الجمهورية الجزائرية الديمقراطية الشعبية

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE SAAD DAHLAB DE BLIDA

Faculté des sciences

Département de Mathématiques

MEMOIRE DE MASTER

En Mathématiques

Option : Modèlisation Stochastiques et Statistique

THÈME :

Analyse des données de survie en présence de risques concurrents à partir d'un modèle de mélange

Réalisé par

KHENNICHE Khadidja et KHETTOU Hannane

Soutenu devant le Jury :

TAMI Omar	Université Blida 1	Président
FRIHI Redhouane	Université Blida 1	Examinateur
RASSOUL Abdelaziz	ENSH de Blida	Promoteur

Juillet 2023

DÉDICACE1

je dédice ce modéste mémoire
A mes parents qu'ont soutenus et encouragés durant ces années d'études.
Qu'ils trouvent ici le témoignage de ma profonde reconnaissance. A mes frères, et Ceux qui ont partagés avec moi tous les moments d'émotion lors de la réalisation de ce travail. Ils m'ont chaleureusement supporté et encouragé tout au long de mon parcours.

A ma famille, mes proches et à ceux qui me donnent de l'amour et de la vivacité,mon binome **Kh. Hanane** m'avoir accompagnée au cours de ces 2 années de master , aussi pour ses conseils et ses encouragements.

A tous mes amis qui m'ont toujours encouragés, et à qui je souhaite plus de succès.

A tous ceux que j'aime.

Kh. Khadidja

DÉDICACE2

Avec l'expression de ma reconnaissance, je dédie ce modeste travail à ceux qui, quelque soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère.

A l'homme, mon précieux offre du dieu, qui doit ma vie, ma réussite et tout mon respect : mon cher papa.

A la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit non àmes exigences et qui n'a épargné aucun effort pour me rendre heureuse : mon adorable ma mère

A ma chère sœur youssra et mon mari qui n'ont pas cessés de me conseiller, encourager et soutenir tout au long de mes études.

Merci pour leurs amours et leurs encouragements. Sans oublier mon binôme khadidja pour son soutien moral, sa patience et sa compréhension tout au long de ce projet .

Kh. Hanane

REMERCIEMENTS

En tout premier lieu, nous tenons à remercier ALLAH, le tout puissant et miséricordieux, qui m'a donné la volonté, la force et la santé de réaliser ce mémoire.
Je tiens aussi à remercier monsieur RASSOUL Abdelaziz de m'avoir offert la possibilité de commencer cette mémoire et qui a accepté de superviser mes premiers pas dans la recherche.

Je désire témoigner de ma reconnaissance envers TAMI Omar et FRIHI Redhouane pour la constance de leur soutien et la justesse des orientations qu'ils ont voulu nous suggérer.
Je remercie aussi tous les membres et personels du département de mathématiques.
Je souhaite enfin exprimer toute ma reconnaissance à ma mère et mes frères pour m'avoir aidés dans tous etude.

Kh. Khadidja
Kh. Hanane

Table des matières

TABLE DES FIGURES

LISTE DES TABLEAUX

ABRÉVIATION ET NOTATIONS

IM :Méthode d'Inférence.
CIF : Fonction d'Incidence Cumulative.
CR : Concurent de Risque.
CPH : Modele de hazard proportionnel de Cox .
PH : Hazard Proportionnel.
HR : Rapport de risque instantané (hazard ratio).
IC : Intervalle de Censure.
FG : Fine et Gray.
CDF : Fonction de Distribution Cumulative.
PDF : Fonction de Densité de Probabilité .
CHF : Fonction de Risque Cumulé.
MLE : Estimation de Vraisemblance Maximale.
FrPH :Frechet Proportional Hazard.
EPH : Exponentielle de hazard proportionnelle .
CPH : Incidence Cumulative des Risques Proportionnels.
MCMC : Chaîne de Markov Monte-Carlo Var : variance.
V :variance pondéré.

ملخص

يعتبر تحليل البقاء منهجًا إحصائيًا يستخدم لدراسة الوقت الذي يستغرقه حدوث حدث معين. يمكن أن يشمل ذلك بقاء المرضى بعد تلقي العلاج الطبي، أو فشل المعدات، أو وفاة شخص، أو أي حدث آخر يكون وقت حدوثه مهمًا. تهدف هذا المذكرة إلى توفير فهم شامل لوظائف البقاء وأنواع الرقابة في سياق تحليل البقاء. تعتبر التعاريف والمفاهيم المقدمة ضرورية لإجراء دراسات قوية وتفسير النتائج بشكل صحيح. تعد هذه المعرفة ذات الصلة بالنسبة للباحثين والمتخصصين العاملين في مجالات مثل الطب والوبائيات والصحة العامة.

نناقش أيضًا استخدام نماذج المخاطر المتنافسة في تحليل البقاء. يتم تطبيق هذه النماذج عندما يمكن حدوث عدة أحداث متنافسة ويمكن أن يمنع حدث واحد حدوث الآخرين. يتمثل الهدف الرئيسي في تقدير تأثير المتغيرات المفسرة على الوقت حتى حدوث حدث معين من بين الأحداث المتنافسة. تقدم المذكرة المبادئ الأساسية لهذه النماذج وتستكشف الأساليب المختلفة للتقدير. كما تسلط الضوء على فوائدها في مجالات البحوث الطبية والوبائيات واقتصاد الصحة.

في الختام، يؤكد هذا المذكرة أهمية نماذج المخاطر المتنافسة في تحليل البقاء وفائدتها في فهم عمليات المرض والنتائج السريرية واتخاذ القرارات الطبية. تسهم نتائج هذا البحث في تقدم المعرفة في مجال تحليل البقاء وتوفر معلومات قيمة لاتخاذ قرارات صحية

Résumé

L'analyse de survie est une méthode statistique utilisée pour étudier le temps nécessaire à la survenue d'un événement spécifique. Cela peut inclure la survie des patients après un traitement médical, la défaillance d'un équipement, le décès d'un individu, ou tout autre événement dont le moment est important. Ce mémoire vise à fournir une compréhension approfondie des fonctions de survie et des types de censure dans le contexte de l'analyse de survie. Les définitions et concepts présentés sont essentiels pour mener des études robustes et interpréter correctement les résultats. Ces connaissances sont pertinentes pour les chercheurs et professionnels travaillant dans des domaines tels que la médecine, l'épidémiologie et la santé publique. on discute aussi sur l'utilisation des modèles de risques compétitifs dans l'analyse de survie. Ces modèles sont appliqués lorsque plusieurs événements concurrents peuvent se produire et qu'un événement peut empêcher la survenue des autres. L'objectif principal est d'estimer l'effet des variables explicatives sur le temps jusqu'à la survenue d'un événement spécifique parmi les événements concurrents. Le mémoire présente les principes fondamentaux de ces modèles et explore les différentes méthodes d'estimation. Il met également en évidence les avantages de leur utilisation dans les domaines de la recherche médicale, de l'épidémiologie et de l'économie de la santé.

A la fin, ce mémoire souligne l'importance des modèles de risques compétitifs dans

l'analyse de survie et leur utilité pour la compréhension des processus de maladie, des résultats cliniques et des décisions médicales. Les résultats de cette recherche contribuent à l'avancement des connaissances dans le domaine de l'analyse de survie et fournissent des informations précieuses pour la prise de décisions en matière de santé.

Abstract

Survival analysis is a statistical method used to study the time until a specific event occurs. This can include the survival of patients after medical treatment, equipment failure, individual death, or any other event where timing is important. This dissertation aims to provide an in-depth understanding of survival functions and censoring types in the context of survival analysis. The definitions and concepts presented are essential for conducting robust studies and properly interpreting results. This knowledge is relevant for researchers and professionals working in fields such as medicine, epidemiology, and public health.

The use of competing risk models in survival analysis is also discussed. These models are applied when multiple competing events can occur, and one event can prevent the occurrence of others. The primary objective is to estimate the effect of explanatory variables on the time until a specific event among competing events. The dissertation presents the fundamental principles of these models and explores different estimation methods. It also highlights the advantages of their use in medical research, epidemiology, and health economics.

In conclusion, this dissertation emphasizes the importance of competing risk models in survival analysis and their utility in understanding disease processes, clinical outcomes, and medical decisions. The findings of this research contribute to advancing knowledge in the field of survival analysis and provide valuable insights for health decision-making.

INTRODUCTION GÉNÉRALE

L'analyse de survie, également connue sous le nom d'analyse de durée de vie ou d'analyse de temps jusqu'à l'événement d'interet, est une méthode statistique utilisée pour étudier la probabilité qu'un événement se produise dans le temps.Cet événement peut être, par exemple, la survie d'un patient après un traitement médical, la défaillance d'un équipement, le décès d'un individu, ou tout autre événement dont le moment d'occurrence est important. Le contexte général de l'analyse de survie est donc l'étude des facteurs qui influencent le temps jusqu'à l'événement d'intérêt. Ces facteurs peuvent être de natures diverses, telles que des caractéristiques démographiques, des variables médicales, des expositions environnementales, des comportements individuels, etc. Aussi il est particulièrement utile lorsque les données d'étude contiennent des individus qui n'ont pas encore subi l'événement d'intérêt à la fin de la période d'observation ou de suivi. Elle permet de prendre en compte ces individus censurés et d'estimer la probabilité de survie ou la fonction de survie, qui représente la probabilité de survie jusqu'à un temps donné.

Les principales méthodes d'analyse de survie incluent le modèle de régression de Cox, qui est une extension du modèle de régression linéaire adapté aux données de survie, et les méthodes non paramétriques telles que l'estimateur de Kaplan-Meier pour estimer la fonction de survie.

L'analyse de survie est utilisée dans de nombreux domaines, tels que la recherche médicale, l'épidémiologie, l'ingénierie, la finance, et d'autres domaines où l'étude des durées de vie et des événements est pertinente.
Les modèles de risques compétitifs, également appelés modèles de risques concurrents ou modèles de risques compétitifs de Cox, sont une extension du modèle de régression de Cox dans le domaine de l'analyse de survie. Ils sont utilisés lorsque plusieurs types d'événements peuvent se produire et que l'occurrence d'un événement empêche la survenue des autres événements. Dans un modèle de risques compétitifs, on s'intéresse à l'estimation des effets des variables explicatives sur le temps jusqu'à la survenue d'un

événement spécifique parmi plusieurs événements concurrents possibles. Par exemple, dans une étude médicale, on pourrait s'intéresser au temps jusqu'à la récidive d'un cancer, au temps jusqu'à la survenue d'une autre maladie ou au temps jusqu'au décès, en prenant en compte le fait que la survenue d'un événement peut empêcher la survenue des autres événements. Le modèle de risques compétitifs de Cox utilise une extension de la fonction de risque de base du modèle de Cox pour chaque événement concurrent. La fonction de risque représente le taux de défaillance instantané à un moment donné, conditionnellement aux événements précédents. Les modèles de risques compétitifs permettent d'estimer les effets des variables explicatives sur chaque événement concurrent, en tenant compte des interactions et des dépendances entre les événements.

L'estimation des paramètres dans les modèles de risques compétitifs se fait généralement par la méthode du maximum de vraisemblance partielle, qui consiste à maximiser la fonction de vraisemblance conditionnelle sur les événements observés.

Les modèles de risques compétitifs sont largement utilisés dans différents domaines de recherche, notamment en épidémiologie, en oncologie, en économie de la santé, où la simultanéité des événements concurrents est d'intérêt pour comprendre les processus de maladies, les résultats cliniques ou les décisions médicales. Les principaux modèles de risques compétitifs sont des extensions du modèle de régression de Cox, adaptés pour tenir compte de la présence d'événements concurrents. Voici quelques-uns des modèles couramment utilisés :

Modèle de risques compétitifs de Fine-Gray : Ce modèle, développé par Gray en 1988 et étendu par Fine et Gray en 1999[1], est largement utilisé dans l'analyse de survie compétitive. Il utilise une approche de sous-modèle de risques proportionnels pour estimer les effets des variables explicatives sur chaque événement concurrent. Il tient compte à la fois de l'événement d'intérêt et des événements concurrents dans l'estimation.
Modèle de risques compétitifs de Prentice, Williams et Peterson : Ce modèle, proposé par Prentice, Williams et Peterson e1981[2], est également basé sur une extension du modèle de régression de Cox. Il utilise une approche basée sur les probabilités conditionnelles pour estimer les effets des variables explicatives sur chaque événement concurrent.

Modèle de risques compétitifs Fine et Gray avec dépendance des événements : Ce modèle, introduit par Fine et Gray en 1999[1], permet de modéliser la dépendance entre les événements concurrents. Il permet d'estimer les effets des variables explicatives sur chaque événement concurrent tout en prenant en compte les interactions et les dépendances entre les événements.

Modèle de risques compétitifs de Beyersmann, Allignol et Schumacher : Ce modèle, proposé par Beyersmann, Allignol et Schumacher en 2012[3], est basé sur une extension du modèle de régression de Cox en utilisant des variables indicatrices pour chaque événement concurrent. Il permet d'estimer les effets des variables explicatives sur chaque événement tout en tenant compte de la présence d'événements concurrents.

Ces modèles, ainsi que d'autres variantes et extensions, sont utilisés pour analyser les données de survie compétitive dans différents domaines de recherche. Le choix du modèle dépendra de la nature des données, de l'objectif de l'étude et des hypothèses spécifiques du chercheur.
Notre mémoire , consérnés trois chapitres à l'analyse de survie cette dérniere revêt une grande importance dans divers domaines tels que la médecine, l'épidémiologie et les sciences sociales. En comprenant les facteurs qui influencent la survie des individus, nous pouvons prendre des décisions éclairées et développer des stratégies appropriées pour améliorer les résultats liés à la survie.
Notre mémoire est reporte en trois chapire. Dans chapitre 1 de notre mémoire se concentre sur les concepts fondamentaux de l'analyse de survie, en mettant l'accent sur les définitions essentielles et les types de censures. Nous expliquons la nature de la variable d'intérêt, qui est le temps jusqu'à la survenue d'un événement, et nous présentons les différents types de censures, tels que la censure à droite, la censure à gauche et la censure aléatoire. Cette introduction est cruciale pour développer une compréhension solide des méthodes d'analyse de survie.

Dans le chapitre 2, nous abordons les modèles de risques concurrents utilisés dans l'analyse de survie. Notre attention se porte notamment sur le modèle de Cox, qui est largement utilisé pour modéliser la relation entre les covariables et la survie, même en présence de censures. Nous expliquons en détail les principes et les applications du modèle de Cox. De plus, nous présentons le modèle de Fine et Gray, qui étend le modèle de Cox en se basant sur les taux de risques des sous-distributions. Cela permet d'estimer les taux de risques spécifiques à des types d'événements particuliers.

Enfin, dans le chapitre 3, nous mettons en pratique les modèles présentés dans le chapitre 2 en réalisant des applications concrètes. Nous utilisons des données réelles ou simulées pour effectuer des ajustements et comparer les résultats dans différents scénarios. Toutes nos analyses ont été effectuées à l'aide du logiciel R, en utilisant divers packages spécifiques à l'analyse de survie. Cette approche nous a permis d'obtenir des informations précieuses sur les relations entre les covariables et les événements d'intérêt, tout en tenant compte des censures potentielles.

CHAPITRE 1

ELÉMENTS DE L'ANALYSE DE SURVIE

1.1 Introduction

Dans ce chapitre, nousdonnons la notion générale des principes fondamentaux de l'analyse de survie, en mettant l'accent sur les définitions essentielles et les principaux types de censure.

L'analyse de survie est une méthode statistique largement utilisée pour étudier la durée de vie ou le temps jusqu'à l'occurrence d'un événement spécifique. Comprendre ces principes est crucial pour mener des analyses de survie adéquates et pour interpréter correctement les résultats obtenus. Nous allons explorer des concepts clés tels que la notion de survie, qui représente la durée de temps avant la réalisation de l'événement d'intérêt, ainsi que le concept de censure, qui se produit lorsque les données ne fournissent pas une information complète sur la durée de survie.

Nous examinerons également les types de censure les plus importants, tels que la censure à droite, la censure à gauche et la censure aléatoire. En acquérant une compréhension solide de ces principes fondamentaux, les chercheurs et les praticiens seront en mesure d'appliquer efficacement l'analyse de survie dans différents domaines d'étude et de recherche, en tirant des conclusions significatives à partir des données de survie.

1.2 Formules en analyse de survie

L'objectif fréquent des études recueillant des données sur la durée avant un événement est d'estimer la probabilité que cet événement se produise après un moment

spécifié. Par exemple, il peut être intéressant de déterminer la probabilité que le temps écoulé entre les épisodes de bronchite soit d'au moins $1 mois$ chez les patients atteints de bronchite chronique. La nature de l'événement peut être soit continue, soit discrèt.

Bien que cette section se concentre principalement sur les situations discrètes, elle présente les formules pour les deux cas. En plus des fonctions statistiques couramment utilisées, telles que la fonction de distribution cumulative (CDF) et la fonction de densité de probabilité (PDF), nous introduisons les concepts spécifiques à l'analyse de survie, tels que la fonction de survie et la fonction de risque.

1.2.1 Distributions continues

Dans l'analyse de survie, la distribution continue est souvent utilisée pour modéliser le temps jusqu'à un événement. L'une des distributions continues les plus couramment utilisées est la distribution exponentielle. Soit T la variable aléatoire pour le temps jusqu'à l'événement, considéré comme continue.

Définition 1.1 *La fonction de survie, notée $S(t)$, est définie comme la probabilité que le temps d'événement dépasse un certain temps t, c'est-à-dire*

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

Si $F(t)$ est le temps jusqu'à l'événement, alors la fonction de survie

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t). \tag{1.1}$$

Définition 1.2 *On défniit la fonction de densité de probabilité (PDF) de la distribution est donnée par la formule :*

$$\begin{aligned} f(t) &= \frac{dF(t)}{dt} \\ &= \frac{d\{1 - S(t)\}}{dt} \\ &= -\frac{dS(t)}{dt}. \end{aligned} \tag{1.2}$$

Définition 1.3 *On appelle fonction de risque cumulé (CDF)la fonction définie par :*

$$F(t) = \int_0^t f(x)dx,$$

la fonction de survie peut être trouvée en intégrant :

$$S(t) = \int_t^{+\infty} f(x)dx.$$

Définition 1.4 *La fonction de risque, également appelée fonction de taux de risque ou fonction de danger (Hazard function en anglais), est une mesure importante utilisée dans l'analyse de survie. Elle représente le taux instantané auquel un événement survient à un moment donné, étant donné que le sujet a survécu jusqu'à ce moment.*
h(t) est définie comme le rapport de la densité de probabilité de l'événement à ce moment t à la probabilité de survie jusqu'à ce moment :

$$h(t) = \lim_{\Delta t \to \infty} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t P(T > t)} \tag{1.3}$$

où, T est une variable aléatoire représentant le temps jusqu'à l'événement,$S(t)$ est la fonction de survie, et Δt est un intervalle de temps très petit.

Cette expression peut être modifiée algébriquement pour obtenir une forme plus acceptable :

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t P(T > t)} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t P(T > t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

La fonction de survie peut être exprimée en termes de fonction de risque en observant que :

$$h(t) = -\frac{d}{dt}\{\log(S(t)\}. \tag{1.4}$$

qui, une fois intégré, devient

$$\log(S(t)) = -\int_0^t h(x)dx.$$

et il s'ensuit que :

$$S(t) = \exp\left(-\int_0^t h(x)dx\right). \tag{1.5}$$

Définition 1.5 *La fonction de risque cumulé est définie comme*

$$H(t) = \int_0^t h(x)d(x).$$

Par conséquent, la fonction de survie peut également être écrite comme :

$$S(t) = \exp(-h(t)). \tag{1.6}$$

Distributions paramétriques courantes

Bien que l'analyse de survie mette souvent l'accent sur les méthodes non paramétriques pour traiter les données de durée de vie, l'utilisation de modèles paramétriques peut s'avérer très utile. Lorsque la distribution proposée est appropriée, l'inférence paramétrique peut être plus efficace que les méthodes qui ne font aucune hypothèse sur la forme de la distribution. Ignorer la forme de la distribution, même si elle est connue, peut conduire à une perte de précision. Plusieurs distributions ou familles de distributions sont fréquemment utilisées dans la recherche pour modéliser les données de temps jusqu'à l'événement. Le choix d'une distribution particulière repose souvent sur des preuves empiriques, où le modèle a été identifié comme décrivant bien les données. Les distributions exponentielle, de Weibull et log-logistique sont couramment utilisées car elles sont relativement faciles à travailler avec des expressions mathématiques définies pour les fonctions de survie et de risque. D'autres distributions telles que la distribution gamma, la distribution gamma généralisée et les distributions log-normales peuvent sembler plus complexes car elles nécessitent des méthodes numériques pour obtenir la fonction de survie. Le tableau (1.2) présente certaines distributions fréquemment utilisées pour décrire les données de durée de vie continue, à titre de référence.

Table 1.1 – Distributions usuelles d'analyse de survie

distribution	f(t)	F(t)	S(t)	h(t)	H(t)
exponentielle	$\lambda e^{-\lambda t}$	1-e$^{-\lambda t}$	$e^{-\lambda t}$	λ	λt
Weibull	$\lambda\theta t^{\theta-1}e^{-\lambda t^{\theta}}$	$1-\exp\{-\lambda t^{\theta}\}$	$\exp\{-\lambda t^{\theta}\}$	$\lambda\theta t^{\Theta-1}$	λt^{θ}
gamma	$\lambda^{\theta}t^{\theta-1}e\{-\lambda t\}$	$\Gamma(\theta,\lambda t)$	$1-\Gamma(\theta,\lambda t)$	$\frac{f(t)}{S(t)}$	$-\log(S(t))$
gamma généralisée	$\frac{\alpha\lambda^{\theta}t^{\alpha\theta-1}\exp\{-\lambda t^{\alpha}\}}{\Gamma(\theta)}$	$(I\Gamma(\theta,\lambda t^{\alpha}))^{\dagger}$	$1-\Gamma(\theta,\lambda t^{\alpha})$	$\frac{f(t)}{s(t)}$	$-\log(S(t))$
log-normal	$\frac{-1/2\left(\frac{\log t-\mu}{\sigma}\right)^2}{\sigma t\sqrt{2\pi}}$	$\phi\left(\frac{\log t-\mu}{\sigma}\right)$	$1-\phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\frac{f(t)}{S(t)}$	$-\log(S(t))$
log-logistique	$\frac{\lambda\theta t^{\theta-1}}{\left(1+\lambda t^{\theta}\right)^2}$	$\Gamma(\frac{\lambda t^{\theta}}{1+\lambda t^{\theta}})$	$\frac{1}{1+\lambda t^{\theta}}$	$\frac{\lambda\theta t^{\theta-1}}{1+\lambda t^{\theta}}$	$\log\left(1+\lambda t^{\theta}\right)$

Γ La fonction gamma incomplète

$$\Gamma(\theta,\lambda t) = \int_0^t \lambda^{\theta}t^{\theta-1}e^{-\lambda t}dt\Gamma(\theta).$$

ϕ La fonction de répartition (CDF) de la distribution normale standard.

1.2.2 Distributions discrètes

En analyse de survie, les distributions discrètes sont utilisées pour modéliser les temps de survie qui sont mesurés de manière discrète, c'est-à-dire à des moments spécifiques. La fonction de probabilité de survie discrète est couramment utilisée pour représenter ces distributions. Elle permet d'estimer la probabilité qu'un événement survienne à un moment précis, en prenant en compte les données discrètes disponibles. Cette fonction est un outil essentiel pour analyser les événements discrets dans le contexte de l'analyse de survie.

Définition 1.6 *La fonction de probabilité de survie discrète est souvent notée $S(t)$ et est définie comme la probabilité qu'un individu survive jusqu'à un temps t donné. Cette fonction est calculée en utilisant la formule suivante : définie comme*

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

Définition 1.7 *La fonction de risque pour les variables aléatoires discrètes de temps jusqu'à l'événement est*

$$S(t) = P(t > T) = \sum_{j \geq 1} P(t_j). \tag{1.7}$$

défini comme :

$$\begin{aligned} h(t_j) &= P(T = t_j | T > t_{j-1}) \\ &= \frac{P(t_j)}{S(t_{j-1})}. \end{aligned}$$
$$pour \quad j = 1, 2, ...$$

Notez que :

$$P(T \geq t_j) = S(t_{j-1}).$$

et depuis

$$p(t_j) = S(t_{j+1}) - S(t_j)$$

(depuis 1.13), il s'ensuit que la fonction de hasard peut s'écrire :

$$h(t_j) = \frac{1 - S(t_j)}{S(t_{j-1})}.$$

	Définition	Autre définition
distribution cumulative	$F(t) = P(T \leq t)$	$F(t) = \int_0^t f(x)dx$
densité de probabilité	$f(t) = \frac{d}{dt}F(t)$	$f(t) = -\frac{d}{dt}S(t)$
survivant	$S(t) = P(T > t)$	$S(t) = 1 - F(t)$
danger	$h(t) = \lim_{\Delta \to 0}\{\frac{p(t \leq T < t+\Delta t \mid T > t)}{\Delta t}\}$	$h(t) = \frac{f(t)}{S(t)}$
danger cumulatif	$H(t) = \int_0^t h(x)dx$	$H(t) = -\log(s(t))$

Table 1.2 – Distributions continues

En utilisant cette relation et le processus d'induction, il est possible d'écrire la fonction de survie en termes de la fonction de risque de la manière suivante :

$$S(t) = \prod_{t_j \neq t}(1 - h(t_j)). \tag{1.8}$$

où **procédure d'induction** fait référence à une méthode ou à un processus permettant d'obtenir des informations générales, des principes ou des conclusions à partir d'exemples spécifiques ou de données spécifiques. Il est souvent utilisé pour tirer des conclusions sur une population plus large à partir d'un échantillon limité.

Les relations entre ces différentes fonctions utilisées dans le l'analyse dans la distribution discet attrayant parce qu'il est intutivement facile à comprendre , en particulier dans paramétres appliques est :

$$H_1(t) = \sum_{tj \neq t} h(t_j) \tag{1.9}$$

Cette définition, cependant, ne préserve pas la propriété $S(t) = \exp H1t$ qui a été vu dans le cas continu. Un suppléant définition qui maintient cette relation est :

$$H_2(t) = \sum_{j \leq t} log(1 - h(t_j) = -log(S(t)) \tag{1.10}$$

Lorsque le danger à chaque instant est faible, ces deux définitions donnent des résultats similaires.

Table 1.3 – Distributions discrètes

	définition	autre définition
cdf	$F(t) = P(T \leq t)$	$F(t) = \sum_{t_j \leq t} p(t_j)$
densité	$p(t_j) = p(T = t_j)$	$p(t_j) = S(t_{j-1}) - S(t_j)$
Survivant	$S(t) = p(T > t)$	$S(t) = \sum_{t_j > t} p(t_j) = \exp(-H_2(t))$
danger	$H_1(t) = \sum_{t_j eqt}$, $h(t_j) = p(t_j)/S(t_{j-1})$	$h(t_j) = p(t_j)/S(t_{j-1})$
danger cumulative	$H_2(t) - \sum_{t_j \leq t} = \log(1 - h(t_j))$	$H_2(t) = -\log S(t)$

Distributions paramétriques courantes

Bien que l'analyse de survie mette souvent l'accent sur les méthodes non paramétriques pour traiter les données de durée de vie, l'utilisation de modèles paramétriques peut s'avérer très utile. Lorsque la distribution proposée est appropriée, l'inférence paramétrique peut être plus efficace que les méthodes qui ne font aucune hypothèse sur la forme de la distribution. Ignorer la forme de la distribution, même si elle est connue, peut conduire à une perte de précision. Plusieurs distributions ou familles de distributions sont fréquemment utilisées dans la recherche pour modéliser les données de temps jusqu'à l'événement. Le choix d'une distribution particulière repose souvent sur des preuves empiriques, où le modèle a été identifié comme décrivant bien les données. Les distributions exponentielle, de Weibull et log-logistique sont couramment utilisées car elles sont relativement faciles à travailler avec des expressions mathématiques définies pour les fonctions de survie et de risque. D'autres distributions telles que la distribution gamma, la distribution gamma généralisée et les distributions log-normales peuvent sembler plus complexes car elles nécessitent des méthodes numériques pour obtenir la fonction de survie. Le tableau (1.2) présente certaines distributions fréquemment utilisées pour décrire les données de durée de vie continue, à titre de référence.

Table 1.4 – Distributions usuelles d'analyse de survie

distribution	f(t)	F(t)	S(t)	h(t)	H(t)
exponentielle	$\lambda e^{-\lambda t}$	1-e$^{-\lambda t}$	$e^{-\lambda t}$	λ	λt
Weibull	$\lambda\theta t^{\theta-1}e^{-\lambda t^{\theta}}$	$1-\exp\{-\lambda t^{\theta}\}$	$\exp\{-\lambda t^{\theta}\}$	$\lambda\theta t^{\Theta-1}$	λt^{θ}
gamma	$\lambda^{\theta}t^{\theta-1}e\{-\lambda t\}$	$\Gamma(\theta,\lambda t)$	$1-\Gamma(\theta,\lambda t)$	$\frac{f(t)}{S(t)}$	$-\log(S(t))$
gamma généralisée	$\frac{\alpha\lambda^{\theta}t^{\alpha\theta-1}\exp\{-\lambda t^{\alpha}\}}{\Gamma(\theta)}$	$(I\Gamma(\theta,\lambda t^{\alpha}))^{\dagger}$	$1-\Gamma(\theta,\lambda t^{\alpha})$	$\frac{f(t)}{s(t)}$	$-\log(S(t))$
log-normal	$\frac{-1/2\left(\frac{\log t-\mu}{\sigma}\right)^2}{\sigma t\sqrt{2\pi}}$	$\phi\left(\frac{\log t-\mu}{\sigma}\right)$	$1-\phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\frac{f(t)}{S(t)}$	$-\log(S(t))$
log-logistique	$\frac{\lambda\theta t^{\theta-1}}{\left(1+\lambda t^{\theta}\right)^2}$	$\Gamma(\frac{\lambda t^{\theta}}{1+\lambda t^{\theta}})$	$\frac{1}{1+\lambda t^{\theta}}$	$\frac{\lambda\theta t^{\theta-1}}{1+\lambda t^{\theta}}$	$\log\left(1+\lambda t^{\theta}\right)$

Γ La fonction gamma incomplète

$$\Gamma(\theta,\lambda t)=\int_0^t \lambda^{\theta}t^{\theta-1}e^{-\lambda t}dt\Gamma(\theta).$$

ϕ La fonction de répartition (CDF) de la distribution normale standard.

1.2.3 Fonction d'incidence cumulative

La fonction d'incidence cumulée (CIF) est le taux de risque à un moment donné, c'est-à-dire le taux d'événements à un temps t qui a été cumulé à partir d'une étape initiale t=0, en prenant en compte les événements survenus entre t=0 et t=t. La formule

de la fonction d'incidence cumulée est donnée par ;

$$h_1(t) = \lim_{\Delta t \to 0} \left(\frac{p_r[t < T \leq t + \Delta t, \delta = 1 | T \geq t]}{\Delta t} \right)^n \tag{1.11}$$

Le CIF est une alternative au taux de risque brut. Cette fonction est définie

$$C_1(t) = \int_0^t h_1(z) S(z) dz$$

Où $h_1(z)$ est le taux de risque de la cause 1 et $S(t)$ est la fonction de survie du temps

1.3 Modèle de censures

Une durée de vie censurée lorsqu'on dispose d'une information limitée sur sa valeur réelle. La censure intervient lorsque le moment exact de l'événement n'est pas connu, mais seulement qu'il se situe dans un intervalle spécifique, généralement en dehors de la période de suivi. On distingue généralement trois types de censure :

1.3.1 Censure à droite

Est utilisé lorsqu'il y a des événements qui peuvent échouer après une certaine période de suivi, mais nous ne disposons que d'informations sur les événements qui se sont produits avant cette période. La formule du modèle de censure à droite est la suivante :

$$f(t) = \lambda(t) \exp(-\lambda(t) T).$$

où :
f(t) la fonction de densité de probabilité de l'événement de censure à droite à l'instant t.
$\lambda(t)$ est le taux d'échec instantané à l'instant t.
T :la durée totale de suivi, c'est-à-dire la période pendant laquelle nous avons collecté des informations sur les événements.

1.3.2 Cencure à gauche

utilisée en analyse de survie lorsqu'on dispose seulement d'informations partielles sur les temps de survie antérieurs à un certain point. Dans ce cas, seuls les individus qui ont survécu jusqu'à cet instant sont inclus dans l'étude, ce qui signifie qu'on n'a pas d'informations sur les temps de survie antérieurs à cet instant. La formule du modèle

de censure à gauche est la suivante :

$$f(t) = \lambda(t)\exp(-\lambda(t)t).$$

où :
f(t) représente la fonction de densité de probabilité (PDF) de l'événement de censure à gauche à l'instant t.
(t) est le taux d'échec instantané à l'instant t, qui peut être une fonction du temps ou une constante.
t est l'instant auquel survient l'événement que nous avons observé.

1.3.3 Censure par intervalle

est une méthode utilisée en analyse de survie lorsque les événements d'intérêt sont partiellement observés dans un intervalle de temps plutôt qu'à des instants précis. Cela signifie que nous disposons d'informations sur les événements qui se sont produits pendant un certain intervalle, mais nous ne connaissons pas les temps de survie exacts au sein de cet intervalle. La formule du modèle de censure par intervalle est la suivante :

$$f(t) = \lambda(t)\exp(-\lambda(t)(T-t))$$

où : f(t) représente la fonction de densité de probabilité de l'événement de censure dans l'intervalle [t, T].
$\lambda(t)$ est le taux d'échec instantané à l'instant t, qui peut être une fonction du temps ou une constante.
T est la durée totale de suivi.
t est l'instant auquel survient l'événement.

1.4 Types de censure

1.4.1 Censure non aléatoire de type I

Définition 1.8 *La censure est dite non aléatoire de type I si, étant donné un nombre positif fixé c et un échantillon $T_1, ..., T_n$, les observations consistent en*

$$X_i = \begin{cases} 1 & si\ T_i \leq c \\ 0 & si\ T_i > c \end{cases}$$

Définition 1.9 *La variable de censure C est définie par la possible non-observation de l'évé-*

nement. Si l'on observe C, et non T, et que l'on sait que $T > C$, on dit qu'il y a une censure à droite. Ce modèle est adapté au cas où l'événement considéré est la date defin d'étude qui est préalablement fixée.

Définition 1.10 *Lorsque nous observons la censure C, et non la durée de survie T, et que l'on sait que $T < C$, un phénomène symétrique au précédent se produit et on dit qu'il s'agit d'une censure à gauche. Ce modèle est adapté au cas où l'incorporation de l'individu dans une étude est conditionnée par un évènement initial.*

1.4.2 Censure de type I

Définition 1.11 *La censure est dite aléatoire de type I si, étant donné un échantillon $T_1, ..., T_n$, il existe une v.a. n-dimensionnelle $(C_1, ..., C_n)$ de $(\mathbf{R}_+)^n$ telle que les observations consistent en $(X_i, \_i)$, où*

$$X_i = T_i C_i = 1_{T_i \leq C_i}$$

où :C_i représente une variable aléatoire C_i dans le contexte de la censure de type I.

1.4.3 Cencure de type II

Ce modèle est typiquement utilisé pour les essais thérapeutiques.
La censure de type *II* se produit si au départ de l'étude, il a été décidé d'observer les heures exactes du première événement. Alternativement, la conception de l'étude peut stipuler que les individus sont suivis pendant une période de temps déterminée, K ; c'est le type I censure.

1.5 Modèles de troncature dans les études de survie

Les troncatures sont différentes des censures car elles impliquent une sélection spécifique dans l'échantillonnage. Lorsqu'une variable X est tronquée par un sous-ensemble potentiellement aléatoire A de l'ensemble des nombres . Les individus de l'échantillon "tronqué" appartiennent tous à A et suivent donc la distribution de T conditionnée par leur appartenance à A. Il est important de ne pas confondre la troncature et la censure. En cas de troncature, une partie des individus (et donc des X_i) ne sont pas observables, et seule un sous-échantillon est étudié, posant ainsi un problème d'échantillonnage. Le biais de sélection est un cas particulier de troncature.

1.5.1 La troncature à droite

A troncature à droite se produit lorsque X n'est observable que si $X < Z$. Cela signifie que seules les valeurs de $X < Z$ sont incluses dans l'analyse, tandis que les valeurs supérieures à Z sont tronquées.

1.5.2 La troncature à gauche

La troncature à gauche se produit lorsque X n'est observable que si X>Z. Dans ce cas, on observe le couple (X, Z) avec$X > Z$. Par exemple, si l'on étudie la durée de vie d'une population à partir d'une cohorte sélectionnée de manière aléatoire, seuls les sujets vivants à la date d'inclusion dans la cohorte pourront être étudiés. Les sujets qui ont survécu jusqu'à cette date sont observables, tandis que ceux qui sont décédés avant ne sont pas inclus dans l'analyse en raison de la troncature à gauche.

1.5.3 La troncature par intervalle

La troncature par intervalle se produit lorsqu'une durée est tronquée à la fois à droite et à gauche. Cela signifie que seules les valeurs de X situées à l'intérieur d'un intervalle spécifié sont incluses, tandis que les valeurs en dehors de cet intervalle ne sont pas prises en compte. Par exemple, dans une étude portant sur des patients d'un registre, seuls les patients diagnostiqués entre deux dates spécifiques seront inclus, tandis que ceux diagnostiqués avant ou après ces dates ne seront pas pris en compte.

1.6 Estimation des fonctions de survie

Une méthode couramment utilisée pour décrire les données de temps d'événement est l'estimation de la limite du produit de la fonction de survie, proposée pour la première fois par Kaplan et Meier (1958) [4].

1.6.1 Estimateur de limite produit

Rappelons que la fonction de survie spécifie la probabilité que le temps de l'événement soit supérieur à t. Dans le cas où il n'y a pas de censure, l'estimateur de limite de produit est identique à l'estimation empirique de la fonction de survie, qui s'obtient simplement en calculant la proportion d'individus ayant vécu l'événement à un certain moment t. Cette approche doit être modifiée lorsque certaines observations sont censurées. Supposons que $t_1, t_2, ..., t_r$ soient les moments de défaillance uniques tels que $t_1 < t_2 < .... < t_r$. Soit d_j le nombre d'événements qui se produisent à l'heure t_j et n_j

le nombre de personnes à risque de subir l'événement immédiatement avant t_j. L'estimateur de limite de produit de la fonction de survie est alors donné par l'équation suivante :

$$\hat{S}(t) = \prod_{t_j \leq t} \frac{n_j - d_j}{n_j}. \tag{1.12}$$

1.6.2 Estimation du maximum de vraisemblance

Nous donnons l'estimation du maximum de vraisemblance de cette expression. La variance de $\hat{S}(t)$ est estimée à l'aide de la formule suivante :

$$\widehat{Var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}. \tag{1.13}$$

1.6.3 Intervalle de confiance

Pour les échantillons de grande taille, $\hat{S}(t)$ suit approximativement une distribution normale. Par conséquent, un moyen simple d'obtenir l'intervalle de confiance à $100(1-\alpha)$ pour $\hat{S}(t)$ est donné par l'équation suivante :

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \sqrt{Var\hat{S}(t)}, \tag{1.14}$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile correspondant au niveau de confiance choisi de la distribution standard (par exemple, si $\alpha = 0,05$, alors $z_{1-\frac{\alpha}{2}} = 1,96$) . Cependant, cet intervalle de confiance peut inclure des valeurs négatives ou des valeurs supérieures à 1. Pour éviter cette difficulté, Kalbfleisch et Prentice (1980) [5]

1.7 Méthodes non paramétriques

Définition 1.12 *On définie les méthodes non paramétriques comme des approches statistiques qui n'imposent pas d'hypothèses spécifiques sur la forme ou la distribution de la population sous-jacente. Contrairement aux méthodes paramétriques, qui supposent une distribution spécifique des données, les méthodes non paramétriques permettent une plus grande flexibilité en s'adaptant aux caractéristiques réelles des données.*

En utilisant des méthodes non paramétriques, on peut analyser les données sans avoir à spécifier explicitement une fonction de densité de probabilité ou une forme fonctionnelle particulière. Ces méthodes sont souvent utilisées lorsque les données ne suivent pas une distribution connue ou lorsque peu d'informations sont disponibles sur la distribution sous-jacente.

1.7.1 Modèles de Kaplan-Meier

Le modèle de Kaplan-Meier est une méthode statistique largement utilisée en analyse de survie pour estimer la fonction de survie d'une population. Il est couramment appliqué dans des études où les temps de survie peuvent varier et où certains individus peuvent être censurés, c'est-à-dire que leur temps de survie n'est pas entièrement observé. La formule de Kaplan-Meier est utilisée pour estimer la fonction de survie au temps donné, en se basant sur les événements de défaillance observés et le nombre d'individus à risque. Ce modèle ne fait pas d'hypothèses sur la distribution sous-jacente des temps de survie et est flexible pour tenir compte de la censure des données. Il permet d'estimer les taux de survie, de comparer les courbes de survie entre différents groupes et d'étudier l'impact de facteurs spécifiques sur la survie.

1.7.2 Tests log-rank et Tests de Wilcoxon

Dans cette section, nous abordons les tests log-rank et les tests de Wilcoxon, qui sont souvent utilisés pour tester l'égalité de deux ou plusieurs fonctions de survie.

1.7.2.1 Test du log-rank

Le test du log-rank, est une procédure non paramétrique couramment utilisée pour comparer les fonctions de survie de deux groupes. Il est basé sur la comparaison des nombres observés et attendus d'événements de défaillance dans les deux groupes.

Dans le cas de la comparaison entre deux groupes, les formules sont les suivantes :
Soit $t_1 < t_2 < ... < t_r$ les temps d'événements uniques et ordonnés pour tous les enregistrements, indépendamment du groupe.
Soient d_{1j} le nombre d'événements de défaillance et n_{1j} les effectifs à risque dans le groupe 1 au temps t_j. De même, d_{2j} et n_{2j} représentent les nombres correspondants pour le groupe 2.

La statistique de test du log-rank est donnée par :

$$U = \sum_{j=1}^{r} (d_{1j} - e_{1j}) \tag{1.15}$$

où $e_{1j} = n_{1j}\frac{d_j}{n_j}$ est le nombre attendu d'événements de défaillance dans le groupe 1 au temps t_j, sous l'hypothèse de survie équivalente entre les deux groupes.

La variance de la statistique U peut être estimée comme suit :

$$\widehat{Var}(U) = \sum_{j=1}^{r} \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \tag{1.16}$$

Le test du log-rank est alors effectué en calculant la statistique :

$$S = \frac{U^2}{\widehat{Var}(U)} \tag{1.17}$$

Si S est significativement différent de zéro, cela suggère une différence significative entre les fonctions de survie des deux groupes.

1.7.2.2 Test général de Wilcoxon pour m groupes

Le test général de Wilcoxon est une extension du test de log-rank pour comparer les fonctions de survie de m groupes. Les formules générales pour m groupes sont les suivantes :

$$u_k = \sum_{j=1}^{r} \left(d_{kj} - n_{kj} \frac{d_j}{n_j} \right), \qquad k = 1, 2, ..., m-1 \tag{1.18}$$

où d_{kj} est la somme des rangs des observations du k-ème groupe pour le j-ème échantillon
n_{kj} est le nombre d'observations dans le k-ème groupe pour le j-ème échantillon.
Les éléments de la matrice de variance-covariance sont donnés par :

$$\sigma_{kk'}^{'} = \sum_{j=1}^{r} \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1} \qquad kk' = 1, 2, ..., m-1 \tag{1.19}$$

où $\sigma_{kk'}^{'}$ est l'élément de la matrice de variance-covariance correspondant aux groupes k et k'.
La statistique de test du test général de Wilcoxon pour m groupes est donnée par :

$$U_W \Sigma_w^{-1} U_w^t \tag{1.20}$$

Sous l'hypothèse nulle, cette statistique suit une distribution du χ^2 avec m-1 degrés de liberté.
Ces tests sont utilisés pour évaluer les différences entre les fonctions de survie dans une étude de survie.
Les statistiques de test et les variances estimées permettent de déterminer la significativité statistique des différences observées entre les groupes.
Dans le test général de Wilcoxon pour m groupes, les éléments de la statistique u_k sont

donnés par :

$$U_k = \sum_{j=1}^{r} \left(d_{kj} - n_{kj} \frac{d_j}{n_j} \right), \qquad k = 1, 2, ..., m-1 \tag{1.21}$$

où :

U_k plutôt une quantité calculée à partir des données observées. d_{kj} est la somme des rangs des observations du k-ème groupe pour le j-ème échantillon, n_{kj} est le nombre d'observations dans le k-ème groupe pour le j-ème échantillon.

d_j est la somme des rangs de toutes les observations pour le j-ème échantillon, et n_j est le nombre total d'observations dans le j-ème échantillon.

La statistique du test général de Wilcoxon pour m groupes est donnée par :

$$U_w \Sigma_w^{-1} U_w^t \tag{1.22}$$

Sous l'hypothèse nulle, cette statistique suit une distribution du χ^2 avec m-1 degrés de liberté.

Exemples de comparaison des tests de log-rank et de Wilcoxon

Comparaison des tests de log-rank et de Wilcoxon pour évaluer les différences entre les fonctions de survie : rôle de la pondération par le nombre d'individus à risque et de la proportionnalité des risques

Table 1.5 – Taux de dangerosité

Temps	Danger Groupe 1	Danger Groupe 2	Danger Groupe 3	Troupe 1 : Groupe 2	Groupe 1 : Groupe 3
1	01	005	0,2	2	0,5
2	02	0101	0,4	2	0,5
3	02	015	0,4	2	0,5
4	03	01	0,6	2	0,5
5	02	005	0,4	2	0,5
6	01	0025	0,2	2	0,5
7	005		0,1	2	0,5

En ce qui concerne le tableau (1.4)donnant un exemple de taux de dangerosité, il illustre une situation où le risque évolue dans le temps pour chaque groupe, tandis que le rapport entre le groupe 1 et le groupe 2 reste constant et entre le groupe 1 et le groupe 3 est de 0,5.

le risque évolue dans le temps pour chaque groupe, le rapport entre le groupe 1 et le groupe 2 est toujours constant (= 2) et entre le groupe 1 et le groupe 3 est de $0,5$.

Notez U est en écrit en terme de groupe 1et que :

$U(groupe1) + U(groupe2) = 0$.

CHAPITRE 2

MODÈLES DE RISQUES CONCCURENTS

2.1 Introduction

Ce chapitre présente une introduction au sujet de la concurrence des risques. Dans l'analyse de survie, les risques compétitifs (ou risques relatifs) font référence à la présence d'événements indépendants du risque d'intérêt qui peuvent influencer la probabilité de l'événement d'intérêt. Ces événements compétitifs sont souvent des événements qui empêchent ou modifient la probabilité de l'événement d'intérêt de se produire.

Lorsque des événements compétitifs surviennent dans une étude de survie, il est important de les prendre en compte car ils peuvent fausser les résultats et l'interprétation des taux de survie. L'omission des événements compétitifs peut conduire à une surestimation de la probabilité de l'événement d'intérêt.

2.2 Définition de risques concurrents

Le risque concurrent est un concept utilisé dans l'analyse de survie pour désigner la présence de multiples événements concurrents qui peuvent affecter le résultat de l'étude de survie. Il fait référence à la possibilité que différents événements, souvent indépendants les uns des autres, puissent se produire simultanément ou à des moments différents chez les individus étudiés.

Dans le contexte de l'analyse de survie, le risque concurrent peut avoir un impact sur la probabilité de survie ou sur le temps jusqu'à l'occurrence de l'événement d'intérêt. Par exemple, dans une étude portant sur la survie des patients atteints de cancer, le risque concurrent peut inclure la mort due à des causes non liées au cancer ou la progression de la maladie vers des stades avancés. Il est essentiel de prendre en compte

les risques concurrents lors de l'analyse de survie, car ils peuvent introduire des biais et affecter les estimations des taux de survie ou des fonctions de survie. Des méthodes statistiques appropriées, telles que les modèles de risques concurrents ou les modèles de compétition, peuvent être utilisées pour tenir compte de ces risques et pour analyser adéquatement les données de survie dans de telles situations.

2.3 Modèles de régression pou les données de survie CR

2.3.1 Modèle à risques proportionnels

Ce modéle représente le contexte théorique pour la construction d'un modèle sous l'hypothèse de risques proportionnels. Le risque pour un individu i avec un ensemble donné de covariables $x_{1i}, x_{2i}, ..., x_{mi}$ peut être décomposé en deux parties : une qui implique le temps mais pas les covariables, et une qui implique les covariables mais pas le temps.
Ainsi,

$$h_i(t) = h_0(t)\exp\{\beta_1 x_{1i} + ... + \beta_m x_{mi}\}. \tag{2.1}$$

où $h_0(t)$ est la fonction de risque de base et $\beta_1, ..., \beta_n$ sont les paramètres ou les coefficients qui doivent être estimés. Comme précédemment, $t_1 < .... < t_r$ sont les points de temps de défaillance ordonnés uniques. Lorsqu'il n'y a pas d'égalité, l'estimation des coefficients $\beta_1, ..., \beta_r$ est obtenue en maximisant la fiabilité partielle.

$$L = \prod_{j=1}^{r} \frac{\exp(\beta_1 x_{1i} + ... + \beta_m x_{mi})}{\sum_{i \in R} \exp(\beta_1 x_{1i} + ... + \beta_m x_{mi})} \tag{2.2}$$

où le produit est pris sur tous les temps de défaillance t_j et R_j représente l'ensemble des individus encore à risque à t_j . La maximisation de la vraisemblance partielle est obtenue par un processus itératif mieux fait sur ordinateur et les estimations $\beta_1, \beta_2, ..., \beta_m$ n'ont pas de forme fermée. Breslow(1974)[6] et Efron (1977)[7] ont suggéré des façons de traiter les observations liées. Des descriptions faciles à suivre de ces méthodes peuvent être trouvées dans Collett (2003)[8] et Therneau et Grambsch (2000)[9]. L'intervalle de confiance pour chaque coefficient $k = 1, 2, ..., m$,

2.3.2 Modèle de Cox

Le modèle de risque proportionnel de Cox est devenu, dans une large mesure, la procédure la plus utilisée pour modéliser la relation des covariables à une survie ou à un autre résultat censuré.

Définition 2.1 *Soit $X_{ij}(t)$ la jième covariable de la ième personne, où $i = 1,...,n$ et $j = 1,...,p$. Il est naturel de penser que l'ensemble des covariables forme un $n*p$ matrice, et nous utilisons X_i pour désigner le vecteur de covariable pour le sujet i, c'est-à-dire la ième ligne de la matrice. Lorsque toutes les covariables sont fixées dans le temps, X_i est juste un vecteur de valeurs des covariables, familier de la régression linéaire multiple. Pour autres ensembles de données une ou plusieurs covariables peuvent varier dans le temps, par exemple un test de laboratoire répété. Nous utilisons X_i à la fois pour les données fixes et variables dans le temps. Processus covariables, employant $X_i(t)$ lorsque l'on souhaite mettre l'accent sur la structure variable dans le temps.*
Le modèle de Cox spécifie le risque pour l'individu i

$$\alpha_J = \alpha_o(t)e^{\beta_0^T}, J = 1,.....j \tag{2.3}$$

où α_o est une fonction non négative non spécifiée du temps appelée la ligne de base hasard, et β est un vecteur colonne $p\times 1$ de coefficients. Les tarifs des événements ne peuvent pas être négatif. Parce que le rapport de risque pour deux sujets avec des vecteurs covariables fixes X_i et X_j est constant dans le temps, le modèle est également connu sous le nom de modèle risques proportionnels.

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)e^{X_i\beta}}{\lambda_0(t)e^{X_j\beta}} = \frac{e^{X_i\beta}}{e^{X_j\beta}} = e^{\beta(X_i - X_j)} \tag{2.4}$$

Définition 2.2 *L'estimation de β est basée sur la fonction de vraisemblance partielle introduite par Cox Pour les données de temps de défaillance non liées, il a la forme :*

$$PL(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left\{\frac{Y_i(t)r_i(\beta,t)}{\sum_j Y_j r_j(\beta,t)}\right\}^{dN_i(t)}. \tag{2.5}$$

où $r_i(\beta,t)$ est le score de risque pour le sujet i,

$$r_i(\beta,t) = \exp[X_i(t)\beta] \equiv r_i(t).$$

Définition 2.3 *Le logarithme de la vraisemblance partielle peut être écrit comme une somme :*

$$\ell(\beta) = \sum_{i=1}^{n}\int_0^\infty \left[Y_i(t)X_i(t)\beta - \log\left(\sum_j Y_j(t)r_j(t)\right)\right]dN_i(t). \tag{2.6}$$

en général, une vraisemblance au sens d'être proportionnelle à la probabilité d'un ensemble de données observés, il peut néanmoins être traité comme une vraisemblance à des fins d'inférence asymptotique.

Définition 2.4 *On définir La dérivation du logarithme de la vraisemblance partielle par*

rapport à β donne le $pX1$ vecteur score, $U(\beta_i(t))$

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} [X_i(s) - \overline{x}(\beta, s)]\, dN_i(s). \tag{2.7}$$

Où $\overline{x}(\beta, s)$ est la moyenne pondérée de X, sur ces observations encore à risque à l'instant s. Où $\overline{x}(\beta, s)$ est défini comme suit :

$$\overline{x}(\beta, s) = \frac{\sum Y_i(s) r_j(s) X_i(s)}{\sum Y_i(s) r_i(s)}$$

avec $Y_i(s), r_i(s)$ comme poids.

Définition 2.5 *On definie la dérivée seconde négative est la matrice d'information $p \times p$ est définie par*

$$I(\beta) = \sum_{i=1}^{n} \int_0^{\infty} V(\beta, s) dN_i(s). \tag{2.8}$$

où $V(\beta, s)$ est la variance pondérée de X au temps s :

$$V(\beta, s) = \frac{\sum_i Y_i(s) r_i(s) [X_i(s) - \overline{x}(\beta, s)]' [X_i(s) - \overline{x}\beta, s)}{\sum_i Y_i(s) r_i(s)}$$

L'estimateur de vraisemblance partielle maximum est trouvé en résolvant l'estimateur partiel équation de vraisemblance :

$$U(\beta) = 0.$$

La solution β est cohérente et asymptotiquement normalement distribuée avec moyenne β vrai vecteur de paramètres, et variance $\{EI\}^{-1}$, l'inverse de la matrice d'information attendue. L'attente exige la connaissance de la distribution de censure même pour les observations avec des échecs observés ; ces informations sont généralement inexistantes. $\beta.dN_i(s)$ représente la dérivée seconde négative.

2.3.3 Modèles de Cox stratifiés

Le modèle de Cox stratifié est une généralisation du modèle de cox usuel , à partir de ce modéle en peut séparé l échentillion l'on J groupes on strates selon les catégories d'une variable explicative .L'hypothèse de proportionalité est toujours vérifiée, a l'in- térieur de ces strates ,la fonction de risque de base diffère mais les covariables Z agissent de la même manière pour les différentes fonctions de risque instantané :

$$\alpha_j(t) = \alpha_{0j}(t) e^{\beta^T}, j = 1, ..., J. \tag{2.9}$$

où $\beta = \beta_0, ..., \beta_k$ est le paramètre de régression commun à toutes les strates et $\alpha_{0j}(t)$,j est la fonction de risque de base spécifique à la strate j. On note par S la variable aléatoire qui renseigne sur la strate de l'individu. Alors de la même façon que dans le cas du modèle de Cox non stratifié, la vraisemblance partielle du n-échantillon $(T_i, \Delta_i, X_i, S_i)$ $1 \leq i \leq n$ partielles à l'intérieur de chaque strate :

$$L^p(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{J} \left\{ \frac{e^{\beta_0^T Z}}{\sum_{K=1}^{n} Y_k(X_i) e^{\beta_0^t Z_i} 1_{(S_K=j)}} \right\} \tag{2.10}$$

où :
$L(\beta)$ est la fonction de vraisemblance.
p est le nombre de strates.
n est le nombre total d'individus.
j représente les strates.
i représente les individus.
β est le vecteur de coefficients de régression.
β_{0j} est le vecteur de coefficients de régression pour la strate .
Z_i est le vecteur de covariables pour l'individu
Y_k est la variable indicatrice de survie de l'individu $k(1sivnement, 0sicensure)$.
S_k est le temps de survie de l'individu

L'estimateur du maximum de vraisemblance partielle est obtenu en maximisant l'expression précédente, ou son logarithme, est donné par :

$$ln(L_n(\beta)) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left[\beta_{0j}^T Z_i - \ln \left(\sum_{k=1}^{n} Y_k e^{\beta_{0j}^T Z_k} 1_{(s_k \geq t_i)} \right) \right] \tag{2.11}$$

Définition 2.6 *Les strates divisent les sujets en groupes disjoints, chacun d'entre eux ayant une fonction de risque de base distincte mais des valeurs communes pour le coefficient. fonction de risque de base distincte, mais des valeurs communes pour le vecteur de coefficient β. Supposons que les sujets $i = 1, ..., n_1$ sont dans la strate 1, les sujets $n_1 + 1, ..., n_1 + n_2$ sont dans la strate 2, et ainsi de suite. Le risque pour un individu i, qui appartient à la strate k. L'analyse des essais cliniques multi-centriques fait souvent appel à la stratification. En raison de la diversité des populations de patients et des schémas d'orientation, les différents centres cliniques participant à l'essai sont susceptibles d'avoir des courbes de survie de base différentes. Les strates jouent un rôle similaire à celui des blocs dans les plans en blocs randomisés analysés par l'analyse de variance à deux voies,par une analyse de variance à deux voies. D'un point de vue informatique, la log-vraisemblance globale devient une somme de*

K :

$$L(\beta) = \sum_{k=1}^{k} l_k(\beta).$$

où : $l_k(\beta)$ est précisément l'équation (2.6), mais additionnée sur les seuls sujets de la strate k. Le vecteur de score et la matrice d'information sont similaires sommes

$$U(\beta) = u_k(\beta).$$

et

$$I(\beta) = I_k(\beta).$$

2.3.4 Modèle de risque proportionnel paramétrique (PH)

Le modèle Cox PH est le plus couramment utilisé pour analyser les données de survie censurées où la distribution de la durée de vie est considérée comme inconnue ou non spécifiée . Le modèle paramétrique à risques proportionnels est la version paramétrique du modèle à risques proportionnels de Cox. Selon Cox (1972) , la fonction de hasard pour la durée de vie T en présence d'un ensemble de covariables $X_1, X_2, ..., X_k$) prend la forme :

$$\begin{aligned} h(x) &= h_0(t) \exp \exp(x^{'} \beta) \\ &= h_0(t) \exp(x_1\beta_1 + x_2\beta_2 + ... + x_k\beta_k). \end{aligned}$$

où $h_0(t)$ désigne la fonction de risque de base au temps t, X désigne le vecteur covariable $k1$ pour une valeur arbitraire individu dans la population et β désigne un vecteur $k1$ de coefficients de régression.
La principale différence entre les deux types de modèles est que la fonction de risque de base est supposée suivre une distribution spécifique lorsqu'un modèle PH entièrement paramétrique est ajusté aux données, alors que le modèle de Cox n'a pas une telle contrainte.
Les coefficients sont estimés par vraisemblance partielle dans le modèle de Cox mais maximum de vraisemblance dans le modèle PH paramétrique. En dehors de cela, les deux types de modèles sont équivalents. Les rapports de risque ont le même l'interprétation et la proportionnalité des dangers est toujours supposée. Un certain nombre de modèles PH paramétriques différents peut être dérivée en choisissant différentes fonction de risques. Les modèles couramment appliqués sont exponentiels et les modèle de Fréchet proposé dans cette étude.

2.3.5 Modèle de hazard proportionnel de Cox (CPH)

Le modèle de régression des hasards proportionnels de Cox (introduit par Cox en 1972) est une méthode d'analyse de survie largement applicable et utilisée pour étudier la relation entre la survie et un ou plusieurs prédicteurs, appelés covariables. Ce modèle permet d'évaluer l'effet des facteurs de risque (variable exploratoire) sur le temps de défaillance par le biais de la fonction de hasard. Dans ce modèle, T représente le temps jusqu'à ce que l'unité subisse une défaillance et Z représente le vecteur observé des covariables. Dans le cadre du modèle CPH, la fonction de risque est définie comme étant la probabilité d'un événement indésirable donné par ;

$$\lambda(t|X) = \lambda_0(t)\exp\{\beta^{'}Z\}. \tag{2.12}$$

La fonction de survie du modèle est donnée par :

$$S(t|Z) = \exp\left(-e^{(\beta^{'}Z)}\Lambda_0(t)\right) = S_0(t)^{\exp(\beta^{'}Z)} \tag{2.13}$$

et la fonction de distribution correspondante à la forme :

$$F(t|Z) = 1 - \exp\left(-\Lambda_0(t)\exp(Z^{'}\beta)\right) \tag{2.14}$$

$\lambda_0(t)$ est le risque de base, β est le vecteur des paramètres de régression, Z est le vecteur des covariables d'une personne.

$$\Lambda_0(t) = \int_0^t \lambda_0(s)ds.$$

est le risque de base cumulatif et $S_0(t)) = \exp(-\Lambda_0(t))$ est la fonction de survie de base.

2.3.6 Modèle de PH exponentiel

La distribution exponentielle est une distribution de probabilité continue avec un seul paramètre inconnu λ. C'est la distribution la plus simple pour les modèles de distribution de durée de vie. La distribution n'est pas assez flexible pour décrire couramment formes de taux de risque rencontrées pour les données de survie. Le pdf, cdf, sf, hrf et chf de la variable aléatoire exponentielle sont, respectivement, comme suit . Soit $X \sim$ exponentielle (λ),

$$f(t) = \lambda e^{-\lambda t}. \tag{2.15}$$

$$F(t) = 1 - e^{-\lambda t}. \tag{2.16}$$

$$S(t) = e^{-H(t)} = e^{-\lambda t}. \tag{2.17}$$

$$h(t) = \lambda. \tag{2.18}$$

$$H(t) = -\log\log S(t) = -\log\log(e^{-\lambda t}) = \lambda t. \tag{2.19}$$

où $\lambda > 0$ est le paramètre d'échelle et $t \geq 0$. Une valeur courte de k indique un faible risque et une longue survie, alors qu'une grande la valeur indique un risque élevé et une courte survie. Pour le modèle PH, le risque exponentiel de base est $h(t) = \lambda$. Ainsi, selon la formulation du cadre PH, le taux de risque pour un individu avec le vecteur covariable X et la fonction de lien $\eta(x)$ est

$$h(t) = h_0(t)\eta(x) = \lambda\eta(x).$$

En appliquant la fonction log-linéaire

$$\eta(x) = \exp(x^{'}\beta).$$

on peut simplifier en

$$h_{EPH}(t) = \lambda \exp\exp(x^{'}\beta) = \lambda \exp(x_1\beta_1 + x_2\beta_2 + ... + x_k\beta_k). \tag{2.20}$$

La distribution exponentielle du hrf dans cette équation, avec le paramètre d'échelle λ, satisfait l'hypothèse PH, comme indiqué par l'expression $(x^{'}\beta)$ Il est important de noter que plusieurs études ont démontré que la distribution exponentielle est insuffisante pour caractériser les données de survie. Cela limite le champ d'application de cette distribution. Voici les distributions de durée de vie alternatives du modèle exponentiel PH. La fonction de survie du modèle de PH exponentiel est :

$$S_{EPH}(t) = \{exp(-\lambda t)\}^{(x^{'}\beta)}. \tag{2.21}$$

Le pdf du modèle exponentiel PH est :

$$f_{EPH}(t) = \lambda. \exp\exp(-\lambda t)(x^{'}\beta)[\exp\exp(-\lambda t)]^{(x^{'}\beta)}. \tag{2.22}$$

Le cdf du modèle exponentiel PH est :

$$F_{EPH}(t) = 1 - [\exp(-\lambda(t)]^{(x^{'}\beta)}. \tag{2.23}$$

Le chf du modèle exponentiel PH est :

$$S_{EPH}(t) = \exp\left(-\lambda t e^{x^{\intercal}\beta}\right), \tag{2.24}$$

2.3.7 Modèle de Fréchet (PH)

La distribution de Fréchet a été initialement proposée par le mathématicien français Maurice Frechet(1878-1973) [10] en 1927. Il s'agit d'une distribution de probabilité continue appartenant aux distributions à queues lourdes et est un cas particulier de la distribution des valeurs extrêmes généralisées lorsque le paramètre de localisation est égal à zéro. Elle est utilisée dans divers domaines, notamment les sciences de la vie, l'ingénierie et l'analyse des événements extrêmes tels que les pluies, la vitesse du vent, les inondations, les tremblements de terre et les tests de durée de vie. Elle est également utilisée pour modéliser les taux d'échec, qui sont couramment utilisés dans la fiabilité et l'analyse des signaux lumi-neux. Les pdf, cdf et chf de la variable aléatoire de Ferchet sont respectivement des flux.
Soit $X \sim Fr(\alpha, \lambda)$.

$$f(t) = \alpha \lambda^{\alpha} t^{-(\alpha-1)} e^{-\left(\frac{\lambda}{t}\right)^{\alpha}}. \tag{2.25}$$

$$F(t) = e^{-\left(\frac{\lambda}{t}\right)^{\alpha}}. \tag{2.26}$$

$$S(t) = 1 - F(t) = 1 - e^{-\left(\frac{\lambda}{t}\right)^{\alpha}}. \tag{2.27}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\alpha \lambda^{\alpha} t^{-(\alpha-1)} e^{-(\frac{\lambda}{t})^{\alpha}}}{1 - e^{(-\frac{\lambda}{t})^{\alpha}}}. \tag{2.28}$$

$$H(t) = -\log\log S(t) = -\log\log\left\{1 - e^{-\left(\frac{\lambda}{t}\right)^{\alpha}}\right\}. \tag{2.29}$$

où $t, \alpha, \beta > 0$. Le paramètre λ est appelé paramètre d'échelle et le paramètre α est appelé paramètre de forme. Ainsi, selon la formulation du cadre PH, le taux de risque pour un individu avec le vecteur covariable x et la fonction de lien $\eta(x)$ est :

$$h(t) \quad = \quad h_0 \eta(t) x) = \frac{\alpha \lambda^{\alpha} t^{-(\alpha-1)} e^{-(\frac{\lambda}{t})^{\alpha}}}{1 - e^{-(\frac{\lambda}{t})^{\alpha}}} \eta(x)$$

En appliquant la fonction log-linéaire $\eta(x) = \exp(x^{'}\beta)$, on peut simplifier en

$$\begin{aligned} h_{FrPH} &= \frac{\alpha\lambda^{\alpha}t^{-(\alpha-1)}e^{-(\frac{\lambda}{t})^{\alpha}}}{1-e^{-(\frac{\lambda}{t})^{\alpha}}}\exp\exp(x^{'}\beta) \\ &= \frac{\alpha\lambda^{\alpha}t^{-(\alpha-1)}e^{-(\frac{\lambda}{t})^{\alpha}}}{1-e^{-(\frac{\lambda}{t})^{\alpha}}}\exp(x_1\beta_1 + x_2\beta_1 + ... + x_k\beta_k). \end{aligned}$$

La distribution de Frechet du hrf dans cette équation, avec le paramètre d'échelle λ et le paramètre de forme β. satisfait l'hypothèse PH, comme le montre l'expression $(x^{'}\beta)$. Voici les distributions alternatives de la durée de vie des Modèle Fréchet PH. La fonction de survie du modèle Frechet PH est :

$$S_{FrPH}(t) = [1 - e^{-(\frac{\lambda}{t})^{\alpha}}]^{(x^{'}\beta)}].$$

La cdf du modèle Frechet PH est :

$$F_{FrPH}(t) = [1 - (e^{(-\lambda t)}]^{(x^{'}\beta)}]. \tag{2.30}$$

Le chf du modèle Frechet PH est :

$$H_{FrPH}(t) = \log\log\left\{1 - e^{-(\frac{\lambda}{t})^{\alpha}}\right\}. \tag{2.31}$$

Le pdf du modèle Frechet PH est :

$$\alpha\lambda^{\alpha}t^{-(\alpha-1)}e^{-(\frac{\lambda}{t})^{\alpha}}1 - e^{-(\frac{\lambda}{t})^{\alpha-1}}. \tag{2.32}$$

La cdf du modèle Frechet PH est :

$$F_{FrPH}(t) = [e^{-(\frac{\lambda}{t})^{(\alpha-1}}]^{(x^{'}\beta)}. \tag{2.33}$$

La chf du modèle Frechet PH est :

$$H_{FrPH} = \log\log\left(1 - e^{-(\frac{\lambda}{t})^{\alpha-1})^{-1}}\right)^{x'\beta}. \tag{2.34}$$

2.3.8 Modèle de Fine et Gray

Fine et Gray (1988)[11] ont proposé un test K-échantillon pour la fonction CIF (Cumulative Incidence Function), ce qui permet une inférence directe sur les fonctions. Le modèle de Fine est un modèle statistique utilisé pour analyser les données de survie en présence de données concurrentes ou compétitives. Il est couramment utilisé pour étudier les taux d'incidence cumulés pour différentes causes de défaillance ou d'événements.

Le modèle de Fine et Gray est une extension du modèle de régression de Cox qui tient compte des causes concurrentes. Il permet d'estimer les effets des covariables sur les taux d'incidence cumulés pour chaque cause spécifique. Contrairement au modèle de régression de Cox, qui est basé sur le taux de risque, le modèle de Fine et Gray est basé sur le taux de risque de sous-distribution. Aussi le modèle Fine-Gray(1988)[11] peut être étendu pour prendre en compte le fait que l'exposition peut également affecter le temps de censure, en autorisant une innovation aléatoire. L'innovation aléatoire suggère un modèle supplémentaire où le temps de censure est déterminé à l'aide d'une fonction de risque proportionnel(Fine et Gray,1999)[1] , qui est définie comme

$$F_j(t, Z_j) = \left(1 - \exp\left\{-\Lambda_0(t).\exp\left(\beta_{0j}.Z_i\right\}\right), i = 1, ..., n. (2.35)$$

où β_{0j} est un vecteur $1 \times p$ de coefficients de régression, lié à la $j^{'}$ème cause, Z_i est un vecteur de $p \times 1$ covariables pour l'individu , et $\Lambda_0(t)$ est un nombre non spécifié, non décroissant ligne de base avec $\Lambda_0(t) = 0$.
Ce modèle ressemble à certains égards à la régression de Cox et a été développé comme un type de modèle de Cox basé sur le taux de risque de sous-distribution.
Plus généralement, toute fonction de lien peut être utilisée pour CIF :

$$F_j(t; Z_j) = h(-\Lambda_0(t), \beta_{0j}, Z_i), i = 1, ..., n. \quad (2.36)$$

De plus, calculez une ligne de base non écroissante $\lambda_0(t)$ et régression coefficient β_{0j} .
Le lien fournit alors le modèle Fine-Gray comme indiqué ci-dessous

$$h_{fg}(a, b, z) = 1 - (\exp(a \exp\{bz\})) \quad (2.37)$$

où b est un coefficient de régression, z est une covariable et a est une ligne de base non décroissante. Fine et Gray (1999) [11] ont pris l'approche de prendre en compte un temps de sous-distribution jusqu'à l'occurrence des risques concurrents pour le type 1 comme :

$$T^{\sim} = \inf\{t > 0, Z_t = 1\}.$$

Si et seulement si $Z_t = 1$ cela équivaut à la durée réelle de l'événement$T$ Sinon, le temps nécessaire car la sous-distribution est infinie. Alors, pour $t \in [0, \infty$, le temps de sous-distribution sera

$$P(T \leq t, Z_T = 1).$$

Aléa de sous-distribution suggéré par Fine et Gray pour l'ajustement d'un modèle de Cox comme :

$$\lambda(t) = -\frac{d}{dt}\log(1 - P(t \leq t, Z_T = 1)) = \frac{P(T > t)}{1 - P(t \leq t, Z_T = 1)}\alpha_{01}(t). \tag{2.38}$$

Le CIF mesure un effet direct sur les événements de type 1 comme :

$$P(T \leq t, Z_T = 1) = 1 - \exp\left\{-\int_0^1 \lambda(u)du\right\}. \tag{2.39}$$

2.3.9 Evaluation du MLE

L'objectif de cette étude est d'évaluer et de dériver les MLE des paramètres pour la méthode FG basée sur CR lorsque les données sont censurées par intervalle. Lorsque les paramètres sont estimées, les performances de ces estimateurs seront évaluées. les covariables seront vérifiées. De plus, pour appliquer et analyser ces inférences procédures sur les données secondaires et les données simulées. Ainsi, les principaux objectifs de ce recherche sont :
1. Estimer les paramètres des modèles FG avec des données censurées par intervalle à l'aide de MLE et méthodes d'imputation.
2. Comparer les performances des méthodes d'imputation médiane pour le modèle FG avec données censurées par intervalle.
3. Comparez les performances de FG avec les données censurées par intervalle et FG avec la droite données censurées via une étude de simulation.
4. Appliquer les techniques proposées à des données médicales réelles.

2.3.10 Modéle CPH pour la sous-distribution

CPH pour la sous-distribution a été présenté par Find-Gray (1999) . Ce modèle est construit sur le modèle de transformation $\log(-log(1-u))$ qui est généralement utilisé avec données de survie. De plus, ce modèle a été estimé en utilisant la sous-distribution

du risque qui est introduit à l'origine par Gray (1988) et donné comme ;

$$\begin{aligned}\lambda_j^*(t,Z) &= \lim_{t\to 0}\frac{1}{dt}P\left(t\le T\le t+dt, C=j|T\le t\cup(T\le t\cap C\neq j),Z\right)\\ &= \frac{dF_j(t,Z)/dt}{1-F_j(t,Z)}\\ &= \frac{-d\log(1-F_j(t,Z))}{dt},\end{aligned} \tag{2.40}$$

où j est la cause d'intérêt de l'échec et λ_j^* est la fonction de risque pour le mauvais variable aléatoire

$$T^* = I(C=j)XT + (1-I(C=j))X\infty.$$

Le temps de panne implicite T^*. De toute évidence, l'ensemble des risques liés au hasard $\lambda_j^*(t,Z)$, est inhabituel, c'est-à-dire que les unités qui sont tombées en panne pour une cause autre que la cause d'intérêt reste indéfiniment dans l'ensemble de risque tant qu'elle n'a pas vécu l'événement qui nous intéresse. Dans une spécification PHs, le risque de sous-distribution est donné comme

$$\lambda_j^*(t,Z) = \lambda_{0j}^*(t)e^{\left(\beta_j' Z\right)}. \tag{2.41}$$

où λ_{0j}^* est une fonction non spécifiée et non négative, et le CIF correspondant est

$$F_j(t,Z) = p(T\le t, C=j|X) = 1-\exp\left(-\Lambda_{0j}^* e^{\beta_j' Z}\right). \tag{2.42}$$

où

$$\Lambda_{0j}^* = \int_0^t \lambda_{0j}^*(s)ds.$$

La vraisemblance totale associée aux données censurées observées est donnée par Kalbfleisch Prentice (1980)[12] comme ;

$$L = \prod_{i=1}^{n_1}\sum_{j\in S_i} p(S_j|T_i, C_i=j, Z_i) f_j(T_i|Z_i)\prod_{n_1+1}^{n_2} S(T_i|Z_i). \tag{2.43}$$

où respectivement n_1 et n_2 représentent le nombre d'échecs et de censures comme type bonnes unités. Comme la relation entre le risque de sous-distribution Λ_j^* et le sous-densité $f_j(t)$ et le CIF a la forme

$$\lambda_j^* = \frac{f_j(t)}{1-F_j(t).} \tag{2.44}$$

et la relation entre la fonction de sous-survie $S_j(t)$ et le CIF $F_j(t)$ a la forme

$$F_j(t) + S_j(t) = p(C = j), \sum_{j=1}^{k} p(C = j) = 1. \tag{2.45}$$

alors (2.52) peut être réécrit comme

$$L = \prod_{i=1}^{n_1} \sum_{j \in S_i} P(S_j|T_i, C_i = j, Z_i) \lambda_j^*(T_i|Z_i) * \prod_{n_1+1}^{n_2} [1 - \sum_{j=1}^{k} F_j(T_j|Z_i)]. \tag{2.46}$$

en plus, en substituant (2.40) et (1.11) dans (2.49) puis la vraisemblance pour le censuré à droite sera comme ;

$$\prod_{i=1}^{n_1} \sum_{j \in S_i} P(S_j|T_i, Z_i) \lambda_{0j}^*(t) e^{-\Lambda_{0j}(t) e^{\beta_j' Z}} (1 - F_j(Ti|Z_i) * \prod_{n_1+1}^{n_2} ([1 - \sum_{i}^{k} 1 - e^{-\Lambda_{0j}(t) e^{\beta_j' Z}}]). \tag{2.47}$$

Maintenant, pour dériver une probabilité pour les données censurées par intervalle (IC), nous devons étendre la probabilité des données RC pour tenir compte des observations censurées par intervalle. En utilisant le relations mentionnées précédemment, la vraisemblance peut être définie comme

$$\begin{aligned} L &= \prod_{i=1}^{n_1} \sum_{j \in S_i} p(S_i|T_i C_i = j, Z_i) \lambda_j^*(T_i|Z_i)(1 - F_j(T_i|Z_i) X \prod_{n_1+1}^{n_2} [1 - \sum_{i=1}^{k}] F_j(T_i|Z_i) \\ X &\prod_{n_2+1}^{n_3} \sum_{j \in S_i} p(S_i|T_i, C_i = j, Z_i)[F_j(R_i|Z_i) - F_j(L_j|Z_i)] \end{aligned} \tag{2.48}$$

à probabilité de volonté censurée par intervalle prend la forme de l'équation (2.53) après en remplaçant (2.12) et (2.13) dans (2.49) par ;

$$\begin{aligned} L &= \prod_{i=1}^{n_1} \sum_{j \in S_i} p(S_i|T_i, C_i = j, Z_i) \lambda_{0j}^*(T_i) e^{\beta_j' Z_i} e^{-\Lambda_{0j}^*(T_i) e^{\beta_j' Z_i}} \\ X &\prod_{i=n_1+1}^{n_2} [1 - \sum_{j=1}^{k} 1 - e^{-\Lambda_{0j}(t) e^{\beta_j' Z}}] \\ X &\prod_{n_2+1}^{n_3} \sum_{j \in S_i} p(S_i|T_i, C_i = j, Z_i)[F_j(R_i|Z_i) - F_j(L_j|Z_i)] \end{aligned} \tag{2.49}$$

en plus, nous prendrons le logarithme de la fonction de vraisemblance des équations (2.48)et (2.49) qui dépend des paramètres inconnus β, les valeurs de Z étant connu. En grand échantillon, la distribution de β peut être approximée par Z une

distribution normale avec le vecteur score, estimée en maximisant la vraisemblance à partir de la dérivée première, et une matrice variance-covariance, estimée à partir de la dérivée seconde de la fonction de vraisemblance. Les coefficients de régression β sont estimés par les valeurs $\hat{\beta}$ qui maximisent la logarithme de la pleine vraisemblance. Les valeurs $\hat{\beta}(\hat{\beta}_1, ..., \hat{\beta}_n)$ sont obtenus en égalant pour mettre à zéro les n premières dérivées de la fonction log de vraisemblance des équations (2.48) et (2.49) par rapport à β. Un processus itératif comme l'algorithme EM ou Newton-Raphson sont adoptés pour résoudre ce système d'équations pour β.

2.4 Modélisation statistique des distributions à priori

Lors de la modélisation statistique des modèles Frechet PH et exponentiel PH, différentes distributions a priori peuvent être utilisées pour les paramètres. Dans cette étude, nous explorons deux types de distributions a priori : la distribution a priori normale pour les covariables et la distribution a priori gamma indépendante pour les paramètres de base des deux modèles. Nous examinons également la distribution postérieure conjointe des paramètres inconnus en utilisant le théorème de Bayes et la fonction de vraisemblance.

2.4.1 Distribution à priori du modèle Frechet PH

Dans le modèle Fréchet PH, différentes distributions a priori sont disponibles pour chaque paramètre. Ces distributions a priori permettent d'introduire des informations préalables sur les paramètres inconnus avant d'observer les données. Bien que des options par défaut soient souvent proposées, l'utilisateur a la possibilité de spécifier explicitement les distributions a priori s'il le souhaite.
Pour le paramètre α, nous supposons une distribution a priori gamma avec les paramètres a_1 et b_1, notée $\alpha \sim \Gamma(a_1, b_1)$. De même, pour le paramètre λ, nous supposons une distribution a priori gamma avec les paramètres a_2 et b_2, notée $\lambda \sim \Gamma(a_2, b_2)$. La fonction de densité de cette distribution a priori est donnée par :

$$p(\alpha) = \frac{b_1^{a_1}}{\Gamma(a_1)} e^{-b_1 \alpha} \alpha^{a_1 - 1}, \quad a_1, b_1, \alpha > 0 \tag{2.50}$$

De même, pour le paramètre λ, nous supposons une distribution a priori gamma avec les paramètres a_2 et b_2, notée $\lambda \sim \Gamma(a_2, b_2)$. La fonction de densité de cette distribution a priori est donnée par :

$$p(\lambda) = \frac{b_2^{a_2}}{\Gamma(a_1)} e^{-b_2 \alpha} \alpha^{a_2 - 1}, a_2, b_2, \lambda > 0.$$

2.4.2 Distribution à priori du modèle exponentiel PH

Nous supposons les priors gamma indépendants pour $p(\lambda) \sim \Gamma(a_2, b_2)$ comme :

$$\Gamma(a_2, b_2) = \frac{b_2^{a_2}}{\Gamma(a_2)} \lambda^{a_2-1} e^{-b_2\lambda}.$$

La fonction de densité de la distribution a priori combinée de tous les paramètres inconnus et des coefficients de régression du modèle exponentiel PH est donnée par

$$p(\beta^{'}) \sim N(a_3, b_3).$$

2.5 Distribution à postériori

La fonction de densité postérieure conjointe des paramètres α, λ et β' du modèle est donnée par : où les termes $p(\alpha)$, $p(\lambda)$ et $p(\beta')$ représentent les distributions à priori des paramètres, et $L(\alpha, \lambda, \beta')$ est la fonction de vraisemblance. La distribution postérieure est obtenue en normalisant cette fonction de densité.

2.5.1 Distribution postérieure du modèle Frechet PH

La fonction de densité postérieure conjointe des paramètres α, λ et β' du modèle Frechet PH, étant donné les données, peut être exprimée en utilisant le théorème de Bayes comme suit :

$$p(\alpha, \lambda, \beta'|x) \propto p(\alpha)p(\lambda)p(\beta')L(\alpha, \lambda, \beta') \tag{2.51}$$

où

$$p(\alpha, \lambda, \beta^{'}|x) \infty p(\alpha)p((\beta^{'})L(\alpha, \lambda, \beta^{'}). \tag{2.52}$$

où les termes $p(\alpha)$, $p(\lambda)$ et $p(\beta')$ représentent les distributions a priori des paramètres α, λ et β' respectivement, et $L(\alpha, \lambda, \beta')$ est la fonction de vraisemblance du modèle Frechet PH. Cette distribution postérieure permet d'obtenir une estimation des paramètres en prenant en compte à la fois les informations a priori et les données observées les deux premiers termes représentent la spécification a priori des paramètres inconnus et sont supposés indépendants et $L(\alpha, \lambda, \beta^{'})$ est la fonction de vraisemblance de l'équation

$$\begin{aligned} L(\alpha, \lambda|t) &= \prod_{i=1}^{n} \left[\frac{\lambda^{\alpha} t^{-(\alpha-1)} e^{-(\frac{\lambda}{t})^{\alpha}}}{1 - e^{-(\frac{\lambda}{t})^{\alpha}}} \exp(x,\beta)] \right]^{\omega_i} \\ &\times \left[\exp\left(- \sum_{i=1}^{n} \log\log(1 - e^{-(\frac{\lambda}{t})^{\alpha})^{-1}} \exp(x,\beta) \right) \right] \end{aligned}$$

$$p(\alpha,\lambda,\beta^{'}|x)\infty\prod_{j=1}^{p}\beta_j\alpha e^{-b_1\alpha}\alpha^{a_2n-1}L(\alpha,\lambda,\beta^{'}). \tag{2.53}$$

2.5.2 Distribution postérieure du modèle exponentiel PH :

La fonction de densité postérieure conjointe des paramètres λ et $\beta^{'}$ du modèle exponentiel PH étant donné que les données peuvent être exprimé en utilisant le théorème de Bayes comme a distribution postérieure du modèle exponentiel PH La fonction de densité postérieure conjointe des paramètres λ et β du modèle exponentiel PH étant donné que les données peuvent être exprimé en utilisant le théorème de Bayes comme

$$p(\lambda\beta^{'}|x \propto p(\lambda)\beta^{'})L(\lambda\beta^{'})$$

où

$$p(\lambda\beta^{'}|x) \propto p(\lambda)p(\beta^{'})L(\lambda\beta^{'})$$

Les deux premiers termes représentent en effet la spécification a priori des paramètres inconnus et sont supposés indépendants. La fonction de vraisemblance correcte pour le modèle exponentiel PH est donnée par :

$$L(\lambda,\beta'|x)=\prod_{i=1}^{n}\left\{\lambda e^{(x_i'\beta')}\right\}^{\omega_i}\exp\left(-\lambda\sum_{i=1}^{n}t_i\exp(x_i'\beta')\right) \tag{2.54}$$

La distribution postérieure conjointe des paramètres λ et β' peut être exprimée comme suit :

$$p(\lambda,\beta'|x)\propto p(\lambda)p(\beta')L(\lambda,\beta'|x) \tag{2.55}$$

$p(\beta')$ est la distribution a priori des coefficients de régression β'. $L(\lambda,\beta'|x)$ est la fonction de vraisemblance du modèle exponentiel PH donnée par l'équation (2.54). où : $p(\lambda)$ est la distribution a priori du paramètre λ et peut être exprimée comme $p(\lambda)\sim\Gamma(a_2,b_2)$, avec la fonction de densité

$$p(\lambda)=\frac{a_2b_2}{\Gamma(a_2)}\lambda^{a_2-1}e^{-b_2\lambda}. \tag{2.56}$$

Dans cette équation, $\Gamma(a_2)$ est la fonction Gamma, a_2 est le paramètre de forme et b_2 est le paramètre d'échelle de la distribution Gamma. La fonction Gamma est définie

comme suit pour un paramètre réel positif x :

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

La distribution marginale des paramètres du modèle et la fonction de densité a posteriori de normalisation sont difficiles à calculer analytiquement et nécessitent généralement des méthodes de simulation telles que les méthodes de Monte Carlo par chaîne de Markov (MCMC) comme Metropolis-Hastings pour obtenir des estimations.

CHAPITRE 3

APPLICATIONS ET ANALYSE DES RÉSULTATS

3.1 Introduction

Ce chapitre porte sur l'analyse de survie et l'utilisation de logiciels dédiés à cette analyse. L'analyse de survie est une méthode statistique utilisée pour étudier le temps jusqu'à l'occurrence d'un événement, comme la survie d'un patient, l'apparition d'une maladie, ou la défaillance d'un équipement. Pour mener ces analyses, plusieurs logiciels sont disponibles, offrant des fonctionnalités avancées pour modéliser et analyser les données de survie.

3.2 Logiciel d'analyse de survie

Il existe plusieurs logiciels d'analyse de survie disponibles pour l'analyse des données de survie. Parmi les logiciels les plus populaires utilisés dans la recherche médicale et la biostatistique, on trouve également le logiciel SAS, le logiciel Stata et le logiciel SPSS et dans étude on utilse logiciel R.
R : est un langage de programmation statistique largement utilisé qui propose de nombreuses bibliothèques dédiées à l'analyse de survie, notamment "survival", "riskRegression", "rms", "cmprsk", etc. Ces packages offrent des fonctions pour effectuer des analyses de survie, ajuster des modèles de survie, estimer des courbes de survie, réaliser des tests de log-rank, etc.

Ces packages offrent des fonctionnalités avancées pour l'analyse de survie, y compris la modélisation de risques proportionnels de Cox, la gestion des données censurées, l'estimation des courbes de survie, les tests de log-rank, l'ajustement de variables covariables, etc. Vous pouvez choisir celui qui convient le mieux à vos besoins en fonction de votre expérience, de vos préférences et des exigences spécifiques de votre ana-

lyse.

3.3 La définition des packages utilisés

3.3.1 Package "survival"

: Le package "survival" est un package en R utilisé pour l'analyse de survie et l'analyse de données censurées. Il fournit des outils statistiques pour modéliser et analyser le temps de survie, la durée jusqu'à un événement, et d'autres variables liées à la survie. Le package "survival" comprend des fonctions pour effectuer des analyses de survie non paramétriques, paramétriques et semi-paramétriques, ainsi que des outils pour la visualisation et l'évaluation des modèles de survie.

3.3.2 Package "cmprsk" :

Le package "cmprsk" est également un package en R utilisé pour l'analyse de survie, mais il se concentre spécifiquement sur l'analyse de données de survie en présence de risques concurrents (ou risques compétitifs). Dans les situations où plusieurs types d'événements peuvent se produire et influencer la survie, ce package permet d'estimer et de modéliser les taux de survie conditionnels à chaque type d'événement, ainsi que les probabilités cumulatives de chaque type d'événement.

3.3.3 Package "rms" :

Le package "rms" (Regression Modeling Strategies) est un package en R développé par Frank E. Harrell Jr[13]. Il fournit des outils pour effectuer des analyses statistiques et de modélisation prédictive. Le package "rms" propose des fonctions pour l'ajustement de modèles de régression, l'estimation des effets, la sélection de variables, la validation des modèles et la création de graphiques diagnostiques. Il est souvent utilisé pour l'analyse de données cliniques et épidémiologiques.

3.3.4 Package "riskRegression" :

Le package "riskRegression" est un autre package en R qui fournit des méthodes avancées pour l'analyse de données de survie. Il se concentre sur les modèles de risques concurrents, les modèles de régression pour des temps de survie dépendants, et les modèles de régression pour des données de survie avec des événements récurrents. Ce package offre des outils pour estimer les taux de risque, effectuer des analyses multi-

variées et ajuster les modèles de régression pour tenir compte de différentes structures de dépendance dans les données de survie.

3.4 Base des données :

A fin de démontrer les méthodes, nous utilisons des données accessibles au public. L'ensemble de données utilisées ici correspond à l'étude sur la maladie de Hodgkin (MH) décrite dans {@pintilie2006compétition}. L'ensemble de données comprend 865 patients diagnostiqués avec stade précoce (I ou II) HD, et qui ont été traités soit par radiothérapie (RT) ou avec radiothérapie et chimiothérapie (CMT).
La maladie de Hodgkin, également appelée lymphome de Hodgkin, est un type de cancer qui affecte le système lymphatique. Voici quelques informations générales sur la maladie de Hodgkin :
Causes : Les causes exactes de la maladie de Hodgkin ne sont pas encore clairement connues. Cependant, certains facteurs de risque ont été identifiés, tels que l'infection par le virus d'Epstein-Barr, des antécédents familiaux de la maladie, une déficience du système immunitaire et l'exposition à certains produits chimiques.
Symptômes : Les symptômes courants de la maladie de Hodgkin comprennent des ganglions lymphatiques enflés (généralement au niveau du cou, des aisselles ou de l'aine), une fièvre inexpliquée, une perte de poids, une fatigue, des démangeaisons, des sueurs nocturnes et des douleurs dans les ganglions lymphatiques après la consommation d'alcool.
Classification : La maladie de Hodgkin est classée en deux grands types : le lymphome de Hodgkin classique (LHC) et le lymphome de Hodgkin à prédominance lymphocytaire nodulaire (LHPLN). Le LHC est le type le plus courant et se divise en quatre sous-types : sclérose nodulaire, cellularité mixte, déplétion lymphocytaire et riches en lymphocytes.
Stades : Le stade de la maladie de Hodgkin est déterminé en fonction de l'extension de la maladie dans le corps. Il existe quatre stades principaux : stade I (la maladie est limitée à un seul groupe de ganglions lymphatiques), stade II (deux ou plusieurs groupes de ganglions lymphatiques du même côté du diaphragme), stade III (ganglions lymphatiques des deux côtés du diaphragme) et stade IV (la maladie s'est propagée à d'autres organes).
Diagnostic : Le diagnostic de la maladie de Hodgkin repose sur une combinaison d'examens médicaux, tels que l'examen physique, l'imagerie médicale (scanner, IRM, PET-CT), la biopsie des ganglions lymphatiques et l'analyse histologique.
Traitement : Le traitement de la maladie de Hodgkin dépend du stade de la maladie et d'autres facteurs individuels. Les options de traitement courantes comprennent la chimiothérapie, la radiothérapie, l'immunothérapie et la greffe de cellules souches.

Les données enregistrées comprennent les covariables suivantes :

- âge : Âge (ans)
- sexe : sexe, F=féminin et M=masculin.
- trtgiven : Traitement administré, RT=Radiation, CMT=Chimiothérapie et radiothérapie
- medwidsi : Taille de l'atteinte du médiastin, N=Non, S=Petit, L=Grand
- extraganglion : maladie extraganglionnaire, Y=maladie extraganglionnaire, N=maladie ganglionnaire
- clinstg : stade clinique, 1=stade I, 2=stade II
- temps : temps jusqu'à l'échec (années) calculé à partir de la date du diagnostic
- statut : 0=censure, 1=rechute et 2=décès.

Nous chargeons et affichons maintenant la structure du jeu de données HD :
library(readr)
Pour procéder à l'analyse, il est important de changer le type de données de sexe, trtgiven, medwidsi et extranod de caractère à facteur. De même, on convertit clinstg du numérique au facteur.

- hd$sex <- as.factor(hd$sex)
- hd$trtgiven <- as.factor(hd$trtgiven)
- hd$medwidsi <- as.factor(hd$medwidsi)
- hd$extranod <- as.factor(hd$extranod)
- hd$clinstg <- as.factor(hd$clinstg)

*M*aintenant, nous explorons le nombre d'événements pour chaque type d'événement :

- r length(which(hd$status==0)) patients censurés,
- r length(which(h$status==1)) avec rechute
- r length(which(hd$status= =2)) qui est décédé.

A partir de maintenant, nous supposons que l'événement d'intérêt est la rechute, c'est-à-dire {status = 1}. A fin de créer un ensemble de test, nous utilisons un échantillonnage stratifié pour partitionner notre données dans :

80% pour l'ajustement
20 % pour le test.

- #require(splitstackshape)
- #set.seed(2022)
- #split_data <- stratified(hd, c("status"), 0.8, bothSets = TRUE)
- #hd_train <- split_data$SAMP1[,-1]
- #hd_test <- split_data$SAMP2[,-1].

Maintenant, nous explorons le nombre d'observations par statut à la fois en apprentissage et en test ensemble :

pander : :pander(table(hd_train))

statut	0 :censure	1 :rechute	2 :deces
nombre d'événement	351	233	108

TABLE 3.1 – Nombre d événements de chaque statut pour l'ajustement

pander : :pander(table(hd_train))

statut	0 :censure	1 :rechute	2 :décès
nombre d'événement	88	58	27

TABLE 3.2 – Nombre d événements de chaque statut pour le test

3.5 Ajustement par le modèle de Cox proportional hasards

Il existe plusieurs packages dans *R* que l'on peut utiliser pour s'adapter à une cause spécifique modèle de régression des risques. A cet effet, séparer les risques proportionnels de Cox. Les modèles sont ajustés pour chaque type d'événement, traitant les autres événements comme observationscensurés . Dans cette section, nous explorons comment nous pouvons effectuer cette tâche avec différents packages *R*.
Nous discutons de ces similitudes et de ces différences : soulignant les aspects importants à considérer dans chaque cas.

3.5.1 Le package survival :

Le modèle peut être ajusté avec la fonction coxph dans le package survival. pour ce faire, il faut d'abord créer un objet de temps de survie avec le Surv fonction. Cet objet sera la variable de réponse dans notre modèle de régression.
À fin d'illustration, nous n'ajustons que le modèle de cause spécifique associé à événements de type 1. Pour ce faire, dans le code ci-dessous, nous avons défini **status==1** dans le champ deuxième argument de **Surv** pour indiquer que nous sommes intéressés par le type d'événement 1.

n= 692
le nombre d évènements= 233 (pour les patients rechute)

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age	0.015879	1.016006	0.004249	3.738	0.000186
sexM	0.087491	1.091432	0.134759	0.649	0.516182
trtgivenRT	0.713903	2.041946	0.194581	3.669	0.000244
medwidsiN	-0.044301	0.956666	0.256368	-0.173	0.862807
medwidsiS	-0.402680	0.668526	0.253404	-1.589	0.112041
extranodY	0.326442	1.386027	0.239225	1.365	0.172384
clinstg2	0.371257	1.449556	0.156389	2.374	0.017600

TABLE 3.3 – Coefficients d' ajusetement par modéle de Cox

ce tableau fournit un résumé des résultats du modèle ajusté. Cela indique que sur le nombre total de patients dans l'ensemble de données d'entraînement 692, un certain nombre 233 d'entre eux ont vécu l'événement d'intérêt, qui dans ce cas est une rechute.

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.0160	0.9842	1.0076	1.025
sexM	1.0914	0.9162	0.8381	1.421
trtgivenRT	2.0419	0.4897	1.3945	2.990
medwidsiN	0.9567	1.0453	0.5788	1.581
medwidsiS	0.6685	1.4958	0.4068	1.099
extranodY	1.3860	0.7215	0.8672	2.215
clinstg2	1.4496	0.6899	1.0669	1.969

TABLE 3.4 – Coefficients de risque par le modéle de Cox

Deuxièmement, il montre les coefficients et les rapports de risque **(exp(coef))** pour chaque covariable, ainsi qu'un test de signification . Les résultats suggèrent que l'âge, le traitement et le stade clinique sont significativement associés avec risque de rechute.

Par exemple, être au stade clinique 2 augmente le risque (taux) de rechute spécifique à la cause par l age18.07 -99.41 par rapport à un patient au stade 1. Il est important de souligner qu'un devrait être prudent lors de l'interprétation de ces effets covariables car cette modélisation stratégie permet seulement de faire des inférences des effets sur danger mais pas sur le pronostic ou la survie (voir plus de détails dans Austin (2016) [14]

Concordance= 0.61 (se = 0.02)
Likelihood ratio test= 37.91 on 7 df, p=3e-06
Wald test = 38.21 on 7 df, p=3e-06
Score (logrank) test = 38.95 on 7 df, p=2e-06

TABLE 3.5 – Résultats des tests de signification du modèle de Cox

Enfin, la sortie rapporte également des statistiques de fitness du modèle.

L'indice de concordance , qui résume la capacité discriminative dans l'échantillon du modèle (une valeur proche de 1 est préférée), et trois tests pour vérifier si (globalement) le modèle est significatif.
Notez que l'estimation naïve des probabilités de survie/risque à partir de ce modèle entraîne une surestimation de la fonction de survie, car les risques concurrents ne sont pas pris en compte.

3.5.2 Le package rms :

Le package **rms** nous permet d'ajuster le modèle de Cox. Ce forfait fournit fonctions utiles pour la validation du modèle et le traçage qui sont bien documentées dans Harrell (2017) [15] .
Avant d'ajuster le modèle, il est important de calculer des statistiques récapitulatives de la covariables qui seront employées. Ces statistiques seront utilisées lors du traçage ou faire des pronostics. Ces statistiques récapitulatives sont calculées à l'aide de (datadist()) fonction.
la fonction "datadist"
Description. Pour un ensemble donné de variables ou un bloc de données, détermine les résumés des variables pour les plages d'effet et de traçage, les valeurs à ajuster et les plages globales pour Predict , plot.
Comme dans la section précédente, nous n'ajustons le modèle spécifique à la cause que pour la première type d'événement. Le code du modèle associé au deuxième type d'événement est analogue, en utilisant **"status == 2" dans l'appel 'Surv()'**. "summary.rms()"

	Model Tests	Discrimination Indexes
obs 692 Events 108 Center 3.5568	LR chi2 159.59 d..f.7 Pr(>chi 2)0.0000 Score chi2 211.82 Pr(>chi 2)0.0000	R2 0.247 R2 (7.692)0.198 R2(7.692)0.757 Dxy 0.631

TABLE 3.6 – Indices de discrimination

est plus utile pour calculer les rapports de risque. Pour continu covariables, les rapports de risque sont calculés par rapport aux quartiles inférieur et supérieur.

Dans ce cas, l'impression du modèle ajusté donne un résultat légèrement différent. Premièrement, la sortie comprend plusieurs indices de discrimination. Noter que, l'indice de concordance peut être récupéré à l'aide de l'indice "Dxy". De plus, notez que ces métriques ignorent la présence de types d'événements concurrents.

	Coef	S.E.	Wald Z	Pr(>|Z|)
age	0.0159	0.0042	3.74	0.0002
sex=M	0.0875	0.1348	0.65	0.5162
trtgiven=RT	0.7139	0.1946	3.67	0.0002
medwidsi=N	-0.0443	0.2564	-0.17	0.8628
medwidsi=S	-0.4027	0.2534	-1.59	0.1120
extranod=Y	0.3264	0.2392	1.36	0.1724
clinstg=2	0.3713	0.1564	2.37	0.0176

TABLE 3.7 – Risque de rechute spésifique

Factor	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95
age	23	43	20	1.555700	0.13734	1.28650	1.824900
Hazard Ratio	23	43	20	4.738300	NA	3.62010	6.201900
sex - F :M	2	1	NA	-0.644930	0.20905	-1.05470	-0.235200
Hazard Ratio	2	1	NA	0.524700	NA	0.34831	0.790420
trtgiven - CMT :RT	2	1	NA	0.062907	0.24193	-0.41127	0.537080
Hazard Ratio	2	1	NA	1.064900	NA	0.66281	1.711000
medwidsi - L :N	2	1	NA	-0.264180	0.40814	-1.06410	0.535770
Hazard Ratio	2	1	NA	0.767830	NA	0.34503	1.708800
medwidsi - S :N	2	3	NA	-0.075774	0.27751	-0.61969	0.468140
Hazard Ratio	2	3	NA	0.927030	NA	0.53811	1.597000
extranod - Y :N	1	2	NA	-0.140510	0.38352	-0.89219	0.611170
Hazard Ratio	1	2	NA	0.868910	NA	0.40976	1.842600
clinstg - 1 :2	2	1	NA	-0.451260	0.22893	-0.89996	-0.002563
Hazard Ratio	2	1	NA	0.636830	NA	0.40659	0.997440

TABLE 3.8 – Coefficients de danger d'une augmentation d'un an

Par exemple, à partir de la sortie ci-dessous, nous pouvons déduire qu'une augmentation de 20 ans de l'âge (passant de 23 à 43), augmente le risque par cause de rechute de 37,38 ce cas correspond à age=15.72 de exp(coef) 0.01 et dont l'interprétation fait référence à l'effet sur le danger d'une augmentation d'un an. Effects Response : Surv(time, status == 2)

De même, pour covariables discrètes, on peut définir le niveau de référence par rapport auquel les rapports de risque sont calculés : Effects Reponse :Surv(times ,status==2)

Ce tableau montre que l'augmentation de risque associée à l'âge et au sexe est respectivement de 10 et 1.172 ans et 0,09. Les effets associés à la bénéficiation du traitement, à la médication et à la présence extra nodal sont respectivement de 0,49, 1,05 et 1,38. La différence entre un stade 1 et un stade 2 est de 0,68.
Nous avons modifié le niveau de référence pour sexe. De plus, on peut définir manuellement les plages dans lesquelles on veut calculer les rapports de risque ; pour allant de 30 à 40 ans, c'est-à-dire qu'une augmentation de 10 ans entraînera une augmenta-

Factor	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95
age	30	40	10	0.777840	0.06867	0.64325	0.912430
Hazard Ratio	30	40	10	2.176800	NA	1.90260	2.490400
sex - M :F	1	2	NA	0.644930	0.20905	0.23520	1.054700
Hazard Ratio	1	2	NA	1.905800	NA	1.26520	2.871000
trtgiven - CMT :RT	2	1	NA	0.062907	0.24193	-0.41127	0.537080
Hazard Ratio	2	1	NA	1.064900	NA	0.66281	1.711000
medwidsi - L :N	2	1	NA	-0.264180	0.40814	-1.06410	0.535770
Hazard Ratio	2	1	NA	0.767830	NA	0.34503	1.708800
medwidsi - S :N	2	3	NA	-0.075774	0.27751	-0.61969	0.468140
Hazard Ratio	2	3	NA	0.927030	NA	0.53811	1.597000
extranod - Y :N	1	2	NA	-0.140510	0.38352	-0.89219	0.611170
Hazard Ratio	1	2	NA	0.868910	NA	0.40976	1.842600
clinstg - 1 :2	2	1	NA	-0.451260	0.22893	-0.89996	-0.002563
Hazard Ratio	2	1	NA	0.636830	NA	0.40659	0.997440

TABLE 3.9 – Augmentation de risque

tion du risque par cause de 17,21 %, comme indiqué ci-dessous.
Le tableau est utilisé pour présenter les résultats d'une analyse statistique qui vise à évaluer l'impact de chaque facteur sur le risque ou le danger, en tenant compte des valeurs basses et élevées ainsi que des intervalles de confiance.
Notez que l'estimation naïve des probabilités de survie à partir de ce modèle entraînera dans la sur estimation de la fonction de survie, car les risques concurrents ne sont pas pris en compte. Voir la section suivante pour une approche valide.
En conclusion l'âge est significativement associé à un risque accru de rechute spécifique, tandis que le sexe (féminin), le traitement administré (RT), et le stade clinique (1 :2) sont associés à une réduction significative du risque de rechute spécifique. Les autres variables (médicaments utilisés, caractéristique spécifique) n'ont pas d'effet significatif sur le risque de rechute spécifique.

3.5.3 Le package riskRegression :

On peut aussi utiliser la fonction CSC de risk CR regression pour ajuster le modèles de régression des risques spécifiques à une cause. Ce package agit comme une interface pour obtenir un objet **CoxPH** ou **CPH** via l'argument **fitter**.
De plus, cette fonction s'adapte simultanément aux modèles des deux événements concurrents lorsque **surv.type = danger**.
Ci-dessous, nous utilisons les mêmes covariables pour les deux types d'événements, mais il est possible d'ajuster différents modèles de régression pour chacun cause (voir exemple ci-dessous).

Notez que dans ce cas, la variable de réponse dans le modèle de régression est obtenu avec la fonction **Hist** disponible dans le Paquet R {prodlim}.
Réponse censurée à droite d'un modèle de risques concurrents Nb d'observations : 692
Modèle

Cause	event	right.censored
1	233	0
2	108	0
unknown	0	351

cause 1
n= 692
number of events= 233

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age	0.015879	1.016006	0.004249	3.738	0.000186 ***
sexM	0.087491	1.091432	0.134759	0.649	0.516182
trtgivenRT	0.713903	2.041946	0.194581	3.669	0.000244 ***
medwidsiN	-0.044301	0.956666	0.256368	-0.173	0.862807
medwidsiS	-0.402680	0.668526	0.253404	-1.589	0.112041
extranodY	0.326442	1.386027	0.239225	1.365	0.172384
clinstg2	0.371257	1.449556	0.156389	2.374	0.017600 *

TABLE 3.10 – Rapports de cotes des facteurs de risque

Affiche les rapports de cotes pour les facteurs de risque inclus dans le modèle de régression logistique multivariable après ajustement pour les covariables. Nous avons observé que l'âge était un prédicteur significatif de la mortalité par cancer du sein à 15 ans, chaque année d'âge supplémentaire augmentant le risque de 1,6%. C'était fourni par trtgivenRT qui avait un rapport de cotes de 2,042, indiquant que le fait d'avoir reçu une radiothérapie était associé à une multiplication par deux du risque de mortalité. Le stade clinique était également associé à un risque accru de mortalité, le stade 4 présentant le risque le plus élevé (OR = 1,334). De plus, nous avons observé que le statut ménopausique n'était pas significativement associé au risque de mortalité (OR = 0,957).

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.0160	0.9842	1.0076	1.025
sexM	1.0914	0.9162	0.8381	1.421
trtgivenRT	2.0419	0.4897	1.3945	2.990
medwidsiN	0.9567	1.0453	0.5788	1.581
medwidsiS	0.6685	1.4958	0.4068	1.099
extranodY	1.3860	0.7215	0.8672	2.215
clinstg2	1.4496	0.6899	1.0669	1.969

TABLE 3.11 – Estimations du risque relatif pour les patients atteints de cancer

Le tableau comprend les variables âge, sexe (avec le masculin comme codé « M »), si le traitement a été administré par radiothérapie (trtgivenRT), si un élargissement médiastinal a été observé à l'imagerie (medwidsiN pour non et medwidsiS pour oui) et si le patient avait un extension extranodale (extranodY pour oui). La dernière colonne est le stade clinique du patient (clinsg2).

Concordance = 0.61 (se = 0.02)
Likelihood ratio test = 37.91 on 7 df, p=3e-06
Wald test = 38.21 on 7 df, p=3e-06
Score (logrank) test = 38.95 on 7 df, p=2e-06

TABLE 3.12 – Résultats statistiques

Il contient les informations suivantes : le coefficient de concordance, le test de la rapport des vraisemblances, le test de Wald et le test de Score (logrank). Ces tests ont été réalisés sur sept variables indépendantes et le rapport p indique le niveau significatif associé aux tests.

Cause :2

n= 692

number of events= 108

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age	0.077784	1.080889	0.006867	11.327	< 2e-16
sexM	0.644928	1.905850	0.209051	3.085	0.00204
trtgivenRT	-0.062907	0.939031	0.241929	-0.260	0.79485
medwidsiN	0.264183	1.302367	0.408144	0.647	0.51745
medwidsiS	0.188405	1.207322	0.404631	0.466	0.64149
extranodY	-0.140512	0.868913	0.383517	-0.366	0.71408
clinstg	2	0.451261	1.570291	0.228931	1.971 0.04871

TABLE 3.13 – Coefficien cts du modèle logistique

Le tableau ci-dessus affiche les coefficients, leurs exposants, les erreurs standard, les scores Z et les probabilités associées à sept variables différentes. Les valeurs les plus élevées se trouvent pour (age) et (sexM), indiquant que ces facteurs sont les plus

influents sur le résultat d'intérêt. Les vitesses précédentes (trtgivenRT) des patients et l'état de l'extranoïde (extranodY) sont les moins significatives car les valeurs z et p sont les plus faibles.
Enfin, (clinstg) revêt une certaine importance mais pas en termes de significativité statistique.

Concordance= 0.816 (se = 0.022)
Likelihood ratio test= 159.6 on 7 df, p=<2e-16
Wald test = 161.9 on 7 df, p=<2e-16
Score (logrank) test = 211.8 on 7 df, p=<2e-16

TABLE 3.14 – Erreurs standard à variables différentes

Maintenant, la sortie affiche des résumés de modèle pour les deux causes. Si nous analysons Cause : 1 les résultats doivent être identiques à ceux obtenus avec **CoxPH** dans la survie package, et sont interprétés de la même manière. Les résultats obtenus en utilisant un ajustemrnt par *"cph"* sont les mêmes que ci-dessus et sont présentés dessous :

	Model Tests	Discrimination
		Indexes
Obs 692 Events 108 Center 3.5568	LR chi2 159.59 d.f. 7 Pr(> chi2) 0.0000 Score chi2 211.82 Pr(> chi2) 0.0000	R2 0.247 R2(7,692)0.198 R2(7,108)0.757 Dxy 0.631

TABLE 3.15 – Ajustement par CPH par rapport à la cause 1

	Coef	S.E.	Wald Z	Pr(>|Z|)
age	0.0778	0.0069	11.33	<0.0001
sex=M	0.6449	0.2091	3.09	0.0020
trtgiven=RT	-0.0629	0.2419	-0.26	0.7948
medwidsi=N	0.2642	0.4081	0.65	0.5175
medwidsi=S	0.1884	0.4046	0.47	0.6415
extranod=Y	-0.1405	0.3835	-0.37	0.7141
clinstg=2	0.4513	0.2289	1.97	0.0487

TABLE 3.16 – Ajustement par CPH par rapport à la cause 2

Supposons que nous souhaitions faire des prédictions de risque au temps $t = 5$ ans pour un nouvel ensemble de patients (c'est-à-dire l'ensemble de données de test *hd_test*).
A titre d'illustration, nous se concentrer sur la prédiction si un événement de type 1

se produira à ce moment-là (pour prédire l'occurrence du deuxième type d'événement, utilisez $cause = 2$ dans le code dessous).
Le package **riskRegression** permet d'obtenir des prédictions de risque absolu à des moments précis.
Pour chaque patient de ce nouvel ensemble, il s'agit de calculer la probabilité d'observer un événement de type 1 d'ici *ans*. Ici, nous définissons **product.limit = FALSE**, pour utiliser l'approximation exponentielle $S(t) = exp(-H_1(t) - H_2(t))$, où $H_j(t)$ désigne le risque cumulé pour cause j au temps t.
Ci-dessous, nous montrons le risque prédit pour les 5 premiers patients de l'ensemble de données de test. Nous voyons comment les deux ajustements **(Coxph ou cph)** donnent le même résultat :

[1]	0.2990793	0.4231062	0.3337163	0.3216817	0.2279138

Table 3.17 – Prédiction risk.CoxPH[1 :5]

[2]	0.2990793	0.4231062	0.3337162	0.3216816	0.2279138

Table 3.18 – Prédiction risk.CPH[1 :5]

Enfin, notez également que **riskRegression** permet d'ajuster différentes régressions modèles pour chaque cause en utilisant les risques proportionnels de Cox.
Dans l'exemple ci-dessous,nous utilisons toutes les covariables pour la cause 1, et nous n'employons que l'âge et le sexe pour la cause 2.
Cause : 1 n= 692
number of events= 233

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age	0.015879	1.016006	0.004249	3.738	0.000186 ***
sexM	0.087491	1.091432	0.134759	0.649	0.516182
trtgivenRT	0.713903	2.041946	0.194581	3.669	0.000244 ***
medwidsiN	-0.044301	0.956666	0.256368	-0.173	0.862807
medwidsiS	-0.402680	0.668526	0.253404	-1.589	0.112041
extranodY	0.326442	1.386027	0.239225	1.365	0.172384
clinstg2	0.371257	1.449556	0.156389	2.374	0.017600 *

Table 3.19 – Ajusement des covariables par la cause 1

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.077	0.9289	1.064	1.089
sexM	1.906	0.5247	1.279	2.840

TABLE 3.23 – Coefficients d'estimation

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.0160	0.9842	1.0076	1.025
sexM	1.0914	0.9162	0.8381	1.421
trtgivenRT	2.0419	0.4897	1.3945	2.990
medwidsiN	0.9567	1.0453	0.5788	1.581
medwidsiS	0.6685	1.4958	0.4068	1.099
extranodY	1.3860	0.7215	0.8672	2.215
clinstg2	1.4496	0.6899	1.0669	1.969

TABLE 3.20 – Ajustement des covariables par la Cause 2

Concordance= 0.61 (se = 0.02)
Likelihood ratio test= 37.91 on 7 df, p=3e-06
Wald test = 38.21 on 7 df, p=3e-06
Score (logrank) test = 38.95 on 7 df, p=2e-06

TABLE 3.21 – Intervalles de confiance à 95 % pour les coefficients de la Cause 2

Cause : 2

number of events= 108

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
age	0.073799	1.076590	0.005933	12.439	< 2e-16
sexM	0.644987	1.905962	0.203404	3.171	0.00152

TABLE 3.22 – Contrôle du Résultat de l'Ajustement de Cox

Concordance= 0.821 (se = 0.02)
Likelihood ratio test= 155.1 on 2 df, p=<2e-16
Wald test = 161.4 on 2 df, p=<2e-16
Score (logrank) test = 209.1 on 2 df, p=<2e-16

TABLE 3.24 – Tests de significativité

3.6 Ajusetement par modéle de FINE-GRAY

Contrairement aux modèles qui utilisent une spécification CPH spécifique à une cause, cette approchepermet des inférences pour les effets covariables à la fois sur la fonction de risque et la fonction de survie due au fait que les événements concurrents

sont traités différemment.
Comme pour les modèles **CPH** spécifiques à une cause, il existe plusieurs packages R pour s'adapter à un modèle de risques de sous-distribution. Ici, nous explorons **cmprsk** et **riskRegression**.

3.6.1 Le package cmprsk

Pour ajuster le modèle, il faut d'abord créer une matrice de conception contenant les covariables d'intérêt, cela peut être fait en utilisant le **model.matrix()** fonction, qui crée des variables fictives pour les prédicteurs discrets.

Intercept	age	sex	trtgivenRT	medwidsiN	medwidsiS	extranodY	clinstg2
1	1 55.00	1	1	1	0	0	0
2	1 51.00	1	1	0	1	1	0
3	1 19.00	1	1	1	0	0	0
4	1 52.23	1	0	1	0	0	1
5	1 41.00	0	1	1	0	0	1
6	1 25.00	0	1	1	0	0	0

TABLE 3.25 – Matrice de conception utilisée pour ajuster un modèle

Notez que nos covariables correspondent maintenant à des indicateurs binaires (1 ou 0) d'un niveau spécifique des variables factorielles d'origine.

Pour utiliser cette conception matrice, nous allons d'abord supprimer l'ordonnée à l'origine. Une fois cela fait, nous pouvons employer la fonction **crr()** pour ajuster le modèle.

	coef	exp(coef)	se(coef)	z	p-value
sexM	0.0571	1.059	0.136	0.420	0.670
trtgivenRT	0.6700	1.954	0.204	3.291	0.001
medwidsiN	0.1128	1.119	0.252	0.447	0.650
medwidsiS	-0.3621	0.696	0.254	-1.425	0.150
extranodY	0.3714	1.450	0.247	1.503	0.130
clinstg2	0.2792	1.322	0.149	1.868	0.062

TABLE 3.26 – Régression des risques concurrents

	exp(coef)	exp(-coef)	2.5%	97.5%
sexM	1.059	0.944	0.811	1.38
trtgivenRT	1.954	0.512	1.311	2.91
medwidsiN	1.119	0.893	0.683	1.84
medwidsiS	0.696	1.436	0.423	1.15
extranodY	1.450	0.690	0.893	2.35
clinstg2	1.322	0.756	0.986	1.77

TABLE 3.27 – limites de confiance du coefficient

Num. cases = 692
Pseudo Log-likelihood = -1462 Pseudo likelihood ratio test = 33.1 on 7 df

TABLE 3.28 – Pseudo Log-likelihood et de Pseudo likelihood ratio test.

La sortie ci-dessus inclut les effets covariables estimés **exp(coef)**, comme ainsi que le seuil de signification à partir duquel nous pouvons déduire que les effets marginaux pour l'âge, le traitement administré et le stade clinique sont importants.
Notez cependant que les rapports de risque estimés ne sont pas identiques à ceux déduits dans vignette spécifique à la cause.

En effet, les risques fixés diffèrent entre ces deux fonctions de danger.
Dans ce contexte, nous pouvons faire des inférences à la fois sur les effets covariables sur le risque de sous-distribution et les effets directionnels des covariables sur le CIF.

Par exemple, le rapport de risque de rechute de la sous-distribution est clinstg2 1.402 lorsqu'un le patient est au stade clinique 2.
De plus, comme le rapport de risque est plus élevé plus d'un, on peut également en déduire que le fait d'être au stade clinique 2 augmente la incidence des rechutes.
Ici, il est important de souligner que nous ne pouvons pas utiliser le l'ampleur des rapports de risque pour faire des déductions sur l'ampleur de l'effet sur la probabilité d'occurrence.
chaque patient (compte tenu de leurs valeurs de covariable) à ces moments d'événement.

En tant que illustration, le code suivant est utilisé pour tracer le CIF estimé pour les événements type 1 pour deux patients dans l'ensemble de données d'entraînement.
Le premier patient a 37 ans âgé, de sexe masculin, sous radiothérapie, pas d'atteinte du médiastin, pas d'extranodal maladie, et au stade clinique 1.
Le deuxième patient a 41 ans, homme, sous rayonnement, pas d'atteinte du médiastin, pas de maladie extra-ganglionnaire, et clinique étape 2.
Le graphique représente la courbe de survie cumulative (CIF) pour un événement de

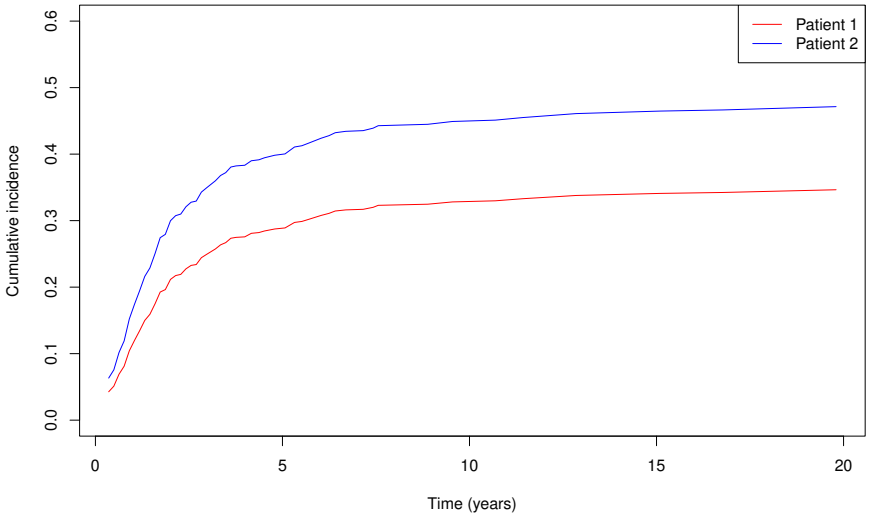

FIGURE 3.1 – La curbe pour CIF pour l'évènement de type I

type 1. Le patient 2 est un homme de 41 ans, sous traitement par rayonnement, sans atteinte du médiastin, sans maladie extra-ganglionnaire, et à l'étape 2 de la maladie. La courbe montre comment la probabilité de survie du patient évolue au fil du temps. Une diminution rapide de la courbe suggère une probabilité plus faible de survie ou une évolution défavorable de la maladie, tandis qu'une courbe plus plate ou ascendante indique une meilleure probabilité de survie ou une réponse positive au traitement.
Cependant, ces résultats sont spécifiques au patient 2 et à l'événement de type 1, et ne peuvent pas être généralisés à d'autres patients ou événements. En utilisant **cmprsk**, il n'est pas simple de faire des prédictions à un point de temps spécifique au-delà des points de temps uniques présents dans les données de formation.

3.6.2 Le package riskRegression

ce package R agit comme un rapport de la fonction **crr** décrite ci-dessus.

Caubukarse	event	right.censored
1	233	0
2	108	0
unknown	0	351

TABLE 3.29 – Réponse censurée à droite d'un modèle de risques concurrents

Contrairement à **cmprsk**, **riskRegression** permet de faire des prédictions du CIF à un point de temps spécifique (par exemple $t=5$ ans) pour un nouvel ensemble de données (par exemple hd_test).
Dessous, nous montrons les résultats pour les 5 premiers sujets de l'ensemble de test.

age"	"sexM"	"trtgivenRT"	"medwidsiN"	"medwidsiS"	"extranodY"	"clinstg2"

Les points de temps sélectionnés sont affichés, utilisez **plot.riskRegression** pour étudier la forme complète.

0.2964653	0.4056692	0.3361023	0.3202368	0.2351294

TABLE 3.30 – FG.prediction[c(1 :5)]

Car le modèle Fine-Gray peut être récupéré lorsqu'un lien log-log complémentaire fonction est utilisée dans un modèle de transformation (gerds2012), nous pouvons utiliser **riskRegression()** avec **link = prop**
riskRegression : modèle de régression des risques concurrents Estimation de l'IPCW. Les poids sont basés sur l'estimation de Kaplan-Meier pour la distribution de censure. Fonction de lien : "proportionnel" produisant des rapports de sous-risque (Fine Gray 1999), Covariables ayant des effets variables dans le temps : Interception (numérique) Les effets de ces variables dépendent du temps. La colonne 'Intercept' est le risque de base où toutes les covariables ont la valeur zéro

	(Intercept)
0.36	"0.0123"
2	"0.0670"
3.8	"0.0906"
6.4	"0.1064"
20	"0.1199"

TABLE 3.31 – Risque de base

Les points de temps sélectionnés sont affichés, utilisez {plot.riskRegression} pour étudier la forme complète. Covariables à effets constants dans le temps :

age	(numeric)
sexM	(numeric)
trtgivenhlRT	(numeric)
medwidsiN	(numeric)
medwidsiS	(numeric)
extranodY	(numeric)
clinstg2	(numeric)

TABLE 3.32 – Coéfficients des covariable à effet constant

Coefficients de régression de la constante de temps :

Factor	Coef	exp(Coef)	StandardError	z	CI_95	Pvalue
age	0.01649	1.01663	0.00431	3.82626	[1.008 ;1.025]	0.0001301
sexM	0.0133	1.0134	0.1419	0.0939	[0.767 ;1.338]	0.9251590
trtgivenRT	0.712	2.037	0.217	3.283	[1.332 ;3.115]	0.0010271
medwidsiN	-0.0687	0.9336	0.2768	-0.2481	[0.543 ;1.606]	0.8040708
medwidsiS	-0.407	0.665	0.274	-1.489	[0.389 ;1.137]	0.1364193
extranodY	0.231	1.259	0.254	0.908	[0.766 ;2.071]	0.3637606
clinstg2	0.339	1.403	0.163	2.082	[1.020 ;1.930]	0.0373014

TABLE 3.33 – Coefficients de régression de la constante de temps

À partir du tableau des coefficients de régression de la constante de temps que vous avez fourni, voici une interprétation générale des résultats :

L'âge (age) est statistiquement significatif avec un coefficient de 0.01649. Cela suggère qu'une augmentation d'une unité dans l'âge est associée à une augmentation d'environ 1.6% du risque de l'événement étudié (la constante de temps), toutes les autres variables étant constantes.

Le sexe masculin (sexM) n'est pas statistiquement significatif, avec un coefficient de 0.0133 et une valeur de p élevée de 0.9251590. Cela indique qu'il n'y a pas de différence significative dans la constante de temps entre les sexes masculin et féminin, toutes les autres variables étant constantes.

Le traitement avec RT (trtgivenRT) est statistiquement significatif avec un coefficient de 0.712. Cela suggère que les patients qui reçoivent ce traitement ont un risque environ deux fois plus élevé (exp(0.712) 2.037) de l'événement étudié, comparés à ceux qui ne le reçoivent pas, toutes les autres variables étant constantes.

Les autres variables, medwidsiN, medwidsiS, extranodY et clinstg2, ne sont pas toutes statistiquement significatives. Cela indique qu'elles n'ont pas d'effet significatif sur la constante de temps, compte tenu des autres variables dans le modèle.

Note : Les valeurs exp(Coef) sont des sous-hazard ratios (Fine Gray 1999). Dans ce cas, **predict()** renvoie le risque aux points de temps observés pour chaque sujet. prediction.FG.prop <- predict(FG.prop, newdata = hd_test[c(1, 2),]) Ci-dessous, nous traçons le **CIF** pour le patient 1 (rouge) et 2 (bleu) dans l'ensemble de test.

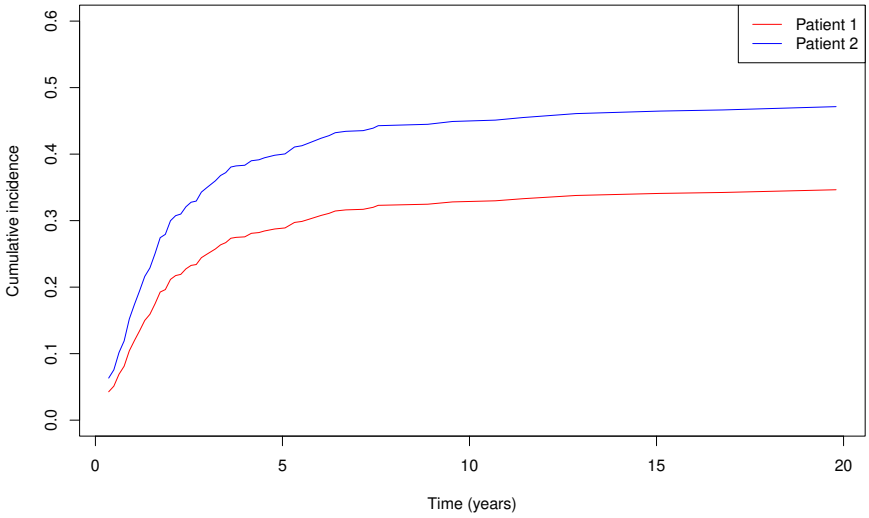

FIGURE 3.2 – La curbe de CIF pour les patients dans l'ensemble de test

La figure représente les courbes de survie cumulative (CIF) pour les patients 1 (en rouge) et 2 (en bleu) dans l'ensemble de test. La courbe rouge montre une diminution plus rapide de la probabilité de survie pour le patient 1 par rapport au patient 2. Cela suggère que le patient 1 présente une évolution plus défavorable de la maladie ou une réponse moins favorable au traitement. La figure permet de visualiser et de comparer les trajectoires de survie des patients, fournissant des informations sur leur pronostic ou l'efficacité des interventions.

CONCLUSION GÉNÉRALE

L'objectif principal de notre travail était de présenter différentes méthodes d'estimation de la fonction de survie. Nous avons utilisé des méthodes telles que l'incidence cumulée et la fonction de survie (CIF) ainsi que des modèles de régression du risque selon la cause tels que Cox, cph, Fine et Gray. Ces méthodes peuvent être utilisées pour effectuer diverses analyses dans le domaine de la survie.

Le modèle de régression de Cox est utilisé pour étudier l'effet des variables explicatives sur le temps jusqu'à la survenue d'un événement spécifique. Son objectif est de déterminer comment ces variables influencent le taux de risque de l'événement au fil du temps, en supposant que l'effet des variables explicatives est proportionnel sur le temps.

Le modèle Fine-Gray, également connu sous le nom de modèle de risques compétitifs, est utilisé lorsque plusieurs types d'événements concurrents peuvent se produire. Son objectif est de comprendre comment les variables explicatives influencent le temps jusqu'à la survenue d'un événement spécifique, en tenant compte de la présence des autres événements concurrents. Il permet d'analyser la probabilité cumulative d'occurrence d'un événement spécifique en présence de la compétition des autres événements.

Le modèle de régression de Cox proportionnelles (CPH) est une extension du modèle de régression de Cox. Son objectif est d'estimer les coefficients de régression associés à chaque variable explicative et de tester leur significativité statistique. Ce modèle suppose également que le rapport des taux de risque est constant au fil du temps, ce qui permet d'évaluer l'effet des variables explicatives de manière proportionnelle sur le temps de survie. Nous avons utilisé des données accessibles au public pour illustrer les méthodes d'estimation. Plus précisément, nous avons utilisé une base de données sur la maladie de Hodgkin, qui comprenait des informations sur des patients diag-

nostiqués avec un stade précoce de la maladie et les traitements qu'ils ont reçus. Les variables enregistrées comprenaient l'âge, le sexe, le traitement administré, la taille de l'atteinte du médiastin, la présence de maladie extra ganglionnaire, le stade clinique, le temps jusqu'à l'échec (mesuré en années) et le statut du patient (censure, rechute ou décès).

Dans cette partie, nous avons chargé cet ensemble de données sur la maladie de Hodgkin et effectué des transformations sur les variables pour les convertir en facteurs afin de faciliter l'analyse. Nous avons également exploré le nombre d'événements pour chaque type de statut (censure, rechute ou décès). Ensuite, nous avons utilisé un échantillonnage stratifié pour partitionner les données en un ensemble d'entraînement (80%) et un ensemble de test (20%). Nous avons présenté le nombre d'observations pour chaque statut dans l'ensemble d'entraînement.
Enfin, les résultats obtenus par Cox, Fine et Gray est indiqué que la méthode de Fin et Gray mieux que celle de Cox proportionnelle Cependant, elle ne fourni pas la précision nécessaire par rapport aux méthodes de Fine et Gray. De plus, les facteurs déterminants étaient l'âge , le sexe , le traitements et le statut du patient . En conclusion, ce travail s'est concentré sur l'examen des estimateurs de données existants pour les données soumises à la censure. Nous avons présenté différentes méthodes d'estimation de la survie et nous avons utilisé la maladie de Hodgkin comme exemple pour applique ces méthodes.
Les résultats obtenus ont montré que la méthode de Fin et Gray était plus efficace, mais que les méthodes de Cox proportionnelle étaient également importantes pour obtenir des informations plus précises. Ce travail ouvre la voie à de futures recherches dans le domaine de l'estimation de la survie et de son application dans d'autres domaines médicaux.

BIBLIOGRAPHIE

[1] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446) :496–509, 1999.

[2] Ross L Prentice, Benjamin J Williams, and Arthur V Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2) :373–379, 1981.

[3] Martin Wolkewitz, Arthur Allignol, Stephan Harbarth, Giulia de Angelis, Martin Schumacher, and Jan Beyersmann. Time-dependent study entries and exposures in cohort studies can easily be sources of different and avoidable types of bias. *Journal of clinical epidemiology*, 65(11) :1171–1180, 2012.

[4] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282) :457–481, 1958.

[5] John D Kalbfleisch and Ross L Prentice. Estimation of the average hazard ratio. *Biometrika*, 68(1) :105–112, 1981.

[6] Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

[7] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359) :557–565, 1977.

[8] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i : basic concepts and first analyses. *British journal of cancer*, 89(2) :232–238, 2003.

[9] Terry M Therneau, Patricia M Grambsch, Terry M Therneau, and Patricia M Grambsch. *The cox model*. Springer, 2000.

[10] Michel Armatte. Maurice fréchet statisticien, enquêteur et agitateur public. *Revue d'histoire des mathématiques*, 7(1) :7–65, 2001.

[11] Stephen B Salter and Frederick Niswander. Cultural influence on the development of accounting systems internationally : A test of gray's [1988] theory. *Journal of international business studies*, 26 :379–397, 1995.

[12] John D Kalbfleisch and RL Prentice. Survival analysis, 1980.

[13] Frank E Harrell, Jr and Frank E Harrell. Binary logistic regression. *Regression modeling strategies : With applications to linear models, logistic and ordinal regression, and survival analysis*, pages 219–274, 2015.

[14] Robert A Cummins and Kenneth C Land. Capabilities, subjective wellbeing and public policy : a response to austin (2016). *Social Indicators Research*, 140 :157–173, 2018.

[15] Megan Bryson. Transforming patriarchy : Chinese families in the twenty-first century, edited by gonçalo santos and stevan harrell, 2017. *Nan Nü*, 20(2) :349–352, 2019.