REPUBLIQUE DU CAMEROUN
Paix – Travail – Patrie
****

UNIVERSITE DE YAOUNDE I
ECOLE NORMALE SUPERIEURE
DEPARTEMENT DE E MATHEMATIQUES
****

REPUBLIC OF CAMEROUN
Peace – Work – Fatherland
****

UNIVERSITY OF YAOUNDE I
HIGHER TEACHER TRAINING COLLEGE
DEPARTMENT OF MATHEMATICS
****

RISQUES COMPETITIFS ET MODELES MULTI-ETATS

Mémoire de DI.P.E.S.II de Mathématiques

Par :

TSAFACK FOTSINGLA Eric Olivier
Licencié en Mathématiques

Sous la direction
Dr.Georges NGUEFACK-TSAGUE
Chargé de cours

BIBLIOTHEQUE UNIVERSITAIRE UYI

Année Académique
2015-2016

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire de Yaoundé I. Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.
Contact : biblio.centarale.uyi@gmail.com

WARNING

This document is the fruit of an intense hard work defended and accepted before a jury and made available to the entire University of Yaounde I community. All intellectual property rights are reserved to the author. This implies proper citation and referencing when using this document.

On the other hand, any unlawful act, plagiarism, unauthorized duplication will lead to Penal pursuits.
Contact: biblio.centarale.uyi@gmail.com

♣ Dédicace ♣

Je dédie ce modeste travail à :

Mon épouse Aurelie et à mes enfants Léonce et Joyce.

♣ Remerciements ♣

Ce mémoire est le fruit d'efforts et de soutiens convergents et je voudrais ici rendre hommage aux principaux acteurs :

1. Je remercie vivement mon Directeur de mémoire ; Dr. Georges NGUEFACK-TSAGUE, pour sa disponibilité, son soutien sans relâche, ainsi que ses conseils ;
2. Mes remerciements vont à l'endroit du Directeur de l'Ecole Normale Supérieure de Yaoundé ; Pr. Nicolas Gabriel ANDJIGA pour l'ambiance mise en place durant ma formation ;
3. Mes remerciements vont aussi à l'endroit de tous les membres du jury qui ont bien voulu consacrer un peu de leurs précieux temps à l'examen de ce travail ; recevez ici toute ma gratitude et ma reconnaissance ;
4. Je tiens aussi à remercier le Chef du Département de Mathématiques ; Pr. DIFO LAMBO et tous les enseignants du Département de Mathématiques de l'Ecole Normale Supérieure de Yaoundé pour leurs encadrements durant ma formation ;
5. Ma gratitude va à l'endroit de mon père ; MBAFOU Richard pour l'amour, la confiance et le soutien qu'il m'a toujours apporté ;
6. Mes remerciements vont aussi à l'endroit de ma défunte mère MANETANE Jeannette pour son amour et son encadrement ;
7. Je voudrais également remercier tous mes frères et soeurs pour leurs soutiens ;
8. Je remercie mes oncles TATSASSI Théodore, TSOPGUE Antoine, MBAFOU Claude, NGUEFACK Romain et ma tante TATSASSI Léonie pour leurs conseils et leurs soutiens ;
9. Je remercie toutes mes grand mères MATSATEUM Sabine et MAWAMBA Pauline pour leur amour et leurs encouragements ;
10. Je trouve là le lieu de remercier aussi mes amis tels que SAMADINE, TCHAPNGA, KALDJOB, FOSSUA, DJOUMBISSIE, DJOUTSOP, MAKONG , WOUNGLI, CLAUD MIH NYANG et tous mes camarades de promotion pour les merveilleux moments que nous avons passés ensemble.

♣ Déclaration sur l'honneur ♣

Le présent travail est une oeuvre originale du candidat et n'a été soumis nulle part ailleurs, en partie ou en totalité, pour une autre évaluation académique. Les contributions externes ont été dûment mentionnées et recensées en bibliographie

Signature du candidat

TSAFACK FOTSINGLA Eric Olivier

♣ Résumé ♣

Les modèles multi-états sont de plus en plus utilisés en analyse de survie. Ils permettent de modéliser les différents états occupés par les sujets au cours du temps. Dans un modèle de risques compétitifs, qui est un cas particulier de modèle multi-états, nous avons un état initial (vivant) et plusieurs états secondaires (mort). En prenant individuellement chaque état secondaire, on définit une fonction de survie. Le problème est d'estimer, en considérant tous les états secondaires, la fonction de survie jointe. Nous avons donné une estimation de la fonction de survie jointe puis nous avons fait quelques applications dans lesquelles nous avons constaté que le séjour des patients aux urgences se voit prolonger lorsqu'ils sont mis sous ventilation et que le risque instantané de décéder aux urgences pour les patients atteints d'une pneumonie est le même que pour les patients n'ayant pas de pneumonie.

Mots-clés :

Données censurées , Fonction d'incidence cumulée ; Fonction de risque cause spécifique ; Risques compétitifs, modèles multi-états.

♣ Abstract ♣

Multistate models are increasingly used in survival analysis. They are used to model the various states occupied by the subjects over time. In a model of competing risks, that is a particular case of multistate model, we have an initial state (living) and several secondary states (death). When taking each secondary state individually, one defines a survival function. The problem is to estimate, by considering all secondary states, the joint survival function. We gave an evaluation of the joint survival function and we made some applications in which we noticed that the stay of the patients to the intensive care unit it seem to be extend when they are exposed under ventilation and the instantaneous risk to die to the intensive care unit for the patients with a pneumonia is the same than those without pneumonia.

keywords

Censored data, the cumulative incidence function, risk function specific cause, competitive risks, multi- state models.

♣ Liste des abréviations et notations ♣

Analyse de survie

n effectif

$\wedge$ minimum

$\lambda\,(.)$ fonction de risque

$\Lambda\,(.)$ fonction de risque cumulée

L vraisemblance

LCox vraisemblance partielle de Cox

T_i durée de survie du sujet i

C_i temps de censure à droite du sujet i

$\widetilde{T}_i$ durée observée du sujet i ($C_i \wedge T_i$)

R(t) Effectif à risque à l'instant t^-

Z_i vecteur des variables explicatives du sujet i

Modèles multi-états

MV maximum de vraisemblance

MVP maximum de vraisemblance pénalisée

X processus (de Markov si pas de précision supplémentaire)

$\alpha_{kl}\,(.)$ intensité de transition d'un modèle multi-état ou illness-death

$\alpha_l\,(.)$ intensité de transition d'un modèle à risques compétitifs associée à la cause l de décès

$A_{kl}(.)$ intensité de transition cumulée d'un modèle multi-état ou illness death

$A_l(.)$ intensité de transition cumulée d'un modèle à risque compétitifs associée à la cause l de décès

$p_{kl}(.,.)$ probabilités de transition de l'état k à l'état l

$p^0_{02}\,(.,.)$, $p^1_{02}\,(.,.)$ probabilités de transiter de l'état 0 à l'état 2 directement/en passant par l'état 1

F_j (t) dans un modèle à risques compétitifs, incidence cumulée associée au décès par la cause j (i.e. probabilité cumulée de décéder de la cause j)

F_{01} (s, t), F_{02} (s, t) dans un modèle illness-death et pour un sujet en l'état 0 au temps s, probabilités d'atteindre respectivement l'état 1 et l'état 2 avant t

$F_{0\bullet}$ (s, t) dans un modèle illness-death et pour un sujet en l'état 0 au temps s, probabilité de sortie de l'état 0 avant t

i.i.d indépendants et identiquement distribués

♣ Table des figures ♣

♣ Table des matières ♣

♣ Introduction générale ♣

Dans de nombreux domaines, décrire l'évolution des phénomènes dans le temps est d'un intérêt capital, en particulier pour aborder les problématiques de la prédiction et de la recherche de facteurs causaux. Les modèles multi-états constituent une alternative intéressante pour modéliser des données c'est pourquoi depuis une quarantaine d'années, les modèles multi-états ne cessent de connaître un intérêt croissant. Ces modèles utilisent la notion d'« état » et de processus pour décrire un phénomène. La notion de processus est utilisée pour représenter les différents états successivement occupés à chaque temps d'observation. En épidémiologie, ils permettent par exemple, de représenter l'évolution d'un patient à travers les différents stades d'une maladie. Après définition des différents stades, les modèles multi-états permettent d'étudier de nombreuses dynamiques complexes. L'étude de ces modèles consiste à analyser les forces de passage (intensités de transition) entre les différents états.

Un nombre important de publications statistiques concerne les modèles multi-états. Cependant, l'application de ces modèles dépasse rarement le cadre des revues spécialisées. Cette situation s'explique en partie, par l'absence de logiciels adaptés et la méconnaissance des méthodes statistiques. La popularité et la richesse des modèles de survie, en particulier du modèle de Cox, dessert l'utilisation de ces modèles dans le domaine appliqué. Il est pourtant des situations où l'étude d'un délai d'apparition d'un événement ne peut apporter qu'une réponse partielle au problème posé. Dans les modèles multi-états les plus simples, l'information sur l'état présent renseigne sur les états précédents : par exemple, les modèles progressifs , les modèles à risques compétitifs , ou encore les modèles de survie qui représentent le cas le plus simple avec uniquement deux états : « vivant » et « décès » . Cependant, dès que le modèle comprend des états réversibles (c'est-à-dire que certains événements sont récurrents), il devient nécessaire de faire des hypothèses sur l'histoire de l'individu. Les modèles de type Markovien sont très utiles car ils supposent que l'information sur les états précédents est résumée par l'état présent. Le terme de modèle multi-états regroupant de nombreuses problématiques biostatistiques, le nombre de publications sur le sujet est très important. On pourra se référer, par exemple, aux

travaux de Saint Pierre (2005), Touraine (2013), Njamen (2014) et Beyersmann et al (2012) qui font le point sur l'état de l'art dans ce domaine. Dans ces modèles de type Markovien, les intensités de transition entre les états peuvent dépendre de différentes échelles de temps, en particulier,

– la durée du suivi (temps depuis l'inclusion dans l'étude),

– le temps depuis la dernière transition (durée dans l'état présent).

Il existe plusieurs possibilités pour définir les intensités de transition $\alpha(t, d)$, où t représente la durée du suivi et d la durée passée dans l'état. Lorsque $\alpha(t, d) = \alpha$, le modèle est dit homogène par rapport au temps t. Lorsque $\alpha(t, d) = \alpha(t)$ le modèle est dit non-homogène.

Dans le cas où les intensités de transition dépendent de la durée du suivi, $\alpha(t, d) = \alpha(d)$, le modèle est semi-Markovien homogène par rapport au temps t. Enfin, lorsque $\alpha(t, d)$ dépend des deux échelles de temps, le modèle est semi-Markovien non-homogène.

Dans certaines applications, la durée du suivi n'est pas toujours l'échelle de temps la mieux adaptée. En effet, le temps calendaire et l'âge peuvent également être considérés comme échelle de temps principale. Par exemple, le temps calendaire peut être adapté quand on considère le risque de contracter une maladie qui a une incidence variant beaucoup, comme l'infection par le VIH dans les années 80. Le choix entre les échelles de temps dépend de ce qui est le plus important dans une application donnée. Plusieurs modèles statistiques sont possibles, on distingue les approches paramétrique, non-paramétrique et semi-paramétrique.

L'approche paramétrique stipule que les intensités de transition appartiennent à une classe particulière de fonctions, qui dépendent d'un nombre fini de paramètres. L'avantage de cette approche est la facilitation attendue de la phase d'estimation des paramètres. L'inconvénient est l'inadéquation pouvant exister entre le modèle retenu et le phénomène étudié.

L'approche non-paramétrique ne nécessite aucune hypothèse sur la forme des intensités de transition et c'est là son principal avantage. L'inconvénient d'une telle approche est la nécessité de disposer d'un nombre important d'observations. En effet, le problème de l'estimation d'un paramètre fonctionnel est délicat puisqu'il appartient à un espace de dimension infinie.

L'approche semi-paramétrique est une sorte de compromis entre les deux approches précédentes. Les intensités de transition appartiennent à une classe de fonctions pour partie dépendant de paramètres et pour partie s'écrivant sous forme de fonctions non-paramétriques. Cette approche est très répandue en analyse de survie au travers du modèle de régression de Cox (Therneau et Grambsch (2000)).

Le modèle peut également faire intervenir un effet aléatoire qui agit de manière multiplicative sur les intensités de transition. Dans les études de survie, ces modèles permettant de tenir compte de la dépendance entre les temps d'évènement sont appelés modèle de fragilité.

Plus généralement, ces modèles permettent de prendre en compte des variables omises dans la modélisation (par exemple, les variables non observées, celles dont les effets sont déjà bien connus ou celles dont il n'est pas certain qu'elles influencent les intensités). Ces modèles sont particulièrement intéressants quand on peut distinguer des groupes d'individus : par exemple, les personnes d'une même famille auront des caractéristiques génétiques communes. Les caractéristiques génétiques étant différentes d'une famille à l'autre, il est intéressant d'avoir un effet aléatoire spécifique à chaque famille.

Si on considère m le nombre total de risques et, pour $j = 1, ..., m, T_{\mathrm{j}}$ la durée jusqu'à l'apparition d'un évènement dû au risque j en l'absence des autres risques, le modèle des risques compétitifs postule que l'on n'observe pas toutes les variables T_j pour $j = 1, ..., m$ mais seulement leur minimum $min(T_1, T_2..., T_{\mathrm{m}})$. Dans ce cas, on supposera toujours que l'on observe en plus une variable η qui prend la valeur$j$ lorsque le minimum observé correspond à un évènement dû au risque j. L'exemple usuel de risques compétitifs est celui de la population humaine qui est soumise à plusieurs causes de mort :

un individu meurt une seule fois et par une seule cause. Les fonctions de répartition spécifiques à une cause donnée correspondant au délai jusqu'à l'avènement d'un évènement d'un type donné permettent de décrire l'évolution d'un risque donné en présence de tous les autres risques. Le terme de risques compétitifs (ou concurrents) se rapporte au domaine de l'analyse de "durée de survie" où, en plus d'un temps (ou délai) d'évènement, on observe aussi un type (ou une cause) uniquement d'évènement.

L'objectif central dans les risques compétitifs est la fonction de risque cause-spécifique d'évènement de type j, qui s'interprète comme la probabilité de survenue de l'évènement de type j dans un interval infinitésimal, sachant que cet évènement ne s'est pas encore produit au début de l'intervalle. D'une manière plus générale, le modèle à risques compétitifs est un cas particulier des modèles multi-états (Commenges (1999)) où à partir d'un état "vivant," les individus peuvent expérimenter m causes d'évènements exclusifs . Les taux de transition (ou intensité) entre chaque état sont des fonctions de risque cause-spécifique. La somme de toutes ces intensités correspond au risque global de quitter l'état "vivant".

Le premier chapitre de ce mémoire est basé sur l'analyse de survie, notion essentielle pour la compréhension des modèles multi-états. Le deuxième chapitre est une présentation des modèles multi-états. Le troisième chapitre est consacré aux risques compétitifs qui est un cas particulier des modèles multi-états. Le quatrième chapitre est consacré aux applications.

Chapitre 1

Analyse de survie

L'analyse de survie est l'étude du délai de survenue d'un évènement d'intérêt. Cet évènement est souvent associé à un changement d'état, communément le passage de l'état « vivant » à l'état « décédé ». Cependant, on s'intéresse souvent à d'autres types de délais que la durée de vie proprement dite : la durée jusqu'à l'apparition d'une maladie, le délai entre la prise d'un traitement et la guérison d'une maladie, la durée de séropositivité sans symptômes de patients infectés par le VIH, ou encore en fiabilité, la durée de fonctionnement d'une machine.

1.1 Notions générales

Soit T la variable aléatoire positive et continue qui représente la durée de survie ou délai, c'est-à-dire la durée écoulée jusqu'à la survenue de l'évènement d'intérêt. Pour définir cette durée, il faut définir une date d'origine qui est généralement propre aux sujets et dont le choix va dépendre de l'évènement d'intérêt. Dans un contexte d'essais cliniques par exemple, si l'on souhaite comparer deux traitements, on choisira la date de mise sous traitement comme date d'origine. Lorsque l'évènement étudié est très dépendant de l'âge, on choisit souvent la date de naissance comme date d'origine et la variable T est alors un âge.

1.1.1 Censure et troncature

La difficulté en analyse de survie réside dans le fait que les données recueillies sont en partie incomplètes.

Censure à droite

Le phénomène le plus souvent à l'origine de ces données incomplètes est la censure à droite. La durée de survie du sujet i, T_i , est dite censurée à droite si le sujet i n'a pas subi l'évènement

à sa date de dernières nouvelles C_i , c'est-à-dire que la seule information dont on dispose est que $T_i > C_i$.

Généralement, dans des données de cohorte, une durée T_i est censurée à droite si le sujet i est :

— perdu de vue : sa surveillance est interrompue alors qu'il n'a pas encore subi l'évènement (pour cause de déménagement par exemple) ;

— exclu vivant : à la date de fin d'étude le sujet n'a pas encore subi l'évènement . De façon formelle, on associe à chaque sujet i la variable aléatoire $\widetilde{T_i}$:

$$\widetilde{T_i} = T_i \wedge C_i$$

qui est la durée réellement observée et un indicateur $\delta_i = 1_{\{T_i \leq C_i\}}$ tel que :

$$\delta_i = \begin{cases} 1 \text{ si la « vraie» durée est observée (dans ce cas } \widetilde{T_i}=T_i) \\ 0 \text{ si la durée est censurée à droite (dans ce cas } \widetilde{T_i}=C_i \text{)} \end{cases}$$

Dans les modèles classiques d'analyse de survie, on fait l'hypothèse que les variables T_i et C_i sont indépendantes, c'est-à-dire que la censure est indépendante de l'évènement. Lorsque cette hypothèse n'est pas vérifiée, par exemple quand la censure est due à un arrêt du traitement ou lorsque les sujets les plus malades et donc les plus à risque de décéder sont perdus de vue, on parle de censure informative.

Censure par intervalle

La durée de survie T_i est dite censurée par intervalle si au lieu de l'observer de façon exacte, la seule information dont on dispose est qu'elle est comprise entre deux dates connues. La censure par intervalle se rencontre généralement dans les études de cohorte lorsque les sujets ne sont pas observés en temps continu mais par intermittence lors de visites. Par exemple, si l'on s'intéresse à l'âge de survenue d'une maladie et que le sujet i est diagnostiqué malade au cours d'une visite, on sait seulement que $T_i \in [L_i , R_i]$ où R_i est l'âge à la visite de diagnostic et L_i est l'âge à la visite précédente. La durée T_i est dite doublement censurée par intervalle lorsqu'elle représente un délai entre deux variables aléatoires censurées par intervalle. On trouve souvent dans la littérature l'exemple où T_i représente le délai entre l'infection par le VIH et le début du SIDA.

Troncature à gauche

La durée de survie T est dite tronquée si son observation est conditionnelle à un autre évènement. La durée de survie T est tronquée à gauche si elle n'est observable qu'à la condition T > A, où A est une variable que l'on suppose indépendante de T . S'il y a troncature à gauche, on n'étudie que le sous-échantillon des sujets dont la durée de survie est supérieure à une certaine valeur. Par exemple, lorsqu'on étudie l'âge de décès et que les sujets ne sont pas suivis depuis leur date de naissance, les données sont tronquées à gauche puisque seuls les sujets vivants à la date d'inclusion sont observables et A représente leur âge à l'inclusion.

1.1.2 Distribution de la durée de survie

On suppose que la durée de survie T est une variable positive ou nulle, et absolument continue. La distribution de T est caractérisée par l'une des cinq fonctions suivantes définies pour t $\geq$ 0, chacune pouvant être obtenue à partir de l'une des autres.

Définition 1.1. La fonction de survie S au temps t est la probabilité de survie jusqu'au temps t :

$$S(t) = P(T > t)$$

Propriété : Soit S la fonction de survie de la variable aléatoire T. On a :

- S(t) $\in$ [0, 1], t > 0 ;
- S est monotone décroissante ; i.e. $t_1 < t_2 \Rightarrow S(t_1) \geq S(t_2)$;
- $S(0^+) = 1$ et $\lim_{t\to\infty} S(t) = 0$.

Définition 1.2. La fonction de répartition F au temps t est la probabilité de subir l'évènement avant le temps t :

$$F(t) = P(T \leq t) = 1 - S(t)$$

Propriété : Soit F la fonction de répartition de la variable T. on a :

- F (t) $\in$ [0, 1], t > 0 ;
- F est continue à droite, i.e $F(t^+) = F(t)$;
- F est monotone non-décroissante, i.e. $t_1 > t_2 \Rightarrow F(t_1) \geq F(t_2)$;
- $F(0^+) = 0$ et $\lim_{t\to\infty} F(t) = 1$.

Définition **1.3.** La densité de probabilité f au temps t représente la probabilité instantanée de subir l'évènement dans un petit intervalle de temps après t :

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = F'(t) = -S'(t)$$

Elle est telle que :

$$F(t) = \int_0^t f(u)\, du$$

Définition **1.4.** La fonction de risque λ au temps t représente la probabilité de subir l'évènement dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'à t :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Propriété : Soit λ la fonction de risque au temps t. On a :

- $\lambda : \mathrm{R}_+ \to \mathrm{R}_+$;
- $\int_0^t \lambda(u)\, du < \infty$, $\forall$ t > 0 mais $\int_0^\infty \lambda(u)\, du = \infty$;
- λ n'est pas nécessairement monotone.

Définition **1.5.** La fonction de risque cumulé est :

$$\Lambda(t) = \int_0^t \lambda(u)\, du = -ln(S(t))$$

Propriété :

$\Lambda(t)$ vaut $+\infty$ lorsque $S(t) = 0$

Remarque 1.1 La fonction de survie peut aussi s'exprimer en fonction du risque cumulé :

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(u)\, du\right\} \qquad (1.1)$$

1.1.3 Estimation non paramétrique

Dans le cas où l'on ne fait pas d'hypothèse a priori sur la distribution de la durée de survie T, les principaux estimateurs non paramétriques sont l'estimateur de Kaplan-Meier de la fonction de survie et l'estimateur de Nelson-Aalen du risque cumulé.

Estimateur de Kaplan-Meier de la fonction de survie

L'estimateur de Kaplan-Meier découle de l'idée suivante : ne pas avoir subi l'évènement à l'instant t, c'est ne pas l'avoir subi juste avant t et ne pas le subir en t. Notons t' le temps « juste avant t » et t" le temps « juste avant t' ». On a :

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(T > t', T > t) \\ &= P(T > t | T > t') P(T > t') \\ &= P(T > t | T > t') P(T > t' | T > t'') P(T > t'') \\ &= ... \end{aligned}$$

En considérant les durées observées (temps d'évènement ou temps de censure) rangées par ordre croissant, $\widetilde{T}_i$, i = 1, . . . , n, et en les supposant distinctes, et avec $\widetilde{T}_0 = 0$, on a :

$$P(T > \widetilde{T}_i) = \prod_{j=1}^{i} p_j \text{ , i = 1,.....,n}$$

où p_j = P (T > $\widetilde{T}_j$ |T>$\widetilde{T}_{j-1}$) est la probabilité de ne pas subir l'évènement dans l'intervalle [$\widetilde{T}_{j-1}$, $\widetilde{T}_j$] sachant qu'il ne s'est toujours pas produit en $\widetilde{T}_{j-1}$.

Considérons $R(\widetilde{T}_j)$ l'effectif à risque à l'instant $\widetilde{T}_j^-$, c'est-à-dire le nombre de sujets n'ayant pas encore subi ni l'évènement ni la censure à droite juste avant $\widetilde{T}_j$. Rappelons que $\delta_j = 1$ si le sujet j a subi l'évènement ($\widetilde{T}_j = T_j$) ; $\delta_j = 0$ si sa durée de survie est censurée à droite ($\widetilde{T}_j = C_j$). Un estimateur naturel pour p_j est :

$$\widetilde{p_j} = \frac{R(\widetilde{T}_j) - \delta_j}{R(\widetilde{T}_j)} = 1 - \frac{\delta_j}{R(\widetilde{T}_j)}$$

et on obtient l'estimateur de Kaplan-Meier :

$$\widehat{S}_{KM}(t) = \prod_{j:\widetilde{T}_j \leq t} \left(1 - \frac{\delta_j}{R(\widetilde{T}_j)}\right)$$

Dans le cas où il y a des ex-aequo, c'est-à-dire que tout les $\widetilde{T}_j$ ne sont pas distincts, on note D($\widetilde{T}_j$) le nombre de sujets subissant l'évènement au temps $\widetilde{T}_j$ et l'estimateur de Kaplan-Meier devient :

$$\widehat{S}_{KM}(t) = \prod_{j:\widetilde{T}_j \leq t} \left(1 - \frac{D(\widetilde{T}_j)}{R(\widetilde{T}_j)}\right)$$

Estimateur de Nelson-Aalen du risque cumulé

Nous avons abordé jusqu'ici l'analyse de survie en considérant la variable aléatoire de durée T mais elle peut aussi être abordée en terme de processus en considérant le processus ponctuel qui vaut 0 tant que l'évènement n'a pas lieu et 1 après. L'estimateur de Nelson-Aalen a été introduit par Aalen (1978) pour généraliser celui de Nelson (1972) aux processus de comptage. Il est donné par :

$$\widehat{\Lambda}(t) = \sum_{j:T_j \leq t} \frac{D(T_j)}{R(T_j)}$$

L'estimateur du risque cumulé a une interprétation moins immédiate que celui de la fonction de survie. Son intérêt réside surtout dans la pente de la courbe correspondante qui estime la fonction de risque λ.

Les estimateurs de Kaplan-Meier et de Nelson-Aalen se généralisent aux données tronquées à gauche mais pas aux données censurées par intervalle car dans ce dernier cas, les temps exacts d'évènements ne sont pas connus.

1.1.4 Modèles paramétriques

Supposons maintenant que la distribution des durées de survie appartient à une famille de loi paramétrique donnée. Bien que chacune des cinq fonctions S , F , f , λ , Λ caractérise la loi de T , on spécifie souvent la forme de la fonction de risque λ qui donne la description la plus intéressante, à savoir celle du futur immédiat du sujet qui n'a pas encore subi l'évènement.

Loi exponentielle

La loi exponentielle, qui ne dépend que d'un paramètre θ, est la seule distribution continue qui admet un risque instantané constant. Pour t $\succeq$ 0 et avec $\theta > 0$:

$$\begin{aligned} \lambda(t) &= \theta \\ S(t) &= e^{-\theta t} \\ f(t) &= \theta e^{-\theta t} \end{aligned}$$

Cette loi est dite « sans mémoire » et est peu adaptée dans le domaine du vivant. Cependant, dans certaines applications, on peut découper le temps en plusieurs intervalles et considérer un

risque constant sur chacun des intervalles (et différent d'un intervalle à l'autre) de façon à obtenir une fonction de risque constante par morceaux.

Loi de Weibull

La loi de Weibull, qui dépend d'un paramètre de forme a et d'un paramètre d'échelle b, admet un risque instantané monotone. C'est une généralisation de la loi exponentielle que l'on retrouve en prenant $a = 1$. Pour t $\succeq 0$ et avec $a > 0$ et $b > 0$:

$$\begin{aligned} \lambda(t) &= a\left(\frac{1}{b}\right)^a t^{a-1} \\ S(t) &= \exp\left\{-\left(\frac{t}{b}\right)^a\right\} \\ f(t) &= a\left(\frac{1}{b}\right)^a t^{a-1}\exp\left\{-\left(\frac{t}{b}\right)^a\right\} \end{aligned}$$

Il existe d'autres lois avec des risques instantanés monotones. Citons notamment la loi Gamma et la loi de Gompertz.

D'autres lois permettent aussi de modéliser des risques instantanés en forme de cloche ($\cup$ ou $\cap$), en particulier la loi de Weibull généralisée.

Vraisemblance

Soit θ le vecteur des paramètres du modèle. Les estimateurs des paramètres sont obtenus en maximisant la vraisemblance détaillée ci-après. En pratique, on utilise des méthodes itératives de type algorithme de Newton-Raphson.

La vraisemblance représente la probabilité d'observer l'échantillon d'après le modèle et est le produit des n contributions individuelles :

$$L = \prod_{i=1}^{n} L_i$$

Soit t_i le temps de participation du sujet i. Dans le cas le plus fréquent de données censurées à droite, la contribution du sujet i à la vraisemblance est :

— $f(t_i;\theta)$ si $\delta_i = 1$ (le sujet i a subi l'évènement)

—$S(t_i;\theta)$ si $\delta_i = 0$ (l'observation du sujet i est censurée à droite)

et la vraisemblance s'écrit :

$$\begin{aligned} L &= \prod_{i=1}^{n} f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i} \\ &= \prod_{i=1}^{n} S(t_i;\theta)\lambda(t_i;\theta)^{\delta_i} \end{aligned}$$

Dans le cas de données tronquées à gauche, la contribution individuelle d'un sujet i dont l'observation est tronquée à gauche en a_i est :

$$L_i = \begin{cases} \frac{f(t_i;\theta)}{S(a_i;\theta)} & si\ \delta_i = 1 \\ \\ \frac{S(t_i;\theta)}{S(a_i;\theta)} & si\ \delta_i = 0 \end{cases}$$

Dans le cas de données censurées par intervalle, la contribution individuelle d'un sujet i dont l'observation est censurée dans l'intervalle $[l_i, r_i]$ est :

$$L_i = \begin{cases} S(r_i;\theta) - S(l_i;\theta)\ sans\ troncature\ à\ gauche \\ \\ \frac{S(r_i;\theta)-S(l_i;\theta)}{S(a_i;\theta)}\ \ avec\ troncature\ à\ gauche \end{cases}$$

Remarque 1.2. Avantages et inconvénients des approches paramétrique et non paramétrique

L'approche paramétrique induit des hypothèses sur la distribution des données mais elle a l'avantage de fournir des estimations en temps continu de n'importe quelle fonction caractérisant la distribution. L'approche non paramétrique fournit des estimations en temps discret et donc les fonctions estimées sont continues par morceaux. La fonction de risque étant la plus intéressante en terme d'interprétation, un lissage a posteriori de l'estimateur de Nelson-Aalen est envisageable en utilisant une méthode à noyau (Ramlau-Hansen (1983)). Une autre approche pour estimer des fonctions lisses sans faire d'hypothèse paramétrique est d'utiliser des fonctions splines (Rosenberg (1995)). Ce type d'approche et l'approche paramétrique, étant basés sur la vraisemblance, ont l'avantage de prendre en compte aisément des données censurées par intervalle.

1.2 Prise en compte des facteurs de risque

Nous nous sommes concentrés jusqu'ici sur le délai jusqu'à la survenue de l'évènement d'intérêt mais l'un des principaux objectifs de l'analyse de survie et de l'épidémiologie est d'évaluer l'impact sur ce délai de facteurs auxquels sont exposés les sujets.

1.2.1 Modèles de régression

Les deux principaux types de modèles qui permettent d'exprimer le risque d'évènement en fonction de variables explicatives sont : Les modèles à temps de vie accélérée et les modèles à risques proportionnels. Dans un modèle à durée de vie accélérée, une variable a pour effet de multiplier la durée de survie par une constante tandis que dans un modèle à risques proportionnels, une variable a pour effet de multiplier le risque instantané par une constante. Le modèle à risques proportionnels s'inscrit plus généralement dans la famille des modèles multiplicatifs. Les modèles additifs constituent une alternative aux modèles multiplicatifs : Au lieu de faire le produit entre la fonction de risque de base et une fonction des variables explicatives, on en fait la somme.

Modèle de Cox

Le modèle de régression le plus largement utilisé est le modèle à risques proportionnels de Cox (1972) :

$$\lambda(t|Z_i) = \lambda_0(t)e^{\beta^T Z_i} \qquad (1.2)$$

où i est l'indice du sujet,

Z_i est le vecteur des variables explicatives,

β est le vecteur des coefficients de régression,

λ_0 est la fonction de risque de base, c'est-à-dire le risque instantané des sujets pour lesquels toutes les variables explicatives sont nulles.

Le risque de survenue d'évènement à l'instant t pour un sujet qui a pour caractéristiques Z_i par rapport à un sujet qui a pour caractéristiques Z_j est :

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = \frac{\lambda_0(t)e^{\beta^T Z_i}}{\lambda_0(t)e^{\beta^T Z_j}} = \frac{e^{\beta^T Z_i}}{e^{\beta^T Z_j}} = e^{\beta^T(\ Z_i - Z_j)}$$

Dans le modèle de Cox, le rapport des risques instantanés est constant au cours du temps. C'est dans ce sens que le modèle de Cox est dit à risques proportionnels. La proportionnalité des risques est une conséquence du modèle mais aussi une hypothèse à vérifier à posteriori.

Si la variable Z_i est une variable quantitative ou qualitative ordonnée à plus de deux modalités, il conviendra aussi de vérifier une hypothèse de log-linéarité. En effet, quand on ajoute une unité à la valeur de Z_i , on multiplie le risque instantané par e^{β} (c'est-à-dire qu'on ajoute β à son logarithme) quelle que soit la valeur de Z_i .

Vraisemblance partielle

Afin d'estimer les paramètres de régression β, l'idée de Cox a été de considérer le risque de base λ_0 comme un paramètre de nuisance en maximisant une vraisemblance dite partielle.

Soit D le nombre de sujets ayant subi l'évènement. Soient T_i , i = 1, . . . , D les différents temps d'évènement supposés distincts et rangés par ordre croissant avec i les indices des sujets correspondant. Notons $R(T_i)$ l'ensemble des sujets encore à risque à l'instant T_i^-.

La probabilité que le sujet i subisse l'évènement en T_i sachant qu'il y a eu un évènement en T_i s'écrit :

$$p_i = \frac{\lambda_0(T_i)e^{\beta^T Zi}}{\sum\limits_{j\in R(T_i)} \lambda_0(T_i)e^{\beta^T Zj}} = \frac{e^{\beta^T Zi}}{\sum\limits_{j\in R(T_i)} e^{\beta^T Zj}}$$

Le produit sur les temps d'évènement des termes p_i qui ne dépendent que du paramètre β définit alors la vraisemblance partielle de Cox :

$$L_{Cox} = \prod_{i=1}^{D} \frac{e^{\beta^T Zi}}{\sum\limits_{j\in R(T_i)} e^{\beta^T Zj}}$$

L'estimateur de β est le vecteur des paramètres qui maximisent cette vraisemblance.

Remarque 1.3.

1. Naturellement, d'autres méthodes n'utilisant pas la vraisemblance partielle de Cox peuvent être utilisées pour estimer les coefficients β dans l'équation (1.2), par exemple en spécifiant un modèle paramétrique pour la fonction de risque de base λ_0 .

2. Par extension, lorsqu'on parle du modèle de Cox, on désigne souvent non seulement l'équation (1.2) mais aussi la méthode d'estimation des effets des facteurs de risque par vraisemblance partielle.

De la même façon que les estimateurs non paramétriques, la vraisemblance partielle de Cox se généralise aux données tronquées à gauche mais pas aux données censurées par intervalle.

1.2.2 Modèles à fragilité

Modèles simples à fragilité

Toutes les variables pertinentes ne sont pas toujours incluses dans un modèle de régression, soit parce qu'elles n'ont pas été mesurées, soit parce qu'elles ne sont pas suspectées être liées à

l'évènement d'intérêt. Ces variables omises peuvent créer une sélection de la population au cours du suivi : Les sujets les plus fragiles subissent l'évènement plus tôt que les autres, entraînant au fil du temps une modification de la structure de la population observée. Ignorer ces variables peut engendrer un biais dans l'estimation de la fonction de risque (sous estimation) . La notion de fragilité traduit le fait que certains individus sont plus susceptibles de subir l'évènement, et donc sont plus « fragiles » que d'autres. La fragilité est représentée par un effet aléatoire qui, comme les variables explicatives, agit de façon multiplicative sur le risque instantané. Le modèle à fragilité est donc une extension du modèle à risques proportionnels de Cox :

$$\lambda_j(t|\omega_j, Z_j) = \lambda_0(t)\omega_j e^{\beta^T Z_j} \qquad (1.3)$$

où j est l'indice du sujet et ω_j est un effet aléatoire spécifique à chaque sujet, appelé variable à fragilité.

Modèles à fragilité partagée

Les modèles à fragilité sont aussi utilisés pour modéliser des données de survie corrélées telles que les données répétées et les données groupées . Dans le cas de données répétées, les individus correspondent aux différentes observations d'un même sujet ; dans le cas de données groupées qui est celui qui nous intéressera par la suite, ils correspondent aux différentes observations des sujets appartenant à un même groupe. Le modèle à fragilité partagée s'écrit :

$$\lambda_{ij}(t|\omega_i, Z_{ij}) = \lambda_0(t)\omega_i e^{\beta^T Z_{ij}} \qquad (1.4)$$

où i est l'indice du groupe, j l'indice du sujet, ω_i un effet aléatoire spécifique au groupe i. Plusieurs distributions peuvent être utilisées pour ω. La plus courante est la loi Gamma car elle possède des propriétés mathématiques entraînant la simplification du calcul de la vraisemblance, ce qui évite de recourir à des méthodes d'intégration gourmandes en temps de calcul. Cette loi a cependant quelques propriétés indésirables (Hougaard, 2000, p. 256). En particulier, la non proportionnalité des risques pourrait avoir une influence plus grande sur les estimations que celle de la corrélation des données. La distribution log-normale est aussi largement utilisée. Contrairement à la loi Gamma, elle ne permet pas de simplification de la vraisemblance mais se révèle très pratique dans un contexte multivarié avec plusieurs effets aléatoires non indépendants.

Dans les cas d'une distribution Gamma ou log-normale, la fragilité ω doit être d'espérance égale à 1 et de variance finie pour des questions d'identifiabilité. La variance est un paramètre à

estimer et exprime l'hétérogénéité des données. Remarquons que l'équation 1.4 peut se réécrire :

$$\lambda_{ij}(t|U_i, Z_{ij}) = \lambda_0(t)e^{\beta^T Z_{ij}+U_i}$$

avec $\omega_i = e^{U_i}$. Dans le cas log-normal, la variable U suit une loi normale $N(0, \sigma^2)$ où σ^2 est le paramètre de variance à estimer.

Tester la significativité de l'effet aléatoire revient à tester si sa variance est significativement non nulle : $H_0 : \sigma^2 = 0$. La valeur 0 étant à la frontière de l'espace des paramètres, le test de rapport de vraisemblance basé sur une distribution asymptotique du χ^2 n'est pas applicable.

Chapitre 2

Modèles multi-états

L'analyse de survie étudie le délai de survenue d'un évènement d'intérêt, ou autrement dit, le délai entre deux états successifs (communément état « vivant » et état « décédé »). Les modèles multi-états, aujourd'hui très populaires, permettent d'étudier des dynamiques plus complexes en utilisant la notion de processus pour représenter l'évolution d'un sujet à travers différents états successifs . En épidémiologie, ils permettent par exemple de modéliser son évolution à travers les différents stades d'une maladie. Un état peut être transitoire ou absorbant lorsqu'on y reste avec une probabilité égale à 1. Les figures 2.1, 2.2 et 2.3 représentent trois exemples simples de modèles multi-états à trois états :

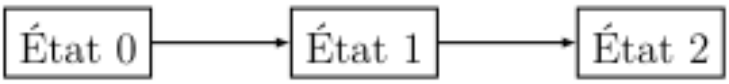

FIGURE 2.1 – Modèle progressif à 3 états

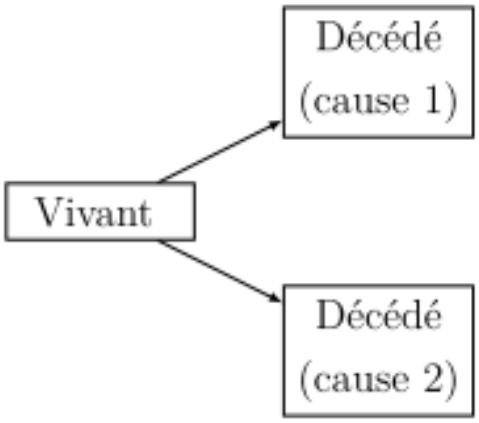

FIGURE 2.2 – Modèle à 2 risques compétitifs

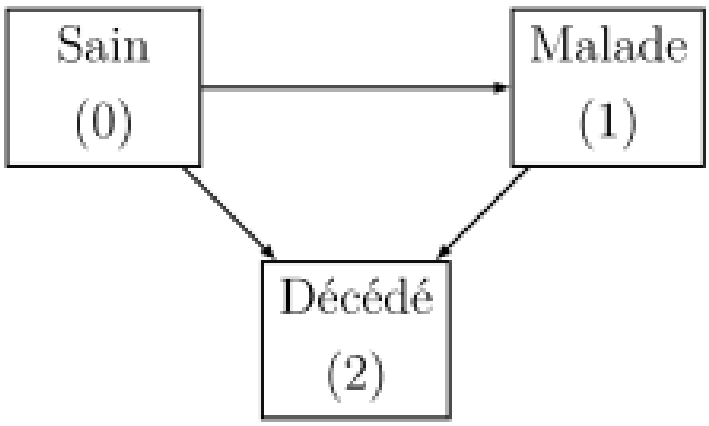

FIGURE 2.3 – Modèle illness-death

Nous avons :

— un modèle progressif : les sujets transitent de l'état initial (0) à l'état absorbant (2) en passant obligatoirement par l'état transitoire (1) (figure 2.1) ;

— un modèle illness-death : les sujets transitent de l'état initial (0) à l'état absorbant (2) directement ou en passant par l'état transitoire (1) (figure 2.3) ;

— un modèle à risques compétitifs (concurrents) : les sujets peuvent transiter vers deux états absorbants qui correspondent typiquement à deux causes de décès (figure 2.2).

Définition 2.1.

Soit $X = \{X(t), t \geq 0\}$ un processus à temps continu et à espace d'états fini i.e. $X(t) \in S = \{0, 1, ..., K\}$ où $K+1$ est le nombre d'états possibles. X peut être caractérisé par les probabilités de transition entre les différents états :

$$p_{kl}(s;t) = P(X(t) = l | X(s) = k\ ; \mathcal{H}_{s^-}) \qquad k, l \in \mathcal{S}$$

où $\mathcal{H}_{s^-}$ représente l'histoire du processus générée par $\{X(u), u < s\}$, ou, par ses intensités de transition :

$$\alpha_{kl}(t) = \lim_{\Delta t \longrightarrow 0} \frac{P_{kl}(t, t + \Delta t)}{\Delta t} \qquad k, l \in \mathcal{S}$$

qui sont le pendant de la fonction de risque dans un modèle de survie. Les intensités de transition cumulées sont définies par :

$$A_{kl}(t) = \int_0^t \alpha_{kl}(u)\, du$$

X est un processus de Markov si l'évolution future du processus dépend uniquement de la connaissance du temps présent s et de l'état en ce temps X(s) (le passé peut être oublié) :

$$p_{kl}(s;t) = P(X(t) = l | X(s) = k) \qquad k, l \in \mathcal{S}$$

On définit un processus de Markov homogène en supposant en plus que les probabilités de transition dépendent uniquement du délai entre chaque observation et non du temps auquel se produisent ces observations i.e. du délai depuis le temps d'origine :

$$p_{kl}(s;t) = p_{kl}(0;t-s)$$

Les intensités de transition sont alors constantes : $\alpha_{kl}(t) = \alpha_{kl}$.

Ce cas particulier est particulièrement intéressant car on peut déduire de l'équation de Chapman-Kolmogorov (donnée par $p_{kl}(s;t) = \sum_{i \in S} p_{ki}(s;u)\, p_{il}(u;t)$) une relation simple entre la matrice des probabilités de transition P et celle des intensités de transition Q (Cox and Miller, 1965).

En effet, de l'équation de Chapman-Kolmogorov, se déduisent des équations différentielles appelées équations arrière (backward) et avant (forward) de Kolmogorov :

$$\frac{\partial P(0,t)}{\partial t} = QP(0,t) \;\; ; \;\; \frac{\partial P(0,t)}{\partial t} = P(0,t)Q$$

Avec la condition initiale $P(0,0) = I$ ($i.e\ p_{kk}(0,0) = 1\ et\ p_{kl}(0,0) = 0, k \neq l$), la solution de l'une ou de l'autre est :

$$P(0,t) = e^{Qt} \qquad (2.1)$$

On peut relâcher un peu l'hypothèse d'homogénéité qui est très forte en considérant un processus de Markov homogène par périodes et donc des intensités de transition constantes par morceaux (constantes sur la même période et différentes d'une période à l'autre). Dans ce cas, on peut encore faire le lien entre intensités et probabilités de transition grâce aux équations de Kolmogorov qui sont alors généralement résolues de façon numérique.

Pour les processus simples, il est possible d'avoir une expression explicite des probabilités de transition en fonction des intensités de transition en utilisant des intégrales . Il est alors possible de considérer un processus de Markov non homogène avec des intensités de transition beaucoup plus flexibles. Un processus de **semi-Markov** est un processus dont l'évolution future dépend du temps auquel la dernière transition a eu lieu en plus du temps présent s et de l'état en ce temps $X(s)$:

$$p_{kl}(s;t) = P(X(t) = l | X(s) = k\ ; t_k) \qquad k, l \in \mathcal{S}$$

où t_k est le temps d'entrée dans l'état k. Par rapport à un processus de Markov, on tient compte en plus du temps passé dans l'état actuel. Les intensités de transition sont de la forme $\alpha_{kl}(t, t - t_k)$. Lorsqu'on ne tient compte que du temps passé dans l'état actuel et plus du temps proprement dit, on obtient un processus de semi-Markov homogène (dans le temps) avec des intensités de transition de la forme $\alpha_{kl}(t - t_k)$.

2.1 Prise en compte des facteurs de risque

Dans un modèle multi-état, les intensités de transition permettent de prendre en compte des variables explicatives de façon similaire à la fonction de risque en analyse de survie. Ainsi, le modèle qui découle du modèle à risques proportionnels de Cox est un modèle à intensités de transition proportionelles :

$$\alpha_{kl}\left(t|Z_{kl}^{(i)}\right) = \alpha_{0,kl}(t)\, e^{\beta_{kl}^T Z_{kl}^{(i)}} \qquad (2.2)$$

où i est l'indice du sujet ,$Z_{kl}^{(i)}$ et β_{kl} les vecteurs des variables explicatives et des coefficients de régression associés à la transition $k \to l$, $\alpha_{0,kl}$ l'intensité de transition de base associée à la transition $k \to l$.

2.2 Modèles et estimation

Dans cette partie , nous ne considérons que des modèles de Markov, c'est-à-dire des modèles qui font l'hypothèse que le processus sous-jacent au modèle est Markovien. De plus, nous considérons des modèles de Markov non homogènes dans le temps, plus plausibles d'un point de vue épidémiologique. Le choix de la méthode d'estimation d'un modèle multi-état et l'ajout éventuel d'hypothèses supplémentaires à l'hypothèse Markovienne dépendent beaucoup du schéma d'observation des données. En effet, on étudie un processus à temps continu mais l'observation des transitions de passage d'un état à l'autre ne se fait pas toujours en temps continu.

2.2.1 Observations en temps continu

Le cas le plus simple est celui où tous les temps de transition sont connus de façon exacte. Intéressons-nous d'abord à un modèle multi-état sans variables explicatives. Grâce au travail fondamental d'Aalen (1975, 1978), les techniques classiques d'analyse de survie ont pu se généraliser aux modèles multi-états dans le contexte de la théorie des processus de comptage (Andersen et Borgan, 1985).

Soit $N_{kl}(t)$ le processus de comptage du nombre de transitions $k \to l$ qui ont eu lieu dans l'intervalle de temps$[0, t]$. Soit $Y_k(t)$ le processus relatif au nombre de sujets à risque au temps t^- pour la transition $k \to l$. L'indice l est omis car le nombre de sujets à risque est le même pour toutes les transitions au départ de l'état k et correspond au nombre de sujets en l'état k au temps t^- . L'estimateur de Nelson-Aalen de la fonction d'intensité cumulée pour la transition $k \to l$, $A_{kl}(t)$, est donné par :

$$\hat{A}_{kl}(t) = \int_0^t \frac{dN_{kl}(u)}{Y_k(u)} = \sum_{t_j \leq t} \frac{m_{kl}(t_j)}{Y_k(t_j)}$$

où les t_j sont les temps d'évènement et m_{kl} est le nombre de transitions $k \to l$ au temps t_j .

Les incréments de l'estimateur de Nelson-Aalen peuvent fournir un estimateur des intensités de transition : $d\hat{A}_{kl}(t) = \frac{dN_{kl}(t)}{Y_k(t)}$. En lissant ces incréments, il est possible d'obtenir des estimations lisses des intensités de transition (Ramlau-Hansen, 1983).

En généralisant l'estimateur de Kaplan-Meier aux modèles de Markov non homogène, Aalen et Johansen (1978) ont proposé un estimateur des probabilités de transition, l'estimateur dit de Aalen-Johansen. Celui-ci peut être également vu comme le produit intégral de la matrice des estimateurs de Nelson-Aalen (Gill et Johansen, 1990).

Intéressons-nous maintenant à un modèle multi-état avec des variables explicatives qui interviennent grâce à des modèles de régression définis par l'équation (2.2). Les estimateurs de Nelson-Aalen et de Aalen-Johansen se généralisent. On peut obtenir des estimations pour les paramètres de régression spécifiques à chaque transition $k \to l$ en considérant autant de modèles de Cox qu'il y a de transitions . Un modèle de Cox sur la transition $k \to l$ doit être estimé sur le sous-échantillon des sujets à risque pour cette transition. Par exemple, dans un modèle illness-death, l'échantillon entier sera utilisé pour estimer β_{01} et β_{02} mais seulement le sous-échantillon des sujets ayant fait la transition $0 \to 1$ au cours du suivi sera utilisé pour estimer β_{12}.

En pratique, le paquet R mstate (De Wreede et al., 2010, 2011) peut être utilisé pour analyser des modèles markoviens multi-états, avec ou sans variables explicatives, lorsque les données sont observées en temps continu. Il permet l'estimation des intensités de transition et des éventuels paramètres de régression mais aussi de faire des prévisions en estimant les probabilités de transition.

Remarque 2.1. Sur l'estimateur de Kaplan-Meier

Il faut être prudent avec l'utilisation de l'estimateur de Kaplan-Meier dans un contexte multi-état. Prenons l'exemple d'un modèle à deux risques compétitifs (voir figure 2.2). La relation directe qui lie la fonction de risque cumulée et la fonction de survie (voir équation 1.1)

n'est plus valable dans ce contexte. Il en résulte que l'estimateur naïf de Kaplan-Meier de la fonction de survie calculé en censurant à droite les sujets ayant subi l'évènement concurrent (à l'évènement d'intérêt) est biaisé . Plus précisément, il sous-estime la fonction de survie. En effet, en analyse de survie, l'évènement d'intérêt est dit « de mortalité toutes causes » , c'est-à-dire que si le suivi était suffisamment long, l'évènement surviendrait avec une probabilité 1. Ce n'est plus le cas dans un contexte de risques compétitifs où la réalisation de l'évènement concurrent empêche la réalisation de l'évènement d'intérêt. La fonction de survie dépend à la fois du risque associé à l'évènement d'intérêt et du risque associé à l'évènement concurrent : $S(t) = e^{-A_1(t)-A_2(t)}$ où A_1 est la fonction de risque cumulé « cause 1-spécifique » et A_2 est la fonction de risque cumulé « cause 2-spécifique ». Elle peut être estimée correctement grâce à l'estimateur de Kaplan-Meier de la fonction de survie dite « globale ». Cet estimateur de Kaplan-Meier est en fait le produit des estimateurs de Kaplan-Meier naïfs associés à chaque évènement.

Sous une hypothèse d'indépendance des deux risques compétitifs, l'estimateur de Kaplan-Meier naïf associé à la cause 1 de décès correspondrait à la survie d'une population hypothétique dans laquelle la cause 2 de décès n'existe pas (et vice versa). Cependant, une telle hypothèse n'est pas vérifiable sur la base des données et est souvent peu crédible d'un point de vue biologique.

Remarque 2.2. Approche « stochastic process » vs. approche « latent failure times »

Dans ce manuscrit, nous nous plaçons toujours dans une approche multi-état avec un processus stochastique sous-jacent. Une autre approche que l'on trouve dans la littérature des modèles à risques compétitifs consiste à considérer des temps d'évènement latents.

1. Plaçons-nous dans le modèle à deux risques compétitifs de la figure (2.2) et considérons les deux variables aléatoires T_1 (temps de décès par la cause 1) et T_2 (temps de décès par la cause 2) telles que :

$$T_j = \inf_{t>0}(X(t) = j) \ , j = 1, 2$$

dont la distribution est donnée par les fonctions « d'incidence cumulée » :

$$F_j(t) = P(T \leq t, D = j)$$

où T = min(T_1, T_2) et D est la cause de décès.

On remarque que F_1 et F_2 ne sont pas des fonctions de répartition ; car en effet au lieu d'avoir $\lim_{t\to\infty} F_j(t) = 1 \quad (j = 1, 2)$, on a plutôt $\lim_{t\to\infty} F_j(t) = P(D = j) < 1 \quad (j = 1, 2)$. Ceci

implique que les variables aléatoires T_1 et T_2 ne sont pas définies correctement car par exemple pour un sujet i tel que $D_i = 2$ on a : $T_1 = \infty$.

En l'absence de risques compétitifs, la distribution de T_j serait donnée par :

$$\widetilde{F}_j(t) = 1 - \exp\{-\int_0^t \alpha_j(u)\, du\}$$

Cette fonction est aussi la distribution de la variable latente mais définie correctement $\widetilde{T}_j$ qui est le temps de décès par la cause j dans un monde dans lequel j serait la seule cause de décès. L'approche « latent failure times » consiste à imaginer l'existence des temps latents $\widetilde{T}_1$, $\widetilde{T}_2$ et de s'intéresser à leur distribution jointe qui est non identifiable à moins de faire certaines hypothèses fortes comme l'indépendance des risques, impossible à vérifier sur la base des données disponibles. Une autre façon de pallier à ce problème de non identifiabilité est de modéliser l'association des temps latents en utilisant une copule . Cependant, ce genre d'approches, faisant intervenir des notions assez évasives, est peu utilisé car suppose des hypothèses non vérifiables sur la base des données existantes et peut entraîner des interprétations hasardeuses.

Remarquons qu'avec une approche multi-état avec processus sous-jacent, aucune hypothèse d'indépendance n'est requise. En effet, c'est parce que la vraisemblance se factorise que chaque risque cause-spécifique $\alpha_1(.)$, $\alpha_2(.)$ peut être analysé séparément en assimilant les décès « autre cause » à des censures à droite .

2. Considérons maintenant le modèle illness-death de la figure (2.3). Dans la littérature, cette configuration illness-death peut être évoquée sous le terme de risques semi-compétitifs. Les risques semi-compétitifs y sont définis de la façon suivante : les sujets peuvent subir deux types d'évènements, un évènement terminal et un évènement non terminal ; l'évènement terminal censure l'évènement non terminal mais pas inversement. Cette description est généralement associée à une analyse basée sur l'estimation d'une fonction de survie jointe de deux temps d'évènements et entretient implicitement l'idée de temps d'évènement latents .

2.2.2 Observations en temps discret : panel data

Le cas le plus délicat est celui où l'observation des différents états se fait en temps discret. Les temps de transition ne sont alors pas connus de façon exacte et le chemin pris pour aller d'un état à l'autre entre deux temps d'observation consécutifs peut être inconnu avec un nombre de chemins possibles pouvant être grand, voire infini. Les modèles multi-états adaptés à de telles données sont souvent désignés dans la littérature sous le terme de Markov models for panel data. Dans ce type de modèles , l'estimation est basée sur la résolution des équations différentielles de Kolmogorov qui permet d'obtenir les probabilités de transition et donc la vraisemblance (qui

s'écrit en fonction des probabilités de transition). C'est pourquoi l'hypothèse d'homogénéité dans le temps du processus est souvent faite bien qu'elle soit très forte. La vraisemblance s'écrit :

$$\prod_{i,j} L_{ij} = \prod_{i,j} ps\left(t_{ij}\right) s\left(t_{i,j+1}\right)\left(t_{ij}, t_{i,j+1}\right) = \prod_{i,j} ps\left(t_{ij}\right) s\left(t_{i,j+1}\right)\left(t_{i,j+1} - t_j\right)$$

où i est l'indice du sujet, j est l'indice correspondant aux différents temps d'observation (j est un entier allant de 1 au nombre de temps d'observation du sujet i), S(t_{ij}) est l'état dans lequel se trouve le sujet i au $j^{ème}$ temps d'observation.

Des solutions analytiques des équations existent cependant aussi pour des intensités de transition non constantes.

— Lorsque les intensités de transition sont constantes par morceaux , la résolution des équations peut être faite sur chaque intervalle de temps. Le modèle obtenu est très flexible mais l'hypothèse de fonctions d'intensité de transition discontinues n'est pas très plausible d'un point de vue biologique. De plus, une hypothèse de temps de séjour distribués de façon exponentielle par morceaux est inhérente à ces modèles.

— En utilisant des modèles de transformation , on peut se ramener à un processus de Markov homogène dans le temps. On peut alors supposer par exemple des distributions de Weibull pour les intensités de transition. Une des limitations de ces modèles est que le ratio des intensités de transition (vers des états différents) doit rester constant dans le temps. Cette hypothèse est parfois peu réaliste. Par exemple, dans un modèle illness-death, on peut s'attendre à ce que le taux de mortalité α_{02} croisse plus rapidement avec l'âge que le taux d'incidence α_{01}. D'autres formes plus flexibles ont été proposées pour les intensités de transition. Titman (2011) propose d'utiliser des méthodes basées sur une solution numérique des équations de Kolmogorov et utilise des B-splines quadratiques pour modéliser les intensités de transition. Cependant, son modèle peut être rapidement limité lorsqu'on y inclut des variables explicatives car alors, autant d'équations différentielles doivent être résolues qu'il y a de valeurs différentes des variables explicatives dans le jeu de données. En pratique, le paquet R msm (Jackson, 2011) peut être utilisé pour l'analyse de modèles multi-états lorsque les données sont observées en temps discret. Les intensités de transition y sont supposées constantes ou constantes par morceaux, et des variables explicatives peuvent être inclues dans des modèles de régression à intensités proportionnelles (équation 2.2).

Cas particuliers :

Pour les modèles progressifs ou hiérarchiques (sans retour en arrière possible), le nombre de chemins possibles entre deux temps d'observation est fini. L'écriture de la vraisemblance

peut être développée et lorsque le nombre d'intégrations est faible il vaut mieux utiliser des méthodes plus flexibles basées sur la vraisemblance. Par exemple, dans un modèle illness-death sans rétablissement possible où les temps de décès sont observés en temps continu et les temps de maladie en temps discret, il y a un ou deux chemins possibles entre deux temps d'observation et les temps de maladie sont seulement censurés par intervalle : Lorsqu'un sujet est observé en l'état 1, on sait que le passage de 0 à 1 a eu lieu dans un intervalle de temps donné.

⋆ ⋆ Chapitre 3 ⋆ ⋆

Risques compétitifs

3.1 Introduction

Le concept des risques compétitifs est apparu au XVIII^{ème} siècle, lorsque Daniel Bernoulli (1760) a étudié l'impact de l'éradication de la variole sur les taux de mortalité en Angleterre. Plus récemment, ce sujet a été l'objet de nombreux débats pour l'estimation de la probabilité d'un évènement particulier en présence des autres évènements, ou bien après la modification ou l'élimination d'un autre évènement. Par exemple, dans le domaine industriel (fiabilité) pour étudier la défaillance d'un élément d'une chaîne de production après la suppression d'une autre source de panne. En démographie, la situation de risque compétitif est observée pour étudier la probabilité d'un couple de se marier ou de vivre en concubinage. En recherche médicale, cette situation est également fréquente dans différents domaines tels qu'en gynécologie pour étudier la probabilité d'accoucher par voie naturelle ou par césarienne (l'accouchement par voie naturelle étant en concurrence avec la césarienne) (Com-nougué (1999)).

Soit un modèle de données de durées de vie dans lequel l'évènement d'intérêt est une panne (ou un décès) imputable à la j^{ème} cause, j $\in$ J = { 1, 2, ..., m } où l'entier non nul m désigne le nombre de causes possibles. Par convention, j = 0 correspond à l'état de fonctionnement (ou de vie) de l'individu observé. On suppose que l'observation s'arrête lorsqu'une panne (ou décès) survient, mais que cette observation peut être censurée à droite de manière non informative. Quelques exemples illustrant cette situation correspondent au cas où l'évènement d'intérêt est dû à une autre cause, ou à un retrait de l'individu de l'étude ou encore à la fin de l'étude. Dans le cas de censure à droite, les temps de panne des individus ainsi que leurs causes ne sont pas connus de l'expérimentateur. Un modèle de données tel que décrit ci-dessus est communément appelé "modèle à risques compétitifs" (ou concurrents) et il est le plus considéré dans les domaines aussi divers tels qu'en contrôle médical, en démographie, en actuariat, en économie ou en fiabilité industrielle. Andersen et al. (1993) est une des références illustrant des détails et techniques mathématiques sur les risques compétitifs dans les applications biomédicales. Par

exemple dans l'étude du SIDA, les risques compétitifs sont : "mort due au Paludisme", "mort due à la Tuberculose" ou "mort due à d'autres causes" et dans ce cas m = 3. La figure 3.1 illustre cet exemple.

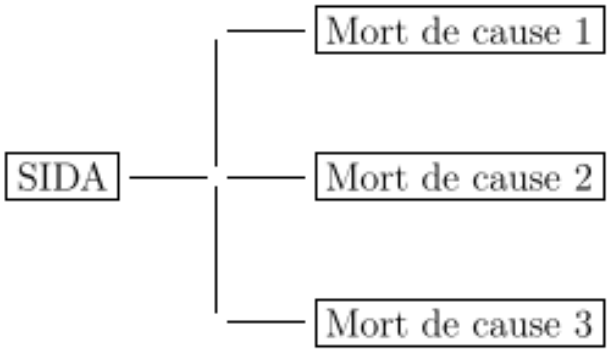

FIGURE 3.1 – Modèle à 3 risques compétitifs

Certains auteurs trouvent pertinent de modéliser les modèles à risques compétitifs sous forme de modèles à variables latentes. C'est le cas de Latouche (2004) qui insiste sur l'existence, l'identifiabilité et l'interpretation des objets dans une modélisation à temps latents. Le modèle qu'il utilise est celui de Fine et Gray (1999) qui est un modèle semi-paramétrique à risque proportionnel de formulation similaire au modèle de Cox, pour la fonction de risque associée à la fonction d'incidence cumulée proposée par Gray (1988) :

$$\alpha_k(t) = -\frac{d}{dt}\log\{1 - F_k(t)\} \qquad (k \in J).$$

Le modèle à risque proportionnel pour la cause numéro k s'écrit :

$$\alpha_k(t;x) = \alpha_{0k}(t)\exp(\gamma x) \ ;$$

où $\gamma > 0$, x le vecteur de covariables et $\alpha_{0k}(t)$ est une fonction continue non spécifiée représentant le risque de base. La fonction de risque associée à la fonction d'incidence cumulée est aussi appelée fonction de risque de sous-répartition.

Récemment, Belot (2009) a présenté les principales méthodes d'analyse statistique utilisées dans le cadre particulier des risques compétitifs : d'une part la méthode d'estimation de la probabilité d'évènement (fonction d'incidence cumulée) et d'autre part, les méthodes pour estimer l'effet des covariables, basées sur la fonction de risque cause spécifique ou sur la fonction de sous-répartition.

3.2 Méthode d'analyse basée sur la fonction de risque cause-spécifique

Plaçons-nous dans une situation où nous étudions m différents types d'événements exclusifs et en concurrence. Les données observées pour un patient i, (i = 1, ..., n) forment un quadruplet $\{ T_i ; j_i ; \delta_i ; x_i \}$ où T_i représente le délai jusqu'au premier évènement (ou à la censure), j_i le type d'évènement (qui n'est pas défini dans le cas où le patient i est censuré), δ_i l'indicatrice d'évènement, égale à 1 si un évènement a été observé en T_i et 0 sinon, et x_i un vecteur de covariables.

3.2.1 Définition des fonctions utilisées dans les risques compétitifs

Pour modéliser les risques compétitifs à partir de ces données observées, nous devons tenir compte du type d'évènement dans la définition de la fonction de risque.

Définition 3.1. La fonction de risque spécifique à l'évènement j (ou taux spécifique) $\lambda_j(t;x)$ est définie par :

$$\lambda_j(t;x) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t,\ \eta = j \ |T \geq t,\ x)}{\Delta t}$$

c'est une fonction estimable à partir des données censurées et η est la variable aléatoire réelle qui indique la cause de la panne (ou décès).

La fonction $\lambda_j(t;x)$ est la fonction de risque instantanée au temps t et spécifique à l'évènement j, sachant le vecteur de covariables x et en présence de tous les autres évènements.

Définition 3.2. La fonction de risque global est définie par :

$$\lambda(t;x) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t,\ |T \geq t,\ x)}{\Delta t} \quad .$$

La fonction de risque globale ne tient pas compte du type d'évènement associé au délai. Tous les délais, quel que soit le type d'évènement associé, sont traités de la même façon dans l'estimation. Ainsi, la fonction de risque globale permet d'estimer la survie sans évènement et s'écrit comme la somme des m fonctions de risques spécifiques :

$$\lambda(t;x) = \sum_{j=1}^{m} \lambda_j(t;x) \ .$$

Définition 3.3. La fonction de survie sans évènement

La survie sans évènement (SSE) est estimée en considérant tous les évènements, quelque soit leur type, comme un seul et même évènement. Dans le cas de m risques compétitifs distincts, la SSE s'écrit :

$$SSE(t,x) = \exp\left(-\int_0^t \sum_{j=1}^{m} \lambda_j(u;x)du\right) \quad .$$

Remarque 3.1 :

La probabilité P(T $\leq$ t, η = j) de subir la cause j avant le temps t en présence des autres évènements se déduit à partir des fonctions de risques spécifiques et de la fonction de risque globale par la formule :

$$P(T \leq t, \eta = j) = \int_0^t \lambda_j(u,x) SSE(u,x)du \quad , \qquad (3.1)$$

qui est appelée fonction d'incidence cumulée de subir une cause du type j.

3.2.2 Définition de la vraisemblance en risque compétitif

Sous l'hypothèse que les patients sont censurés de manière non-informative, l'écriture de la vraisemblance des n données observées s'obtient à partir des fonctions de risques spécifiques et de la fonction de risque globale définies précédemment (voir Kalbfleisch et Prentice, (2002))

$$\prod_{i=1}^{n} \left(\lambda_{ji}\left(t_i, x_i\right)\right)^{\delta_i} \exp\left(-\int_0^{t_i} \lambda\left(u, x_i\right) du\right) \quad . \qquad (3.2)$$

Après manipulation algébrique, cette fonction de vraisemblance peut également s'écrire sous la forme :

$$\prod_{j=1}^{m} \left(\prod_{i=1}^{n} \left(\lambda_j\left(t_i, x_i\right)\right)^{\delta_{ji}} \exp\left(-\int_0^{t_i} \lambda_j\left(u, x_i\right) du\right)\right) \quad , \qquad (3.3)$$

où δ_{ji} est l'indicatrice d'évènement de type j pour le patient i.

De plus, l'indicatrice δ_{ji} permet de ne pas prendre en compte les fonctions de risques qui ne sont pas impliquées dans la contribution à la vraisemblance du patient i. Notons que la contribution à la vraisemblance d'un patient censuré est égale à la probabilité d'être en vie et indemne de tout autre évènement jusqu'à ce délai.

Grâce à l'écriture (3.3) de la vraisemblance en fonction des $\lambda_j(t;x)$, tous les outils statistiques définis pour estimer une fonction de risque "unique" sont utilisables. En effet, d'après

l'écriture de la vraisemblance en produit des $\lambda_j(t;x)$, si l'analyste s'intéresse à un évènement en particulier, il peut réaliser l'analyse en censurant tous les suivis des patients qui ne subissent pas l'évènement étudié, et utiliser les outils classiques d'analyse de survie sur ce "nouveau" jeu de données.

3.2.3 Contrainte de non-identifiabilité et hypothèse d'indépendance

Traditionnellement, la théorie des risques compétitifs était basée sur la définition de temps "latents" d'évènements (i.e. qui ne sont pas tous observés) . Considérons un échantillon de N patients qu'on suppose exposés à m risques compétitifs. On note T_j le délai jusqu'au premier évènement due à la cause j = 1, ..., m. On définit la variable aléatoire T par T = min (T_1 , ..., T_m). On peut remarquer que T = T_j si la cause de la panne est la cause j = 1, ..., m. Soit x le vecteur de covariables et soit η la variable aléatoire réelle qui indique la cause de la panne.

Le problème des risques compétitifs est d'estimer la fonction de survie jointe (appelée également "multiple decrement function") :

$$Q\left(t_1\ ,t_2,\ .....,\ t_m\right) = P\left(T_1 > t_1,\ T_2 > t_2, ..., T_m > t_m\right)$$

Ce modèle présente l'intérêt de spécifier la distribution jointe de l'ensemble des m lois de probabilités associées aux m variables aléatoires T_j ; j = 1, ..., m.

Selon ce modèle de régression linéaire, la probabilité de survie (quelle que soit la cause) au delà du temps t est définie par :

$$P\left(T > t\right) = Q\left(t\ ,t,\ .....,\ t\right) = P\left(T_1 > t,\ T_2 > t, ..., T_m > t\right)\ .$$

La fonction de risque spécifique à l'évènement j (telle que définie précédemment) s'exprime en fonction de la survie jointe Q par :

$$\begin{aligned} \lambda_j(t) &= \lim_{\Delta t \to 0} \frac{P(t < T_j < t + \Delta t\ |T > t)}{\Delta t} \\ &= \frac{-\partial \log\left(Q\left(t_1\ ,t_2,\ .....,\ t_m\right)\right)}{\partial t_j} |\ t_1\ = t_2 = \ ..... = t_m = t \end{aligned} \tag{3.4}$$

A noter ici que l'évènement $t < T_j < t+\Delta t$ est équivalent à l'évènement $t < T < t+\Delta t, \eta = j$ et que cette fonction de risque est estimée conditionnellement à être en vie et indemne des causes étudiées au temps t. C'est à dire que la fonction de risque définie à partir de la fonction de survie jointe est bien égale à la fonction de risque définie au paragraphe précédent.

Les temps d'évènements T_j n'étant pas tous observés (puisqu'on observe uniquement le minimum), la fonction Q ne peut être estimée de façon empirique, à moins de faire certaines hypothèses. Ainsi, cette fonction est dite non identifiable. De même, la distribution marginale du temps latent T_j associée à la fonction de survie nette $S_j(t)$ n'est pas identifiable. La survie nette (également appelée survie marginale) est définie par :

$$S_j(t) = P(T_j > t) = Q(0, ..., t, ...0) \quad ,$$

et s'interprète comme la survie des patients à la cause j sous l'hypothèse que seule cette cause agit sur la population, et que la suppression des autres causes n'a pas modifié les distributions associées aux temps latents T_j . La fonction de risque associée à cette survie nette est la fonction de risque nette spécifique à l'évènement j, notée ici λ_j^m . Cette fonction de risque est associée au temps d'évènement T_j , sous l'hypothèse d'avoir éliminé toutes les autres causes et s'écrit :

$$\begin{aligned} \lambda_j^m(t) &= \lim_{\Delta t \to 0} \frac{P(t < T_j < t + \Delta t \mid T_j > t)}{\Delta t} \\ &= \frac{-\partial \log(S_j(t))}{\partial t} \\ &= \frac{-\partial S_j(t)}{\partial t} \mid S_j(t) \ . \end{aligned}$$

Sous l'hypothèse d'indépendance entre les différents temps d'évènements T_j , il y a égalité entre le taux spécifique brut et le taux spécifique net. Sous cette même hypothèse, la fonction de survie jointe Q s'exprime comme le produit des survies nettes de chaque évènement :

$$Q(t_1, t_2, ..., t_m) = \prod_{j=1}^{m} (S_j(t)) \ .$$

Ainsi, l'hypothèse d'indépendance entre les m temps latents permet d'identifier la fonction de survie marginale.

L'égalité entre le taux spécifique brut et le taux spécifique net se prouve facilement dans le cas de deux évènements concurrents (la généralisation à m évènements s'en déduit simplement). En effet :

$$\begin{aligned} P(t < T_1 < t+\Delta t \,|T>t) &= P(t<T_1<t+\Delta t \,|\, \{T_1>t, T_2>t\}) \\ &= \frac{P(\{t<T_1<t+\Delta t\}\ , \{T_1>t\}, \{T_2>t\})}{P(\{T_1>t\}, \{T_2>t\})} \\ &= \frac{P(\{t<T_1<t+\Delta t\}\ , \{T_1>t\}).P(\{T_2>t\})}{P(T_1>t).P(T_2>t)} \\ &= \frac{P(\{t<T_1<t+\Delta t\}\ , \{T_1>t\})}{P(T_1>t)} \\ &= P(\{t<T_1<t+\Delta t\}\ |\{T_1>t\}) \quad . \end{aligned}$$

En prenant la limite quand Δt tend vers 0, on obtient :

$$\lambda_1(t) = \lambda_j^m(t) \quad .$$

⋆ ⋆ Chapitre 4 ⋆ ⋆

Applications

4.1 Risques compétitifs

L'ensemble de données fait partie du paquet de R. Il contient un sous-échantillon aléatoire de 747 patients de la SIR 3 (propagation des infections nosocomiales et résistant pathogènes) étude de cohorte à l'hôpital universitaire Charité à Berlin, en Allemagne, en vue de l'évaluation prospective des données pour examiner l'effet de l'infection nosocomiale en soins intensifs (Wolkewitz et al., 2008). L'ensemble de données contient des informations sur l'état de la pneumonie à l'admission aux urgences, le temps passé aux urgences et le résulat à l'issue du séjour qui est soit la mort à l'hôpital soit la sortie en vie de l'hopital. La pneumonie est une infection grave, soupçonnée d'augmenter la mortalité aux urgences.

On modélise les données en prenant 1 pour les patients atteints de pneumonie à l'admission aux urgences, et 0 pour ceux n'ayant aucune pneumonie. L'état d'un patient à la fin de la période d'observation est désigné par 1 si le patient sort en vie, et 2 si le patient décède. On utilise 0 pour désigner les patients étant encore aux urgences à la fin de l'étude. Ces patients sont dits censurés à droite. La durée du séjour d'un patient est en jours.

Il y avait 97 patients atteints de pneumonie à l'admission aux urgences. Au total, 657 patients sont sortis en vie, 76 patients sont décédés, et 14 patients étaient encore aux urgences à la fin de l'étude. Parmi les patients qui sont décédés, 21 avaient une pneumonie à l'admission. L'ensemble de données est un exemple des risques compétitifs dans lequel nous étudions le temps jusqu'à la fin du séjour et l'état de sortie qui est soit en vie (ou décharge) soit décédé à l'hôpital.

4.1.1 Estimation non paramétrique

Le but de l'étude de la pneumonie est d'étudier l'impact de la pneumonie présents à l'admission aux urgences sur la mortalité. Comme la pneumonie est une maladie grave, nous devrions

nous attendre à ce que plus de patients meurent d'une pneumonie que sans pneumonie. Ici la mort est la manifestation d'intérêt et la décharge (sortie en vie) est l'évènement en compétition.

Nous étudions d'abord les risques spécifiques causes cumulatifs en utilisant R. Nous utilisons la même matrice des valeurs logiques indiquant les types de transition possibles dans les risques compétitifs.

Nous avons également recodé les données de telle sorte que l'évènement d'intérêt, la mort, corresponde à l'état 1 et l'état 2 est l'évènement en compétition :

Ensuite, nous calculons les risques cumulatifs de cause-spécifique pour la mort et la décharge, respectivement, et stratifions par le statut de la pneumonie à l'admission.

Une parcelle personnalisée des estimations Nelson-Aalen est affichée dans la figure 4.1.

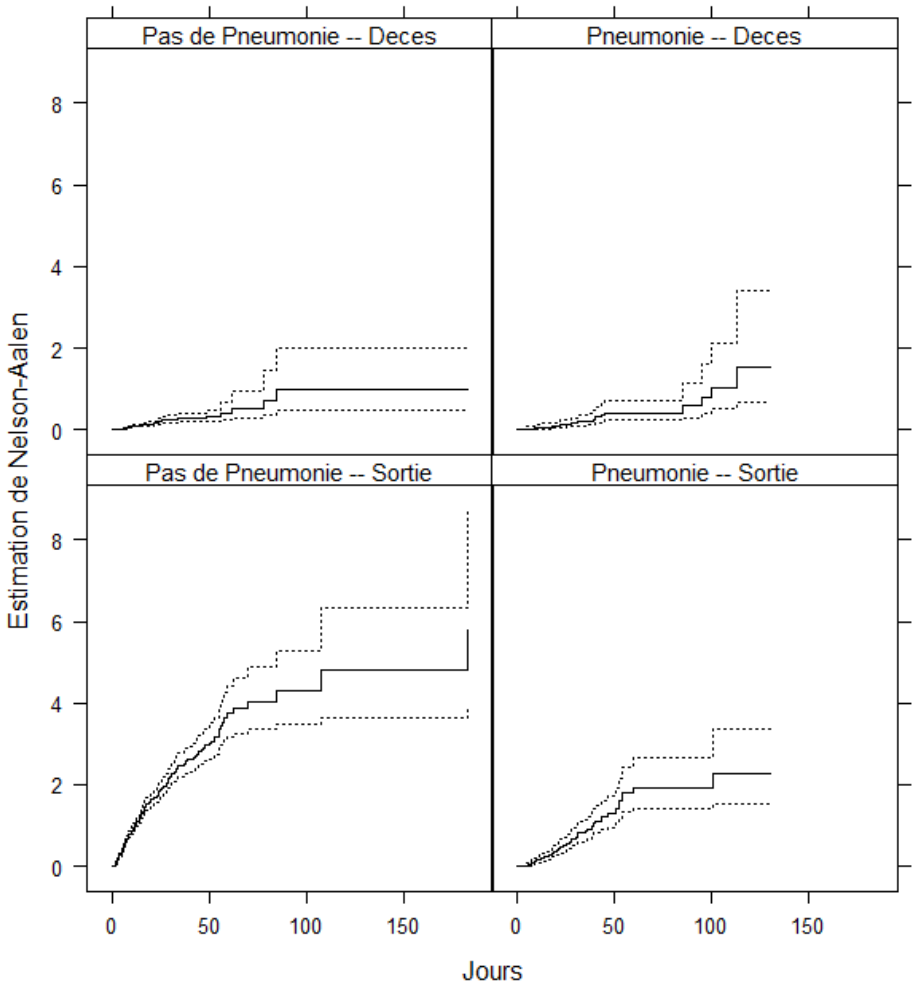

FIGURE 4.1 – Les données hospitalières. Rangée du haut : Estimateur de Nelson-Aalen $\widehat{A}_{01}(t)$ du risque cumulatif de mort. Rangée du bas : Estimateur de Nelson-Aalen $\widehat{A}_{02}(t)$ du risque cumulatif de décharge.

Notons que la pneumonie semble n'avoir aucun effet sur le danger de mort. cependant, cela ne signifie pas que la pneumonie n'a pas d' effet sur la mortalité. La raison est que la pneumonie semble réduire le risque de décharge. Cela implique :

1. La pneumonie semble réduire le risque de toutes causes de fin de séjour dans l'unité de soins intensifs.
2. Les patients atteints d'une pneumonie à l'admission restent plus longtemps dans l'unité de soins intensifs. Pendant ce séjour prolongé, ils sont exposés à un risque de mort essentiellement inchangé.
3. En conséquence, plus de patients atteints de pneumonie meurent que de patients sans pneumonie.

   En conclusion, on observe que la pneumonie augmente la probabilité de mourir à l'hôpital, mais ne semble pas avoir effet sur la probabilité quotidienne de mourir à l'hôpital.

Ceci est un phénomène typique des risques compétitifs. Parce qu'il y a plus d'un risque agissant sur un individu, on ne peut pas prédire à l'aide d'un seul risque quel sera l'évolution future de l'individu.

4.1.2 Risque proportionnel

Nous considérons les données hospitalières que nous avons analysés dans la première partie.

Rappelons, en particulier, que d'après l'estimateur de Nelson-Aalen des risques cumulatifs avec causes spécifiques la pneumonie à l'admission augmente la mortalité hospitalière par une diminution de l'effet sur le risque de décharge en vie, alors que le risque instantané de décès à l'hôpital est laissé essentiellement inchangé. Le but de la présente analyse est de réexaminer cet constat via le modèle de Risques proportionnel avec cause spécifique.

Nous avons tout simplement fait deux modèles de Cox différents.

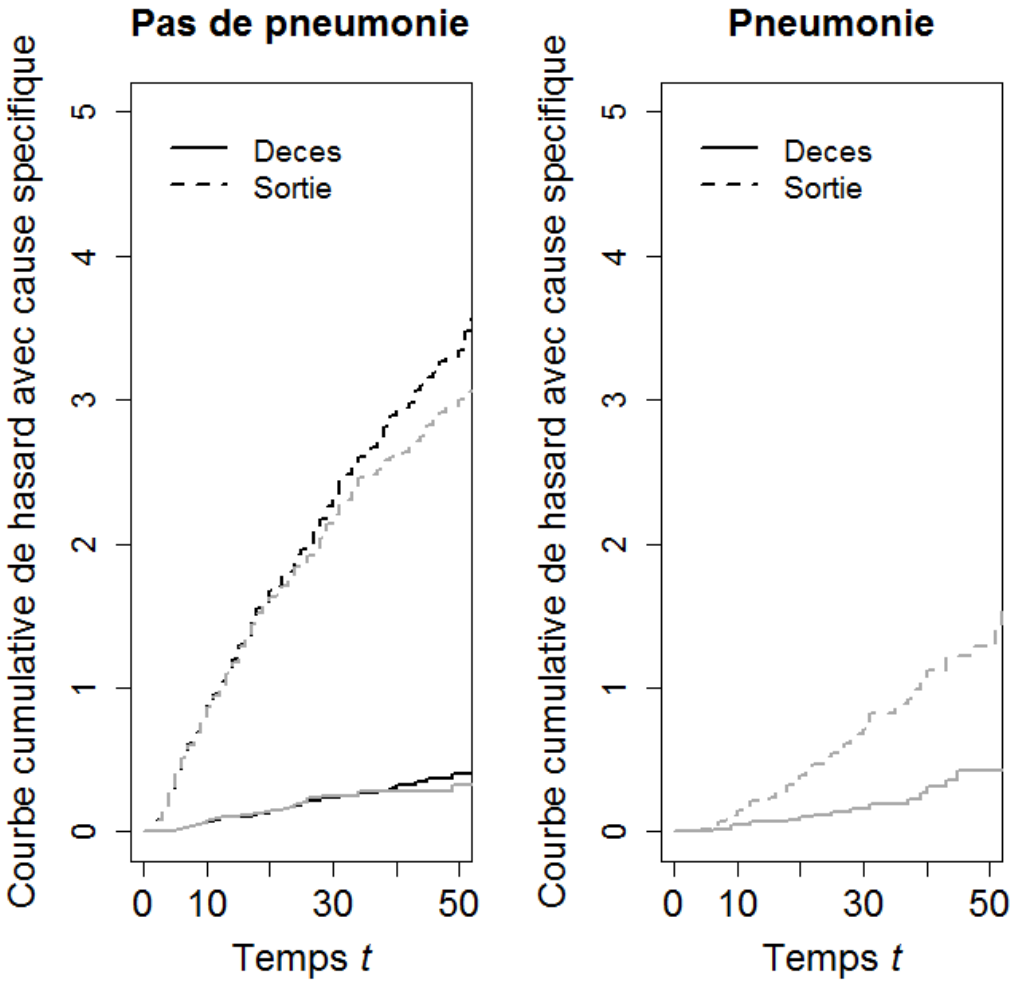

FIGURE 4.2 – Les données hospitalières. lignes grises : estimateurs Nelson-Aalen de la fonction cumulative de risques avec causes spécifiques. Lignes noires : estimateurs Breslow du Risque cumulé avec cause spécifique (tracé de gauche) et risque cumulatif basé sur un modèle d'estimation pour les patients atteints de pneumonie (graphique de droite).

la figure 4.2 montre un tracé personnalisé des estimateurs de Nelson-Aalen basé sur un modèle GETHER avec les estimateurs Breslow (pour pas de pneumonie) et l'estimateur de risques cumulatifs (pour une pneumonie à l'admission). Nous avons restreint l'axe du temps sur [0, 50], car la plupart des évènements se produisent dans cet intervalle de temps. Nous constatons que toutes les courbes de risques spécifiques de décès sont en bon accord, tout comme les estimateurs de base pour le risque cumulatif de décharge. Cependant, les estimateurs respectifs du risque cumulatif de décharge pour les patients atteints de pneumonie ne sont pas en accord, ce qui indique que l'effet de la pneumonie sur le risque de décharge ne peut pas suivre un modèle de risque proportionnel avec cause spécifique.

4.2 Modèles multi-états

L'ensemble des données fait partie du paquet de R et contient un autre sous-échantillon aléatoire de 747 patients de l'étude SIR 3 décrit plus haut. L'ensemble des données contient des informations sur les temps de ventilation et le temps de séjour aux urgences. Il y a aussi des variables informatives supplémentaires sur l'âge et le sexe.

sir.cont est un exemple d'un modèle maladie $\rightarrow$ mortalité avec récupération, parce que la ventilation peut être allumée ou étteinte pendant le séjour à l'hôpital. En outre, un patient peut être considéré soit sur ventilation soit sans ventilation lors de l'admission aux urgences.

Les évènements sont représentés dans sir.cont selon le modèle multi-état. La mise sous ventilation est représentée par la transition $0 \rightarrow 1$, comme indiqué dans les colonnes et la mise hors ventilation est représentée par la transition $1 \rightarrow 0$.

Les heures des évènements sont dans le temps de la colonne. L'heure représentant la fin de séjour a la valeur 2 dans la colonne. L'entrée est «cens» pour les heures des événements censurés.

Le but de la présente analyse est d'étudier, en utilisant une estimation non paramétrique, l'effet de la ventilation sur la durée du séjour aux urgences. Nous le faisons en estimant les risques cumulatifs pour la fin du séjour.

Une caractéristique importante de ces données est que la ventilation peut apparaître ou disparaître pendant le séjour à l'hôpital. En outre, les patients peuvent être sur ventilation ou non lors de l'admission aux urgences. L'état 0 représente «pas de ventilation», l'état 1 représente «ventilation», et la fin du séjour est modélisée par des transitions dans l'état 2.

La figure 4.3 affiche les estimateurs Nelson-Aalen $\widehat{A}_{02}(t)$ et $\widehat{A}_{12}(t)$ log-transformés des intervalles de confiance.

On constate que La ventilation prolonge la durée du séjour aux urgences. Ceci est dû au fait que les patients ventilés sont censés exiger des soins supplémentaires, conduisant à un long séjour aux urgences.

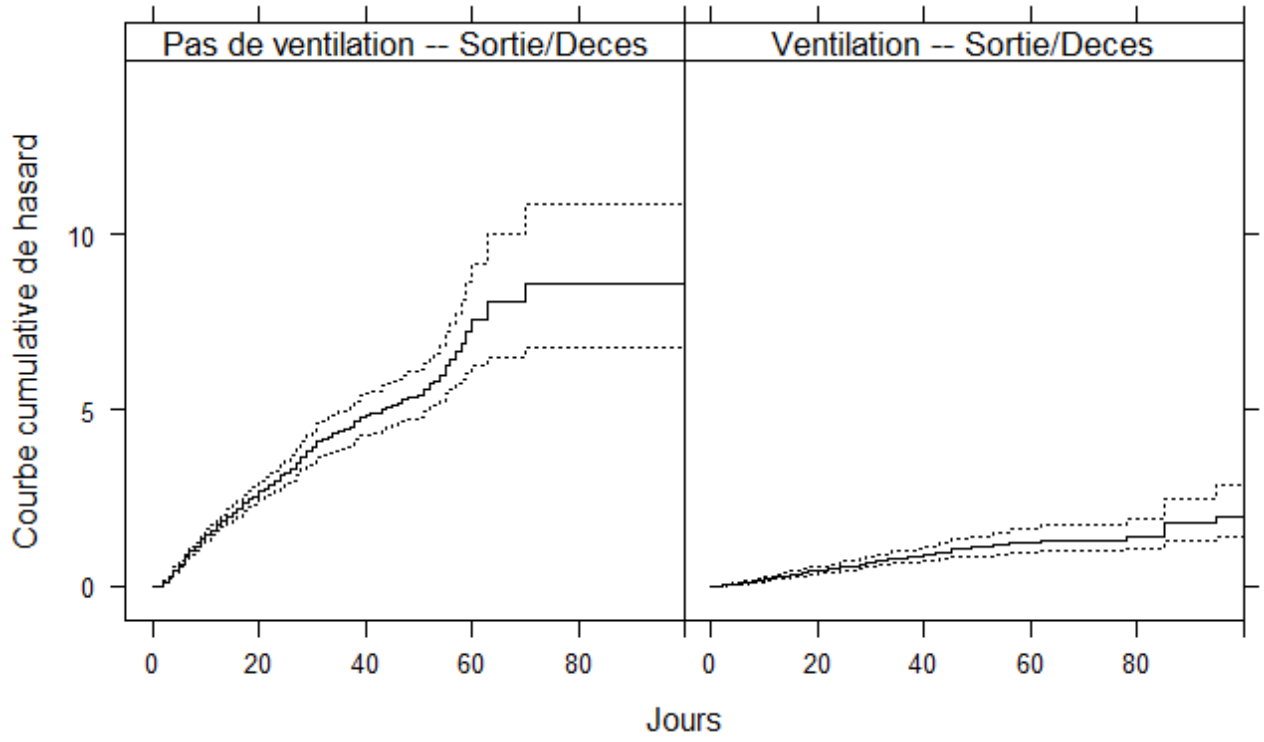

FIGURE 4.3 – Les données de ventilation. Estimateur de Nelson-Aalen pour les transitions sans ventilation → fin de séjour (à gauche) et ventilation → fin de séjour (à droite).

4.3 Implication didactique

Les notions de risques compétitifs et de modèles multi-états nous sont utiles dans le cadre des enseignements à plus d'un titre.

Premièrement au niveau individuel le fait pour moi de travailler sur ce thème m'a permi de renforcer mes capacités en statistiques, m'offrant ainsi la possibilité d'être plus efficace en situation d'enseignement-apprentissage, toute chose qui contribuerait à une meilleure compréhension des statistiques par les élèves.

Dans un second plan les notions de risques compétitifs et de modèles multi-états nous permettent de voir les statistiques sous un nouvel angle, un peu plus pratique, notament avec la prise en compte des données censurées lors d'une étude et l'utilisation du logiciel R pour les simulations. Toute chose qui concoure à nous rapprocher un peu plus de la réalité. A ce titre il est pertinent de prendre en compte ces notions lors des différentes études statistiques en milieu scolaire telles que : Le décrochage scolaire, l'échec scolaire en milieu urbain et en milieu rural, le redoublement scolaire des filles et des garçons, l'impact de la durée des enseignements sur la qualité des résultats.

♣ Conclusion ♣

Notre travail portait sur les risques compétitifs et les modèles multi-états. Il a été réparti en quatre parties ; dans la première partie nous avons abordé la notion d'analyse de survie qui est un préambul à la notion de risques compétitifs et qui est également un model multi-état ; dans cette partie nous avons mis en exergue les différentes fonctions utilisées en analyse de survie et leurs propriétés. Dans la seconde partie nous avons abordé la notion de modèles multi-états ; ici nous avons présenté quelques modèles multi-états et quelques fonctions utilisées dans ce cadre notament les probabilités de transitions entre états et les intensités de transitions entre états. Dans la troisième partie nous avons abordé la notion de risques compétitifs qui est un cas particulier de modèles multi-états ; cette notion vient pallier aux limites observées dans l'analyse de survie ; dans cette partie nous avons présenté quelques fonctions utilisées dans l'étude des risques compétitifs. Dans la dernière partie nous avons procédé à une application dabord dans le cadre des risques compétitifs et en suite dans le cadre des modèles multi-états ceci pour mettre en exergue la pertinence de l'utilisation de ces différentes notions. Nous avons également, dans cette dernière partie, montré l'implication didactique des risques compétitifs et des modèles multi-états.

En perspective, nous comptons donner une estimation de la fonction de survie jointe Q en risques compétitifs lorsque les temps d'évènements sont dépendants. Nous comptons également étudier le phénomène de décrocharge scolaire à l'aide des risques compétitifs.

♣ Bibliographie ♣

[1] Aalen, O. O. (1975). Statistical inference for a family of counting processes. PhD thesis, Department of Statistics, University of California, Berkeley.

[2] Aalen, O. O. et Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics, 5 , pages 141–150.

[3] Andersen, P. K. et Borgan, O. (1985). Counting process models for life history data : a review. Scandinavian Journal of Statistics, 12 , pages 97–158.

[4] Andersen P. K., Borgan Ø., Gill R. D. et Keiding N. (1993). Statistical models based on counting processes. Springer-Verlag.

[5] Andersen, P. K., Geskus, R. B., de Witte, T., et Putter, H. (2012). Competing risks in epidemiology : possibilities and pitfalls. International journal of epidemiology, 41(3) , pages 861–870.

[6] Belot, A.(2009). Modélisation flexible des données de survie en présence de risques concurrents et apports de la méthode du taux en excès. Thèse de doctorat. Université de la Méditerranée.

[7] Bernoulli D.(1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. Histoires et Mémoires de l'Académie Royale des Sciences. pages 1-45.

[8] Beyersmann, J. Allignol, A. et Schumacher, M. (2012). Competing Risks and Multistate Models with R. Springer, New York.

[9] Com-nougué C.(1999). Estimation des risques associés à des événements multiples. Revue d'Epidémiologie et de Santé Publique, 47, pages 75-85.

[10] Commenges, D.(1999). Risques compétitifs et modèles multi-états en épidémiologie. Revue d'Epidémiologie et Santé Publique, 47, pages 605 - 611.

[11] Cox, D. R. et Miller, H. D. (1965). The Theory of Stochastic Processes. Chapman an Hall.

[12] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), pages 187–220.

[13] De Wreede, L. C., Fiocco, M., et Putter, H. (2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. Computer methods and programs in biomedecine, 99(3), pages 261–274.

[14] De Wreede, L. C., Fiocco, M., et Putter, H. (2011). mstate : An R package for the analysis of competing risks and multi-state models. Journal of Statistical Software, 38(7) , pages 1–30.

[15] Fine, J. P. et Gray, R. J.(1999). A proportional hazards model for subdistribution of competing risk. Journal of the American Statistical Association. 94(446), pages 496 - 509.

[16] Gill, R. D. et Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. The annals of statistics, pages 1501–1555.

[17] Gray R.J.(1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. The Annals of statistics 116, pages 1141 - 1154.

[18] Hougaard P. (2000). Analysis of multivariate survival data. Springer – Statistics for biology and health.

[19] J. D. Kalbfleisch, Ross L. Prentice.(2002). The statistical analysis of failure time data 2nd edn, New York : John Wiley, New York.

[20] Jackson, C. H. (2011). Multi-state models for panel data : the msm package for R. Journal of Statistical Software, 38(8) , pages 1–29.

[21] Janssen, J. et Limnios, N. (1999). Semi-Markov models and applications. Kluwer Academic Publishers Dotrecht.

[22] Latouche, A.(2004) Modèles de régression en présence de compétition. Thèse de doctorat. Université de Paris 6.

[23] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. Technometrics, 14(4) , pages 945–966.

[24] Njamen Njomen,D.A. (2014). Inférence non-paramétrique en Risques Compétitifs.Thèse de doctorat. Université de Yaounde I, pages 90-103.

[25] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. Annals of Statistics, 11 , pages 453–466.

[26] Rosenberg, P. S. (1995). Hazard function estimation using B-splines. Biometrics, pages 874–887.

[27] Saint-Pierre, P. (2005). Modèles multi-états de type markovien et application à l'asthme. Thèse de doctorat. Université de Montpellier 1.

[28] Titman, A. C. (2011). Flexible nonhomogeneous Markov models for panel observed data. Biometrics, 67(3) , pages 780–787.

[29] Touraine, C. (2013). Modèles illness-death pour données censurées par intervalle : Application à l'étude de la démence.Thèse de doctorat. Université de Bordeaux 2, pages 11-31.

♣ Annexe ♣

4.4 Liste des commandes de R ayant produit la figure 4.1

```
###Donnees Peumonia####
###
data(sir.adm)
id <- sir.adm$id
from <- sir.adm$pneu
to <- ifelse(sir.adm$status==0,"cens",sir.adm$status+1)
times <- sir.adm$time
dat.sir <- data.frame(id,from,to,time=times)
# Possible transitions
tra <- matrix(ncol=4,nrow=4,FALSE)
tra[1 :2,3 :4] <- TRUE
na.pneu <- mvna(dat.sir,c("0","1","2","3"),
tra,"cens")
#Figure 4.1
if(require("lattice")) {
xyplot(na.pneu,tr.choice=c("0 2","1 2","0 3","1 3"),
aspect=1,strip=strip.custom(bg="white",
factor.levels=c("Pas de Pneumonie – Sortie",
"Pneumonie – Sortie",
"Pas de Pneumonie – Deces",
"Pneumonie – Deces"),
par.strip.text=list(cex=0.9)),
scales=list(alternating=1),xlab="Jours",
ylab="Estimation de Nelson-Aalen")
```

4.5 Liste des commandes de R ayant produit la figure 4.2

```
###Donnees Peumonia####
###
data(sir.adm)
id <- sir.adm$id
from <- sir.adm$pneu
to <- ifelse(sir.adm$status==0,"cens",sir.adm$status+1)
times <- sir.adm$time
dat.sir <- data.frame(id,from,to,time=times)
# Possible transitions
tra <- matrix(ncol=4,nrow=4,FALSE)
tra[1 :2,3 :4] <- TRUE
na.pneu <- mvna(dat.sir,c("0","1","2","3"),
tra,"cens")
to <- ifelse(sir.adm$status == 0, "cens",
ifelse(sir.adm$status == 1, 2, 1))
my.sir.data <- data.frame(id = sir.adm$id, from = 0, to,
time = sir.adm$time, pneu = sir.adm$pneu)
fit.pneu.01 <- coxph(Surv(time, to == 1) ~pneu, my.sir.data)
fit.pneu.02 <- coxph(Surv(time, to == 2) ~pneu, my.sir.data)
summary(fit.pneu.02)
summary(fit.pneu.02)
#Figure 4.2
tra <- matrix(FALSE, ncol = 3, nrow = 3)
dimnames(tra) <- list(c("0", "1", "2"), c("0", "1", "2"))
tra[1, 2 :3] <- TRUE
#Base hazard
## no pneumonia
my.nelaal.nop <- mvna(my.sir.data[my.sir.data$pneu == 0, ],
c("0", "1", "2"), tra, "cens")
## with pneumonia
my.nelaal.p <- mvna(my.sir.data[my.sir.data$pneu == 1, ],
c("0", "1", "2"), tra, "cens")
```

```
a01.0 <- basehaz(fit.pneu.01, centered=FALSE)
a02.0 <- basehaz(fit.pneu.02, centered=FALSE)
split.screen(figs=c(1,2))
screen(1)
plot(c(0, 50), c(0, 5), xlab = expression(paste(Temps, " ", italic(t))),
ylab = "Courbe cumulative de hasard avec cause specifique", type = "n", axes = FALSE,
main = "Pas de pneumonie", cex.main = 1.5, cex.lab = 1.5)
axis(1, at=seq(0, 50, 10), cex.axis=1.5)
axis(2, at=seq(0, 5, 1), cex.axis=1.25)
box()
lines(a02.0$time, a02.0$hazard, type="s", lwd=2, lty=2)
lines(a01.0$time, a01.0$hazard, type="s", lwd=2)
lines(my.nelaal.nop, conf.int = FALSE, col = rep("darkgray", 2),
lty = c(1, 2), lwd = 2)
legend(0,5,c("Deces", "Sortie"), lty=1 :2,bty="n",
cex=1.2, lwd=2)
screen(2)
plot(x=c(0, 50), y=c(0, 5), xlab=expression(paste(Temps, " ", italic(t))),
ylab="Courbe cumulative de hasard avec cause specifique", type="n", axes=F,
main="Pneumonie", cex.main=1.5, cex.lab=1.5)
axis(1, at=seq(0, 50, 10), cex.axis=1.5)
axis(2, at=seq(0, 5, 1), cex.axis=1.25)
box()
#lines(a02.0$time, hr2 * a02.0$hazard, type="s", lwd=2, lty=2)
#lines(a01.0$time, hr1 * a01.0$hazard, type="s", lwd=2)
lines(my.nelaal.p, conf.int = FALSE, col = rep("darkgray", 2),
lty = c(1, 2), lwd = 2)
legend(0,5,c("Deces", "Sortie"), lty=1 :2,bty="n", cex=1.2, lwd=2)
close.screen(all.screens=TRUE)
```

4.6 Liste des commandes de R ayant produit la figure 4.3

```
####Non Parametric######
```

```
#A little modification of the data to avoid entry times equal to exit times, which throws an error
data(sir.cont)
sir.cont <- sir.cont[order(sir.cont$id, sir.cont$time), ]
for (i in 2 :nrow(sir.cont)) {
if (sir.cont$id[i]==sir.cont$id[i-1]) {
if (sir.cont$time[i]==sir.cont$time[i-1]) {
sir.cont$time[i-1] <- sir.cont$time[i-1] - 0.5
}
}
}
#Definition of matrix of logical values specifying possible transitions :
tra.ventil <- matrix(FALSE, 3, 3, dimnames =
list(c("0", "1", "2"), c("0", "1", "2")))
tra.ventil[1, c(2, 3)] <- TRUE
tra.ventil[2, c(1, 3)] <- TRUE
#Estimation of cumulative transition hazards :
mvna.ventil <- mvna(sir.cont, c("0", "1", "2"),
tra.ventil, "cens")
#Figure 4.3
xyplot(mvna.ventil, tr.choice = c("0 2", "1 2"),
aspect = 1, strip = strip.custom(bg = "white",
factor.levels = c("Pas de ventilation – Sortie/Deces",
"Ventilation – Sortie/Deces"),
par.strip.text = list(cex = 1.1)),
scales = list(alternating = 1),xlab = "Jours",ylab = "Courbe cumulative de hasard", xlim = c(-5, 100))
```