappa = 0$, la distribution est de type Weibull et le paramètre γ représente le logarithme du risque relatif pour une augmentation d'une unité de $x_{h,v_{h,r}}$. Le risque sera décroissant si $0 < \alpha \leq 1$, constant si $\alpha = 1$ et croissant si $\alpha > 1$. En sélectionnant $\kappa = 1$, on obtient une distribution log-logistique et γ représente le changement du logarithme du rapport des incidences cumulées. Ce risque décroît pour $0 < \alpha \leq 1$. Si $\alpha > 1$, le risque augmente jusqu'au temps $[(\alpha-1)/\beta]^{1/k}$ puis décroît vers 0. Pour $\alpha = 1,5$ et $\kappa = 0,5$, le risque augmente rapidement vers un plateau, puis décroît lentement. La fonction de risque (9.36) couvre donc un grand nombre de situations.

Posons $\mathbb{1}_{\{s\}}$ l'indicatrice égale à 1 si la condition s est respectée, et 0 sinon. Alors, la fonction de risque cumulée, au temps $t \in \{v_{h,r}, r = 0, ..., n_h\}$ pour le h*ième* sujet, est égale à :

$$\begin{aligned}\Lambda(t|x_{h[t]}) &= \sum_{r=0}^{n_h-1}\left[\mathbb{1}_{\{v_{h,r+1}=t\}}\int_{v_{h,r}}^{v_{h,r+1}}\{\alpha\beta_{h,v_{h,r}}u^{\alpha-1}\}/\{(1+\beta_{h,v_{h,r}}v_{h,r}^{\alpha})^{\kappa}\}du\right] \\ &= \sum_{r=0}^{n_h-1}\mathbb{1}_{\{v_{h,r+1}=t\}}[(1+\beta_{h,v_{h,r}}v_{h,r+1}^{\alpha})^{1-\kappa} - (1+\beta_{h,v_{h,r}}v_{h,r}^{\alpha})^{1-\kappa}]/[1-\kappa] \end{aligned} \tag{9.37}$$

Les covariables sont directement associées au temps de survie. Le modèle est donc de type accéléré (voir équation 1.17). La fonction (9.37) n'est pas définie pour $\kappa = 1$. Ce problème sera pris en compte dans l'optimisation de la logvraisemblance, définie à partir de l'expression (9.35) :

$$log(\mathcal{V}) = \sum_{h=1}^{n}\delta_h[log(\lambda(t_h|x_{h[t_h]}))] - \Lambda(t_h|x_{h[t_h]}) \tag{9.38}$$

Dans ce nouveau contexte, la sensibilité est définie comme la proportion d'individus pour lesquels le marqueur est supérieur au seuil c au temps d, parmi ceux ayant subi l'événement entre le temps d et t depuis l'inclusion $(d < t)$, :

$$\begin{aligned} se(c,t,d) &= P(X_d > c|D(t) = 1, D(d) = 0) \\ &= P(X_d > c|d < T \leq t) \end{aligned} \tag{9.39}$$

où X_d est la variable aléatoire associée au marqueur au temps d. Parallèlement, la sensibilité est égale à :

$$\begin{aligned} sp(c,t,d) &= P(X_d \leq c | D(t)=0, D(d)=0) \\ &= P(X_d \leq c | T > t) \end{aligned} \tag{9.40}$$

Il s'agit de la probabilité d'obtenir un test pronostique négatif au temps d sachant qu'aucun événement ne se produit jusqu'au temps t ($d < t$). En notant $f_{X|d}(x_d)$, la densité du marqueur X sachant que la mesure est réalisée au temps d depuis la greffe, alors les définitions (9.39) et (9.40) peuvent être développées comme suit. Premièrement, intéressons-nous à la sensibilité.

$$\begin{aligned} se(c,t,d) &= P(X_d > c, d < T \leq t)/P(d < T \leq t) \\ &= \int_c^\infty \int_d^t f(u,x_d) du dx_d \Big/ \int_{\mathbb{R}} \int_d^t f(u,x_d) du dx_d \\ &= \int_c^\infty \int_d^t f(u|x_d) f_{X|d}(x_d) du dx_d \Big/ \int_{\mathbb{R}} \int_d^t f(u|x_d) f_{X|d}(x_d) du dx_d \end{aligned}$$

En reprenant la définition de la fonction de risque cumulée (9.37), on a :

$$\begin{aligned} se(c,t,d) &= \Big\{ \int_c^\infty [exp(-\Lambda(d|x_d)) - exp(-\Lambda(t|x_d))] f_{X|d}(x_d) dx_d \Big\} \\ &\times \Big\{ \int_{\mathbb{R}} [exp(-\Lambda(d|x_d)) - exp(-\Lambda(t|x_d))] f_{X|d}(x_d) dx_d \Big\}^{-1} \\ &= \Big\{ \int_c^\infty [1 - exp(-\Lambda(t|x_d))] f_{X|d}(x_d) dx_d \Big\} \\ &\times \Big\{ \int_{\mathbb{R}} [1 - exp(-\Lambda(t|x_d))] f_{X|d}(x_d) dx_d \Big\}^{-1} \end{aligned} \tag{9.41}$$

où

$$\Lambda(t|x_d) = [(1 + exp(\theta + \gamma x_d) t^\alpha)^{1-\kappa} - (1 + exp(\theta + \gamma x_d) d^\alpha)^{1-\kappa}]/[1-\kappa]$$

Deuxièmement, appliquons des développements similaires à la spécificité.

$$\begin{aligned} sp(c,t,d) &= P(X_d \leq c, T > t)/P(T > t) \\ &= \int_{-\infty}^c \int_t^\infty f(u,x_d) du dx_d \Big/ \int_{\mathbb{R}} \int_t^\infty f(u,x_d) du dx_d \\ &= \int_{-\infty}^c \int_t^\infty f(u|x_d) f_{X|d}(x_d) du dx_d \Big/ \int_{\mathbb{R}} \int_t^\infty f(u|x_d) f_{X|d}(x_d) du dx_d \\ &= \int_{-\infty}^c exp(-\Lambda(t|x_d)) f_{X|d}(x_d) dx_d \Big/ \int_{\mathbb{R}} exp(-\Lambda(t|x_d)) f_{X|d}(x_d) dx_d \end{aligned} \tag{9.42}$$

Remarquons que dans le cas où $d = 0$, la sensibilité (9.41) et la spécificité (9.42) correspondent aux expressions (9.17) et (9.18). Pour calculer ces nouvelles quantités, il est nécessaire d'estimer la fonction $f_{X|d}(x_d)$, c'est à dire la distribution du marqueur X en

fonction du temps d depuis l'inclusion. Les données étant répétées sur un même individu, nous utiliserons un modèle linéaire marginal [109, 110, 111] pour approcher cette fonction :

$$x_{h,v_{h,r}} = \eta_0 + \eta_1 min(v_{h,r}, a) + \eta_2 (v_{h,r} - a) \mathbb{1}_{\{v_{h,r} > a\}} + \epsilon_{h,v_{h,r}} \tag{9.43}$$

avec $a = 6$ ans et $\epsilon = (\epsilon_{1,0}, \epsilon_{1,1}, ..., \epsilon_{1,n_1}, ..., \epsilon_{n,v_{n,nn}})$ les résidus du modèle. Cette régression linéaire à deux pentes, avant et après 6 mois, est justifiée dans l'article de Giral et al. [31]. Pour prendre en compte la dépendance des observations répétées sur un même sujet, on suppose que ϵ suit une loi normale centrée. La matrice de variance-covariance suppose que les mesures issues de deux individus différents possèdent une covariance nulle et que les termes diagonaux représentent la variance résiduelle, notée σ^2. Nous supposons l'intensité de la corrélation, entre les mesures issues d'un même sujet, fonction de leur proximité temporelle. Plus deux mesures sont rapprochées, plus leur covariance sera forte. Une structure de corrélation auto-régressive d'ordre 1 est utilisée :

$$cor(\epsilon_{h,r}, \epsilon_{h,r+k}) = \rho^k \quad k = 0, 1, ...$$

où k représente l'écart ordinal entre les deux mesures. Par exemple, si les deux mesures sont contiguës alors $k = 1$. ρ est le paramètre de corrélation. Ce modèle linéaire marginal est estimé par la méthode des moindres carrés généralisés (*gls* sous le logiciel R).

Comme pour les autres méthodes précédemment abordées, la question centrale reste l'estimation du seuil c. La fonction de coût est définie par :

$$C(c, t, d) = C_{FP} n_{FP}(c, t, d) + C_{FN} n_{FN}(c, t, d)$$

où $n_{FP}(c, t, d)$ et $n_{FN}(c, t, d)$ correspondent, respectivement, aux effectifs de faux positifs et de faux négatifs issus du test $X_d > c$, au temps d, pour prédire un événement entre d et t $(d < t)$. Ainsi, si n_d représente l'effectif de patients à risque au temps d :

$$\begin{aligned} C(c, t, d) &= n_d \{C_{FP} P(T > t, X_d > c) + C_{FN} P(d < T \leq t, X_d \leq c)\} \\ &\propto k \int_c^{\infty} \int_t^{\infty} f(u|x_d) f_{X|d}(x_d) du dx_d \\ &+ \int_{-\infty}^{c} \int_d^{t} f(u|x_d) f_{X|d}(x_d) du dx_d \\ &\propto k \int_c^{\infty} exp(-\Lambda(t|x_d)) f_{X|d}(x_d) dx_d \\ &+ \int_{-\infty}^{c} [1 - exp(-\Lambda(t|x_d))] f_{X|d}(x_d) dx_d \end{aligned} \tag{9.44}$$

Le seuil optimal, minimisant la fonction (9.44), peut être obtenu par dérivation :

$$\partial C(c, t, d)/\partial c = -k exp(-\Lambda(t|c)) f_{X|d}(c) + [1 - exp(-\Lambda(t| = c))] f_{X|d}(c)$$

En remplaçant les termes précédents par leur estimation, on obtient :

$$1/(1 + k) = exp(-\Lambda(t|\tilde{c}))$$

9.3 Pronostic d'un échec et clairance de la créatinine

9.3.1 Les données

Les données sont issues de la cohorte DIVAT. Pour éviter une population trop hétérogène, nous limitons l'inclusion aux critères suivants : patients majeurs, transplantations de janvier 1996 à septembre 2006 (gel de la base de données) et greffes de rein uniquement. L'échantillon est ainsi composé de 688 patients.

L'événement étudié est le temps écoulé entre la date de greffe et la date de retour en dialyse ou de décès. 82 individus décèdent ou retournent en dialyse, soit presque 12%. A la date de greffe, les patients sont en moyenne âgés de 47,8 ans (ET=14,8), alors que les donneurs sont âgés en moyenne de 45,8 ans (ET=16,3). 61,6% des receveurs sont des hommes, contre 63,7% chez les donneurs. Chaque patient est suivi régulièrement, que ce soit pour des visites programmées (périodicité d'un an obligatoire) ou pour des visites plus aléatoires. En moyenne, un patient est vu 48,5 fois, avec un délai moyen entre visites de 1,5 mois. La moitié des visites est espacée de moins de 20 jours. La moyenne de CL à un an est de 51,3 ml/min (ET=19,9).

9.3.2 Une mesure à un an - Aucun ajustement

La méthode simple de Kaplan-Meier (section 9.1.1.) et la méthode paramétrique (section 9.2.1.) ont été utilisées. Heagerty et al. [106] ont montré peu de différences entre les deux approches non-paramétriques.

L'augmentation de la valeur de CL est un facteur protecteur (alors que les développements précédents considèrent une augmentation du risque avec celle du marqueur). De plus, la distribution de CL est normalisée par une transformation racine carrée. Nous avons donc choisi d'étudier le marqueur $X = -\sqrt{CL}$, la CL étant mesurée un an après la greffe.

Le tableau (9.2) synthétise les estimations des fonctions $S(t|x)$ et $f_X(x)$. Le marqueur X suit une loi normale de moyenne -7,07 ml/min (ET = 1,24). La courbe ROC, ainsi obtenue (équations 9.17 et 9.18), est représentée par la figure (9.1) pour $c \in [-10,0; -3,0]$. Elle représente le pouvoir prédictif de la CL à un an pour des échecs se produisant dans les cinq ans après la greffe. Cette courbe est comparée avec celle obtenue par la méthode de KM (équations 9.7 et 9.8). L'estimation de la survie par KM nécessite de conserver un nombre d'événements consistant dans chacun des deux groupes définis par le seuil c. Les courbes ROC pour t petit sont alors difficilement estimables étant donné le faible nombre d'événements précoces. La courbe non-paramétrique de la figure (9.1) est ainsi calculée avec $c \in [-8,5; -3,5]$.

	Coef.	ET.	Test Wald	p-value
Modèle de survie : $\lambda(t\|X=x)=\lambda_0(t)exp(\beta x)$				
σ	0,27	0,16		
ν	1	.		
θ	1	.		
β	0,88	0,10	8,74	<0,0001
$f_X(x)\sim\mathcal{N}(\mu,\varphi),\ E[X]=\mu$				
μ	-7,07	0,03		
φ	1,24	0,02		

TAB. 9.2 – Modèles paramétriques pour la CL mesurée à un an (sans ajustement)

En considérant la méthode non-paramétrique comme référence, il semble que la méthode paramétrique sous-estime légèrement le pouvoir discriminant de CL à un an pour les échecs se produisant dans les 5 ans après la greffe. Cependant, les deux méthodes restent proches. Les valeurs des aires sous la courbe (AUC) sont de 0,78 pour la méthode paramétrique et de 0,80 pour la méthode non-paramétrique. En se référant aux critères de Hosmer et Lemeshow, ces AUC illustrent le très bon pouvoir pronostique de la CL à un an. Notons que les valeurs limites, présentées en introduction, sont définies dans le cadre de tests diagnostiques, où le *gold standard* est mesuré en même temps que le marqueur. Notre problématique est de type pronostique, ce qui renforce l'idée que ces AUC sont élevées. La CL à un an offre une excellente prédiction des échecs se produisant dans les 5 ans qui suivent la transplantation.

L'intérêt des développements réside aussi dans la définition du seuil c optimal. Pour cela, le coût d'un faux positif relativement au coût d'un faux négatif doit être défini. En transplantation, les marqueurs ont pour objectif premier d'alerter d'un éventuel échec de la greffe, même si cela sous-entend un nombre de faux positifs plus important. On préfère donc classer un patient dans le groupe à risque d'échec même si ce dernier ne subissait pas d'échec, plutôt que de le classer dans le groupe à faible risque à tort. Autrement dit, on a $C_{FN} > C_{FP}$ et donc $k < 1$. Toujours à 5 ans, en considérant un faux positif 5 fois moins important qu'un faux négatif ($k = 0,2$), le seuil optimal est de 25 ml/min. Il passe à 33 ml/min pour un poids égal à 10 ($k = 0,1$).

Notre approche temps-dépendante permet d'étudier le pouvoir pronostique d'un marqueur sur l'apparition d'un échec avant un temps t fixé. Le seuil optimal de décision n'est donc pas seulement fonction du poids entre faux positifs et faux négatifs. Comme le souligne l'égalité (9.21), ce seuil varie selon le temps du pronostic. Cet ensemble de seuils optimaux, selon le poids et le temps, est présenté dans la figure (9.2). Comme nous l'avons vu précédemment, le seuil optimal pour $t = 5$ ans et pour $k = 0,1$ est de 33,0

ml/min. Pour une inférence à 10 ans, ce seuil passe 44,5 ml/min. En terme plus pratique, si un patient possède une valeur de CL à un an inférieure à 44,5 ml/min, il sera considéré comme à risque d'échec avant son $10ième$ anniversaire de greffe. Cependant, si on s'intéresse aux échecs plus précoces, avant le $5ième$ anniversaire de greffe par exemple, cet individu sera considéré comme à risque d'échec si sa valeur de CL est inférieure à 33,0 ml/min. Plus ce réservoir de CL est important, plus le temps avant l'échec est lointain. Le test, basé sur 44,5 ml/min pour prédire un événement dans les 10 ans, possède une sensibilité égale à 0,76 et une spécificité égale à 0,66. Le test basé, sur 33,0 ml/min pour prédire un événement dans les 5 ans après la greffe, possède une sensibilité égale à 0,48 et une spécificité égale à 0,87.

9.3.3 Une mesure à un an - Ajustement sur l'incompatibilité

Comme nous venons de le voir, l'information apportée par CL à un an se révèle importante dans la prédiction de la réussite de la greffe. Or, certains facteurs peuvent venir modifier l'interprétation d'un tel marqueur. Considérons deux catégories de greffes selon le nombre d'incompatibilités HLA (A+B+DR), $y = 1$ si au moins 4 incompatibilités sont observées et $y = 0$ sinon. Le tableau (9.3) montre l'importance de ce facteur d'ajustement ($p = 0,0155$). Nous avons aussi retenu son interaction avec la CL à un an ($p = 0,0248$). Pour tenir compte de y dans l'estimation de la distribution de CL, nous supposons toujours que $-\sqrt{CL}$ suit une loi normale de moyenne μ et de variance φ, avec $\mu = \alpha_0 + \alpha_1 y$. La paramètre α_1 est égal à 0,18 et est significativement différent de 0 ($p = 0,0075$). Il semble donc que le nombre d'incompatibilités modifie à la fois l'effet de la CL sur le temps de survie et la valeur moyenne de la CL.

En considérant la sensibilité (9.17) et la spécificité (9.18), ajustées sur un tiers facteur Y, les courbes ROC correspondantes sont présentées dans la figure (9.3). Le pouvoir discriminant de CL à un an apparaît nettement supérieur pour les receveurs présentant au moins 4 incompatibilités. Sans ajustement, l'AUC était égale à 0,78. Elle augmente à 0,87 pour les greffons ayant au moins 4 incompatibilités et diminue à 0,74 pour les greffons ayant moins de 4 incompatibilités.

De même que le pouvoir de discrimination se trouve modifié, le seuil de décision est différent selon le nombre d'incompatibilités. En considérant un rapport de 10 entre la gravité d'un faux négatif et d'un faux positif, la figure (9.4) montre que le seuil de décision optimal reste égal à 33,0 ml/min pour les greffes présentant au moins 4 incompatibilités, alors qu'il approche les 35,3 ml/min pour les autres greffes. D'un point de vue clinique, cette distinction de seuils selon la catégorie d'incompatibilités n'est pas forcément informative.

A partir de l'équation (9.21), la figure (9.5) montre la variation du seuil de décision avec le temps. Pour prédire des événements très précoces, c'est-à-dire identifier une population très fragilisée, la tendance sera de choisir un seuil de CL faible. Par exemple, dans la

	Coef.	ET.	Test Wald	p-value
Modèle de survie [a] : $\lambda(t\|X=x,Y=y)=\lambda_0(t)exp(\beta x+\gamma y+\alpha xy)$				
σ	0,60	0,45		
ν	1	.		
θ	1	.		
β	0,73	0,12	6,15	<0,0001
γ	3,29	1,36	2,42	0,0155
α	0,53	0,24	2,24	0,0248
$f_{X\|y}(x)\sim\mathcal{N}(\mu,\varphi),\ E[X]=\mu=\alpha_0+\alpha_1 y$				
α_0	-7,15	0,04		
α_1	0,18	0,07	2,67	0,0075
φ	1,24	0,02		

[a] $y=1$ si le nombre d'incompatibilités est supérieur ou égal à 4 et $y=0$ sinon

TAB. 9.3 – Modèles paramétriques pour la CL mesurée à un an (avec ajustement)

population des transplantés ayant au moins 4 incompatibilités, le sous-groupe à faible risque d'échec dans les 10 ans suivant la greffe est constitué des patients dont la valeur de CL à un an est supérieure à 43,3 ml/min. Pour les autres, ce seuil est de 46,9 ml/min. Si la discrimination porte sur les événements se produisant dans les 5 ans après la greffe, ces seuils sont, respectivement, de 35,5 et 33,1 ml/min. Les décisions sont peu différentes entre les deux sous-groupes d'incompatibilités.

9.3.4 Deux mesures à 3 et 12 mois - Méthode simplifiée

A partir des résultats précédents, il semble que la CL à un an constitue un marqueur important pour le pronostic d'un échec. Les développements à venir consistent à évaluer si ce test pronostique est amélioré en prenant en compte deux mesures de CL. Pour que X_0 et X_1 soient deux facteurs de risque et pour que leur distribution jointe soit proche d'une loi Normale bivariée, le choix suivant a été réalisé : X_0 est l'opposé de la racine carrée de la CL à 3 mois et X_1 représente son évolution à 12 mois, de sorte que $x_0=-\sqrt{CL(3\,mois)}$ et $x_1=-\sqrt{CL(12\,mois)}-x_0$.

La méthode simplifiée suppose que les individus aient survécu jusqu'au temps de la seconde mesure, autrement dit t_1 constitue l'origine du temps de survie. Le test est dit positif si $X_0>c_0$ et $X_1>c_1$. Pour le calcul de la sensibilité (9.22) et de la spécificité (9.23), il est nécessaire d'estimer les fonctions f_{X_0,X_1} et $S(t|x_0,x_1)$, avec $t>t_1$. Celles-ci sont résumées dans le tableau (9.4). La CL moyenne évolue peu entre t_0 et t_1 ($\mu_1=0,02$). De plus l'évolution est peu corrélée avec la valeur à t_0 ($\rho=-0,35$). Les deux marqueurs,

x_0 et x_1, constituent bien deux facteurs de risque d'échec de la transplantation.

	Coef.	ET.	Test Wald	p-value
Modèle de survie : $\lambda(t\|x_0,x_1)=\lambda_0(t)exp(\beta_1x_0+\beta_2x_1)$, $(t>t_1)$				
σ	0,64	0,45		
ν	1	.		
θ	1	.		
β_1	0,75	0,11	6,77	<0,0001
β_2	1,04	0,12	8,85	<0,0001
Distribution du couple (X_0,X_1) *:* $f_{X_0,X_1}(x_0,x_1)$ *voir équation (9.24)*				
μ_0	-7,10	0,03		
σ_0	1,25	0,02		
μ_1	0,02	0,02		
σ_1	0,87	0,01		
ρ	-0,35	0,02		

TAB. 9.4 – Modèles paramétriques pour la CL répétée à 3 et 12 mois (sans ajustement)

De manière classique, le pouvoir pronostique d'un test basé sur un seul marqueur est représenté par une courbe ROC. Dans l'approche actuelle, deux marqueurs composent le test. Le pouvoir prédictif de ce test est alors représenté par une aire ROC. Plus la limite supérieure de cette aire est proche de l'angle supérieur gauche du repère, plus le pouvoir pronostique du test sera important. L'aire ROC est représentée par la figure (9.6). Les deux courbes ROC, pour les marqueurs x_0 et x_1, correspondent respectivement à l'approche bivariée pour $c_1 = inf(X_1) = -3$ et pour $c_0 = inf(X_0) = -12$. En effet, si c_0 est très petit, on a $P(X_0 > c_0, X_1 > c_1|D(t) = 1)$ qui tend vers $P(X_1 > c_1|D(t) = 1)$, d'où $se(c_0, c_1, t) \simeq se(c_1, t)$. De la même manière, on obtient $sp(c_0, c_1, t) \simeq sp(c_1, t)$. Ces courbes sont comprises dans la partie basse de l'aire ROC, montrant ainsi l'apport de l'analyse conjointe des deux marqueurs.

Les seuils de décision optimaux sont ceux qui minimisent la fonction de coût (9.25). Pour $k = 0,1$, le couple de seuils optimaux $(\tilde{c}_0, \tilde{c}_1)$ est égal à (-6,71, 0,17). Ainsi, on conclura qu'un individu est à risque d'échec avant son $5ième$ anniversaire de greffe, si x_0 est supérieur à -6,71 et si x_1 est supérieur à 0,17. Ces seuils de décisions correspondent à une sensibilité égale à 0,34, à une spécificité égale à 0,90 et à un coût de 0.0417. En termes plus médicaux, le test $X_0 > c_0$ et $X_1 > c_1$ est équivalent à tester si $CL(3mois) < min(c_0, (c_1 - \sqrt{CL(12\ mois)})^2)$. Cette région critique peut être explicitement définie dans une figure (voir méthode complète).

Pour être exhaustif, la seconde règle de décision doit être étudiée, le test est alors positif si $X_0 > c_0$ ou si $X_1 > c_1$. Les estimations (9.26) et (9.27) de la sensibilité et de

la spécificité sont alors utilisées. La minimisation de la fonction de coût (9.28) retourne $(\tilde{c}_0, \tilde{c}_1) = (-5,29, 1,37)$, correspondant à une sensibilité égale 0,34, à une spécificité égale à 0,88 et à un coût égal à 0,0442. Le coût relatif à la première règle de décision est plus faible. La première règle sera donc préférée par rapport à la seconde.

9.3.5 Deux mesures à 3 et 12 mois - Méthode complète

Les résultats de la section 3.3. supposent que tous les individus aient survécu jusqu'au temps t_1 de la seconde mesure de CL. L'information apportée par le couple (X_0, X_1) concerne donc le pronostic d'événements se produisant après t_1. La méthode complète permet de prendre en compte l'information apportée par X_0 pour tous les échecs pouvant se produire entre t_0 et t_1. Les estimations supplémentaires de $S(t|x_0)$ et de $f_{X_0}(x_0)$ sont nécessaires par rapport à la section précédente. Elles sont présentées dans le tableau (9.5).

	Coef.	ET.	Test Wald	p-value
Modèle de survie : $\lambda(t\|x_0) = \lambda_0(t)exp(\beta_1 x_0),\ (t > t_0)$				
σ	1,70	0,87		
ν	1	.		
θ	1	.		
β_1	0,55	0,08	7,00	<0,0001
$f_{X_0}(x_0) \sim \mathcal{N}(\mu, \varphi),\ E[X] = \mu$				
μ	-7,12	0,03		
φ	1,29	0,02		

TAB. 9.5 – Modèles paramétriques pour la CL répétée à 3 mois (sans ajustement)

Pour la première règle, l'aire ROC, correspondant à la sensibilité (9.29) et à la spécificité (9.23), est présentée dans la figure (9.7). Là encore, cette représentation montre l'augmentation d'information apportée par la prise en compte conjointe de X_0 et de X_1. La minimisation de la fonction de coût (9.33), représentée par la figure (9.8), permet d'obtenir $(\tilde{c}_0, \tilde{c}_1) = (-7,52,\ 0,57)$. Ces seuils sont assez différents de ceux obtenus par la méthode simplifiée. Ils correspondent à une sensibilité de 0,44, une spécificité de 0,89 et un coût de 0,0443.

Pour la seconde règle de décision, le calcul de la sensibilité et de la spécificité pour tous les couples possibles de X_0 et de X_1 est représenté par l'aire ROC (9.9). Par rapport à l'aire ROC relative à la première règle de décision (figure 9.7), cette seconde règle privilégie la sensibilité à la spécificité. Les seuils de décision optimaux sont égaux à (-5,73, 1,47), associés à une sensibilité de 0,14, à une spécificité de 0,95 et à un coût de 0,05193. Par

comparaison des coûts, on préférera la première règle de décision. Elle est schématisée par la figure (9.10).

9.3.6 Clairance variable au cours du temps

Ce paragraphe généralise les approches précédentes. L'intérêt est de savoir à tout temps du suivi d'un individu s'il est à risque d'échec ou non. Pour ne pas compliquer les résultats, nous considérons uniquement le pouvoir pronostique de la CL mesurée au temps d pour les échecs se produisant avant le $10ième$ anniversaire de la greffe, avec $d < 10$ ans. Afin de toujours normaliser le marqueur et d'obtenir un facteur de risque, le marqueur d'intérêt est égal à l'opposé de la racine carrée de la CL. Les estimations, issues de la maximisation de la vraisemblance (9.38) du modèle de survie, sont présentées dans le tableau (9.6). Le marqueur reste un facteur de risque d'échec, quelque soit le temps auquel la mesure est réalisée. Le modèle marginal est aussi résumé dans ce tableau et montre une diminution faible de la CL après le $6ième$ mois de la greffe. Remarquons enfin la forte corrélation entre les mesures rapprochées d'un même individu ($\rho = 0,85$).

	Coef.	ET.	Test Wald	p-value
Modèle de survie : $\lambda(v_{h,v_{h,r}}\|x_{h,v_{h,r}}) = \{\alpha\beta_{h,v_{h,r}}v_{h,r}^{\alpha-1}\}/\{(1+\beta_{h,v_{h,r}}v_{h,r}^{\alpha})^{\kappa}\}$				
α	0,56	0,07		
κ	-0,39	0,41		
θ	3,26	0,56		
γ	0,91	0,10	8,99	<0,0001
Modèle marginal (voir équation 9.43)				
η_0	-7,08	0,02		
η_1	-0,10	0,05	-1,91	0,0562
η_2	0,05	0,01	6,73	<0,0001
ρ	0,85			
σ	1,31			

TAB. 9.6 – Modèles paramétriques pour la CL comme marqueur temps-dépendant

A partir de ces résultats, la sensibilité (9.41) et la spécificité (9.42) peuvent être estimées, en fixant le temps de mesure d et le seuil de décision c. La figure (9.11) montre deux courbes ROC, illustrant le pouvoir prédictif de la CL mesurée à 1 et 9 ans, pour tous les seuils c possibles. Les deux AUC correspondantes sont proches, même si l'information apportée par le marqueur à 9 ans semble légèrement supérieure.

En ce qui concerne la minimisation de la fonction de coût (9.44), en fixant $k = 0,1$ et $t = 10$ ans, le seuil optimal de décision dépend alors uniquement du temps auquel la

mesure de CL est réalisée. La figure (9.12) illustre cette minimisation et montre que le coût issu de la décision est d'autant plus faible que la valeur du marqueur est mesurée tardivement. Ce résultat reste cohérent avec la réalité.

Le résultat principal est illustré par la figure (9.13). Il permet de répondre à notre objectif initial, à savoir identifier à chaque visite l'état de gravité du patient en fonction de la valeur de la CL. Toujours en supposant $k = 0,1$ et $t = 10$ ans, un sujet sera considéré à risque d'échec à l'inclusion si sa valeur de CL est inférieure à 58,0 ml/min, à un an si sa CL est inférieure à 52,8 ml/min, à deux ans si sa CL est inférieure à 49,7 ml/min, etc.

9.4 Discussion

Nous avons pu dans ce dernier chapitre décrire une nouvelle approche permettant de définir l'état de santé d'un patient en fonction d'un marqueur pronostique. Selon le problème posé, le test pronostique peut être réalisé à l'inclusion, à un temps fixé ou même tout au long du suivi. Les développements présentés doivent être considérés comme des résultats préliminaires, mais la méthode semble très prometteuse. Pour une interprétation plus précise par le praticien, il semble par exemple intéressant de continuer les développements pour le calcul des valeurs pronostiques positives et des valeurs pronostiques négatives.

L'avantage principal de l'approche est d'être réellement basée sur une décision clinique. Les conséquences, c'est-à-dire les erreurs issues de cette décision, sont en effet mesurées en terme de faux négatifs et de faux positifs, leurs poids n'étant pas forcément équilibrés dans la décision ($k \neq 1$). La méthode alternative, présentée dans le chapitre 6, est basée sur le maximum de vraisemblance. Outre le problème d'estimation de la valeur du seuil (la fonction de vraisemblance n'est pas différentiable au seuil), cette méthode ne considère pas la problématique spécifique de la prise de décision.

Dans l'ensemble des modèles utilisés dans cette thèse, les fonctions de risque de base ont été choisies distribuées selon une loi de Weibull généralisée. Les derniers développements de la section 9.2.4. introduisent une nouvelle distribution basée sur le papier récent de Sparling, Younes et Lachin [108]. Elle permet aussi l'ajustement d'une fonction de risque non-monotone, le modèle restant accéléré. Nous aurions pu, cependant, conserver nos choix précédents de modélisation. Il sera intéressant de comparer les deux approches.

Pour respecter la logique de la construction d'un modèle multi-états répondant à une structure de type aggravations/échecs, la méthode présentée ici aurait dû faire l'objet du premier chapitre. Ceci illustre bien l'amélioration continue des outils méthodologiques au service de la recherche clinique. La perspective à court terme est donc d'utiliser cette nouvelle méthode de classification pour justifier une nouvelle structure multi-états.

Une autre application concrète serait la construction d'un score composite, prenant en

compte plusieurs facteurs de risque, qu'ils soient cliniques ou biologiques. En effet, nous avons pu montrer le très bon pouvoir pronostique de la CL, mais il pourrait être amélioré en prenant en compte d'autres facteurs d'ajustement. Ce score peut, par exemple, être égal à la somme des valeurs observées des facteurs de risque (significativement associées au temps de survie) pondérés par leur coefficient de régression respectif.

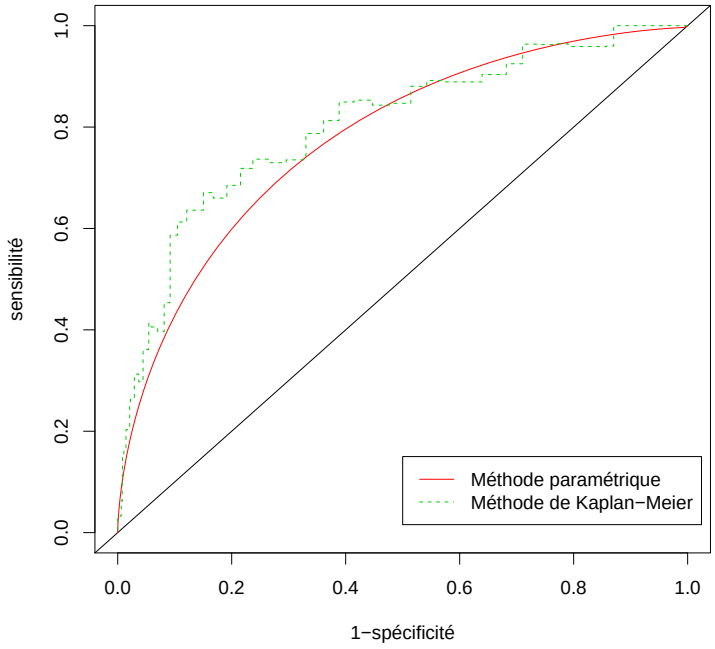

FIG. 9.1 – Courbes ROC paramétrique et non-paramétrique, pour un pronostic à 5 ans à partir de la CL à un an

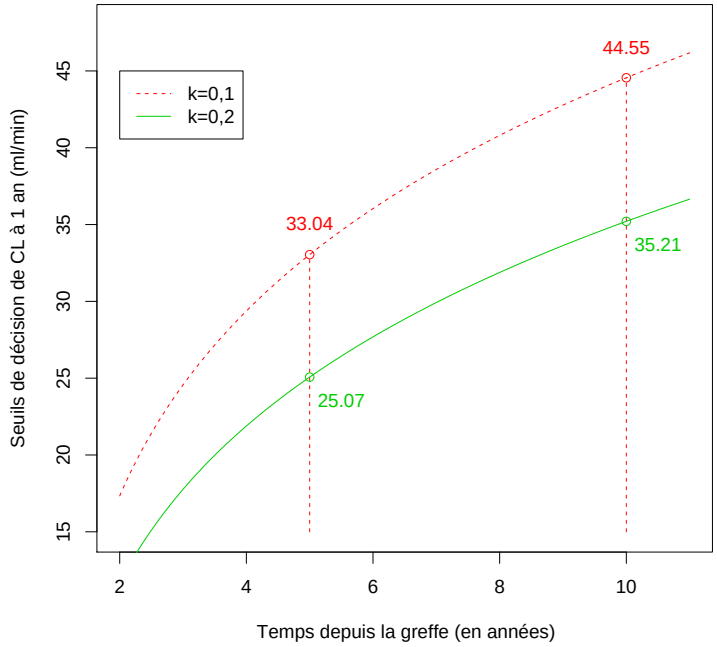

FIG. 9.2 – Seuils optimaux en fonction du poids et du temps (sans ajustement)

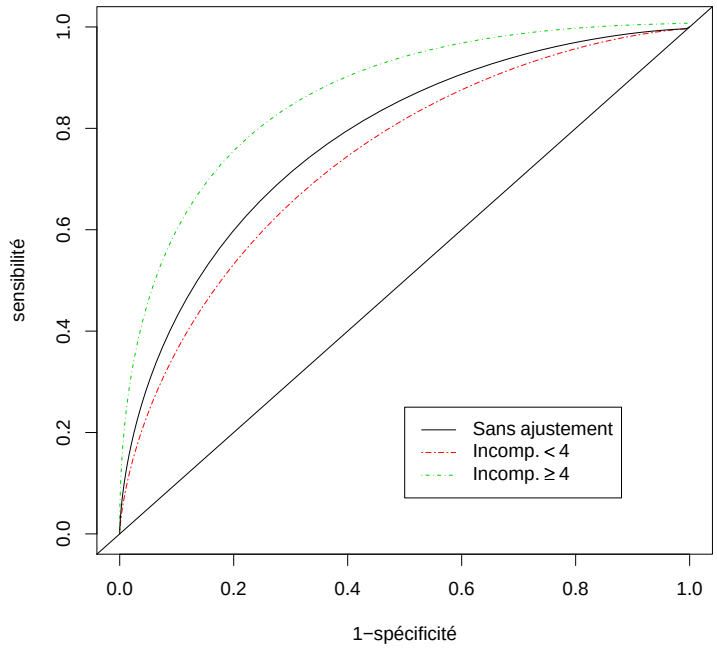

FIG. 9.3 – Courbes ROC paramétriques ajustées sur le nombre d'incompatibilités, pour un pronostic à 5 ans à partir de la CL à un an

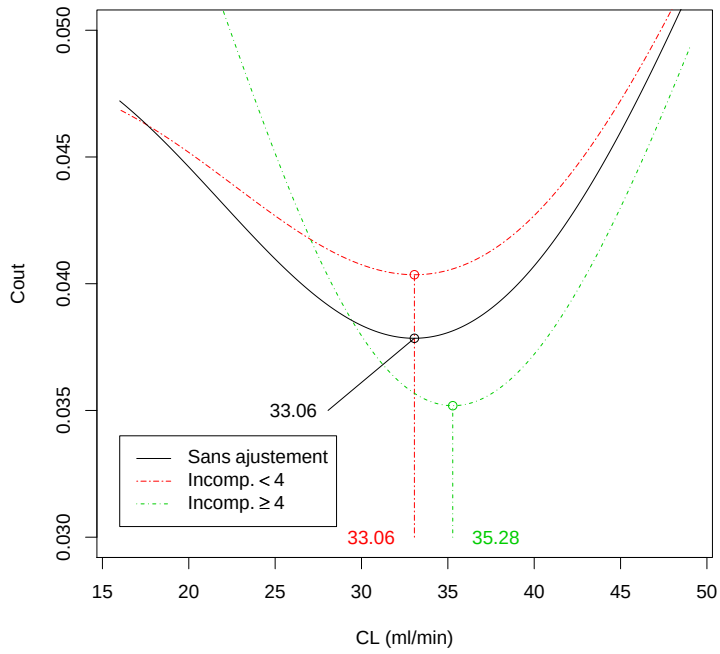

FIG. 9.4 – Minimisation de la fonction de coût relative à CL à un an pour les événements à 5 ans en fonction du nombre d'incompatibilités ($k = 0, 1$)

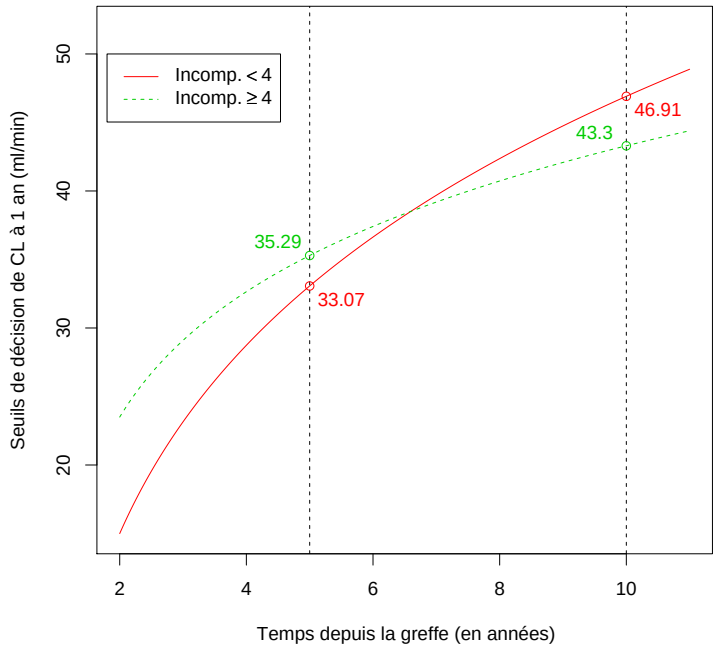

FIG. 9.5 – Seuils optimaux de CL à un an en fonction du nombre d'incompatibilités et du temps de pronostic

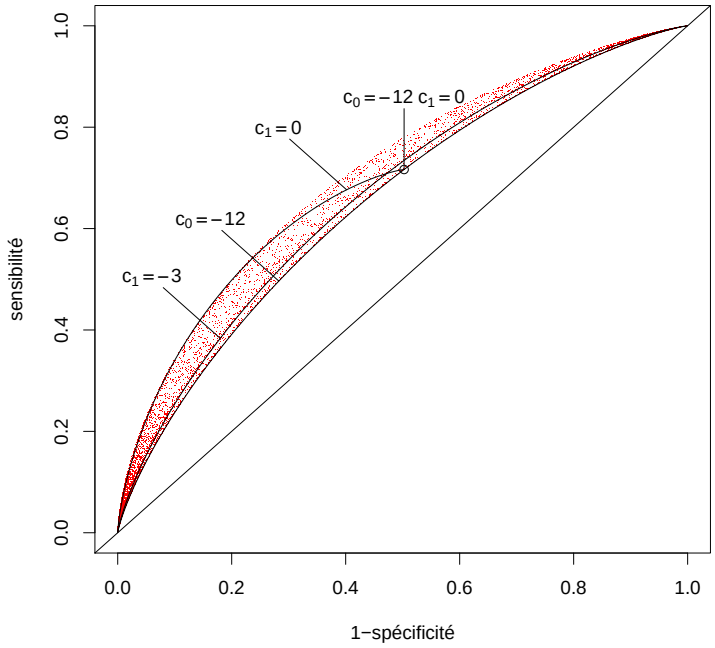

FIG. 9.6 – Aire ROC à partir des 2 mesures du marqueur (méthode simplifiée et première règle de décision), pour un pronostic à 5 ans

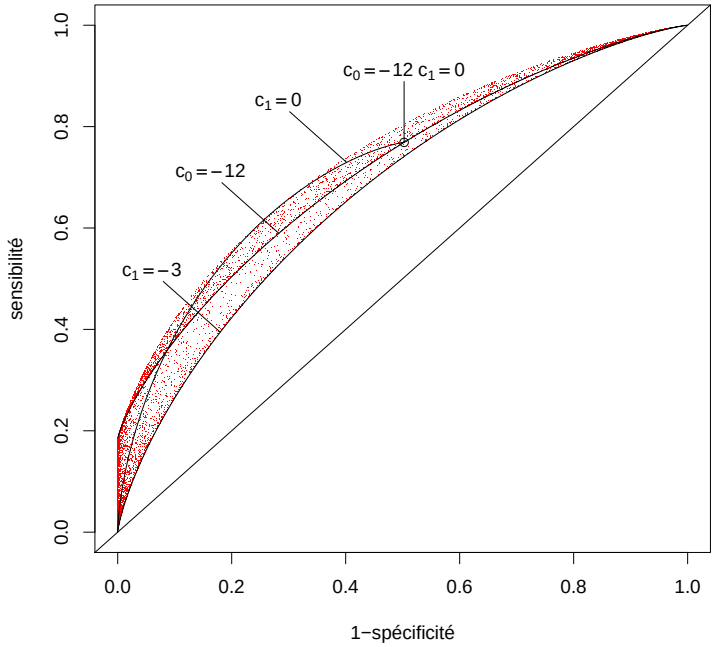

FIG. 9.7 – Aire ROC à partir des 2 mesures du marqueur (méthode complète et première règle de décision), pour un pronostic à 5 ans

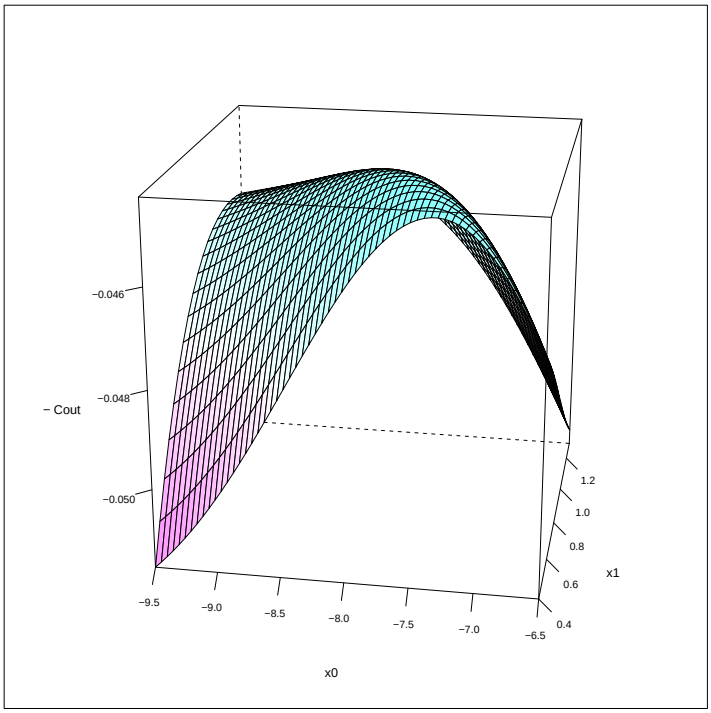

FIG. 9.8 – Minimisation de la fonction de coût en fonction de c_0 et c_1 (méthode complète et première règle de décision), pour un pronostic à 5 ans

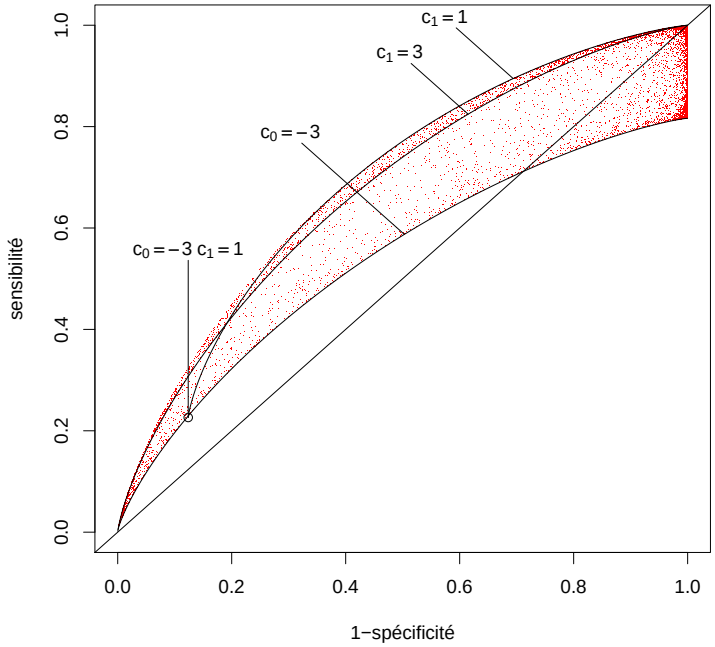

FIG. 9.9 – Aire ROC à partir des 2 mesures du marqueur (méthode complète, seconde règle de décision et $k = 0, 1$), pour un pronostic à 5 ans

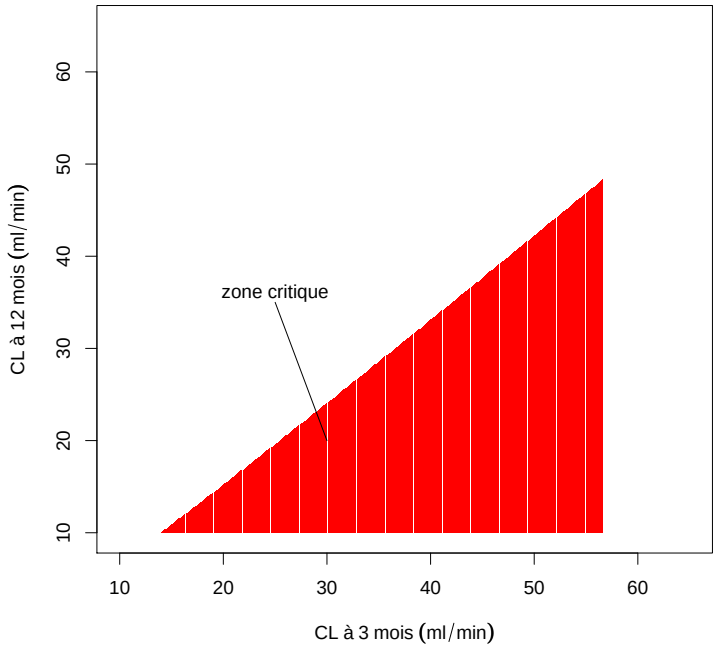

FIG. 9.10 – Zone de décision critique : le sujet est à risque d'échec avant le $5ième$ anniversaire de greffe (méthode complète, première règle de décision et $k = 0,1$)

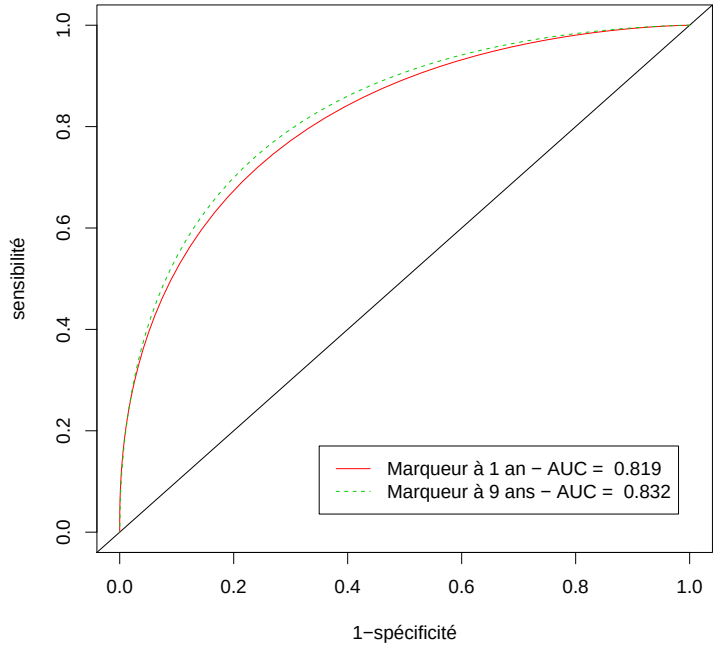

FIG. 9.11 – Courbes ROC paramétriques pour la CL prise en compte comme marqueur temps-dépendant (échecs avant le $10ième$ anniversaire)

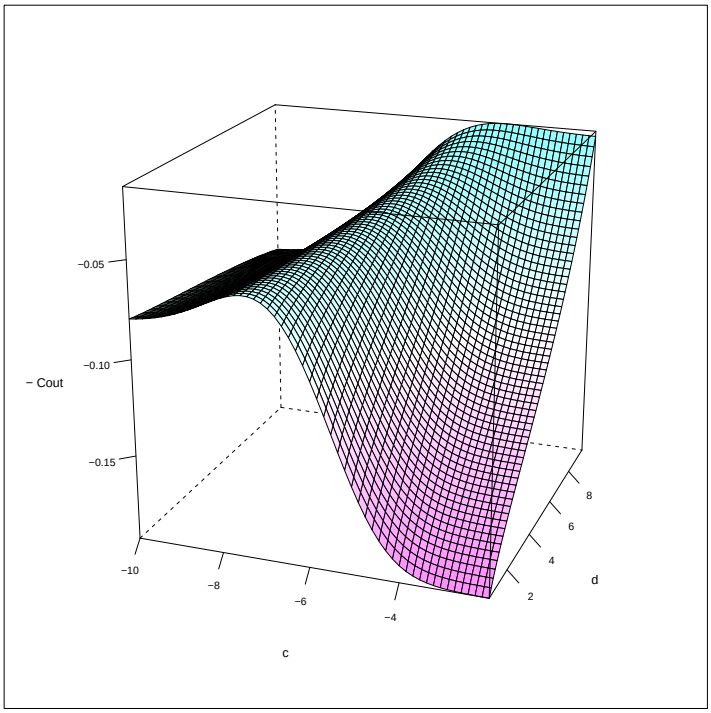

FIG. 9.12 – Minimisation de la fonction de coût en fonction du temps de mesure de CL et du seuil de décision ($k = 0,1$ et $t = 10$ ans)

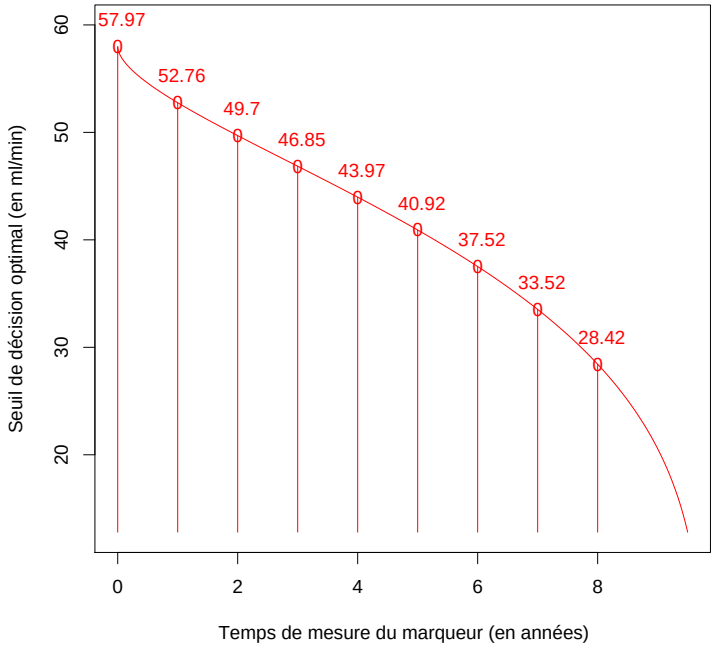

FIG. 9.13 – Seuils optimaux de décision de la CL prise en compte comme marqueur temps-dépendant ($k = 0,1$ et $t = 10$ ans)

Discussion et perspectives

Discussion générale

Le thème central de cette thèse a été le développement de modèles multi-états pour l'analyse de l'évolution de patients atteints d'une pathologie chronique. Ces modèles ont été appliqués à deux cohortes : les patients atteints du VIH et les patients greffés rénaux. La cohorte liée à la transplantation a néanmoins été plus souvent la base de réflexion.

Même si la plupart des travaux se consacrent à l'utilisation de modèles markoviens, nous avons choisi d'investir le principe d'une régression basée sur les durées dans les états. Nous avons pu montrer l'intérêt de ce type d'approche en médecine, où le temps passé dans un certain état de santé constitue un facteur important de l'évolution future des patients.

Le modèle, initialement défini, est basé sur le papier de Perez-Ocon et Ruiz-Castro [20]. Il s'agit du point de départ à partir duquel le modèle semi-markovien est défini. Une première partie du modèle est caractérisée par une chaîne de Markov qui gère la trajectoire du processus. Sa particularité réside dans la matrice des probabilités de transition : la probabilité de rester dans l'état est nulle (exception faite pour les états dits absorbants). La seconde partie gère les distributions des temps d'attente qui peuvent ainsi être explicitement définies par les lois usuelles en analyse de survie (exponentielle, weibull, log-logistique, etc.). Les covariables sont prises en compte dans cette partie du modèle. Elles sont en effet supposées proportionnelles aux fonctions de risque des temps d'attente. L'adaptation de ce modèle aux différentes problématiques médicales étudiées a été réalisée en plusieurs étapes.

La première généralisation a été l'introduction de la distribution Weibull généralisée. Elle permet d'ajuster des fonctions de risque non-monotones. Les différentes applications ont permis de mesurer l'intérêt de ce type de distributions [63]. Dans le dernier chapitre, consacré aux tests pronostiques, une autre distribution généralisant les lois usuelles est définie à partir de l'article de Sparling et al. [108]. L'inconvénient majeur des approches paramétriques étant les hypothèses quant aux distributions utilisées, le développement de ces lois flexibles est important.

Une seconde difficulté, à l'utilisation du semi-Markov, est la censure par intervalle des observations lorsque l'état de santé du patient n'est connu qu'à certaines dates de son suivi. La méthode d'estimation du modèle a été adaptée pour les données issues de ce type de cohorte [90]. L'approche paramétrique est particulièrement adaptée à ce type de données incomplètes, en utilisant par exemple des produits de convolution.

Un troisième point important a été la proposition de méthodes alternatives d'introduction des covariables. Pour les facteurs associés aux temps d'attente dans les états, nous avons systématiquement testé graphiquement la proportionnalité des risques. Pour les covariables ne respectant pas cette hypothèse, l'ajout d'interactions avec la durée dans l'état et l'utilisation d'une méthode inspirée des modèles de vie accélérée ont permis d'améliorer leur modélisation. Nous avons proposé parallèlement d'introduire certains facteurs explicatifs dans la chaîne de Markov. Ainsi, l'effet des covariables n'est pas seulement associé aux durées dans les états mais aussi aux trajectoires du processus.

La quatrième extension importante du modèle initial a été l'introduction de termes aléatoires. Ils permettent de modéliser une structure de dépendance des observations. Pour l'analyse de la dynamique des patients atteints du VIH, la répétition des transitions pour un même sujet a été prise en compte à travers un effet aléatoire individuel par transition [105]. Même si ces fragilités n'ont pas été retenues comme significatives, la méthode possède l'intérêt d'avoir testé si l'hypothèse d'indépendance était valide. L'estimation du modèle est basée sur l'utilisation des transformées de Laplace. Ce type de fragilité a aussi été introduit dans le modèle concernant les transplantés rénaux pour modéliser l'hétérogénéité due au biais période. La définition du modèle ne permettant pas l'utilisation des transformées de Laplace, une méthode d'estimation numérique a été proposée.

A ce niveau des développements, deux échelles de temps ont été considérées : la durée dans l'état et le temps calendaire. Le temps depuis l'origine du suivi n'est pas pris en compte dans la dynamique du processus. En effet, le modèle suppose la stationnarité des forces de transition. Pour tester si cette hypothèse n'est pas abusive, une statistique de test de type Pearson est définie en adaptant le travail de Aguirre-Hernandez et Farewell proposé dans le cadre markovien [97]. La distribution de cette statistique est obtenue à partir d'un échantillonnage par bootstrap semi-paramétrique.

L'ensemble de ces points concerne l'adaptation du modèle semi-markovien aux données cliniques. Cependant, l'idée implicite est que la structure multi-états semble déjà définie. Comme tout postulat, il est le plus souvent oublié alors qu'il conditionne l'ensemble des résultats. En transplantation par exemple, aucun état de gravité intermédiaire n'est parfaitement défini. Certains spécialistes discutent encore de la classification de certains échecs de la transplantation. Le marqueur principal de l'activité rénale est la clairance de la créatinine. La difficulté est donc d'établir une règle de décision à partir de cette biologie pour catégoriser l'évolution d'un patient. Deux approches ont été décrites. La première est basée sur la définition d'un seuil de diminution de clairance à partir duquel le patient entre dans un état à fort risque de rejet. Ce seuil est calculé par maximisation de la

vraisemblance partielle d'un modèle de Cox où l'état de gravité est considéré comme une covariable temps-dépendante. De plus, cette méthode permet de diminuer la fluctuation à court terme du marqueur par lissage. Cependant, deux inconvénients sont à noter. Premièrement, la fonction de vraisemblance n'est pas différentiable au seuil, son estimation est réalisée à partir d'une grille de valeurs prédéfinies. Deuxièmement, le calcul ne prend pas en compte les conséquences réelles de la décision. La seconde méthode est justement basée sur les erreurs de classification. En se plaçant dans le contexte de tests pronostiques, le poids d'un faux positif par rapport à celui d'un faux négatif est pris en compte dans une fonction de coût. La minimisation de cette fonction permet l'identification d'un seuil optimal. Pour aboutir à ce type de conclusions, plusieurs développements des méthodes basées sur la notion de courbe ROC sont proposés. Ces développements, réalisés en fin de thèse, devront faire l'objet d'approfondissements.

Limites et perspectives

Les dernières remarques illustrent bien que les développements méthodologiques, initiés dans cette thèse, ne constituent qu'une étape. A partir de la nouvelle définition de la structure multi-états, illustrée par la figure (9.13), il serait intéressant de mettre à jour l'estimation du modèle semi-markovien correspondant. La vraisemblance (6.17) et le test d'adéquation (7.1) sont directement applicables à ce dernier schéma.

D'un point de vue méthodologique, nombreuses sont les limites abordées dans les discussions des différents chapitres. Le point central est peut-être le postulat paramétrique de tous les modèles. Même si nous avons eu le souci constant d'utiliser des formes de distributions générales et flexibles, il serait intéressant de comparer les résultats obtenus avec des méthodes non-paramétriques. Ces dernières possèdent cependant d'autres inconvénients, en particulier dans l'instabilité des estimations. Certains auteurs, comme Joly et Commenges [112, 27] ou Alvarez [113], proposent une approche intermédiaire, basée sur l'ajustement de la fonction de risque de base par splines.

De plus, notons qu'en présence de censures par intervalle, toutes les structures multi-états sont unidirectionnelles. Les transitions sont dirigées vers les événements terminaux, sans retour possible. Cependant, il n'est pas évident qu'un patient ne puisse pas retrouver un état de bonne santé suite à une aggravation de son pronostic. La difficulté, lorsqu'un individu n'est pas observable pendant un intervalle de temps donné, serait alors de prendre en compte la probabilité qu'il puisse réaliser une infinité d'allers et de retours. L'emboîtement de produits de convolution pourrait être une solution si le nombre d'intégrales n'est pas trop important. Il faudrait donc définir un délai incompressible pendant lequel un individu ne peut pas subir plus d'une transition et disposer d'une base de données où les intervalles de censure ne sont pas trop importants. Ceci devrait donc être possible.

Concernant ensuite la statistique de test permettant d'évaluer le respect de la station-

narité du processus, la limite principale est le regroupement dans le tableau de contingence des transitions entre les deux états de gravité et des transitions vers un événement absorbant. Ce regroupement est nécessaire puisque seuls les temps d'entrée dans un état absorbant sont disponibles exactement. Les transitions de l'état 1 à l'état 2 sont toutes censurées par intervalle, les bornes de ces intervalles pouvant quasiment être considérées différentes entre patients. Le classement de ces transitions resterait possible si les intervalles de censure étaient homogènes, c'est-à-dire si les dates de visite étaient les mêmes pour tous les individus. Les fonctions de risque n'étant pas constantes au cours du temps, choisir le milieu de l'intervalle comme temps de transition est peu satisfaisant. La reconstruction des trajectoires exactes, en choisissant les temps de transition les plus probables en fonction du modèle initialement estimé, serait peut-être envisageable.

Enfin, il est nécessaire de développer la dernière partie de ce document concernant le pouvoir pronostique de la clairance de la créatinine. Plusieurs pistes de travail semblent importantes à suivre. Premièrement, un score pourrait être construit en prenant en compte dans son calcul d'autres facteurs de risque ou protecteurs d'un échec. Ce score, défini à l'inclusion ou dépendant du temps, serait par définition plus informatif que la clairance de la créatinine étudiée seule. Deuxièmement, dans la méthode considérant le marqueur continuellement évolutif au cours du temps, nous avons inclu deux temps dans la décision : le temps où le marqueur est mesuré (d) et le temps de pronostic (t). Comme nous avons pu le montrer, le marqueur est d'autant plus intéressant que d est proche de t ($d < t$). Ce résultat n'est pas surprenant. L'objectif étant de définir à n'importe quelle visite l'état du patient, il sera intéressant de faire varier le temps de pronostic souhaité en fonction du temps auquel le marqueur est mesuré. Par exemple, si l'objectif est de prédire les échecs dans les 5 ans suivant la mesure de clairance, les effectifs de faux positifs et de faux négatifs seraient alors calculés à partir des expressions suivantes, à remplacer dans la fonction de coût (9.44).

$$n_{FP}(c,d) = nP(T > (d+5), X_{h,d} > c) \text{ et } n_{FN}(c,d) = nP(d < T \leq d+5, X_{h,d} \leq c)$$

Le temps de pronostic est bien variable en fonction du temps de mesure du marqueur.

Pour conclure cette thèse, il me semble important de souligner qu'une part importante de ce travail reste à venir. En effet, l'intérêt premier des développements méthodologiques abordés reste l'inférence médicale qui en découle. Les résultats doivent être validés et interprétés plus précisément par les cliniciens, en rapport avec la littérature. Des efforts de vulgarisation, de communication et de valorisation doivent ainsi être entrepris à court terme.

Annexe A

Logvraisemblance du modèle semi-markovien

L'hypothèse de semi-proportionnalité des risques (2.22) s'écrit :

$$\lambda_{ij}(x, z_{ij}) = \lambda_{0,ij}(x) exp(\beta_{ij}^T z_{ij})$$

De plus, pour une distribution Weibull généralisée (1.11), la fonction de risque de base précédente est définie par :

$$\lambda_{ij}(t) = \frac{1}{\theta_{ij}}\Big(1 + (\frac{t}{\sigma_{ij}})^{\nu_{ij}}\Big)^{\frac{1}{\theta_{ij}}-1} \frac{\nu_{ij}}{\sigma_{ij}}\Big(\frac{t}{\sigma_{ij}}\Big)^{\nu_{ij}-1} \quad \forall\ \nu_{ij}\ ,\ \sigma_{ij}\ ,\ \theta_{ij} > 0$$

Ainsi, à partir de la formulation générale de la logvraisemblance (4.6), on obtient la logvraisemblance à maximiser :

$$\begin{aligned}
ln\mathcal{V} &= \sum_h \{\gamma_{0X_{h,1}} + \beta_{0X_{h,1}} z_{h,0X_{h,1}} - ln(\sum_{i=1}^{c} exp(\gamma_{0i} + \beta_{0i} z_{h,0X_{h,1}})) \\
&+ \sum_{ij} \sum_{X_{h,r}=i, X_{h,r+1}=j} \{\delta_{h,r}^E [ln\ P_{ij} - ln\ \theta_{ij} + (\frac{1}{\theta_{ij}} - 1) ln(1 + (\frac{d_{h,r}}{\sigma_{ij}})^{\nu_{ij}}) \\
&+ ln\ \nu_{ij} + (\nu_{ij} - 1) ln\ d_{h,r} - \nu_{ij} ln\ \sigma_{ij} + \beta_{ij} z_{h,ij} \\
&+ exp(\beta_{ij} z_{h,ij})(1 - (1 + (\frac{d_{h,r}}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}})] \\
&+ \delta_{h,r}^I [ln\ P_{ij} + ln(exp(1 - (1 + (\frac{d_{h,r}^0}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}})^{exp(\beta_{ij} z_{h,ij})} \\
&- exp(1 - (1 + (\frac{d_{h,r}^1}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}})^{exp(\beta_{ij} z_{h,ij})})]\} \\
&+ \sum_{ij} \sum_{X_{h,r}=i} \{\delta_{h,r}^R [ln(\sum_{j \neq i} P_{ij} exp(1 - (1 + (\frac{d_{h,r}^0}{\sigma_{ij}})^{\nu_{ij}})^{\frac{1}{\theta_{ij}}})^{exp(\beta_{ij} z_{h,ij})})]\}\}
\end{aligned} \tag{A.1}$$

où $\gamma_{03} = \beta_{03} = 0$.

Annexe B

Racines et poids des polynômes de Legendre

Racines u_q	Poids w_q	Racines u_q	Poids w_q
-0,9739065	0,06667134	0,1488743	0,29552422
-0,8650634	0,14945135	0,4333954	0,26926672
-0,6794096	0,21908636	0,6794096	0,21908636
-0,4333954	0,26926672	0,8650634	0,14945135
-0,1488743	0,29552422	0,9739065	0,06667134

TAB. B.1 – Racines et poids du $10 ième$ polynôme de Legendre.

Racines u_q	Poids w_q	Racines u_q	Poids w_q
-0,99689348	0,007968192	0,05147184	0,102852653
-0,98366812	0,018466468	0,15386991	0,101762390
-0,96002186	0,028784708	0,25463693	0,099593421
-0,92620005	0,038799193	0,35270473	0,096368737
-0,88256054	0,048402673	0,44703377	0,092122522
-0,82956576	0,057493156	0,53662415	0,086899787
-0,76777743	0,065974230	0,62052618	0,080755895
-0,69785049	0,073755975	0,69785049	0,073755975
-0,62052618	0,080755895	0,76777743	0,065974230
-0,53662415	0,086899787	0,82956576	0,057493156
-0,44703377	0,092122522	0,88256054	0,048402673
-0,35270473	0,096368737	0,92620005	0,038799193
-0,25463693	0,099593421	0,96002186	0,028784708
-0,15386991	0,101762390	0,98366812	0,018466468
-0,05147184	0,102852653	0,99689348	0,007968192

TAB. B.2 – Racines et poids du 30*ième* polynôme de Legendre.

Annexe C

Effets fixes du modèle avec biais période

Transition	Variable	Coef.	ET	p-value
$1 \rightarrow 2$	Intercept	2,12	0,65	0,0011
$1 \rightarrow 2$	Délai de reprise	0,76	0,37	0,0400
$1 \rightarrow 2$	Age du donneur	-2,17	0,66	0,0010
$1 \rightarrow 3$	Intercept	-2,64	0,54	0,0001
$2 \rightarrow 3$	Intercept	1,15	0,34	0,0008
$2 \rightarrow 3$	Incompatibilités A+B+DR	0,95	0,48	0,0463

TAB. C.1 – Coefficients de régression associés à la chaîne de Markov

	σ_{ij}		ν_{ij}		θ_{ij}	
Transition	Estim.	ET	Estim.	ET	Estim.	ET
$1 \rightarrow 2$	7,25	1,72	0,86	0,05	1	.
$1 \rightarrow 3$	79,33	112,28	1	.	1	.
$1 \rightarrow 4$	77,66	40,84	1	.	1	.
$2 \rightarrow 3$	12,67	3,21	1	.	1	.
$2 \rightarrow 4$	6,51	4,23	1	.	1	.

TAB. C.2 – Paramètres associés aux lois d'attente dans les états

Transition	Variable	Coef.	ET	p-value
$1 \rightarrow 2$	Traitement d'induction	0,27	0,16	0,0948
$1 \rightarrow 2$	Sexe du receveur	-0,26	0,13	0,0521
$1 \rightarrow 2$	Age du donneur	0,77	0,26	0,0031
$1 \rightarrow 3$	Ischémie froide	5,63	1,49	0,0002
$2 \rightarrow 3$	Incompatibilité A+B+DR	0,88	0,28	0,0019
$2 \rightarrow 3$	PRA	1,09	0,35	0,0017
$2 \rightarrow 3$	PRA $\times$ d	-0,47	0,22	0,0309
$2 \rightarrow 4$	Délai de reprise	2,00	0,62	0,0012
$2 \rightarrow 4$	Sexe du receveur	1,50	0,66	0,0227
$2 \rightarrow 4$	Sexe du receveur $\times$ d	-4,23	1,18	0,0004
$2 \rightarrow 4$	Sexe du receveur $\times$ d^2	1,28	0,31	0,0001

TAB. C.3 – Coefficients de régression associés aux temps de séjours

Bibliographie

[1] Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society : Series B*, 34 :187–220, 1972.

[2] Aalen O and Husebye E. Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine*, 10 :1227–1240, 1991.

[3] Hougaard P. *Analysis of Multivariate Survival Data*. Springer, 2000.

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

[5] Beck RJ and Paucker SG. The Markov process in medical prognosis. *Medical Decision Making*, 3 :419–458, 1983.

[6] Aalen OO and Johansen S. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5 :141–50, 1978.

[7] Kousignian I, Abgrall S, Duval X, Descamps D, Matheron S, and 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. *Journal of Acquired Immune Deficiency Syndromes*, 34 :50–57, Sep 2003.

[8] Alioum A, Leroy V, Commenges D, Dabis F, and 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'epidemiologie clinique du SIDA en Aquitaine. *Epidemiology*, 9 :605–612, 1998.

[9] Aalen O, Farewell VT, Angelis D, Day N, and Gill N. 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.

[10] Mauskopf J. Meeting the NICE Requirements : A Markov Model Approach. *Value Health*, 3 :287–287, Jul 2000.

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

[12] Saint-Pierre P, Combescure C, Daures JP, and Godard P. The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine*, 22 :3755–3770, Dec 2003.

[13] Klein JP, Keiding N, and Copelan EA. Plotting summary predictions in multistate survival models - probabilities of relapse and death in remission for bone-marrow transplantation patients. *Statistics in Medicine*, 12 :2315–2332, 1993.

[14] Keiding N, Klein JP, and Horowitz MM. Multi-state models and outcome prediction in bone marrow transplantation. *Statistics in Medicine*, 20 :1871–1885, 2001.

[15] Commenges D. Multi-state models in epidemiology. *Lifetime data analysis*, 5 :309–321, 1999.

[16] Commenges D. Risques compétitifs et modèles multi-états en épidémiologie. *Revue d'Epidémiologie et de Santé Publique*, 47 :605–611, 1999.

[17] Hougaard P. Multi-state Models : A Review. *Lifetime Data Analysis*, 5 :239–264, 1999.

[18] Com-Nougué C, Guérin S, and Rey A. Estimation des risques associés à des événements multiples. *Revue Epidemiologie et de Santé Publique*, 47 :75–85, 1999.

[19] Marshall G and Jones RH. Multi-state models and diabetic retinopathy. *Statistics in Medicine*, 14 :1975–1983, 1995.

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

[21] Dabrowska DM, Sun G, and Horowitz MM. Cox Regression in a Markov Renewal Model : an application to the analysis of bone transplant data. *Journal of the American Statistical Association*, 89 :867–877, 1994.

[22] Whitmire JK, Asano MS, Murali-Krishna K, Suresh M, and Ahmed R. Long-Term CD4 Th1 and Th2 Memory following Acute Lymphocytic Choriomeningitis Virus Infection. *Journal of virology*, 72 :8281–8288, 1998.

[23] Grivel JC, Malkevitch N, and Margolis L. Human Immunodeficiency Virus Type 1 Induces Apoptosis in CD4+ but Not in CD8+ T Cells in Ex Vivo-Infected Human Lymphoid Tissue. *Journal of virology*, 74 :8077–8084, 2000.

[24] Berzuini C and Allemani C. Effectiveness of potent antiretroviral therapy on progression of human immunodeficiency virus : Bayesian modelling and model checking via counterfactual replicates. *Journal of the Royal Statistical Society : Series C*, 53 :633–650, 2004.

[25] Longini IM, Clark WS, Byers RH, Ward JW, Darrow WW, Lemp G, and Hethcote HW. Statistical analysis of the stages of HIV Infection using a Markov model. *Statistics in Medicine*, 8 :831–843, 1989.

[26] Gentleman RC, Lawless JF, Lindsey JC, and Yan P. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13 :805–821, 1994.

[27] Joly P and Commenges D. A penalized likelihood approach for a progressive three-state model with censored and truncated data : Application to AIDS. *Biometrics*, 55 :887–890, Sept 1999.

[28] Pradier C, Puglièse P, Caissotti C, Martini S, Pueyo B, and Dellamonica P. Dossier médical informatisé pour les patients atteints d'infection par le VIH (ADDIS). *Médecine et maladies infectieuses*, 28 :291–295, 1998.

[29] Puglièse P, Cuzin L, Enel P, Agher R, Alfandari S, Billaud E, Druard P, Duvivier C, Perez M, Salmi D, and Pradier C. Le projet NADIS 2000 : Développement d'un Dossier Médical Informatisé pour les patients VIH, VHC ou VHB. *Presse médicale*, 7 :299–303, 2003.

[30] Alejandro V, Scandling JD, Sibley RK, Dafoe D, Alfrey E, Deen W, and Myers BD. Mechanisms of filtration failure during postischemic injury of the human kidney. A study of the reperfused renal allograft. *Journal of Clinical Investigation*, 95 :820–831, 1995.

[31] Giral M, Nguyen JM, Karam G, Kessler M, Bruno Hurault de Ligny B, Buchler M, Bayle F, Meyer C, Foucher Y, Martin ML, Daguin P, and Soulillou JP. Impact of Graft Mass on the Clinical Outcome of Kidney Transplants. *Journal of the American Society of Nephrology*, 16 :261–268, 2005.

[32] Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, and Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine : a new prediction equation. Modification of Diet in Renal Disease Study Group. *Annals of Internal Medicine*, 130 :461–70, 1999.

[33] Gjertson DW, Dabrowska DM, Cui X, and Cecka JM. Four Causes of Cadaveric Kidney Transplant Failure : A Competing Risk Analysis. *American Journal of Transplantation*, 2 :84–93, 2002.

[34] Nankivell BJ, Borrows RJ, Fung CL, O'Connell PJ, Allen RD, and Chapman JR. The natural history of chronic allograft nephropathy. *New England Journal of Medicine*, 349 :2326–2333, 2003.

[35] Schnuelle P, Yard BA, Braun C, Dominguez-Fernandez E, Schaub M, Birck R, Sturm J, Post S, and van der Woude FJ. Impact of donor dopamine on immediate graft function after kidney transplantation. *American Journal of Transplantation*, 4 :419–426, 2004.

[36] Perico N, Cattaneo D, Sayegh MH, and Remuzzi G. Delayed graft function in kidney transplantation. *Lancet*, 364 :1814–1827, 2004.

[37] Giral M, Bertola JP, Foucher Y, Villers D, Bironneau E, Blanloeil Y, Karam G, Daguin P, Lerat L, and Soulillou JP. Effect of brain-dead donor resuscitation on delayed graft function : results of a monocentric analysis. *Transplantation*, 83 :1174–1181, 2007.

[38] Giral-Classe M, Hourmant M, Cantarovich D, Dantal J, Blancho G, Daguin P, Ancelet D, and Soulillou JP. Delayed graft function of more than six days strongly

decreases long-term survival of transplanted kidneys. *Kidney International*, 54 :972–978, 1998.

[39] Neugarten J, Srinivas T, Tellis V, Silbiger S, and Greenstein S. The effect of donor gender on renal allograft survival. *Journal of the American Society of Nephrology*, 7 :318–324, 1996.

[40] Zeier M, Dohler B, Opelz G, and Ritz E. The effect of donor gender on graft survival. *Journal of the American Society of Nephrology*, 13 :2570–2576, 2002.

[41] Mackenzie HS, Azuma H, Rennke HG, Tilney NL, and Brenner BM. Renal mass as a determinant of late allograft outcome : insights from experimental studies in rats. *Kidney International Supplement*, 52 :S38–42, 1995.

[42] Oudar O, Elger M, Bankir L, Ganten D, Ganten U, and Kriz W. Differences in rat kidney morphology between males, females and testosterone-treated females. *Renal Physiology and Biochemistry*, 14 :92–102, 1991.

[43] Nyengaard JR and Bendtsen TF. Glomerular number and size in relation to age, kidney weight, and body surface in normal man. *Anatomical Record*, 232 :194–201, 1992.

[44] Vereerstraeten P, Wissing M, De Pauw L, Abramowicz D, and Kinnaert P. Male recipients of kidneys from female donors are at increased risk of graft loss from both rejection and technical failure. *Clinical Transplantation*, 13 :181–186, 1999.

[45] Campos EF, Tedesco-Silva H, Machado PG, Franco M, Medina-Pestana JO, and Gerbase-DeLima M. Post-transplant anti-HLA class II antibodies as risk factor for late kidney allograft failure. *American Journal of Transplantation*, 10 :2316–2320, 2006.

[46] Terasaki PI. The UNOS Scientific Renal Transplant Registry-1991. *Clinical Transplantation*, pages 1–11, 1991.

[47] Suthanthiran M and Strom TB. Renal transplantation. *New England Journal of Medicine*, 331 :365–376, 1994.

[48] Cicciarelli J and Cho Y. HLA matching : univariate and multivariate analyses of UNOS Registry data. *Clinical Transplantation*, pages 325–333, 1991.

[49] Susal C and Opelz G. Kidney graft failure and presensitization against HLA class I and class II antigens. *Transplantation*, 73 :1269–1273, 2002.

[50] Terasaki PI. Humoral theory of transplantation. *American Journal of Transplantation*, 3 :665–673, 2003.

[51] Opelz G. Non-HLA transplantation immunity revealed by lymphocytotoxic antibodies. *Lancet*, 365 :1570–1576, 2005.

[52] de Fijter JW, Mallat MJ, Doxiadis II, Ringers J, Rosendaal FR, Claas FH, and Paul LC. Increased immunogenicity and cause of graft loss of old donor kidneys. *Journal of the American Society of Nephrology*, 12 :1538–1546, 2001.

[53] de Fijter JW. The impact of age on rejection in kidney transplantation. *Drugs Aging*, 22 :433–449, 2005.

[54] Reutzel-Selke A, Filatenkov A, Jurisch A, Denecke C, Martins PN, Pascher A, Jonas S, Pratschke J, Neuhaus P, and Tullius SG. Grafts from elderly donors elicit a stronger immuneresponse in the early period posttransplantation : a study in a rat model. *Transplant Proc*, 37 :382–383, 2005.

[55] Meier-Kriesche HU, Ojo AO, Cibrik DM, Hanson JA, Leichtman AB, Magee JC, Port FK, and Kaplan B. Relationship of recipient age and development of chronic allograft failure. *Transplantation*, 70 :306–310, 2000.

[56] Meier-Kriesche HU, Cibrik DM, Ojo AO, Hanson JA, Magee JC, Rudich SM, Leichtman AB, and Kaplan B. Interaction between donor and recipient age in determining the risk of chronic renal allograft failure. *Journal of the American Geriatrics Society*, 50 :195–197, 2002.

[57] Bradley BA. Rejection and recipient age. *Transplant Immunology*, 10 :125–132, 2002.

[58] Meier-Kriesche HU, Ojo A, Hanson J, Cibrik D, Lake K, Agodoa LY, Leichtman A, and Kaplan B. Increased immunosuppressive vulnerability in elderly renal transplant recipients. *Transplantation*, 69 :885–889, 2000.

[59] Moreso F, Ortega F, and Mendiluce A. Recipient age as a determinant factor of patient and graft survival. *Nephrology Dialysis Transplantation*, 19 Suppl 3 :16–20, 2004.

[60] Sternberg M and Satten G. Discrete time nonparametric estimation for chain of events data subject to interval censoring and truncation. *Biometrics*, 55 :514–522, 1999.

[61] Satten G and Sternberg M. Fitting Semi-Markov Models to interval-Censored Data with Unknown Initiation Times. *Biometrics*, 55 :507–513, 1999.

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

[63] Foucher Y, Mathieu E, Saint-Pierre P, Durand JF, and Daures JP. A Semi-Markov model based on generalized Weibull distribution with an illustration for HIV disease. *Biometrical Journal*, 47 :825–833, 2005.

[64] McCullagh P. *Generalized Linear Models*. Chapman and Hall, 1983.

[65] Abrahamowicz M, Mackenzie T, and Esdaile JM. Time-dependent hazard ratio : modelling and hypothesis testing with application in Lupus Nephritis. *Journal of the American Statistical Association*, 91 :1432–1439, 1996.

[66] Orbe J, Ferreira E, and Nunez-Anton V. Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in medicine*, 21 :3493–3510, 2002.

[67] Lawless JF. *Statistical Models and Methods for Lifetime Data*. Wiley, New York, 1982.

[68] Stablein DM, Carter WH, and Novak JW. Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*, 2 :149–159, 1981.

[69] Bagdonavicius V and Nikulin M. *Accelerated Life Models*. Chapman and Hall/CRC, 2002.

[70] Odell P, Andersen PK, and D'Agostino R. Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull-Based Accelerated Failure Time Model. *Biometrics*, 48 :951–959, 1992.

[71] Klein JP. Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm. *Biometrics*, 48 :795–806, 1992.

[72] Therneau TM, Grambsch PM, and Pankratz VS. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 1 :156–175, 2003.

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

[74] Gill RD. Nonparametric estimation based on censored observations of a Markov Renewal Process. *Zeitschrift Fr Wahrscheinlichkeits Theorie Verwandte Gebiete*, 53 :97–116, 1980.

[75] Byrd RH, Lu P, Nocedal J, and Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16 :1190–1208, 1995.

[76] Kaplan EL and Meier P. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53 :457–481, 1958.

[77] Escolano S, Golmard JL, Korinek AM, and Mallet A. A multi-state model for evolution of intensive care unit patients : prediction of nosocomial infections and deaths. *Statistics in Medicine*, 19 :3465–3482, 2000.

[78] Yau CL and Huzurbazar AV. Analysis of censored and incomplete survival data using flowgraph models. *Statistics in Medicine*, 21 :3727–3743, 2002.

[79] Ripatti S, Gatz M, Pedersen NL, and Palmgren J. Three-State Frailty Model for Age at Onset of Dementia and Death in Swedish Twins. *Genetic Epidemiology*, 24 :139–149, 2003.

[80] Sun J. Variance estimation of a survival function for interval-censored survival data. *Statistics in medicine*, 20 :1249–1257, 2001.

[81] Chi Y and Tseng CH. Comparison of Several Relative Risk Estimators with Interval-Censored Data. *Biometrical Journal*, 44 :197–212, 2002.

[82] Commenges D. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11 :162–182, 2002.

[83] McCullagh P. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society : Series B*, 42 :109–142, 1980.

[84] Roodnat JI, Mulder PG, Rischen-Vos J, van Riemsdijk IC, van Gelder T, Zietse R, IJzermans JN, and Weimar W. Proteinuria after renal transplantation affects not only graft survival but also patient survival. *Transplantation*, 72 :438–444, 2001.

[85] Fernandez-Fresnedo G, Plaza JJ, Sanchez-Plumed J, Sanz-Guajardo A, Palomar-Fontanet R, and Arias M. Proteinuria : a new marker of long-term graft and patient survival in kidney transplantation. *Nephrology Dialysis Transplantation*, 19 Suppl 3 :47–51, 2004.

[86] Kay R. Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, 26 :227–237, 1977.

[87] Hess RH. Graphical methods for assessing violations of the proportional hazards assumption in cox regression. *Statistics in Medicine*, 14 :1707–1723, 1995.

[88] Satten GA and Longini IM. Markov chains with measurement error : Estimating the true course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistics - Journal of the Royal Statistical Society Series C*, 45 :275–295, 1996.

[89] Jackson CH and Sharples LD. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, 21 :113–128, 2002.

[90] Foucher Y, Giral M, Soulillou JP, and Daures JP. A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. *Statistics in Medicine*, 2007. En révision.

[91] Saporta G. *Probabilités, analyse des données et statistique*. Technip, Paris, 1990.

[92] S Hariharan, MA McBride, WS Cherikhn, CB Tolleris, BA Bresnahan, and CP Johnson. Post-transplant renal function in the first year predicts long-term kidney transplant survival. *Kidney International*, 62 :311–318, 2002.

[93] Hastie TJ and Tibshirani RJ. *Generalized additive models*. Chapman and Hall, London, 1990.

[94] de Boor C. *A Practical Guide to Splines*. Springer, New York, 1978.

[95] Duchateau L and Janssen P. Penalized Partial Likelihood for Frailties and Smoothing Splines in Time to First Insemination Models for Dairy Cows. *Biometrics*, 60 :608–614, 2004.

[96] Abramowitz M and Stegun IA. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, Inc., New York, 1972.

[97] Aguirre-Hernandez R and Farewell VT. A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medicine*, 21 :1899–1911, 2002.

[98] Bishop YMM, Fienberg SE, and Holland PW. *Discrete Multivariate Analysis : Theory and Practice*, chapter 7. MIT, Cambridge, 1975.

[99] Kalbfleisch JD and Lawless JF. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80 :863–871, 1985.

[100] Stavola de BL. Testing departures from time homogeneity in multistate Markov processes. *Applied Statistics*, 37 :242–250, 1988.

[101] Lawless JF and Babineau D. Models for interval censoring and simulation-based inference for liftime distributions. *Biometrika*, 93 :671–686, 2006.

[102] Hosmer DW and Lemeshow D. *Applied Logistic Regression.* Wiley, New York, 1989.

[103] Efron B and Tibshirani RJ. *An Introduction to the Bootstrap.* Chapman and Hall, New York, 1993.

[104] Bouleau N. *Probabilités de l'Ingénieur, variables aléatoires et simulation.* Hermann, Paris, 2002.

[105] Foucher Y, Saint-Pierre P, Daures JP, and Durand JF. A semi-Markov frailty model for multistate and clustered survival data. *Far East Journal of Theoretical Statistics*, 19 :185–201, 2006.

[106] Heagerty PJ, Lumley T, and Pepe SP. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, 56 :337–344, 2000.

[107] Akritas MG. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, 22 :1299–1327, 1994.

[108] Sparling YH, Younes N, and Lachin JM. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics*, 7 :599–614, 2006.

[109] Liang KY and Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 :12–13, 1986.

[110] Diggle P, Liang KY, and Zeger S. *Analysis of longitudinal data.* Oxford Science Publications, Oxford, 1994.

[111] Pinheiro JC and Bates DM. *Mixed-Effects Models in S and S-PLUS.* Springer, 2000.

[112] Commenges D and Joly P. Multi-state model for dementia, institutionalization and death. *Communications in Statistics A*, 33 :1315–1326, 2004.

[113] Alvarez E. Smoothed nonparametric estimation in window censored semi-Markov processes. *Journal of statistical planning and inference*, 131 :209–229, 2005.

Résumé

L'étude de l'évolution du pronostic de santé d'un patient constitue un domaine important en recherche clinique. Récemment, le développement des modèles multi-états a permis d'étudier cette dynamique en prenant en compte plusieurs états de santé. Dans ce manuscrit, nous utilisons plus particulièrement les modèles semi-markoviens. Ce type de processus distingue les temps de séjour dans les états et les trajectoires des transitions, contrairement à l'approche markovienne classique. Nous avons proposé plusieurs adaptations pour pouvoir appliquer ce type de modèle : la censure par intervalle, le choix des distributions des temps d'attente et l'introduction des covariables. Un test d'adéquation est aussi proposé pour vérifier l'hypothèse de stationnarité. Enfin, une méthode originale, incluant la théorie des courbes ROC, est présentée pour définir des états de santé pertinents au regard du pronostic. Ces développements sont principalement appliqués à une cohorte de patients greffés rénaux (base de données DIVAT).

Abstract

The study of the evolution of a patient constitutes an important field in clinical research. Recently, the development of the multi-state models allows to study this dynamics by taking into account several health states. In this manuscript, we use the semi-markovian models. This type of process distinguishes the durations in the states and the trajectories of the transitions, contrary to the traditional markovian approach. We proposed several adaptations to apply this type of model: the interval-censoring, the choice of the distributions of the durations and the introduction of the covariates. A goodness-of-fit statistic is also proposed to check the stationnarity assumption. Lastly, an original method, including the theory of the ROC curves, is presented to define relevant health states. These developments are mainly applied to kidney transplant recipient follow-up (DIVAT database).