Analyse de survie dans le cadre d'une étude en oncologie, le modèle de Cox et ses alternatives

Mickaël Hartweg

MASTER de statistique
Université de Strasbourg
Du 1er Février au 31 Juillet 2013

Rapport de stage : Analyse de Survie dans le cadre d'une étude en oncologie, le modèle de Cox et ses alternatives.

Hartweg Mickaël

Département de Biostatistiques
Quintiles Benefit France
RUE JEAN-DOMINIQUE CASSINI
67400 Illkirch-Graffenstaden France

Directrice : Madame Geneviève Jehl
Responsable du master : Mme Armelle Guillou
Maitre de stage : Monsieur Nicolas Girard
Référent lors du stage : Monsieur Nicolas Poulain

Table des matières

Remerciement et Introduction

Nous tenons à remercier Mme Geneviève Jehl pour nous avoir accepté au sein de l'équipe de biostatistique de Quintiles - Illkirch. Nous remercions également notre tuteur et mentor lors de ce stage, monsieur Nicolas Girard qui nous a guidé tout du long. Nous remercions toute l'équipe de biostatistique pour son accueil chaleureux et ses précieux conseils. Nous remercions pour finir toute l'équipe enseignante du master de statistique pour leur disponibilité et leur réponses aux questions que nous nous sommes posés durant le stage.

Ce rapport a pour but de présenter le travail effectué et les enseignements tirés lors du stage de fin d'étude d'une durée de 6 mois réalisé au sein de l'entreprise Quintiles. Le temps de stage à été répartis en deux principales activitées.

La première a été la découverte du métier de biostatisticien en CRO : depuis la réception du protocole jusqu'à la livraison des différentes sorties constituant l'analyse à proprement parler, nous avons pu participer à chaques étapes de la rédaction d'un rapport d'étude en oncologie.
L'entreprise Quintiles, créée en 1982 par le biostatisticien Dennis Gillings, compte aujourd'hui prêt de 22000 employés et est présente dans plus de 50 pays à travers le monde. Ce succès est du principalement à un maintient permanent de la meilleure qualité de service possible. Cela se traduit dans le département biostatistique par une double revue complète de tout les éléments de l'étude par un biostatisticien senior, ainsi qu'au codage SAS des analyses par deux biostatisticiens en parrallèle et de manière indépendante (afin de confronter les résultats et mettre en avant d'éventuelles erreurs).

FIGURE 1 – Le lieux de stage : Quintiles - Illkirch (France)

Dans le cadre de notre stage, nous avons été convié à suivre en premier lieux le cursus de formation

que chaque biostatisticien est ammené à suivre lors de son arrivée dans la société : il s'agit d'un ensemble de présentations, exercices de mises en situation et tests présentant le cadre général de l'étude pharmacologique, les lois en vigueurs (en Europe et aux Etats Unis principalement), les GCP (pour Good Clinical Practices) , une description du contenu de la conférence ICH (pour International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use) et une initiation aux analyses médicales et au vocabulaire courant dans le domaine de l'oncologie.
En parrallèle nous avons suivis une formation au logiciel SAS qui nous a permise de mettre en pratique nos connaissances et nous exercer à la programmation sur une étude factice, ainsi qu'à tout les autres logiciels dont un biostatisticien se sert fréquement au sein de Quintiles.
Après cette période d'apprentissage, nous avons écrit des codes SAS permettant de résumer les informations présentes dans différents jeux de données afférant à plusieurs études en cours (tables de statistiques descriptives et listings) à partir de modèles déja rédigé, puis nous avons réalisé ces modèles de tables à destinations des programmeurs et biostatisticien pour une étude en particulier. Nous avons également écrit des programmes permettant la conversion de jeu de données d'un format standard SDTM (pour Study Data Tabulation Model [1]) vers un format spécifique (CDISC ADaM Basic Data Structure for Time-to-Event Analyses [2]).

La deuxième partie du stage a eu pour but de permettre la réutilisation des connaissances acquises lors du Master et du stage de première année (au sein du FDM : Freiburg Center for Data Analysis and Modeling) , dans le contexte d'une étude clinique en oncologie. Pour cela nous avons analysé un jeu de données à l'aide d'un outil vu en cours d'analyse de survie (le modèle de Cox), puis nous avons étudié des solutions/méthodes alternatives existantes pour résoudres les différents problèmes rencontrés : les donnéess manquantes, les évènements concurrents et le non respect de l'hypothèse de risques proportionnels nécessaire à l'estimation des paramètres (dans le cadre du modèle de Cox).

Ce rapport présente les outils et méthode mathématiques utilisées pour les analyses : les bases de l'analyse du temps de survie et du modèle de Cox , les bases de la théorie des risques concurrents,le modèle de Fine et Gray ainsi que les bases du modèle de regression sur les quantiles. Pour éviter toute redites nous renverrons le lecteur vers les articles/cours utilisés lorsqu'aucune explication complémentaire ne sera fournie, et au contraire tacherons de détailler et expliquer au mieux les étapes de calculs lorsque celles-ci ne le sont pas dans le document de référence.
Le rapport présente ensuite une application de ces méthodes sur un jeu de données issus de l'analyse de survie dans le domaine de l'oncologie (dont nous détaillerons le principe dans le chapitre Analyses 2).

1. Source : CDISC, "Study Data Tabulation Model","http ://www.cdisc.org/sdtm"
2. Source : CDISC, "ADaM","http ://www.cdisc.org/adam"

Chapitre 1

Outils mathématiques employés

1.1 Notations choisies et vocabulaire

Dans ce qui suit nous présenterons sommairement les outils mathématiques utilisés durant le stage. Les notations suivantes seront adoptés durant toute cette partie :

- i=1...I les individus observés.
- i'=1...I' les individus observés et non censurés triés par odre croissant de temps observé (on a donc I'≤I).
- j=1...J les j évènements concurrents.
- k=1...K les K covariables mesurées pour un patient en particulié.
- X la variable aléatoire représentant le temps observé, δ l'indicatrice de censure et Z la matrice des covariables de taille I(K+1) (comprennant l'intercept).
- T variable aléatoire représentant le temps jusqu'à l'évènement d'interet pour les patients non censuré , et T' sa version ordonnée croissante.
- C variable aléatoire représentant le temps jusqu'à censure pour les patients censurés.
- $M_{i'}$ le nombre de morts observés en t'_i
- $R_{i'}$ le nombre de sujets ni morts ni censurés juste avant t'_i (sujets dit "à risques").
- $z_{1-\frac{\alpha}{2}}$ le quantile en $1-\frac{\alpha}{2}$ de la loi normale centrée réduite.
- Y la variable aléatoire qui indique le nombre de patients présentant une réponse moléculaire positive à la fin de la période d'observation. Y_i la variable aléatoire binaire indiquant l'occurence d'une réponse moléculaire positive pour le patient i durant la période d'observation.
- p la proportion de réponse moléculaire positive observée parmis les patients.
- γ la demi-longueur d'un intervalle centré autour de l'estimation ponctuelle à laquelle il se réfere.

Et les abréviations/désignations suivantes :

- v.a. l'abréviation pour variable aléatoire.
- IC : intervalle de confiance
- L(.) la loi d'une variable aléatoire
- 1(.) la fonction indicatrice

- CRO : Contract Research Organization. Société de service qui propose toutes sortes de prestations aux entreprises pharmaceutiques (recueil, structuration et analyse statistiques des données. Redaction de rapport pour soumission aux autorités).
- Oncologie : domaine médical englobant tout les types de cancers.
- Hématologie : étude de la composition du sang.
- Cytologie : étude des cellules
- LMC : Leucémie myéloïde chronique : cancer affectant les cellules présentent dans le sang (globules blancs/rouges, plaquettes, cellules immatures)

1.2 Estimation d'une proportion

L'analyse des données recquise par le client est axée autour de l'estimation d'une proportion. Nous présentons ici rapidement quelques méthodes proposée par SAS, ainsi que la méthode employée par le biostatisticien (du client) en charge de l'étude pour estimer la taille d'échantillon nécessaire lors de la planification expérimentale.

1.2.1 Méthode approchée (approximation normale, Wald)

C'est la méthode qui a été choisie pour effectuer le dimensionnement et les estimations lors de l'étude. Soit Y_i, I variables aléatoires de loi de Bernouill(p) indépendantes. En appliquant le théorème central-limite on peut écrire :

$$L\left(\frac{\sum_{i=1}^{n}(Y_i) - E(\sum_{i=1}^{n} Y_i)}{\sqrt{Var(\sum_{i=1}^{n} Y_i)}}\right) \to N(0,1) \; lorsque \; n \to +\infty \tag{1.1}$$

Or $Y = \sum_{i=1}^{n} Y_i$ suit une loi Binomiale(n,p) d'espérance np et de variance $np(1-p)$. Ce qui nous donne en remplaçant dans la formule (1.1) :

$$L\left(\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}}\right) \to N(0,1) \; lorsque \; n \to +\infty \tag{1.2}$$

Et ainsi, en factorisant par n dans (2.2) on obtient :

$$L\left(\frac{n * (\overline{Y_n} - p)}{n * \sqrt{\frac{1}{n} * p(1-p))}}\right) = L\left(\frac{(\overline{Y_n} - p)}{\sqrt{\frac{1}{n} * p(1-p))}}\right) \to N(0,1) \; lorsque \; n \to +\infty \tag{1.3}$$

Avec $\overline{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
On estime p par $\widehat{p} = \frac{y}{n}$. Ou y est l'obervation de Y sur l'échantillon. En procédant par plug-in dans (1.3) on obtient $p \in \widehat{p} \pm z_{1-\frac{\alpha}{2}} * \sqrt{\frac{\widehat{p}*(1-\widehat{p})}{n}}$

En inversant cette formule pour n, en supposant une valeur pour p et une valeur pour la précision de l'intervalle de confiance (demi-longueur) notée γ (ici $\gamma = z_{1-\frac{\alpha}{2}} * \sqrt{\frac{\widehat{p}*(1-\widehat{p})}{n}}$) on obtient la formule de dimmensionnement utilisée dans l'étude :

$$n_{requis} = \left(\left(\frac{\gamma}{z_{1-\frac{\alpha}{2}}}\right)^2 * \left(\frac{1}{p(1-p)}\right)\right)^{-1} \tag{1.4}$$

1.2.2 Méthode de Wilson

La formule de l'intervalle de Wilson pour une proportion est :

$$\frac{\widehat{p} + \frac{z_{1-\frac{\alpha}{2}}}{2n}}{1 + \frac{(z_{1-\frac{\alpha}{2}})^2}{n}} \pm z_{1-\frac{\alpha}{2}} \left(\frac{\sqrt{\widehat{p}(1-\widehat{p}) + \frac{z_{1-\frac{\alpha}{2}}}{4n}}}{1 + \frac{(z_{1-\frac{\alpha}{2}})^2}{n}}\right) \tag{1.5}$$

L'idée est de pondérer l'estimation de p en mélangeant l'estimation classique $\widehat{p}$ avec $\frac{1}{2}$, considérée comme la meilleure estimation en absence d'informations (plus le nombre d'observation est important moins on va accorder de poid à $\frac{1}{2}$). En inversant cette formule pour n, en supposant une valeur pour p et une valeur pour la précision de l'intervalle de confiance (demi-longueur) notée γ on obtient la formule de dimmensionnement associé à cette méthode (en résolvant une équation du second degrée en n) :

$$
\begin{aligned}
z_{1-\frac{\alpha}{2}}\left(\frac{\sqrt{p(1-p)+\frac{z_{1-\frac{\alpha}{2}}}{4n}}}{1+\frac{(z_{1-\frac{\alpha}{2}})^2}{n}}\right)=\gamma \leftrightarrow \frac{4np(1-p)+(z_{1-\frac{\alpha}{2}})^2-\left(\frac{\gamma}{z_{1-\frac{\alpha}{2}}}\right)^2*4n^2}{4n^2}=0 \\
\leftrightarrow n=\frac{(z_{1-\frac{\alpha}{2}})^2}{2\gamma^2}*\left(p(1-p)+\sqrt{(p(1-p))^2-\gamma^2}\right) \\
si\ (4*p(1-p))^2-16*\left(\frac{\gamma}{z_{1-\frac{\alpha}{2}}}\right)*(z_{1-\frac{\alpha}{2}})^2>0
\end{aligned}
\tag{1.6}
$$

1.2.3 Méthode d'Agresti-Coull

Il s'agit de l'intervalle obtenu par approximation normale non pas pour la variable $\frac{Y}{n}$ mais pour $\frac{Y+2}{n+4}$ (ajout de deux succès et deux échecs.). Cette idée vient de l'observation de l'expression de l'intervalle de confiance obtenu par la méthode de Wilson en posant $z_{1-\frac{\alpha}{2}} \approx 2$ et $\widehat{p}=\frac{y}{n}$. On obtient $\widehat{p_2}=\frac{n\widehat{p}+2}{n+4} \approx \frac{n\widehat{p}+z_{1-\frac{\alpha}{2}}}{n+(z_{1-\frac{\alpha}{2}})^2}$

1.2.4 Méthode de Clopper-Pearson

Soit Y une v.a. suivant une loi Binomiale(θ, I). L'idée est de résoudre les équations en θ (ou $\theta \in [0;1]$).

$$
P(Y \geq y)=\sum_{i=y}^{n}\left(\begin{array}{c} n \\ i \end{array}\right)\theta^i(1-\theta)^{n-i}=\frac{\alpha}{2} \tag{1.7}
$$

Et

$$
P(Y \leq y)=\sum_{i=0}^{y}\left(\begin{array}{c} n \\ i \end{array}\right)\theta^i(1-\theta)^{n-i}=\frac{\alpha}{2} \tag{1.8}
$$

Ou n est le nombre total de patients, y le nombre de patients présentant le critère dont on veut estimer la proportion et α la précision souhaitée. Notons l_y et u_y les solutions respectives à y fixé de ces équations. Alors l'intervalle de confiance pour θ avec I fixé est $[l_y; u_y]$.
Cette méthode, idéale en apparence, a en réalité une probabilité de couverture (probabilité que la vraie valeur soit comprise dans l'intervalle) toujours supérieure ou égale à la probabilité souhaité (en général 95 %), fournissant des intervalles de confiances plus larges que nécessaire. Cela provient du fait que la fonction de densité d'une loi binomiale n'est pas continue (loi discrète), ce qui pose des problèmes lors de la résolution numérique des équations.

1.3 Analyse de survie.

1.3.1 Théorie de la survie et fonctions usuelles.

On cherche à modéliser la distribution d'une variable aléatoire T représentant un temps de survie. Le terme décès sera synonyme par la suite d'occurrence de l'événement d'intérêt , ainsi le « décès » sera l'occurrence d'une réponse moléculaire positive (pour plus de précision quand à la définition d'une réponse moléculaire se réferer à la partie Analyses).

La loi de T est déterminée de manière unique par sa fonction de répartition :

$$
F(t)=P(T \leq t) \text{ avec } t \geq 0 \tag{1.9}
$$

Aussi appelée **fonction d'incidence cumulée**, représentant la probabilité de décès avant l'instant t.

De manière équivalente aux autres lois de probabilité, on peut déduire de la fonction d'incidence cumulée une fonction de densité telle que :

$$f(t) = F(t)' = \lim_{h \to 0} \left(\frac{F(t+h) - F(t)}{h} \right) \tag{1.10}$$

L'aire sous la courbe entre deux points de l'axe des abscisses t_1 et t_2 se traduit par la probabilité de décès entre ces deux instants.

Nous introduisons aussi une fonction directement calculée à partir de celles citées ci-dessus, qui modélise la probabilité de décès à l'instant t pour un patient sachant que ce patient a « survécu » jusqu'à l'instant t (« Survivre » signifie ne pas subir de décès au sens présenté plus haut). C'est la fonction de **risque instantané** :

$$\lambda(t) = P(T = t | T \geq t) = \lim_{h \to 0} \left(\frac{P(t \leq T \leq t+h | T \geq t)}{h} \right) = \frac{f(t)}{1 - F(t^-)} \tag{1.11}$$

Et sa version cumulée :

$$\Lambda(t) = \int_0^t \lambda(u) du \tag{1.12}$$

(Attention à l'interprétation, il s'agit d'un risque cumulé et non d'une probabilité, en effet cette « somme continue » des probabilités conditionnelles en t peut être supérieure à 1.).

Remarque : ces fonctions sont fortement liées entre elle, comme le prouve ces relations :

$$\begin{aligned}
\Lambda(t) &= \int_0^t \lambda(u) du \\
&= \int_0^t \frac{f(u)}{1 - F(u^-)} du \\
&= \int_0^t \frac{-(1 - F(u^-))'}{1 - F(u^-)} du \\
&= -[ln(1 - F(u))]_0^t = -ln(1 - F(t)) \\
&\iff 1 - F(t) = exp(-\Lambda(t))
\end{aligned}$$

1.3.2 Théorie de la censure.

En réalité les donnéess recueillies ne sont pas toutes des réalisations de la variable aléatoire T. En effet l'observation peut s'arrêter alors qu'aucun décès n'a été observé pour certains patients. Néanmoins la durée d'observation est reportée puisqu'elle contient l'information suivante : « le patient i n'est pas décedé pendant la période d'observation », il s'agit d'un patient censuré.

On introduit les notations suivantes pour traduire ce type observations (appelée observations censurées à droite) : On observe en réalité la variable aléatoire X=min(T,C) représentant la durée jusqu'à l'occurrence du premier événement parmi deux possible : le décès et la censure. Avec T la v.a. mesurant le temps jusqu'au décès et C la variable mesurant le temps jusqu'à la censure. Pour indiquer quel événement s'est produit, on introduit également l'indicatrice de censure $\delta = 1(T \leq C)$ qui vaut 0 en cas de censure et 1 si la durée d'intérêt est observée.

Au final les donnéess pour le patient i se présenteront sous la forme des triplés (x_i, δ_i, z_i) réalisation du vecteur aléatoire (X, Δ, Z) ou x représente la durée observée, δ la variable binaire de censure et z le vecteur contenant les donnéess personelles du patient, appelée covariables.

1.3.3 Estimation des fonctions usuelles en présence de censure aléatoire droite.

L'estimateur de Nelson-Aalen du risque cumulé $\Lambda(t)$ en présence de censure aléatoire droite (avec les notations présentées en début de document) :

$$\widehat{\Lambda}_n(t) = \sum_{\substack{i'=1 \\ T'_i \leq t}}^{I'} \frac{M_{i'}}{R_{i'}} \tag{1.13}$$

On peut estimer sa variance et utiliser un résultat de convergence en loi pour en déduire des intervalles de confiances ponctuels en t pour $\widehat{\Lambda_n(t)}$. Mais nous ne nous attarderons pas dessus puisqu'aucune estimation de cette grandeur ne sera présentée.

L'estimateur de Kaplan-Meier de l'incidence cumulée en présence de censure aléatoire droite est défini par :

$$1 - \widehat{F}_n(t) = \sum_{\substack{i'=1 \\ T'_i \leq t}}^{I'} \left(1 - \frac{M'_i}{R'_i}\right) = \widehat{S}_n(t) \tag{1.14}$$

Une méthode de construction d'intervalles de confiance ponctuels est présentée dans le cour d'analyse de survie de Mme Geffray Ségolène. Il existe également une méthode de construction de bande de confiance dont les bornes sont matérialisées par deux courbes et assurant une probabilité $1 - \alpha$ que la"vraie" courbe soit tout entière comprise entre ces bornes (ce qui n'est pas le cas pour les intervalles de confiance ponctuels mis bout à bout. Pour plus de détails quand aux méthodes utilisées se référer à [7]).
Cette méthode se base sur la théorie des processus, que l'on peut présenter comme des fonctions continues aléatoires. Les auteurs se basent alors sur des résultats de convergence vers des processus particuliés (dans les cas les plus simples des processus Gaussiens, généralisation continue des vecteurs Gaussiens. Pour plus de précision se référer au cour [1] consultable sur internet.) pour déterminer les bandes de confiances.

1.3.4 Les évènements compétitifs, théorie

Les évènements compétitifs sont tout les évènements entrainant un arrêt des observations (exception faite de l'évènement d'interêt). Dans la théorie présenté précédement, ces évènements sont considérés comme faisant partie intégrante de la censure aléatoire droite. Or nous avons vu que pour fournir les estimations des différentes fonctions, il faut supposer que cette censure est indépendante de l'évènement d'interêt. Cette hypothèse n'est pas valable dans tout les cas.

Il faut alors modifier la modélisation des données. On observe en théorie les couples (x_i, j_i) qui représentent le sujet i où x est une réalisation de X (temps jusqu'au premier évènement parmis les J+1 possible avec la censure). et j réalisation de D la v.a. représentant le premier évènement subit (parmis les J+1 évènements possibles).

De cette nouvelle modélisation découle directement de nouvelles fonctions :
La loi du couple (T, D) est déterminée par l'ensemble des fonctions d'incidence cumulée (cumulative incidence function) :

$$F_j(t) = P(T \leq t \cap D = j) \tag{1.15}$$

Ce sont des sous fonctions de répartition, c'est à dire que :

- $\lim_{t \to \infty} F_j(t) = P(D = j) \leq 1$
- $\sum_{j=1}^{J} F_j(t) = P(T \leq t) = F(t)$ la fonction de répartition de T présentée en (1.9).

On notera $f_j(t)$ la ("sous")-densité associé à $F_j(t)$ telle que :

$$f_j(t) = F_j(t)' = \lim_{h\to 0}\left(\frac{F_j(t+h) - F_j(t)}{h}\right) = \lim_{h\to 0}\left(\frac{P((t \leq T \leq t+h) \cap d = j)}{h}\right) \tag{1.16}$$

Et on remarque de la même façons que :

- $\sum_{j=1}^{J} f_j(t) = f(t)$ la densité associée à T (1.10)

On introduit ensuite les fonctions de risques spécifiques comme la probabilité instantanée de mourrir en t de cause j sachant que l' on a survécu toutes causes confondues jusqu' en t.

$$\begin{aligned}
\lambda_j(t) &= \lim_{h\to 0}\left(\frac{P((t \leq T \leq t+h) \cap d = j | T \geq t)}{h}\right) \\
&= \lim_{h\to 0}\left(\frac{P((t \leq T \leq t+h) \cap d = j \cap T \geq t)}{P(T \geq t)} * \frac{1}{h}\right) \\
&= \frac{f_j(t)}{1 - F(t^-)}
\end{aligned} \tag{1.17}$$

A partir de cette fonction est déduite la fonction de risque spécifique cumulée :

$$\Lambda_j(t) = \int_0^t \lambda_j(u)du \tag{1.18}$$

Ces fonctions ne sont néanmoins pas satisfaisantes pour interpreter correctement les résultats. Gray (1988) [6] a alors introduit la fonction de risque associée a la fonction d'incidence cumulée à la cause j :

$$\gamma_j(t) = \frac{f_j(t)}{1 - F_j(t^-)} = -\frac{\delta}{\delta t} ln(1 - F_j(t))$$

En remplaçant $f_j(t)$ et $F_j(t)$ par leur définition :

$$\begin{aligned}
&= \lim_{h\to 0}\left(\frac{\frac{P((t \leq T \leq t+h) \cap D = j)}{h}}{1 - P(T < t \cap D = j)}\right) \\
&= \lim_{h\to 0}\left(\frac{P(((t \leq T \leq t+h) \cap D = j) \cap (T \geq t \cup (T \leq t \cap D \neq j)))}{h * P(T \geq t \cup (T \leq t \cap D \neq j))}\right) \quad (*) \\
&= \lim_{h\to 0}\left(\frac{P((t \leq T \leq t+h) \cap D = j | (T \geq t \cup (T \leq t \cap D \neq j)))}{h}\right)
\end{aligned} \tag{1.19}$$

(*) En remarquant que :

$$\begin{aligned}
1 - P(T < t \cap D = j) &= P(T \geq t \cup D \neq j) \\
&= P(T \geq t \cup ((T > t \cup T \leq t) \cap D \neq j)) \\
&= P(T \geq t \cup ((T > t \cap D \neq j) \cup (T \leq t \cap D \neq j))) \\
&= P((T \geq t \cup (T > t \cap D \neq j)) \cup (T \leq t \cap D \neq j)) \\
&= P(T \geq t \cup (T \leq t \cap D \neq j))
\end{aligned}$$

(*)Et que :

$$\begin{aligned}
&P((t < T \leq t+h) \cap D = j) \cap (T > t \cup (T \leq t \cap D \neq j))) \\
&\quad = P((((t < T \leq t+h) \cap D = j) \cap (T > t)) \cup (((t < T \leq t+h) \cap D = j) \cap (T \leq t \cap D \neq j))) \\
&\quad = P(((t < T \leq t+h) \cap D = j) \cup \emptyset) \\
&\quad = P((t < T \leq t+h) \cap D = j)
\end{aligned}$$

On notera que cette fonction correspond à la fonction de risque instantanée de la variable aléatoire (impropre) $T_j^* = 1(d=j)T + 1(d \neq j)\infty$ en effet :

$$\lambda_{T_j^*}(t) = \lim_{h\to 0} \frac{P(t \leq T^* \leq t+h | T^* \geq t)}{h} \tag{1.20}$$
$$= \lim_{h\to 0} \frac{P(t \leq (1(D=j)T + 1(D\neq j)) \leq t+h | (1(D=j)T + 1(d\neq j)) \geq t)}{h}$$
$$= \lim_{h\to 0} \frac{P([(t \leq T \leq t+h) \cap (D=j)] \cup [(t \leq \infty \leq t+h) \cap (D \neq j)] | (T \geq t \cap D = j) \cup (\infty \geq t \cap D \neq j)}{h}$$
$$= \lim_{h\to 0} \frac{P([(t \leq T \leq t+h) \cap (D=j)] \cup [\emptyset \cap (D \neq j)] | (T \geq t \cap D = j) \cup (\Omega \cap D \neq j)}{h}$$

Car t, t+h $< \infty$, donc l'ensemble {t $< \infty <$ t+h} est l'ensemble vide. Et l'ensemble $\{\infty >$ t$\}$ est l'univers Ω.

$$= \lim_{h\to 0} \frac{P((t \leq T \leq t+h) \cap (D=j) | (T \geq t \cap D = j) \cup (D \neq j))}{h}$$
$$= \lim_{h\to 0} \frac{P((t \leq T \leq t+h) \cap (D=j) | ((T \geq t) \cup (D \neq j)) \cap (D = j \cup D \neq j))))}{h} \quad (*)$$
$$= \lim_{h\to 0} \frac{P((t \leq T \leq t+h) \cap (D=j) | ((T \geq t) \cup (D \neq j)))}{h}$$
$$= \gamma_j(t) \tag{1.21}$$

Cela nous suggère une transformation de nos données : le temps mesuré devient T si d=1 (l'évènement d'interet), $+\infty$ sinon. (astuce utilisée lors du stage de M1, permettant d'utiliser une fonction estimant un modèle de Cox sur des données correctement modifiées pour estimer un modèle de Gray).

Une dernière fonction intéressante est la fonction $CP_j(t)$ (pour conditionnal probability) introduite par Peppe et Mori (1993) [13] , correspondant à la probabilité conditionelle de mourrir avant t de cause j sachant que l'on à survécus jusqu'en t aux autres causes de déces :

$$CP_j(t) = \frac{F_j(t)}{1 - \sum_{\substack{u=1 \\ u\neq j}}^{J} F_u(t)} \tag{1.22}$$

L'estimation de cette fonction et des intervalles de confiances ponctuels associés est explicitement recquise dans le protocole de l'étude.

La prise en charge de la censure aléatoire droite (constituée des évènements pour qui l'hypothèse d'indépendance est probable) s'effectue de manière similaire à celle précédemment présentée.
On pose X=min($T_j, T_1..T_{j-1}, T_{j+1}..T_J$,C) ou T_j la durée jusqu'à l'évènement d'intéret (ici le j-ième), T_k avec $k = 1...j-1, j+1...J$ le durée jusqu'à l'évènements compétitif k et C la durée mesurée jusqu'à la censure.
On pose D=k si min($T_j, T_1..T_{j-1}, Tj+1..T_J, C) = T_k$.
On pose Δ=0 si min($T, T_1..T_{j-1}, Tj+1..T_J$,C)=C et Δ=1 sinon.
On observe donc les triplés $(x_i, \delta_i * j_i, z_i)$ réalisation des variables aléatoire $(X, \Delta * D, Z)$
où $\delta * j = \begin{cases} 0 & \text{si le sujet est censuré} \\ j & \text{si le sujet meurt de cause j} \end{cases}$

1.3.5 Estimation des fonctions spécifiques en présence de censure aléatoire droite.

L'incidence cumulée spécifique

Nous utiliserons l'estimateur de l'incidence cumulée spécifique proposé par Marubini et Vasecchi ([5] page 338 équation 10.11) et détaillerons l'obtention de l'estimateur de sa variance, nécessaire à la

construction de l'intervalle de confiance ponctuel associé ([5] page 341 équation 10.12). L'estimateur s'obtient en sommant l'estimateur de $f_j(t)$ obtenus par plug-in dans $f_j(t) = \lambda_j(t) * (1 - F(t))$ (3.9) et s'écrit :

$$\widehat{F}_j(t) = \sum_{\substack{t \leq T'_i \\ T'_i \leq t}}^{I'} (1 - \widehat{F}_n(t)) * \frac{M_{i'j}}{R'_i} \tag{1.23}$$

$$1 - \widehat{F}_n(t) = \sum_{\substack{i'=1 \\ T'_i \leq t}}^{I'} \left(1 - \frac{M'_i}{R'_i}\right) = \widehat{S}_n(t) \tag{1.24}$$

Avec $1-\widehat{F}_n(t)$ l'estimateur de Kaplan Meier toutes causes de décès confondues (ie avec $M_{i'} = \sum_j M_{i'j}$ (3.6), et $\frac{M_{i'j}}{R'_i}$ l'estimateur de $\lambda_j(t_{i'})$ (avec $M_{i'j}$ le nombre de mort de cause j en T'_i et $R_{i'}$ le nombre de sujet à risque en T'_i, c'est à dire ni censurés ni décédés d'une autre cause que j).

A partir de l'écriture de cet estimateur, on peut démontrer que l'estimateur de Kaplan-Meier surestime l'incidence cumulée si il est utilisé en présence d'évènements compétitifs. En effet, si l'on s'interesse à la cause de décès 1 (choisie parmis les J possibles) on a :

$$F_1(t) = \int_0^t S(u^-)\lambda_1(u)du \quad \text{estimée par (1.23) avec j=1}$$

Avec $S(u) = exp(-\Lambda(t)) = exp(-(\sum_{j=1}^{J} \Lambda_j(t)))$

$$\begin{aligned} &= \int_0^t exp(-(\sum_{j=1}^{J} \Lambda_j(u^-)))\lambda_1(u)du \\ &= \int_0^t exp(-(\sum_{\substack{j=1 \\ j\neq 1}}^{J} \Lambda_j(u^-)))exp(-\Lambda_1)\lambda_1(u)du \end{aligned}$$

On note $exp(-(\sum_{\substack{j=1 \\ j\neq 1}}^{J})) = C(u)$ et on remarque que $C(u) \leq 1$

Si à présent on considère les J-1 autres évènements compétitifs comme des censures non informatives, on suppose qu'il n'y a qu'une seule cause de décès, la 1. Donc :

$$S'(u) = exp(-(\sum_{j=1}^{1} \Lambda_j(t))) = exp(-\Lambda_1(t)) \tag{1.25}$$

Dans ce cas de figure :

$$\begin{aligned} F'_1(t) &= \int_0^t S(u^-)\lambda_1(u)du \\ &= \int_0^t exp(-\Lambda_1)\lambda_1(u)du \quad \text{estimée par } 1 - \widehat{S}(t) \text{ ou } \widehat{S}(t) \text{ est l'estimateur de Kaplan Meier} \\ &\geq \int_0^t C(u^-)exp(-\Lambda_j)\lambda_j(u)du \text{ (car } C(u) \leq 1 \ \forall u) \quad \text{estimée par (1.23)} \end{aligned}$$

Donc le complémentaire à 1 de l'estimateur de Kaplan Meier appliqué à un contexte d'évènements compétitifs (mais en le négligeant) estime une fonction qui n'est pas l'incidence cumulée spécifique recherchée,

mais une fonction supérieure ou égale. (strictement supérieure dès lors que $\sum_{\substack{z=1 \\ z\neq j}}^{J} \Lambda_z(u^-)) > 0$), d'ou la surestimation.

Nous détaillons en annexe (A.1) les étapes de calculs nécessaires pour déterminer l'estimateur de la variance de l'incidence cumulée spécifique, et donnerons en partie Analyses une estimation de l'incidence cumulée spécifique à l'évènement d'intêret accompagnée de ses intervalles de confiances ponctuels.

La fonction CP_j

L'estimateur de CP_j est obtenu par plug-in dans la définition (1.22) (Peppe et Mori [?]) :

$$\widehat{CP}_j = \frac{\widehat{F}_j(t)}{1 - \widehat{F}_g(t)} \tag{1.26}$$

Avec j l'evenement d'interet et g l'union de tout les autres évènements compétitifs.
Les auteurs proposent un résultat de convergence assymptotique pour en dériver intervalles de confiances et tests. On a :

$$\sqrt{n}\left(\widehat{CP}_j(t) - CP_j(t)\right) \to N(O, \sigma^2(t)) \tag{1.27}$$

Avec :

$$\widehat{\sigma^2(t)} = \left(\frac{(\widehat{S}(t))^2}{(1-\widehat{F}_g(t))^4}\right) \sum_{\substack{k=1 \\ t_k \leq t}}^{n'} \left(\frac{(1-\widehat{F}_g(t_k))^2 M_{jk} + (\widehat{F}_j(t_k))^2 M_{gk}}{\frac{R_k(R_k-1)}{n}}\right) \tag{1.28}$$

1.3.6 Le modèle de Cox.

Il s'agit d'une méthode pour modéliser la fonction de risque instantanée $\lambda(t)$. On pose :

$$\lambda(t|Z_i) = \lambda_0(t) * exp(Z_i^{'} * \beta). \tag{1.29}$$

Avec $\lambda_0(t)$ la fonction de risque instantanée associée au patient de référence. β le vecteur des paramètres et Z_i le vecteur covariable pour l'individus i. On remarque qu'au travers de cette modélisation, on cherche a imposer au hasard instantanée le fait d'être proportionnel d'un individus à un autre. Cette hypothèse permet de simplifier le problème d'estimation, mais doit être vérifiée post-modélisation.
De prime abord seul le paramètre β nous importe, c'est lui qui nous permettra de quantifier l'importance de l'effet de la covariable à laquelle il est rattaché. De plus le parametre $\lambda()$ est une fonction non paramétrique (quelconque) de $R^+ \to [0;1]$, donc le nombre de points à estimer est infinis. Pour estimer uniquement le paramètre β, on utilise le maximum de vraisemblance partielle en présence de censure aléatoire droite (pour plus de détail se réferer à Cox 1975 [2] et en particulié à l'exemple 2).

On se place dans le contexte de censure aléatoire droite présenté précédement. On dispose de réalisation (x,δ,z) du triplé (X,Δ, Z) et l'on souhaite maximiser la vraisemblance, qui pour rappel est le produit des valeurs prises par la fonction de densité f (dans le cadre de l'analyse de survie elle est aussi appelée fonction d'incidence) pour chaque patient en chaques points de l'interval de temps.
Si δ=1 alors :

$$\begin{aligned} f_{X,\Delta|Z}(x,\delta) &= \lim_{h\to 0}\left(\frac{P(x \leq X \leq x+h, \delta = 1|Z)}{h}\right) \\ &= \lim_{h\to 0}\left(\frac{P(x \leq X \leq x+h, C \geq T|Z)}{h}\right) \end{aligned}$$

Et on sait que $x = min(t, c), \delta = 1(T \leq C)$ donc $\delta = 1 \rightarrow x = t \leq c$

$$\begin{aligned}
&= \lim_{h \to 0} \left(\frac{P(x \leq T \leq x + h, C \geq x|Z)}{h} \right) \\
&= \lim_{h \to 0} \left(\frac{P(x \leq T \leq x + h|Z) * P(C \geq x|Z)}{h} \right) \text{ par } \perp \text{ de T et C} \\
&= f_{T|Z}(x) * S_{C|Z}(x)
\end{aligned}$$

Et ainsi : $f_{X,\Delta,Z}(x, \delta, z) = f_{X,\Delta|Z}(x, \delta) * f_Z(z) = f_{T|Z}(x) * S_{C|Z}(x) * f_Z(z)$ lorsque $\delta = 1$

De la même manière :
Si δ=0 alors :

$$\begin{aligned}
f_{X,\Delta|Z}(x, \delta) &= \lim_{h \to 0} \left(\frac{P(x \leq X \leq x + h, \delta = 0|Z)}{h} \right) \\
&= \lim_{h \to 0} \left(\frac{P(x \leq X \leq x + h, T \geq C|Z)}{h} \right)
\end{aligned}$$

Et on sait que $x = min(t, c), \delta = 1(T \leq C)$ donc $\delta = 0 \rightarrow x = c \leq t$

$$\begin{aligned}
f_{X,\Delta|Z}(x, \delta) &= \lim_{h \to 0} \left(\frac{P(x \leq C \leq x + h, T \geq x|Z)}{h} \right) \\
&= \lim_{h \to 0} \left(\frac{P(x \leq C \leq x + h|Z) * P(T \geq x|Z)}{h} \right) \text{ par } \perp \text{ de T et C} \\
&= f_{C|Z}(x) * S_{T|Z}(x)
\end{aligned}$$

Et ainsi : $f_{X,\Delta,Z}(x, \delta, z) = f_{X,\Delta|Z}(x, \delta) * f_Z(z) = f_{C|Z}(x) * S_{T|Z}(x) * f_Z(z)$ lorsque $\delta = 0$

Au final :

$$\begin{aligned}
f_{X,\Delta,Z}(x, \delta, z) &= (f_{T|Z}(x) * S_{C|Z}(x) * f_Z(z))^{\delta} * (f_{C|Z}(x) * S_{T|Z}(x) * f_Z(z))^{1-\delta} \ \forall \delta \\
&= (f_{T|Z}(x) * S_{C|Z}(x))^{\delta} * (f_{C|Z}(x) * S_{T|Z}(x))^{1-\delta} * f_Z(z) \ \forall \delta
\end{aligned}$$

Ceci résoud le problème de la prise en compte de la censure dans la détermination du maximum de vraisemblance. Voyons maintenant comment transformer la vraisemblance découlant de l'expression ci-dessus et faisant intervenir à la fois β et $\lambda_0()$ en une vraisemblance partielle ne faisant intervenir que β.
Commençons par rappeler que :

$$\begin{aligned}
\lambda_{T|Z}(x) &= \lambda_0(x) exp(z'\beta) \\
S_{T|Z}(x) &= exp(-\Lambda_{T|Z}(x)) \\
f_{T|Z}(x) &= \lambda_{T|Z}(x) S_{T|Z}(x)
\end{aligned}$$

Et notons que :

$$\begin{aligned}
\Lambda_{T|Z}(x) &= \int_0^x \lambda_{T|Z}(u)du = \int_0^x \lambda_0(u) exp(z'\beta)du = exp(z'\beta) \int_0^x \lambda_0(u)du \\
&= \Lambda_0(x) exp(z'\beta)
\end{aligned}$$

On en déduis que :

$$\begin{aligned} f_{T|Z}(x) &= \lambda_{T|Z}(x)S_{T|Z}(x) \\ &= \lambda_{T|Z}(x)exp(-\Delta_{T|Z}(x)) \\ &= \lambda_0(x)exp(z'\beta)exp(-(\Lambda_0(x)exp(z'\beta))) \end{aligned}$$

En remplaçant dans la densité qui nous intéresse :

$$\begin{aligned} f_{X,\Delta,Z}(x,\delta,z) &= (f_{T|Z}(x) * S_{C|Z}(x))^{\delta} * (f_{C|Z}(x) * S_{T|Z}(x))^{1-\delta} * f_Z(z) \\ &= \left(\lambda_0(x)exp(z'\beta)exp(-(\Lambda_0(x)exp(z'\beta))) * S_{C|Z}(x)\right)^{\delta} * \left(f_{C|Z}(x) * exp(-\Lambda_{T|Z}(x))\right)^{1-\delta} \\ &* f_Z(z) \\ &= (\lambda_0(x)exp(z'\beta)exp(-(\Lambda_0(x)exp(z'\beta))) * S_{C|Z}(x))^{\delta} * (f_{C|Z}(x) * exp(-(\Lambda_0(x)exp(z'\beta)))^{1-\delta} \\ &* f_Z(z) \\ &= \lambda_0(x)^{\delta}exp(z'\beta)^{\delta}exp(-(\Lambda_0(x)exp(z'\beta)))^{\delta} * S_{C|Z}(x)^{\delta} * f_{C|Z}(x)^{1-\delta} \\ &* exp(-(\Lambda_0(x)exp(z'\beta))^{1-\delta} * f_Z(z) \\ &= \lambda_0(x)^{\delta}exp(z'\beta)^{\delta}exp(-(\Lambda_0(x)exp(z'\beta))) * S_{C|Z}(x)^{\delta} * f_{C|Z}(x)^{1-\delta} * f_Z(z) \end{aligned}$$

Sous hypothèse de censure non informative et en supposant que Z ne dépend pas de $(\beta, \lambda_0())$ on peut s'affranchir de la partie de l'expression qui ne dépend ni de β ni de $\lambda_0()$, le maximum de cette nouvelle fonction étant atteint pour les même valeurs de $(\beta, \lambda_0())$ que la précédente

$$\propto \lambda_0(x)^{\delta}exp(z'\beta)^{\delta}exp(-(\Lambda_0(x)exp(z'\beta)))$$

La vraisemblance à maximiser serai alors $l = \prod_{i=1}^{I} \lambda_0(x_i)^{\delta_i}exp(z_i'\beta)^{\delta_i}exp(-(\Lambda_0(x_i)exp(z_i'\beta)))$ avec (x_i, δ_i, z_i) le triplé observé pour le patient i. En introduisant un terme intermédiaire, on peut réecrire cette densité ainsi :

$$\begin{aligned} l &= \prod_{i=1}^{I} \lambda_0(x_i)^{\delta_i}exp(z_i'\beta)^{\delta_i}exp(-(\Lambda_0(x_i)exp(z_i'\beta))) \\ &= \prod_{i=1}^{I} \left(\lambda_0(x_i)exp(z_i'\beta) * \left(\frac{\sum_{m\in R_{x_i}} \lambda_0(x_i)exp(z_j'\beta)}{\sum_{m\in R_{x_i}} \lambda_0(x_i)exp(z_j'\beta)}\right)\right)^{\delta_i} exp(-(\Lambda_0(x_i)exp(z_i'\beta))) \\ &= \prod_{i=1}^{I} \left(\frac{\lambda_0(x_i)exp(z_i'\beta)}{\sum_{m=1}^{I} \lambda_0(x_i)exp(z_m'\beta) * 1(T_i \leq T_m)}\right)^{\delta_i} * \left(\sum_{m=1}^{I} \lambda_0(x_i)exp(z_m'\beta) * 1(T_i \leq T_m)\right)^{\delta_i} \\ &* exp(-(\Lambda_0(x_i)exp(z_i'\beta))) \end{aligned}$$

Cette écriture se simplifie en factorisant par $\lambda_0(x_i)$

$$\begin{aligned} &= \prod_{i=1}^{I} \left(\frac{exp(z_i'\beta)}{\sum_{m=1}^{I} exp(z_m'\beta) * 1(x_i \leq x_m)}\right)^{\delta_i} * \left(\sum_{m=1}^{I} \lambda_0(x_i)exp(z_m'\beta) * 1(x_i \leq x_m)\right)^{\delta_i} \\ &* exp(-(\Lambda_0(x_i)exp(z_i'\beta))) \end{aligned}$$

En utilisant la théorie de la vraisemblance partielle, on montre que si l'on s'interesse uniquement au paramêtre β, il est quasiment équivalent de maximiser cette vraisemblance ou de maximiser la parenthèse de gauche uniquement (voir Cox 1975 [2] et Tsiatis 1981 [18]). On aboutit finalement à la vraissemblance partielle de Cox utilisée pour estimer le paramètre β :

$$l_{partielle} = \prod_{i=1}^{I} \left(\frac{exp(z_i'\beta)}{\sum_{m=1}^{I} exp(z_m'\beta) * 1(x_i \leq x_m)}\right)^{\delta_i} \tag{1.30}$$

A maximiser en β pour obtenir l'estimation $\widehat{\beta}$.

1.3.7 Le modèle de Fine et Gray

Il s'agit d'un modèle reposant sur la définition suivante :

$$\gamma_j(t|Z_i) = \gamma_0(t) * exp(Z_i^{'} * \beta). \tag{1.31}$$

Avec $\gamma_0(t)$ la fonction de risque (associé à la fonction d'incidence cumulée spécifique à l'évènement j) associée au patient de référence. β le vecteur des parametres et Z_i le vecteur covariable pour l'individus i. De la même façons que pour le modèle de Cox on cherche à imposer au risque instantané le fait d'être proportionnel d'un individus à un autre.
Nous avons montré (1.21) que moyennant un changement de variable aléatoire adéquat, on pouvait calculer $\gamma_j(t)$ pour le trio (X, j, δ) comme le risque instantanée λ_j associé au trio(X^*, j, δ) en posant $X^* = T$ si T est observée et $X^* = \infty$ si T n'est pas observée (en conservant la notation T pour la variable aléatoire qui représente le temps jusqu'à l'évènement d'intêret).
Nous avons donc estimé les paramêtres d'un modèle de Cox sur cette variable aléatoire transformée en utilisant la méthode détaillée par Benjamin Esterni [3] et implémentée dans la macro SAS PHSREG (par Georg Heinze, 2011-2012).

1.3.8 Le modèle Quantiles.

Ce modèle repose une approche totalement différente des deux précédents puisqu'on ne cherche plus à modéliser l'incidence cumulée (ou une fonction qui en est dérivée comme le risque instantané dans le modèle de Cox et le risque instantané spécifique à la fonction d'incidence cumulée de la cause j pour le modèle de Gray) en fonction du temps et des covariables, mais le temps $(=Q(\tau))$ en fonction de l'incidence cumulée $(=\tau)$ et des covariables. Le modèle fournit une estimation du temps nécessaire pour qu'un certain pourcentage choisis de la population ait subis l'évènement d'intêret, en fonction de caractéristiques de cette population (nous pourrons par exemple comparer le temps nécessaire pour que 50 pourcents de la population masculine présente une réponse au traitement avec le temps nécessaire pour que 50 pourcents de la population féminine présente la même réponse.). Le pourcentage sera appelé dans la suite quantile (ou τ), et le temps nécessaire pour que 100τ pourcents de la population ait subis l'évènements d'intêret sera appelé $Q(\tau)$
En conservant les notations des chapitres précédent on a :

$$Q_T(\tau) = \inf_b(F_T(b) \geq \tau) \tag{1.32}$$

$$= F_T^{-1}(\tau) \text{ avec } 0 < \tau < 1 \text{ lorsque F est inversible} \tag{1.33}$$

Une manière naturelle de calculer les valeurs de cette fonction serait de considérer une statistique d'ordre pour T (ranger de manière croissante les valeurs) et d'estimer $Q(\tau)$ comme étant égale à la $\lceil n * \tau \rceil$-ieme valeur ordonnée, avec n le nombre de valeur différentes observée pour T et $\lceil . \rceil$ la partie entière supérieure. Néanmoins pour des raisons pratiques, une autre définition à été introduite par Roger Koenker [10] :

$$Q_T(\tau) = \underset{a}{\text{argmin}}\ E[((\tau - 1(T - a < 0))(T - a)] \tag{1.34}$$

Cette definition donne $Q_T(\tau)$ comme la valeur qui minimise l'espérance de la v.a. $((\tau - 1(T - a < 0))(T - a)$. En effet :

$$\begin{aligned} E[((\tau - 1(T - a < 0))(T - a)] &= E[\tau(T - a) - (T - a)1(T - a < 0)] \\ &= E[\tau(T - a)] - E[(T - a)1(T - a < 0)] \text{ par linéarité de l'espérance} \\ &= \tau(E[T] - a) - \int_{-\infty}^{+\infty}(u - a)1(u - a < 0)f_T(u)du \\ &= \tau(E[T] - a) - \int_{-\infty}^{a}(u - a)f_T(u)du \end{aligned}$$

Qui est une fonction dérivable en a avec :

$$\frac{d}{da}\left(E[((\tau - 1(T-a<0))(T-a)]\right) = \frac{d}{da}\left(\tau(E[T]-a) - \int_{-\infty}^{a}(u-a)f_T(u)du\right)$$
$$= \frac{d}{da}\left(\tau(E[T]-a)\right) - \frac{d}{da}\left(\int_{-\infty}^{a}(u-a)f_T(u)du\right) = -\tau - \frac{d}{da}\left(\int_{-\infty}^{a}uf_T(u)du - \int_{-\infty}^{a}af_T(u)du\right)$$
$$= -\tau - \frac{d}{da}\left(\int_{-\infty}^{a}uf_T(u)du\right) + \frac{d}{da}\left(\int_{-\infty}^{a}af_T(u)du\right) = -\tau - (af_T(a)) + \left(\int_{-\infty}^{a}f_T(u)du + af_T(a)\right)$$
$$= -\tau - (a-a)f_T(a) + \int_{\infty}^{a}f_T(u)du = F_T(a) - \tau$$

Donc un extremum local de la fonction est atteint en $a = F_T^{-1}(\tau) = Q_T(\tau)$, la valeur qui annule $\frac{d}{da}\left(E[((\tau - 1(T-a<0))(T-a)]\right)$. C'est bien un minimum car $F_T(a) - \tau$ est négative avant $F^{-1}(\tau)$ et positive après (voir le schéma 1.1, ou a $=F^{-1}(\tau)$).

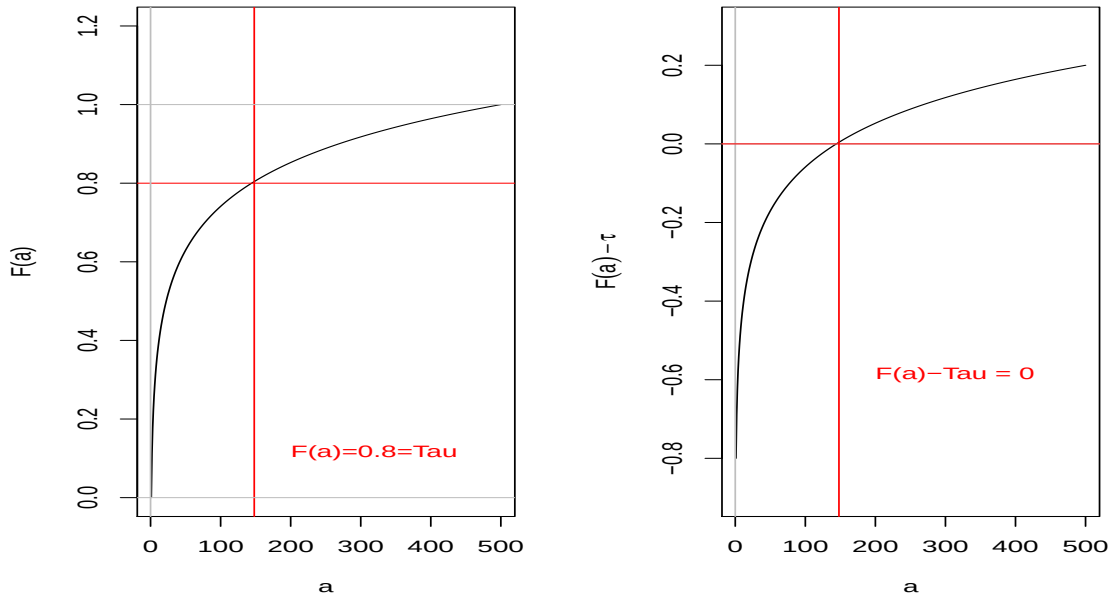

FIGURE 1.1 – Schema explicatif avec τ=0.8

Dans le cadre d'une regression, on peut étendre cette fonction pour prendre en compte l'influence des covariables en considérant l'espérance conditionnellement au vecteur des covariables du patient.

$$Q_{T|Z}(\tau|Z=z) = \underset{a}{\operatorname{argmin}}\ E[[\tau - 1(T-a<0)](T-a)|Z=z] \tag{1.35}$$

Pour être en accord avec la publication de Roger Koenker et alléger les notations, on notera dans la suite $\rho_\tau(u) = u(\tau - 1(u<0))$, fonction qui est appliquée en T-a dans toutes les lignes ci-dessus.
Le modèle dont les paramètres sont estimés par la fonction du package R quantreg [11] est le suivant :

$$Q_{T|Z=z}(\tau|Z=z) = z'\beta(\tau) \tag{1.36}$$

Sans censure, on estime $\beta(\tau)$ en utilisant l'estimateur des moments pour l'espérance et en remplaçant "a" par sa forme paramétrique (dépendant de z et β) :

$$\widehat{\beta}(\tau) = \underset{b}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} [\rho_\tau(T_i - z_i' b)] \right) \tag{1.37}$$

Dans un contexte de censure aléatoire droite, deux estimateurs ont été proposés, l'un basé sur l'estimateur de Kaplan-Meier de l'incidence cumulée et l'autre basé sur l'estimateur de Nelson-Aalen du risque instantané (voir [10] : chapitre 4 : "Random Censoring"). De la même façon des résultats de convergence assymptotique (voir [10] : chapitre 5 "Some One-sample Asymptotics") sont disponible, permettant de construire des tests et des intervalles de confiance ponctuels. La variance de $\widehat{\beta}(\tau)$ est estimée par bootstrap dans le package R utilisé ("quantreg" [11]).

Chapitre 2

Description des données et choix des analyses.

2.1 Le contexte de collecte des données

les donnéess disponible proviennent d'une étude sur un nouveau traitement contre la LMC (leucémie myéloïde chronique). Cette maladie est provoquée par la présence de chromosomes de Philadelphie dans les globules blancs, résultat de l'échange de matériel génétique (nommé Translocation) entre deux chromosomes (le n^o 9 et le n^o 22). Dans ce chromosome (n^o 22-) apparait un gène mutant (nommé gène BCR-ABL), résultat de la fusion du gène BCR provenant du chromosome 9 avec le gène ABL provenant du chromosome 22. Ce gène produit à son tour une molécule favorisant la production de globules blancs déficient au détriment des globules blancs sains, provoquant leur remplaçement progressif dans l'organisme et des problèmes imunologique.

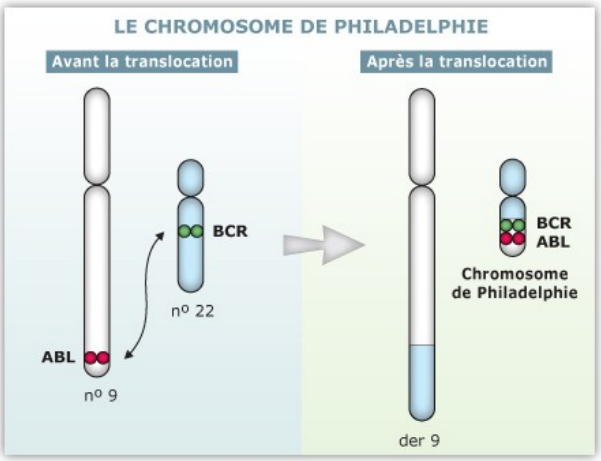

FIGURE 2.1 – Schéma explicatif de la translocation [1]

Deux méthodes d'évaluation classique de la réponse au traitement (RH : réponse hématologique et Rcy : réponse cytogénétique) sont évaluées en addition d'une méthode innovante (RM : réponse moléculaire) :

- La réponse hématologique est évaluée en analysant la composition d'un échantillon de sang prélevé chez le patient (globules blancs, plaquettes, etc...).Des comptages anormalement hauts ou anormalement bas sont des indices annonciateurs d'une LMC.

1. Source : CML Society, "Comprendre la LMC","http ://cmlsociety.org/comprendre-la-lmc/ ?lang=fr"

- La réponse cytogénétique est représentée par le pourcentage de globules blancs déficients parmi les globules blancs prélevés dans un échantillon de moelle osseuse. La difficulté réside en l'identification des globules anormaux.
- La réponse moléculaire est représentée par le pourcentage de gènes BCR-ABL détectés par rapport au nombre de gène ABL dans un échantillon de moelle osseuse.
  Cette analyse est beaucoup plus précise que les précédentes et sera utilisé comme indicateur de réponse au traitement.
  Plus en détail le comptage est effectué sur deux échantillons et deux cas de figures se présentent :
  1. Si aucun gènes BCR-ABL ne sont détectés (dans les deux échantillons) et que le nombre de gènes analysés est suffisant nous concluons à une réponse moléculaire positive, si le nombre de gène analysé n'est pas suffisant on peut retirer un troisième échantillon pour venir appuyer la conclusion faite au regard des deux autres analyses.
  2. Si des gènes défficients sont détectés nous calculons le taux moyen de gènes BCR-ABL sur les deux échantillons, si ce taux moyen est inférieur à un taux de référence et le nombre de gène analysés par échantillons suffisant (suffisant relativement à une valeur de référence), on conclu également à une réponse moléculaire négative.

2.2 But de l'étude et extension lors du stage

Le but de l'étude est de confirmer le taux de réponse moléculaire positive au seuil 4.0 (seuil qui définit les valeurs de références) observé dans de précédentes études impliquant le nouveau traitement. Il ne s'agit donc pas d'une étude à titre comparatif entre deux traitements, mais il s'agit de confirmer un taux observé. L'étude est dimensionnée dans le but de calculer une estimation de ce taux avec un intervalle de confiance au seuil alpha de 95% de largeur $\pm$3,3% (intervalle bilatéral et équilibré centré autour de l'estimation ponctuelle). Le calcul de la taille d'échantillon et l'IC sont basés sur l'approximation normale de la loi Binomiale.
Pour décrire l'évolution du taux au cours du temps, la courbe de survie (la courbe complémentaire à 1 de l'incidence cumulée) pour la population doit être estimée en utilisant l'estimateur de Kaplan-Meier.

Or l'hypothèse de censure indépendante nécessaire à l'utilisation de cet estimateur n'est pas plausible ici (présence d'événements concurrents à une réponse moléculaire, par exemple l'occurrence d'un effet secondaire grave du traitement) et cette non-plausibilité de l'hypothèse ne peut pas être négligé sans évaluer son impact sur les estimations. En effet l'estimateur de Kaplan Meier classique, en présence de risques compétitifs, est biaisé et surestime la courbe d'incidence cumulée spécifique à la réponse moléculaire au seuil 4.0 (Une preuve est fournie section 1.3.5 dans le chapitre traitant de l'estimation de l'incidence cumulée spécifique). Cette surestimation de la courbe d'incidence cumulée entrainant une surestimation de l'efficacité du traitement, elle ne peut pas être négligé.
Il faut donc utiliser une méthode d'estimation de l'incidence cumulée en présence d'événements concurrents.

En addition de l'estimation adéquate de la fonction de survie à une réponse moléculaire, nous réaliserons des modélisations (modélisation du temps de survie à une réponse moléculaire positive au seuil 4.0) en appliquant et sans appliquer la théorie des événements concurrents pour évaluer l'impact du non respect de cette hypothèse.

Les modèles de références (modèles de Cox) seront tout d'abord sélectionnés et ajustés sur les donnéess brutes présentant des valeurs manquantes, puis une analyse de sensibilité aux valeurs manquantes sera effectuée à l'aide des même modèles évalués sur des jeux de données obtenus par imputation multiple dont les estimations seront combiné en utilisant les règles de Rubin (Little et Rubin, 1987).

2.3 Description des données disponibles

Le jeu de donnée est composé de 751 patients, atteints de LMC et satisfaisants à tous les critères d'inclusion de l'étude (âge supérieure à 18 ans, etc..).

Les variables quantitatives sont :

- **COUNTRY** : indique le pays dans lequel le patient à été recruté (parmis les 24 possibles). Les effectifs sont inhomogènes, cette variable présente peu d'intêret dans la modélisation.

- **SEX** : le sexe du patient. L'étude comprend 444 hommes et 307 femmes.

- **RACE** : variable indiquant l'origine du patient. Dans l'étude 96% des patients sont "caucasiens".

- **WHO** : score de santé rudimentaire mis en place par la World Health Organization, allant de 0 à 5, chaques chiffres correspondant à un état de santé apparent (5 étant le plus grave). Les patients de l'étude présentent 3 scores WHO :
  - 0 (Sans symptômes apparent, actifs : 591 patients)
  - 1 (Présence de symptômes apparent mais aucune gêne dans la vie de tout les jours 145 patients)
  - 2 (Présence de symptômes apparent et gênant, mais moins de 50 % du temps allité : 15 patients).

- **CMLTRT** : indique si le patient a subi un traitement contre la LMC avant d'être inclus dans l'étude. Les catégories sont :
  - Traité à l'IMATINIB moins d'un mois (50 patients)
  - Traité à l'IMATINIB entre un et deux mois (59 patients)
  - Traité à l'IMATINIB plus de deux mois (39 patients)
  - Traité avec un traitement alternatif (388 patients)
  - Aucun traitement (215 patients).

  Cette variable est très importante car il est évident que la réponse au traitement va varier entre un patient qui suis déja un traitement et un patient qui n'en a jamais suivis.

- **AGE** : Les patients sont âgés de 18 à 86 ans (18 ans étant la limite inférieure imposée lors du recrutement) pour une moyenne de 51 ans et une médiane de 53 ans.

- **SOKAL** : Un score prognostique de la durée de survie établi pour des patients atteint de CGL (chronic granulocytic leukemia). La formule de calcul est la suivante :

$$RR = 0.0016(ge-43.4)+0.0345(spleen-7.51)+0.188\left((\frac{platelets}{700})^2-0.563\right)+0.0887(blasts-2.10) \tag{2.1}$$

  Avec
  - RR = risque relatif
  - spleen = taille de la rate du patient en cm
  - platelet = nombre de plaquette sanguine dans un échantillon de taille standard
  - blasts = quantité représentant le nombre de globules blancs immatures (pour plus de précision, se réferer au schéma 2.2) .

  Selon les recommendations (voir [17]) les patients peuvent être regroupés de la manière suivante : Risque faible : RR$<$0.8 et Risque élevé : RR$>$1.2.

- **EURO** : Score beaucoup plus complexe (la formule est présentée dans la publication [9]), mais conçu pour évaluer le risque de décès lors d'une intervention de chirurgie cardiaque. Il s'agit en fait du prédicteur linéaire obtenu à partir d'une régression logistique sur la variable de mortalité en chirurgie cardiaque, composé entre autre de l'âge, du sexe ... (pour une liste complète se référer à la publication [9]). Il n'est pas spécifique à la LMC mais est utilisé comme indicateur de risque général.
- **EUTOS** : Score conçu comme un prédicteur de la rémission complète cytogénétique (CCgR : aucun globules blancs déficients détectés dans les résultat d'analyses, et ce sur une période donnée) pour les patients atteints de LMC. Il s'agit à nouveau du prédicteur linéaire obtenu par régression logistique sur la variable de CCgR à partir d'un certain nombre de covariables (se référer à [8] pour

une liste des variables.)

- **WEIGHT** : poids du patient au début de l'étude. Si nécessaire, il sera possible de récupérer le poid du patient à différents moments de l'étude (données présentes dans un autre jeu de donnée de la base).

- **PB blasts % at diagnosis** : pourcentage de globules immatures dans l'échantillon de sang lors de l'entrée dans l'étude (comptage rammené à la taille de l'échantillon).

- **PB eosinophils % at diagnosis** : pourcentage d'Eosinophiles dans l'échantillon de sang lors de l'entrée dans l'étude (comptage rammené à la taille de l'échantillon).

- **PB basophils % at diagnosis** : pourcentage de basophiles dans l'échantillon de sang lors de l'entrée dans l'étude (comptage rammené à la taille de l'échantillon).

- **Platelets at diagnosis** : pourcentage de plaquettes (ou Thrombocytes) dans l'échantillon de sang lors de l'entrée dans l'étude (comptage rammené à la taille de l'échantillon).

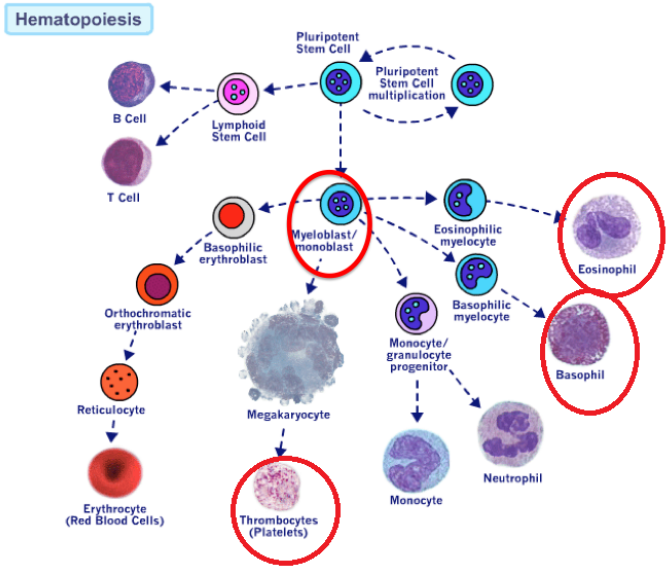

FIGURE 2.2 – Les différents types de globules, ceux considérés dans l'étude sont entourés en rouge [2]

- **Spleen size at diagnosis** : taille de la rate en cm (l'hypertrophie de la rate est un des symptômes de la LMC).

- **TIME** : Temps jusqu'au premier événement listé. C'est la variables nomée X dans les modèles de survies.

- **DESCEVENT** : description du premier évènement observé dans la liste d'évènement pré-définie.(figure 2.4). Dans un soucis de clarté, nous avons fait le choix de regrouper ces évènements en classes (figure 2.5).
  Nous y retrouvons l'estimation brute du taux de réponse moléculaire au seuil 4.0 (30,63%). Les évènements clairement indépendants de la réponse moléculaire (problemes administratifs, choix d'un nouveau traitement, déviation du protocole d'étude) sont inclus dans la censure.

2. Source : About.com , "Leukemia and Lymphoma","http ://lymphoma.about.com/od/glossary/ss/Blast-Cells.htm"
3. Source : Arcagy.org , "Info-Cancer","http ://www.arcagy.org/infocancer/localisations/hemopathies-malignes-cancers-du-sang/myelome-multiple/maladie/les-organes-lymphoides-peripheriques.html"

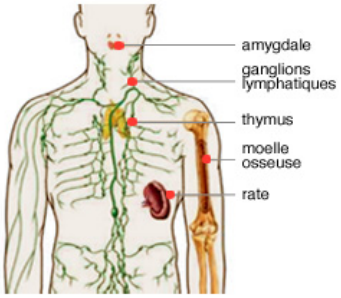

FIGURE 2.3 – Localisation de la rate [3]

desc_event	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Abnormal laboratory value(s)	3	0.40	3	0.40
Abnormal test procedure result(s)	3	0.40	6	0.80
Administrative problems	2	0.27	8	1.07
Adverse Event(s)	77	10.25	85	11.32
Death	3	0.40	88	11.72
Disease progression	12	1.60	100	13.32
Lost to follow-up	5	0.67	105	13.98
MR40	230	30.63	335	44.61
New cancer therapy	5	0.67	340	45.27
Protocol deviation	7	0.93	347	46.21
Subject withdrew consent	23	3.06	370	49.27
censor	381	50.73	751	100.00

FIGURE 2.4 – Tableau récapitulatif des évènements présents dans le jeu de données

	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Effets secondaires grave	77	10.25	77	10.25
Censure	395	52.60	472	62.85
Données biologique anormales	18	2.40	490	65.25
Décès	3	0.40	493	65.65
MR40	230	30.63	723	96.27
Perte du suivis	28	3.73	751	100.00

FIGURE 2.5 – Tableau récapitulatif des groupes créés à partir des évènements présent dans le jeu de données

2.4 Les valeurs manquantes

Nous remarquons qu'il y a de nombreuses valeurs manquantes dans le jeu de donnée. Si rien n'est fait pour prendre en compte ces manquants, la procédure PHREG utilisée pour obtenir les estimations des modèles de survie se contente de supprimer les patients concernés (cette méthode pouvant induire

un biais).

Nous commençons par explorer le "missing data pattern", c'est à dire la répartition des valeurs manquantes pour s'assurer que celle-ci ne forment pas une structure particulière. Suivant les recommandations (publication [16] et formation interne à Quintiles [15]), nous avons inclus toutes les covariables qui seront présentes dans les modèles (voir partie Modélisation 3.3). Les variables non-incluses sont celles ne présentant aucun intêret ni pour les modèles ni pour l'explication des manquants, et les variables quantitatives (puisque la méthode s'appuie sur la loi normale multi-variée, loi qui s'applique à des variables continues non bornée).

Missing Data Patterns									
Group	WEIGHT	EURO	logPCTBLA1N	sqrtPCTEOS1N	BAP1N	PLT1N	SIZ1N	Freq	Percent
1	X	X	X	X	X	X	X	665	88.55
2	X	.	X	X	X	X	.	26	3.46
3	X	.	X	X	X	.	X	2	0.27
4	X	.	X	X	X	.	.	2	0.27
5	X	.	X	.	X	X	X	3	0.40
6	X	.	X	.	.	X	.	3	0.40
7	X	.	.	X	X	X	X	6	0.80
8	X	.	.	X	X	X	.	7	0.93
9	X	.	.	.	.	X	X	1	0.13
10	X	.	.	.	.	X	.	2	0.27
11	X	.	.	.	.	.	X	1	0.13
12	X	.	.	.	.	.	.	5	0.67
13	.	X	X	X	X	X	X	26	3.46
14	.	.	X	X	X	X	.	2	0.27
Transformed Variables: WEIGHT EURO									

FIGURE 2.6 – Répartition des valeurs manquantes

Aucun mécanisme particulier ne se dessine pour expliquer les manquants : on suppose que les donnéess manquantes sont MAR (missing at random). Cette hypothèse est nécessaire pour utiliser la méthode d'imputation multiple, qui se déroule en deux étapes :

- On impute les valeurs manquantes en utilisant un algorithme MCMC (Monte Carlo Markov Chain) :
  1. Nous tirons aléatoirement des valeurs pour les manquants selon une loi normale multivariée dont les paramêtres sont estimés à partir des données observée : $Z^0_{miss} \rightarrow p(Z_{miss}|Z_{obs}, \theta^0)$ avec Z_{miss} les valeurs manquantes,Z_{obs} les valeurs observées et θ^0 les paramêtres initiaux de la loi multi-normale estimés uniquement sur les valeurs observée.
  2. Nous estimons θ^1 à partir des observations complétées à l'étape précédente : (Z_{obs}, Z^0_{miss}), en supposant que $\theta^1 \rightarrow p(\theta|Z_{obs}, Z^0_{miss})$ ou p représente la loi non-informative de Jeffrey. (pour plus de précisions, se rapporter au cours d'Analyse Bayésienne M1 de Monsieur Dortet).
  3. Puis nous recommençons successivement les étapes 1 et 2 (en remplaçant θ^t et Y^t_{miss} par leur versions en t+1), et ce jusqu'à ce que la chaine "converge", c'est à dire jusqu'à ce que l'écart entre deux estimations successives de θ soit inférieure à un seuil prédéfini.

  Ces étapes fournissent un jeu de donnée complet, et doivent être répétées N fois (avec N suffisamment grand) afin de produire N jeu de données complétés pour que le tirage aléatoire de l'étape 1 ai bien parcouru l'ensemble de la distribution (nécessaire pour correctement estimer la variabilité due à l'imputation multiple).

- Nous repetons ensuite les analyses sur les N jeu de données, puis nous combinons (Procédure MIANALYZE de SAS) les résultats obtenus en tenant compte à la fois de la variabilité induite par le modèle et de la variablilité induite par la méthode d'imputation multiple de la manière suivante :

  1. L'estimation combiné des coefficients du modèle (les $\widehat{\beta}_k$) est la moyenne empirique des N estimations fournies par les modèles estimés sur les N jeu de données imputés. C'est à dire :

$$\widehat{\beta}_k = \frac{1}{N}\sum_{n=1}^{N}\widehat{\beta}_{kn} \tag{2.2}$$

     Avec $\widehat{\beta}_k$ le k-ième coefficient du prédicteur linéaire combiné et $\widehat{\beta}_{kn}$ le k-ième coefficient du prédicteur linéaire estimé sur le n-ième jeu de données imputé.

  2. La variance combinée des coefficients du modèle ($\mathrm{Var}(\beta_k)$) est estimée en combinant la variance intra- et inter-jeu de données imputées.

     - La variance intra est la moyenne empirique des variances obtenues pour les N jeu de données :

$$\widehat{Var}_{intra}(\beta_k) = \frac{1}{N}\sum_{n=1}^{N}\widehat{Var}(\beta_{kn}) \tag{2.3}$$

     - La variance inter est la variance corrigée des variances obtenues pour les N jeu de données.

$$\widehat{Var}_{inter}(\beta_k) = \frac{1}{N-1}\sum_{n=1}^{N}(\widehat{Var}(\beta_{kn}) - \widehat{Var}_{intra}(\beta_k))^2 \tag{2.4}$$

     - La variance totale est alors :

$$\widehat{Var}_{tot}(\beta_k) = \widehat{Var}_{intra}(\beta_k) + (1+\frac{1}{N})\widehat{Var}_{inter}(\beta_k) \tag{2.5}$$

       Le coefficient $(1+\frac{1}{N})$ est utilisé pour accorder plus d'importance à la variance inter-jeu de données si peu de jeu sont imputés (plus d'incertitude car moins d'informations quand à la variablilité du au processus d'imputation si N est petit).

D'autres méthodes sont disponibles pour effectuer la combinaison de quantitées différentes (statistiques de tests par exemple).

Chapitre 3

Estimations du taux de réponse moléculaire, estimations non-paramétriques et modélisations

3.1 Estimation ponctuelle du taux de réponse moléculaire

Le but principal de l'étude est de confirmer le taux de réponse moléculaire comme premier évènement durant les 24 mois d'observation pour des patients atteints de LMC recevant le nouveau traitement. La méthode choisie dans le protocole pour estimer ce taux, ainsi que pour calculer l'intervalle de confiance, est l'approximation par la loi Normale de la loi Binomiale. Puisque cette analyse est l'analyse principale de l'étude, le nombre de patients nécessaires a l'étude à été calculé en inversant les formules d'estimations de ce taux (voir partie outils mathématique (1.4)).
En supposant un taux de 25% à 24 mois, et souhaitant obtenir un IC de plus ou moins 3,3% avec un seuil de confiance alpha de 95%, nous obtenons 661 patients requis par la méthode d'approximation normale, et 656 par la méthode de Wilson. Il est aussi supposé que 15% des patients récrutés environ ne correspondront pas aux critères d'inclusions dans l'étude, ce qui ramène à respectivement 760 et 754 patients le nombre de patients à recruter pour atteindre les objectifs. Au final 751 patients ont été observé sur 820 recrutés (d'où un taux de rejet de seulement 8,4 /La procédure FREQ de SAS produit les résultats suivants pour l'estimation du taux de réponse moléculaire :

MR40	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	230	30.63	230	30.63
0	521	69.37	751	100.00

Binomial Proportion for event_binom = 1	
Proportion	0.3063
ASE	0.0168

Type	95% Confidence Limits	
Wald	0.2733	0.3392
Wilson	0.2743	0.3401
Agresti-Coull	0.2743	0.3402
Jeffreys	0.2741	0.3399
Clopper-Pearson (Exact)	0.2734	0.3406

Test of H0: Proportion de MR40 = 0.25	
ASE under H0	0.0158
Z	3.5605
One-sided Pr > Z	0.0002
Two-sided Pr > |Z|	0.0004

FIGURE 3.1 – Estimation du taux de réponse moléculaire par différentes méthodes.

Remarque : comme on pouvait s'y attendre au vue de la méthode employée, les intervalles de Wilson et d'Agresti-Coull sont quasi-identiques (puisqu'Agresti-Coull correspond à Wilson avec l'approximation $1.96 \approx 2$).
L'objectif de produire des intervalles de confiances à $\pm 3{,}3\%$ est atteint, puisque $(0.3392\text{-}0.2733)/2 = 0.3295$. Néanmoins comme l'illustre le test de H0 :$p_{MR4.0}$=0.25, l'hypothèse posée lors du calcul de la

taille de l'échantillon (on supposais que p=0.25) n'est pas fondée (on rejette H0 avec une p-value de 0.0004). Au final l'erreur est rattrapée par les 90 patients supplémentaires recrutés.
Comme annoncé lors de la présentation de la méhode exacte (méthode de Clopper-Pearson, chapitre 1.2.4), l'intervalle fournit est plus large que les autres (demi-longueur = 0.0336), mais assure une probabilité de couverture (probabilité que la vrai valeur soit comprise dans cet intervalle) d'au moins 0.95%.
Basé sur ces estimations, on peut conclure que la proportion de patients suivant le nouveau traitement qui présenteront une réponse moléculaire au seuil 4.0 sur une période de 24 mois est comprise entre 27,34 et 34,06 pourcents (avec un seuil alpha de 0.05).
Une cinquième méthode est proposée par la procédure de SAS mais n'a pas été détaillée dans la partie mathématique de ce rapport, il s'agit de l'intervalle de crédibilité bayésien associé au taux (intervalle de Jeffrey).

3.2 Estimations des fonctions de survie

Nous employons l'estimateur de Kaplan et Meier pour estimer la courbe de survie en présence de censure indépendante aléatoire droite. La courbe résultante est présentée avec l'intervalle de confiance ponctuel et les bandes de confiances de Hall-Wellner (figure 3.2)

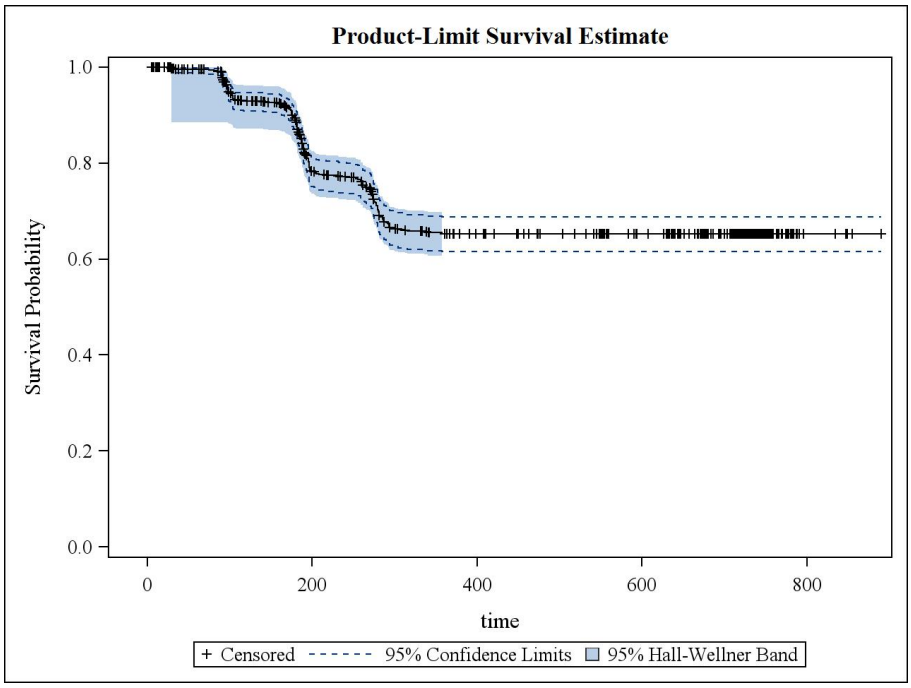

FIGURE 3.2 – Courbe de survie à la réponse moléculaire estimée par l'estimateur de Kaplan-Meier.

Pour chaques valeurs de t (axe des abscisses) cette courbe représente la probabilitée qu'un patient atteint de LMC et suivant le nouveau traitement n'ai pas de réponse moléculaire au seuil 4.0 après t-jours (à compter du premier jour de prise du traitement). La bande de confiance est la région dans laquelle la courbe se trouve avec un seuil de confiance de 95 pourcents.

Nous estimons également la courbe de survie en présence de risques compétitifs (les risques pris en compte sont ceux décrit précédements.) (figure 3.3). les donnéess nécessaires sont fournies par la macro %COMPRISK (macro basée sur l'article [4]). On présente également ces deux courbes accompagnées de leurs intervalles de confiances ponctuels respectifs (figure 3.4) calculés à la main puisqu'aucune procédure de SAS ne permet d'obtenir d'intervalles de confiance ponctuels pour l'estimateur de l'incidence cumulée en présence de risques compétitifs.

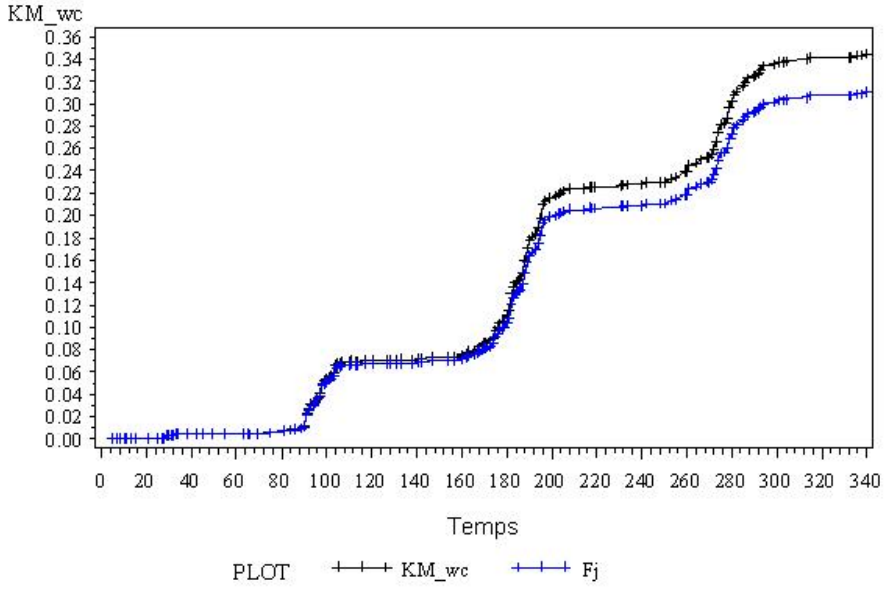

FIGURE 3.3 – Courbes d'incidences cumulées estimées en prenant en compte ou non les évenments concurrents à la réponse moléculaire.

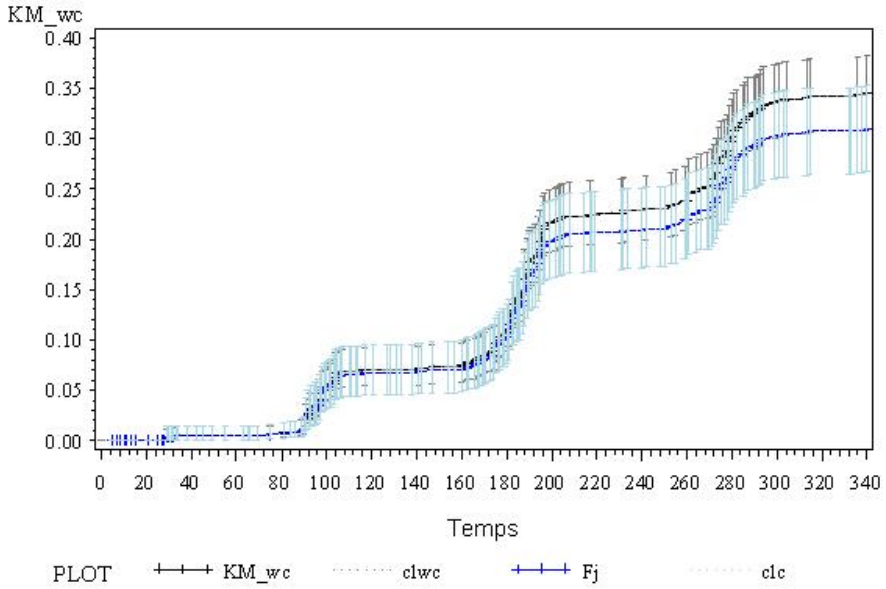

FIGURE 3.4 – Courbes d'incidences cumulées estimées en prenant en compte ou non les évènements concurrents à la réponse moléculaire, avec intervalles de confiance ponctuels.

Comme on peut le voir, la probabilité de survie estimée par les deux méthodes diffère peu (probabilité de survie à 340 jours estimée : 0.65 pour la méthode classique contre 0.69 pour la méthode prenant en compte l'influence des évènements compétitifs.), de plus les intervalles de confiance ponctuels ne sont jamais mutuellement exclusifs. On ne peut pas conclure que les deux courbes estimées sont différentes.

Intéressons nous à la fonction CP à présent, qui réprésente le rapport en t de la fonction d'incidence cumulée spécifique à la réponse moléculaire divisée par la fonction de survie à tout les autres types d'évènements (effets secondaires, décès, sortie de l'étude non indépendante du traitement, etc..). (figure

3.5). Les estimations sont fournies par la macro SAS % CUMINC.

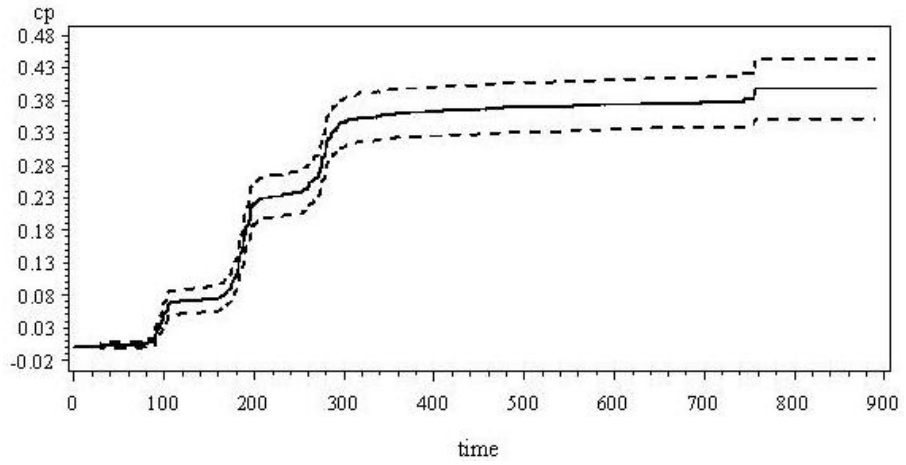

FIGURE 3.5 – Fonction CP estimée pour la réponse moléculaire accompagnée des intervals de confiances ponctuels.

A 340 jours, $\widehat{CP_{MR40}} = 0.36 = \frac{\widehat{F_{MR40}}}{\widehat{S}_{autrecauses}}$ avec $\widehat{F_{MR40}}(340) = 0.31$ (chiffres fournits par la macro SAS % CUMINC, implémentation des formules données dans la publication [13]). On retrouve bien l'estimation complémentaire à celle de la survie en présence de risques compétitifs présentée ci-dessus. En effet : $0.31 + 0.69 = 1$.
Cette courbe s'interprete pour chaques valeurs de t comme la probabilité qu'un patient suivant le nouveau traitement présente une réponse moléculaire au seuil 4.0 durant les t-jours (à compter du premier jour de traitement) sachant qu'il n'a pas subit d'effets secondaires graves (suite au traitement), que ses données biologiques sont restées stables et qu'il n'a pas stoppé le traitement.

3.3 Modèles de Cox

Nous cherchons à présent à modéliser la probabilité instantanée (à l'instant t) qu'un patient présente une réponse moléculaire positive au seuil 4.0 **sachant** que celui ci n'en a pas présenté auparavant (décrite par le risque instantané $\lambda(t)$) à l'aide d'un modèle de Cox (voir partie outils mathématique chapitre 1.3.6). Cela nous permettra d'identifer les covariables influentes et de quantifier cette influence sur le risque instantanné.
Nous rappelons que ce modèle nécéssite l'hypothèse de censure aléatoire (qui est discutable ici) et de courbes de risques instantanés proportionnelles d'un patient à l'autre.
Nous utilisons la méthode de sélection suivante : nous partons du modèle complet (avec ou sans interactions d'odre 2) et nous retirons un à un les effets non significatifs (au sens du test de Wald de significativité des coefficients) en recalculant les coefficients à chaque fois qu'un effet est retiré. Nous suivons aussi la recommandation de ne retirer un effet que si toutes les interactions associées l'on été. La validité des hypothèses est testée pour les modèles candidats grâce à la méthode implémentée dans SAS ([12]) , et les remèdes conseillés sont appliqués.

Les deux modèles retenu suite aux étapes de sélection (selon la méthode détaillée ci-dessus) sont :
- SEXE + CMLTRT + PCTBLA1N + PCTEOS1N + PLT1N + SIZ1N (modèle 1)
- SEX+ WEIGHT+ CMLTRT+ EURO + WEIGHT*EURO + PCTBLA1N +WEIGHT*PCTBLA1N +PCTBLA1N*CMLTRT +PCTEOS1N +BAP1N + WEIGHT*BAP1N + PLT1N + PCTEOS1N*PLT1N +SIZ1N
  (modèle 2).

Pour une description précise de chaques covariables, nous renvoyons au chapitre 2.3.

Evaluation du modèle 1 :
Obtenu en retirant un à un les effets non-significatifs (selon le test de Wald) à partir du modèle complet sans interactions.
Le premier test fournit par la méthode de vérification (H0 : bonne forme fonctionelle de la covariable dans le modèle contre H1 : mauvaise forme fonctionelle [1]) révèle une mauvaise forme fonctionnelle pour les effets PCTBLA1N (p-value=0.009) et PCTEOS1N (p-value=0.0330). De plus le test d'hypothèse de hasards proportionnels (H0 : risques proportionnels contre H1 : risques non proportionnels [2]) rejette H0 pour les effets CMLTRT et SIZ1N, il faut envisager de stratifier le modèle selon ces variables (nous choisissons de commencer par CMLTRT).
Nous choisissons de constituer 3 groupes : les patients ayant déja suivis un traitement à base d'Imatinib, les patients ayant suivis un autre type de traitement et ceux n'ayant pas suivis de traitement (au regard des résultats du test de risques proportionnels en choisissant successivement chaque modalités comme référence).

The PHREG Procedure

Supremum Test for Functional Form				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
PCTBLA1N	18.0021	1000	11111	0.0090
PCTEOS1N	16.4631	1000	11111	0.0330
PLT1N	15.0048	1000	11111	0.0640
SIZ1N	5.0294	1000	11111	0.6330

FIGURE 3.6 – Vérification de la forme fonctionnelle : tableau de résultats des tests

La stratification résout le problème de non proportionalité en dissociant les courbes de risques pour les patients de chacun des groupes (les autres covariables passant alors toutes le test avec une p-value >0.05). Mais les effets PCTBLA1N,PCTEOS1N et PLT1N ne présentent toujours pas une forme fonctionnelle adéquate.
On cherche donc des transformations appropriées pour ces trois effets en suivant les recommandations fournies dans la documentation de la procédure PHREG [3] Nous choisissons la transformation en log pour PCTBLA1N et PLT1N et la transformation en racine pour PCTEOS1N.

Le modèle numéro 1 après application des remèdes aux violations d'hypothèses est donc :

- SEX+ log(PCTBLA1N) + sqrt(PCTEOS1N) + log(PLT1N) + SIZ1N
- Strates : CMLTRT (regroupé).

1. Pour plus de détail sur ce test nous renvoyons vers la publication [12] chapitre 2.3 :"Checking the functional form of a covariate" page 561
2. Pour plus de détail sur ce test nous renvoyons vers la publication [12] chapitre 2.5 :"Checking the proportionnal hazards assumption" page 562
3. Source :"*http : //support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug$_{p}$hreg$_{s}$ect057.htm*" Output 66.12.3 "Typical Cumulative Residual Plot Patterns"

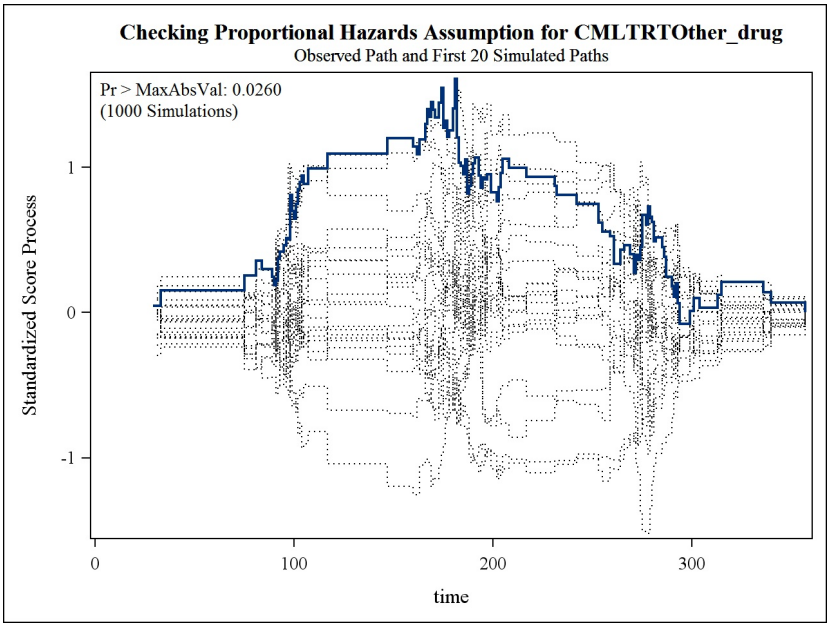

FIGURE 3.7 – Vérification de l'hypothèse de risque instantanés proportionnels : tracé d'une trajectoire pour le processus estimé avec le jeu de donnée contre 20 trajectoires pour ce même processus sous hypothèse de proportionnalité des risques pour CMLTRT

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
SEX	Female	1	0.29758	0.14012	4.5101	0.0337	1.347	1.023	1.772	Sex Female
logPCTBLA1N		1	-0.06259	0.01208	26.8486	<.0001	0.939	0.917	0.962	
sqrtPCTEOS1N		1	0.20212	0.08664	5.4421	0.0197	1.224	1.033	1.451	
logPLT1N		1	0.44765	0.12095	13.6988	0.0002	1.565	1.234	1.983	
SIZ1N		1	-0.06894	0.02176	10.0371	0.0015	0.933	0.894	0.974	Spleen size at diagnois

FIGURE 3.8 – Résultat des estimations pour les parametres du modele de Cox numéro 1.

Résultat : au regard des estimations on peut conclure que dans le cadre de ce modèle :

- Les femmes ont significativement plus de risque (entre 1.023 et 1.772 fois) que les Hommes de présenter une réponse moléculaire au seuil 4.0 à un instant t en suivant le nouveau traitement. Un pourcentage d'eosinophile et de plaquette élevé augmentent aussi ce risque.
- Au contraire un pourcentage de cellules immatures élevées et une rate de taille supérieure à la moyenne auront tendance à faire diminuer ce risque (sans surprise pour la taille de la rate, puisque c'est un facteur prognostique connu de la LMC).
- Pour finir la comparaison des baselines entre les différentes strates (attention à l'échelle) nous permet de voir que le risque cumulé de réponse moléculaire est plus important chez les patients n'ayant pas encore suivis de traitement contre la LMC que chez les autres patients (risque cumulée à 340 jours de 0,47 contre 0,24 (autres traitements) et 0,35 (imatinib)). Une explication à ce phénomène serait que les patients inclus dans l'étude ayant déja suivis un traitement sans avoir présenté de réponse moléculaire présente une LMC plus "resistante" en moyenne que ceux n'ayant jamais suivis de traitement. (les patients pré-traité et peu atteints ayant probablement déja eu une réponse moléculaire positive avant le début de l'étude et n'étant pas recrutés pour cette raison)

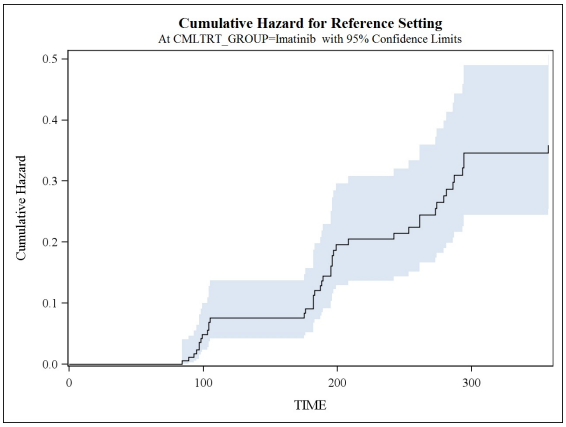

FIGURE 3.9 – Risque cumulé pour le patient de référence (un homme avec des données biologique moyenne) ayant suivis un traitement à l'imatinib avant de commencer le nouveau traitement)

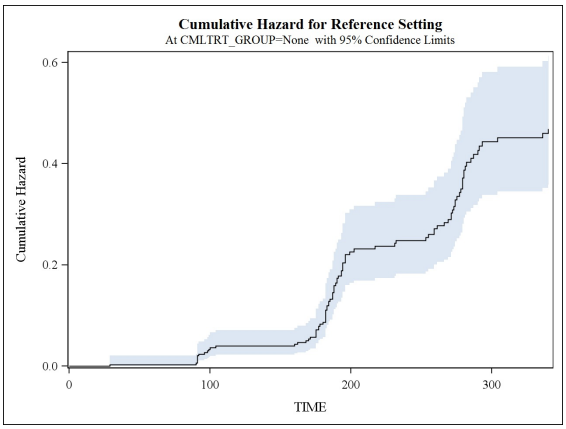

FIGURE 3.10 – Risque cumulé pour le patient de référence (un homme avec des données biologique moyenne) n'ayant suivis aucun traitement avant de commencer le nouveau traitement)

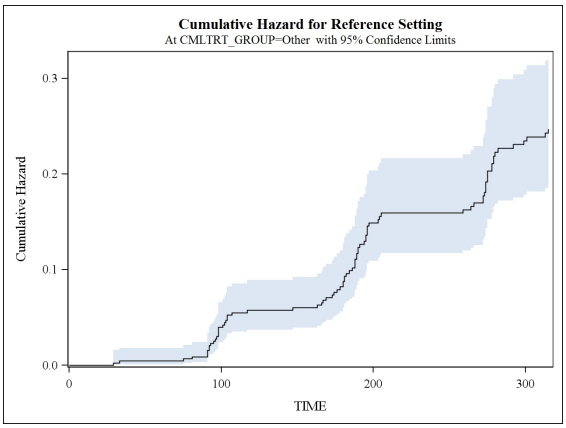

FIGURE 3.11 – Risque cumulé pour le patient de référence (un homme avec des données biologique moyenne) ayant suivis un traitement (différent de l'Imatinib) avant de commencer le nouveau traitement)

On évalue l'impact des observations manquantes (voir partie : description des données) dans ce modèle grâce à la méthode de l'imputation multiple. Comme on peut le voir les estimations sont relativement stable : aucun changement n'est à noter dans les conclusions faites ci-dessus (les intervalles de confiance

excluent tous 0).

Parameter Estimates							
Parameter	**Estimate**	**Std Error**	**95% Confidence Limits**		**DF**	**Minimum**	**Maximum**
SEX	0.290727	0.135230	0.02568	0.55577	5.8E7	0.264212	0.318855
logPCTBLA1N	-0.061335	0.011790	-0.08444	-0.03823	1.58E6	-0.066860	-0.055427
sqrtPCTEOS1N	0.176848	0.085785	0.00871	0.34498	1.6E7	0.148927	0.200449
logPLT1N	0.533880	0.116875	0.30481	0.76295	4.8E7	0.500409	0.556273
SIZ1N	-0.065732	0.021223	-0.10733	-0.02413	134422	-0.085419	-0.045853

FIGURE 3.12 – Résultat des estimations pour les parametres du modele de Cox numéro 1 après combinaison des résultat obtenus sur 1000 jeux de donnée imputés.

Critiques :
Ce modèle ne prend pas en compte les interactions d'ordre 2.
L'hypothèse de risques proportionnels est discutable pour l'effet SIZ1N, en effet le test ne rejette pas l'hypothèse H0 mais avec une p-value de 0.05 soit juste à la frontière.
De plus il a fallu stratifier le modèle pour être en adéquation avec les hypothèses posées.

Evaluation du modèle 2 :
Ce modèle est obtenu en retirant les effets non significatifs selon le test de Wald (H0 : nullité du coefficient Beta), à partir du modèle complet avec interactions d'ordre 2.
On remarque l'apparition des effets du Poids, du score EURO et de leurs interactions, masqués dans le modèle 1. De même l'introduction du poids permet de mettre en évidence un effet de l'interaction entre le poid et le pourcentage de cellules immatures, et entre le poids et le pourcentage de Basophile. Pour terminer l'effet du précédent traitement sur le pourcentage de cellules immatures fait lui aussi son apparition.

Amélioration :
L'hypothèse de risques proportionnels n'est pas respectée par les effets CMLTRT et SIZ1N.
Nous procédons tout d'abord à une stratification par CMLTRT. Cela conduit à retirer l'effet principal CMLTRT du modèle, pour cela nous devons retirer l'interaction l'impliquant : PCTBLA1N*CMLTRT.
Le modèle 2_b prend donc en compte la stratification (selon les même groupes de pré-trraitement que le modèle 1) et se voit retirer l'interaction ci-dessus.
Dans l'évaluation du modèle 2_b , nous voyons que l'hypothèse de risques proportionnels reste fausse pour SIZ1N. Il faut donc également stratifier le modèle par cette variable. Pour éviter une surmultiplication des courbes de risques instantané de base (SIZ1N représente la taille de la rate chez le patient, c'est donc une variable continue) , nous choisissons de dichotomiser la variable en deux groupes, les patients ayant une rate de taille inférieure à 10cm et les patients ayants une rate de taille supérieure ou égale à 10cm (10 cm étant une taille moyenne attendue), conduisant au modèle 2_c.
Le modèle 2_c ne contient que des effets respectant l'hypothèse de risques proportionnels. Néanmoins les formes fonctionnelles des effets WEIGHT, PCTBLA1N et PCTEOS1N sont mauvaises (p-value resp 0.007, 0.009 et 0.005). Nous explorons donc diverses transformations possibles et sélectionnons les plus appropriées.

Supremum Test for Functional Form				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
WEIGHT	14.9559	1000	11111	0.0090
EURO	10.0149	1000	11111	0.2650
WEIGHTEURO	12.0729	1000	11111	0.0740
PCTBLA1N	16.3577	1000	11111	0.0080
WEIGHTPCTBLA1N	17.2238	1000	11111	0.0080
PCTEOS1N	19.2263	1000	11111	0.0100
BAP1N	7.5243	1000	11111	0.5310
WEIGHTBAP1N	10.4676	1000	11111	0.2910
PLT1N	11.1704	1000	11111	0.2810
PCTEOS1NPLT1N	19.1969	1000	11111	0.0070

FIGURE 3.13 – Résultat des tests de vérification de forme fonctionelle pour le modèle 2_c

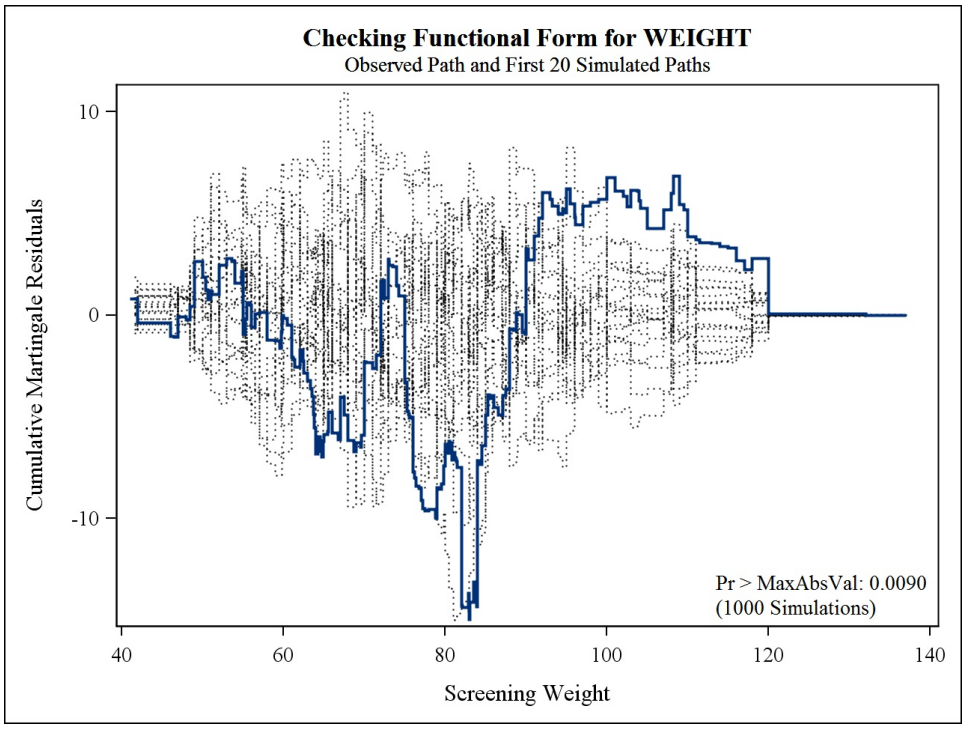

FIGURE 3.14 – Résultat des tests de vérification de forme fonctionelle pour le modèle 2_c : tracé d'une trajectoire pour le processus estimé non-concluant pour la variable Weight

Aucun trio de transformation pour ces trois effets ne rend le test de forme fonctionnelle positif. Le problème pourrait provenir de l'effet WEIGHT pour qui aucune transformation usuelle ne semble convenir. Puisque les informations sont disponibles, nous allons essayer d'introduire cet effet non plus comme un effet fixe dans le temps mais comme un effet dépendant du temps : à chaque instant t la valeur de la covariable WEIGHT pour un patient sera son dernier poid connus, et plus le poid lors de l'entrée de l'étude. Cette modélisation de la variation du poid n'est pas très réaliste (saut de poid instantané lors d'une nouvelle pesée) mais est la plus simple à mettre en place.

Cela nous conduit au modèle 3_d de la forme :

SEX+WEIGHT(t)+EURO+WEIGHT(t)*EURO+log(PCTBLA1N)+WEIGHT(t)*log(PCTBLA1N)
+sqrt(PCTEOS1N)+BAP1N+WEIGHT(t)*BAP1N+PLT1N+sqrt(PCTEOS1N)*PLT1N.

Qui n'a malheureusement pas de meilleures propriétés que le modèle sans dépendance au temps, et qui de plus n'est plus vérifiable par la méthode fournie dans SAS (en terme d'adéquations aux hypothèses, les effets inclus dans le modèle vérifié devant être fixe) . Nous nous arreterons donc au conclusions du modèle sans dépendance temporelle.

Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX	Female	1	0.39536	0.15688	6.3514	0.0117	1.485	1.092	2.019
WEIGHT		1	0.00852	0.01217	0.4899	0.4840	.	.	.
EURO		1	0.00239	0.0008039	8.8340	0.0030	.	.	.
WEIGHT*EURO		1	-0.0000332	0.0000105	10.0485	0.0015	.	.	.
logPCTBLA1N		1	-0.03136	0.06352	0.2438	0.6215	.	.	.
WEIGHT*logPCTBLA1N		1	-0.0004899	0.0008323	0.3465	0.5561	.	.	.
sqrtPCTEOS1N		1	0.52639	0.15060	12.2176	0.0005	.	.	.
BAP1N		1	-0.28635	0.10754	7.0897	0.0078	.	.	.
WEIGHT*BAP1N		1	0.00423	0.00142	8.9353	0.0028	.	.	.
PLT1N		1	0.00198	0.0004595	18.5556	<.0001	.	.	.
sqrtPCTEOS1N*PLT1N		1	-0.0006988	0.0002405	8.4392	0.0037	.	.	.

FIGURE 3.15 – Estimations des coefficients pour le modèle de Cox 2 après application des remèdes.

Critiques :
Nous voyons pour ce modèle les limitations inhérentes au modèle de Cox et ses hypothèses. L'hypothèse de risques proportionnels est plausible en générale, mais peut ne pas être satisfaite pour un certain nombre de covariables. La technique de stratification du modèle est possible pour y remédier, mais est limité par le nombre de baseline qui devient rapidement exhubérent(surtout lors d'une stratification par une variable continue : on ne peux pas stratifier une infinité de fois, il faut donc dichotomiser la variable, en perdant de l'information) rendant l'interprétation compliquée. De plus les estimations fournies reposent sur la maximisation de la vraisemblance partielle de Cox, qui comme nous l'avons montré repose elle même sur l'hypotèse de censure aléatoire indépendante de T.
Nous allons donc changer de modèle pour prendre en compte les évènements concurrents. Ce sera l'objet du paragraphe suivant.

3.4 Modèles de Fine et Gray

On modélise les donnéess dans un contexte de risques compétitifs. On a déja vu que ce changement de modélisation entraine une rectification de l'estimation non paramétrique de l'incidence cumulée, puisque l'estimateur classique de Kaplan et Meier se trouve être biaisé ici.
On modélise à présent la fonction de risque instantanée associé à la sous fonction d'incidence cumulée spécifique à la réponse moléculaire (présentée dans la partie outils mathématiques, équation (1.19)).
Pour mettre en relief les différences d'estimations, nous avons choisis de conserver les modèles précédents (modèles 1 et 2), ce sont donc les mêmes covariables qui sont inclusent dans le prédicteur. Une autre façon d'aborder le problème aurait été de resélectionner des prédicteurs à partir de ce nouveau modèle et de comparer les prédicteurs entre eux (mais alors les estimations obtenues n'auraient plus été comparable).
Les résultats des modèles de Gray estimés par la macro PHSREG (Georg Heinze, 2011-2012) sont les suivants :

Analysis of Maximum Likelihood Estimates										
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
SEX	Female	1	0.29758	0.14012	4.5101	0.0337	1.347	1.023	1.772	Sex Female
LOGPCTBLA1N		1	-0.06259	0.01208	26.8486	<.0001	0.939	0.917	0.962	
SQRTPCTEOS1N		1	0.20212	0.08664	5.4421	0.0197	1.224	1.033	1.451	
LOGPLT1N		1	0.44765	0.12095	13.6988	0.0002	1.565	1.234	1.983	
SIZ1N		1	-0.06894	0.02176	10.0371	0.0015	0.933	0.894	0.974	Spleen size at diagnois

FIGURE 3.16 – Estimations des coefficients pour le modèle de Gray 1.

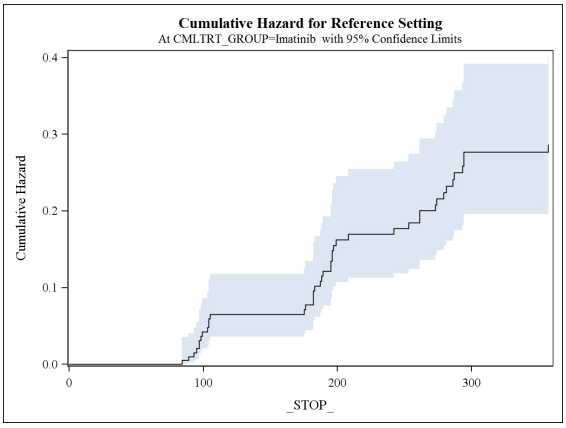

FIGURE 3.17 – Risque cumulé spécifique à la réponse moléculaire pour le patient de référence (un homme avec des données biologiques moyennes) ayant suivis un traitement à L'imatinib avant de commencer le nouveau traitement)(modèle 1)

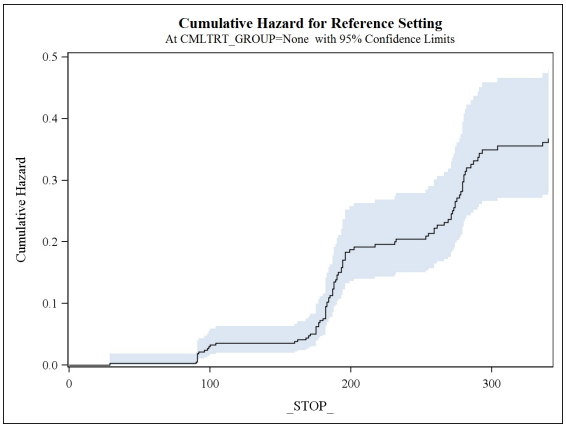

FIGURE 3.18 – Risque cumulé spécifique à la réponse moléculaire pour le patient de référence (un homme avec des données biologiques moyennes) n'ayant suivis aucun traitement avant de commencer le nouveau traitement)(modèle 1)

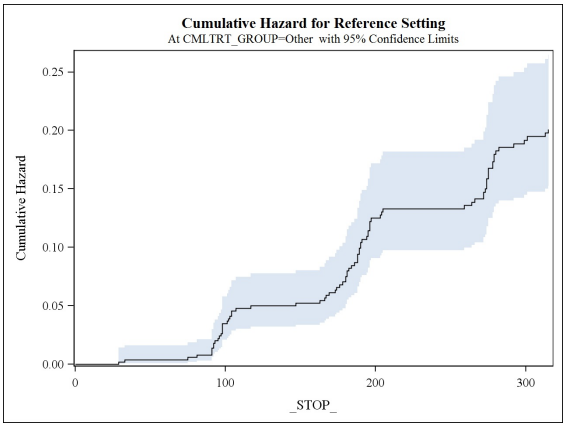

FIGURE 3.19 – Risque cumulé spécifique à la réponse moléculaire pour le patient de référence (un homme avec des données biologiques moyennes) ayant suivis un traitement (différent de l'Imatinib) avant de commencer le nouveau traitement)(modèle 1)

Analysis of Maximum Likelihood Estimates									
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX	Female	1	0.43459	0.15668	7.6939	0.0055	1.544	1.136	2.099
WEIGHT		1	0.00593	0.01229	0.2329	0.6294	.	.	.
EURO		1	0.00245	0.0008078	9.2170	0.0024	.	.	.
WEIGHT*EURO		1	-0.0000356	0.0000105	11.4952	0.0007	.	.	.
logPCTBLA1N		1	-0.01635	0.05301	0.0951	0.7578	.	.	.
WEIGHT*logPCTBLA1N		1	-0.0004901	0.0006933	0.4996	0.4797	.	.	.
sqrtPCTEOS1N		1	0.50604	0.15558	10.5793	0.0011	.	.	.
BAP1N		1	-0.33276	0.10847	9.4111	0.0022	.	.	.
WEIGHT*BAP1N		1	0.00488	0.00142	11.8110	0.0006	.	.	.
PLT1N		1	0.00181	0.0004718	14.7431	0.0001	.	.	.
sqrtPCTEOS1N*PLT1N		1	-0.0006331	0.0002547	6.1776	0.0129	.	.	.

FIGURE 3.20 – Estimations des coefficients pour le modèle de Gray 2.

Observations :

Pour le modèle 1 :

Les effets des covariables sur γ sont tous significatifs. On observe de légers glissement des intervalles de confiance (par exemple pour le sex : IC pour l'odd-ratio de [1.023 ;1.772] avec Cox contre [1.082 ;1.876] pour Gray, soit une légere augmentation de l'écart entre homme et femme). Ces glissements ne changent pas les conclusions, mais seulement l'importance de ces effets.

Pour le modèle 2 :

La significativité générale du modèle est toujours bonne (test de $H0 : \beta = (0....0)$ *contre* $H1 : \beta \neq (0.....0)$) et les conclusions des tests de Wald restent inchangées. Les estimations des paramètres sont très proches de celles obtenues pour le modèle de Cox.

La faible sensibilité des estimations des modèles à l'utilisation ou non de la théorie des risques compétitifs nous encourage à appliquer une méthode alternative, la regression quantiles, considérant elle aussi que les donnéess sont censurées aléatoirement à droite (ce qui n'est pas très handicapant dans notre cas comme nous venons de le voir lors de cette analyse de sensibilité) et surtout ne posant plus d'hypothèse de risques proportionnels. On s'affranchit donc du second point problématique soulevé dans l'analyse par les modèles de Cox (surmultiplication des stratifications pour s'assurer du respect de l'hypothèse de

risques proportionnels.).

3.5 Modèles de regression Quantiles

On reprend les mêmes covariables et on modélise cette fois ci non plus l'influence de celles ci sur le temps jusqu'à la première réponse moléculaire mais sur les quantiles de la fonction de répartition de la variable aléatoire représentant le temps jusqu'a la première réponse moléculaire. Cette modélisation considère notre problème sous un angle totalement différent, permettant de s'affranchir de l'hypothèse de risques proportionnels et de ne pas s'inquieter de la non plausibilité de l'hypothèse de censure aléatoire indépendante (dont la violation n'a qu'un faible impact sur les estimation comme on l'a montré dans la partie 3.4).
La version de SAS disponible (9.2) ne nous permet pas d'utiliser ce logiciel pour réaliser les estimations. Nous avons donc utilisé R ([14]) et le package quantreg ([11]) et effectué les estimations par la méthode de Portnoy (estimation des coefficients d'un modèle de regression Quantile en présence de censure aléatoire droite inspirée de l'estimateur de Kaplan Meier de l'incidence cumulée, publication [10] chapitre 4.2).
Les résultats sont présenté sous la forme de graphes indiquants, pour chaques effets inclus dans le modèle, la valeur du coefficient β qui lui est rattaché (dans le modèle : $Q_{log(T)|Z_i}(\tau|Z_i) = z_i'\beta(\tau)$,avec z_i le vecteur des covariables associée au patient i, auquel on ajoute en premiere position la valeur 1 pour figurer l'intercept, dont le coefficient β_0 pourra s'interpreter comme la valeur moyenne de $Q_{log(T)|Z_i}(\tau|Z_i)$).
Ces estimations s'interpretent de la manière suivante :
Un effet significativement supérieur à 0 indique qu'une augmentation de la covariable en question (ou bien le fait d'être placé dans le groupe en question dans le cas de variables qualitatives) augmente le temps nécessaire pour que $(100 * \tau)\%$ de la population concernée présentent une réponse moléculaire positive.
Ainsi comme on peut le voir sur la figure 3.21 une augmentation du pourcentage de cellules immmatures (PCTBLA1N) va allonger significativement le temps associé à tout les quantiles, donc diminuer l'espérance de présenter une réponse moléculaire.
Ces estimations sont accompagnée d'intervalles de confiance ponctuels en τ (malheureusement présentés comme des bandes bleus continues, et non des intervalles ponctuels dans les figures crées par la fonction présente dans le package) calculés à partir de résultats de convergence assymptotique présentés dans la publication [10] au chapitre 5.

Bien entendu ces grandeurs ne sont pas directement comparables aux estimations présentées pour les modèle de Cox et Gray, mais pour se donner une idée les auteurs du package ont intégré une méthode de transformation de l'estimation de l'effet des covariables sur le risque instantanné dans un modèle de Cox (coefficient de proportionnalité entre les courbes de risques instantannées du patient de référence et un patient présentant une différence d'une unité dans le cas d'une covariable quantitative, ou d'une classe différentes dans le cas d'une covariable qualitative.) en effet sur la fonction Quantile (tout en restant sous les hypothèses du modèle de Cox). Les résultats sont matérialisés par les courbes en rouge sur la figure 3.21

La figure 3.22 présente une estimation de la fonction Quantile de référence pour le patient dit médian : les valeurs de ses covariables continues sont les valeurs médianes calculée sur l'échantillon. Ses variables qualitatives sont les variables choisies commme références, c'est à dire un Homme ayant suivis un traitement (autre qu'à l'Imatinib) avant de commencer le nouveau traitement.

Les résultats viennent appuyer les conclusions faites dans les chapitres précédents, en effet on remarque que :

- Malgré des effets non-significatifs au regard des tests pour la variable CMLTRT, on voit très clairement que les courbes ne sont pas proportionnelles. On est à présent en mesure d'expliquer plus finement le phénomène qui nous avait ammené à effectuer une stratification.
- Une variable peu avoir un effet fortement changeant au cours du temps, voir un effet positif sur les premiers quantiles puis négatif sur les suivants (par exemple le fait d'appartenir au groupe de traitement "Imatinib depuis moins d'un mois" dans les modèles 1 et 2).

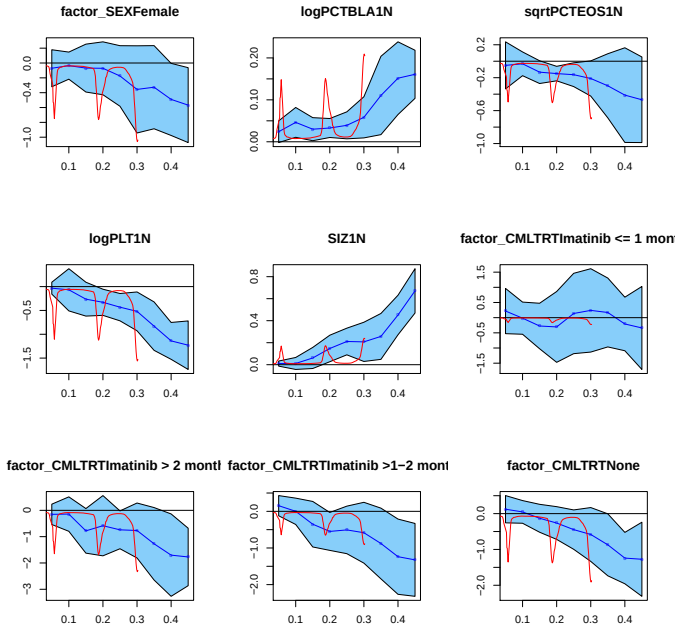

FIGURE 3.21 – Estimations des coefficients pour le modèle de regréssion Quantile 1.

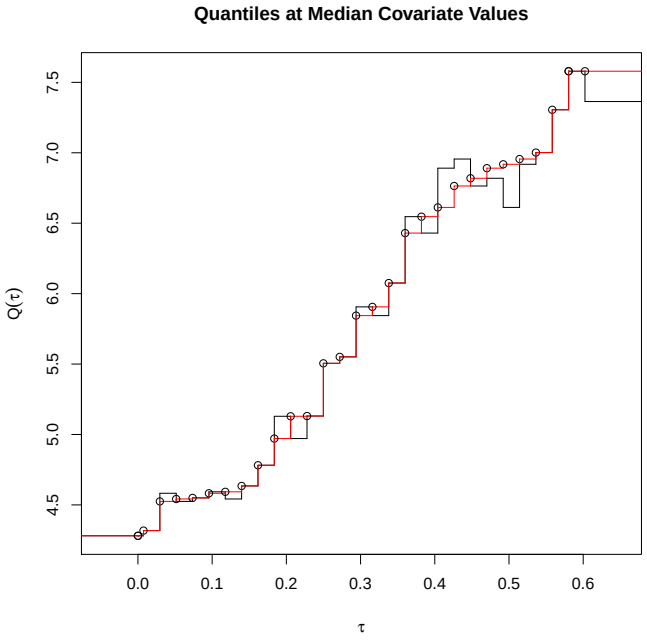

FIGURE 3.22 – Estimations de la fonction Quantile du patient médian pour le modèle de regréssion Quantile 1.

Conclusion et discussion

Le modèle de Cox est le modèle le plus utilisé en recherche clinique, dès lors qu'il s'agit de prendre en compte l'influence de covariables sur la survie des patients. Ce modèle repose sur deux hypothèses essentielles :

- Il postule que les observations dites "censurées" le sont de manière indépendante du temps de survie. Par exemple un patient quittant l'étude pour des raisons strictement personnelles n'ayant aucun lien avec l'étude. Hors la pluspart du temps les patients dits censuré sont tout ceux dont nous n'avons pas pu observer l'évènement d'intêret, pour des raisons qui sont loin d'être indépendantes du temps de survie(décès relatif au traitement par exemple). Le modèle présente alors des estimations biaisées car utilisées dans un contexte qui n'est pas le bon. La solution dans ce cas est de se rammener, comme nous l'avons vu, à la théorie des évènements compétitifs, qui estime de manière équivalente au modèle de Cox l'influence d'un certain nombre de variables choisies sur le risque instantanné (associé à la sous-fonction d'incidence cumulée spécifique à l'évènement d'intêret) , mais en tenant compte de l'influence des évènements concurrents. Un certain nombre de macro SAS existent déja pour estimer ce type de modèle.

- Le modèle de Cox se base également sur la supposition que l'effet d'une covariable sur le risque instantanné sera un effet multiplicatif constant au cours du temps. Ainsi les courbes de risques instantanné de tout les patients se doivent d'être proportionnelles les unes par rapport aux autres. Une telle supposition interdit à une des variables d'avoir un effet sur le risque instantané changeant au cour du temps (changement d'intensité ou même de signe).
  Une méthode consiste à dissocier les cas (la stratification), en estimant une courbe de risque instantannée de référence pour chaques groupes de covariables incriminées. La méthode étudiée dans ce rapport présente une alternative à la stratification. On estime l'influence des covariables non plus de manière générale sous une hypothèse qui peut se réveler infondée, mais en différents points de la courbe, autorisant cette fois ci une covariable à avoir un effet changeant au cours du temps. Le modèle ainsi posé présente la courbe des Quantiles, c'est à dire le temps nécessaire pour qu'un certain pourcentage de la population soit "décédé". L'intérêt de regresser sur les Quantiles et non directement sur un certain nombre d'instant t est que les Quantiles sont indépendants de la durée de l'étude. On peut donc décider d'estimer le modèle en τ égal à 0.5 sur un jeu de donnée provenant d'une étude sur 100 jours et sur un autre jeu de donnée provenant d'une étude sur 10 000 jours, et comparer les résultats (la médiane reste la médiane).

Une méthode de regression sur les Quantiles prennant en compte la censure aléatoire droite et les évènements concurrents serait utile (basée sur l'estimateur de l'incidence cumulée spécifique ?) . En effet les méthodes actuelles (méthode de Portnoy basée sur l'estimateur de Kaplan-Meier, ou la méthode plus récente de Peng et Huang basée sur l'estimateur de Nelson-Aalen) supposent elles aussi la présence de censure aléatoire droite comme seul évènement pouvant entrainer l'arrêt des observations.

Annexe A

A.1 Estimateur de la variance de la fonction d'incidence cumulée en présence de risques compétitifs.

En conservant les notations posée dans la partie Outils Mathématiques on a :

$$\widehat{F}_j(t) = \sum_{t \leq T'_i} ((\widehat{S}(t) * \widehat{\lambda}_j(t))$$

On peut écrire la variance ainsi :

$$Var(\widehat{F}_j(t_i)) = Var(\sum_{t \leq t'_i} ((\widehat{S}(t) * \widehat{\lambda}_j(t))) \tag{A.1}$$

En rappelant que les T'_i sont les instants de décès distincts ré-ordonnés, et que i' est le numéro du T'_i-ème instant de décès on a :

$$\begin{aligned} &= Var(\sum_{k \leq i'} ((\widehat{S}(t_k) * \widehat{\lambda}_j(t_k))) \\ &= \sum_{k=1}^{i'} Var((\widehat{S}(t_k) * \widehat{\lambda}_j(t_k))) + 2 \sum_{k=1}^{i'-1} \sum_{b=k+1}^{i'} cov((\widehat{S}(t_k) * \widehat{\lambda}_j(t_k))), (\widehat{S}(t_b) * \widehat{\lambda}_j(t_b)))) \end{aligned} \tag{A.2}$$

On rappelle que $\widehat{S}(t_k) = \prod_{b=1}^{k-1}(1 - \frac{M_b}{R_b})$ et que $\widehat{\lambda}_j(t_k) = \frac{M_{jk}}{R_k}$

On souhaite utiliser la delta-méthode pour obtenir une expression de cette variance. On accepte que $M_{jk}, M_1 ..... M_{k-1}$ sont non corrélés (Dinse et Larson 1986).
On pose : $g(M_{jk}, M_1, ... M_{k-1}) = \prod_{b=1}^{k-1}(1 - \frac{M_b}{R_b}) * \frac{M_{jk}}{R_k}$
Alors :

$Var(g(M_{jk}, M_1, ... M_{k-1})) =$

$$\begin{pmatrix} \frac{\delta g}{\delta M_{jk}} \\ \frac{\delta g}{\delta M_1} \\ ... \\ \frac{\delta g}{\delta M_{k-1}} \end{pmatrix}^T * \begin{pmatrix} Var(M_{jk}) & Cov(M_{jk}, M_1) & ... & Cov(M_{jk}, M_{k-1}) \\ Cov(M_{jk}, M_1) & Var(M_1) & ... & ... \\ ... & ... & ... & Cov(M_{jk}, M_{k-1}) \\ Cov(M_{jk}, M_{k-1}) & ... & Cov(M_{jk}, M_{k-1}) & Var(M_{k-1}) \end{pmatrix} * \begin{pmatrix} \frac{\delta g}{\delta M_{jk}} \\ \frac{\delta g}{\delta M_1} \\ ... \\ \frac{\delta g}{\delta M_{k-1}} \end{pmatrix}.$$

Et car les M sont tous non corrélé deux à deux :

$$\begin{pmatrix} \frac{\delta g}{\delta M_{jk}} \\ \frac{\delta g}{\delta M_1} \\ ... \\ \frac{\delta g}{\delta M_{k-1}} \end{pmatrix}^T * \begin{pmatrix} Var(M_{jk}) & 0 & ... & 0 \\ 0 & Var(M_1) & ... & ... \\ ... & ... & ... & 0 \\ 0 & ... & 0 & Var(M_{k-1}) \end{pmatrix} * \begin{pmatrix} \frac{\delta g}{\delta M_{jk}} \\ \frac{\delta g}{\delta M_1} \\ ... \\ \frac{\delta g}{\delta M_{k-1}} \end{pmatrix}.$$

Donc

$$Var(g(M_{jk}, M_1, ...M_{k-1})) = \left(\frac{\delta g}{\delta M_{jk}}\right)^2 Var(M_{jk}) + \sum_{a=1}^{k-1}\left(\frac{\delta g}{\delta M_a}\right)^2 Var(M_a) \tag{A.3}$$

Avec

$$\frac{\delta g}{\delta M_{jk}} = \left(\prod_{z=1}^{k-1}\left(1-\frac{M_z}{R_z}\right)\right)\frac{1}{R_k} = \frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{M_{jk}}$$

$$\frac{\delta g}{\delta M_a} = \frac{M_{jk}}{R_k}\left(\prod_{\substack{z=1\\z\neq a}}^{k-1}\left(1-\frac{M_z}{R_z}\right)\right)\left(\frac{-1}{R_a}\right) = \frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{\left(1-\frac{M_a}{R_a}\right)}\left(\frac{-1}{R_a}\right) = \frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{(R_a-M_a)}(-1)$$

$$M_{jk} \to Bin(R_k, \frac{M_{jk}}{R_k}, R_k * (\frac{M_{jk}}{R_k}) * (1-\frac{M_{jk}}{R_k})) \Rightarrow Var(M_{jk}) = R_k * (\frac{M_{jk}}{R_k}) * (1-\frac{M_{jk}}{R_k}) = \frac{M_{jk}(R_k-M_{jk})}{R_k}$$

$$M_k \to Bin(R_k, \frac{M_k}{R_k}, R_k * (\frac{M_k}{R_k}) * (1-\frac{M_k}{R_k})) \Rightarrow Var(M_k) = R_k * (\frac{M_k}{R_k}) * (1-\frac{M_k}{R_k}) = \frac{M_k(R_k-M_k)}{R_k}$$

En remplaçant dans (A.3) on obtient :

$$\begin{aligned}
Var(g(M_{jk}, M_1, ...M_{k-1})) =& \ Var((\widehat{S}(t_k) * \widehat{\lambda}_j(t_k))) \\
&= \left(\left(\prod_{z=1}^{k-1}\left(1-\frac{M_z}{R_z}\right)\right)\frac{1}{R_k}\right)^2\left(\frac{M_{jk}(R_k-M_{jk})}{R_k}\right) \\
&+ \sum_{a=1}^{k-1}\left(\left(\frac{M_{jk}}{R_k}\left(\prod_{\substack{z=1\\z\neq a}}^{k-1}\left(1-\frac{M_z}{R_z}\right)\right)\left(\frac{-1}{R_a}\right)\right)^2\left(\frac{M_a(R_a-M_a)}{R_a}\right)\right) \\
&= \left(\frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{M_{jk}}\right)^2\left(\frac{M_{jk}(R_k-M_{jk})}{R_k}\right) + \sum_{a=1}^{k-1}\left(\frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{(R_a-M_a)}(-1)\right)^2\left(\frac{M_a(R_a-M_a)}{R_a}\right) \\
&= (\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k))^2\left(\frac{R_k-M_k}{M_{jk}R_k} + \sum_{a=1}^{k-1}\left(\frac{M_a}{(R_a-M_a)R_a}\right)\right)
\end{aligned} \tag{A.4}$$

(A.4) nous donne une première partie de l'expression de (A.2)

Et on procède de la même façon pour déterminer une expression de

$$cov((\widehat{S}(t_{k-1}) * \widehat{\lambda}_j(t_k))), (\widehat{S}(t_{b-1}) * \widehat{\lambda}_j(t_b))))$$

On pose :

$$g(M_{kj}, M_1, ...M_{k-1}) = \left(\prod_{z=1}^{k-1}\left(1-\frac{M_z}{R_z}\right)\right)\frac{M_{jk}}{R_k}$$

$$f(M_{bj}, M_1, ...M_{k-1}, M_k.....M_{b-1}) = \left(\prod_{z=1}^{b-1}\left(1-\frac{M_z}{R_z}\right)\right)\frac{M_{jb}}{R_b}$$

Avec : $k = 1....(i'-1)$ et $b = (k+1)....i'$.
Et $Cov(M_{jz}, M_{jz'}) = 0$ si $z \neq z'$, donc $Cov(M_z, M_{jz'}) = 0$ si $z \neq z'$ et $Cov(M_z, M_{z'}) = 0$ si $z \neq z'$
Il reste au final :

$$\begin{aligned}
&Cov(g(M_{kj}, M_1, ...M_{k-1}), f(M_{bj}, M_1, ...M_{k-1}, M_k.....M_{b-1})) \\
&= \left(\frac{\delta g}{\delta M_{jk}}\right)\left(\frac{\delta f}{\delta M_k}\right)cov(M_{jk}, M_k) + \sum_{a=1}^{k-1}\left[\left(\frac{\delta g}{\delta M_a}\right)\left(\frac{\delta f}{\delta M_a}\right)Var(M_a)\right]
\end{aligned} \tag{A.5}$$

Et

$$\frac{\delta g}{\delta M_{jk}} = \left(\prod_{z=1}^{k-1}\left(1-\frac{M_z}{R_z}\right)\right)\frac{1}{R_k} = \frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{M_{jk}}$$

$$\frac{\delta f}{\delta M_k} = \frac{M_{jb}}{R_b}\left(\prod_{\substack{z=1\\z\neq k}}^{b-1}\left(1-\frac{M_z}{R_z}\right)\right)\left(\frac{-1}{R_k}\right) = \frac{\widehat{S}(t_{b-1})\widehat{\lambda}_j(t_b)}{\left(1-\frac{M_k}{R_k}\right)}\left(\frac{-1}{R_k}\right) = \frac{\widehat{S}(t_{b-1})\widehat{\lambda}_j(t_b)}{(R_k-M_k)}(-1)$$

$$Cov(M_{jk},M_k) = Cov(M_{jk},\sum_{z=1}^{J}M_{zk}) = \sum_{z=1}^{J}Cov(M_{jk},M_{zk}) = \sum_{\substack{z=1\\z\neq j}}^{J}Cov(M_{jk},M_{zk}) + Var(M_{jk})$$

$$= \frac{M_{jk}(R_k-M_k)}{R_k}$$

$$\frac{\delta g}{\delta M_a} = \frac{M_{jk}}{R_k}\left(\prod_{\substack{z=1\\z\neq a}}^{k-1}\left(1-\frac{M_z}{R_z}\right)\right)\left(\frac{-1}{R_a}\right) = \frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{(R_a-M_a)}(-1)$$

$$\frac{\delta f}{\delta M_a} = \frac{M_{jb}}{R_b}\left(\prod_{\substack{z=1\\z\neq k}}^{b-1}\left(1-\frac{M_z}{R_z}\right)\right)\left(\frac{-1}{R_a}\right) = \frac{\widehat{S}(t_{b-1})\widehat{\lambda}_j(t_b)}{(R_a-M_a)}(-1)$$

$$Var(M_a) = \frac{M_a(R_a-M_a)}{R_a}$$

En remplaçant dans (A.5) on obtient :

$$\begin{aligned}
&Cov(g(M_{kj},M_1,.....M_{k-1}),f(M_{bj},M_1,....M_{k-1},M_k.....M_{b-1}))\\
&= Cov((\widehat{S}(t_{k-1})*\widehat{\lambda}_j(t_k))),(\widehat{S}(t_{b-1})*\widehat{\lambda}_j(t_b))))\\
&= \left(\frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{M_{jk}}\right)\left(\frac{\widehat{S}(t_{b-1})\widehat{\lambda}_j(t_b)}{(R_k-M_k)}(-1)\right)\left(\frac{M_{jk}(R_k-M_k)}{R_k}\right)\\
&+\sum_{a=1}^{k-1}\left(\frac{\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)}{(R_a-M_a)}(-1)\right)\left(\frac{\widehat{S}(t_{b-1})\widehat{\lambda}_j(t_b)}{(R_a-M_a)}(-1)\right)\left(\frac{M_a(R_a-M_a)}{R_a}\right)\\
&= \widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)*\widehat{S}(t_{b-1})\widehat{\lambda}_j(t_b)*\left(\frac{-1}{R_k}+\sum_{a=1}^{k-1}\frac{M_a}{(R_a-M_a)R_a}\right)
\end{aligned} \tag{A.6}$$

Au final, en remplaçant (A.6) et (A.4) dans (A.2) on obtient la formule de la variance de l'estimateur :

$$\begin{aligned}
Var(\widehat{F}_j(t_i)) &= Var(\sum_{t\leq t'_i}((\widehat{S}(t)*\widehat{\lambda}_j(t)))\\
&= \sum_{k=1}^{i'}Var((\widehat{S}(t_k)*\widehat{\lambda}_j(t_k))) + 2\sum_{k=1}^{i'-1}\sum_{b=k+1}^{i'}cov((\widehat{S}(t_k)*\widehat{\lambda}_j(t_k))),(\widehat{S}(t_b)*\widehat{\lambda}_j(t_b))))\\
&= \sum_{k=1}^{i'}\left((\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k))^2\left(\frac{R_k-M_k}{M_{jk}R_k}+\sum_{a=1}^{k-1}\left(\frac{M_a}{(R_a-M_a)R_a}\right)\right)\right)\\
&+2\sum_{k=1}^{i'-1}\sum_{b=k+1}^{i'}\left(\widehat{S}(t_{k-1})\widehat{\lambda}_j(t_k)*\widehat{S}(t_{b-1})\widehat{\lambda}_j(t_b)*\left(\frac{-1}{R_k}+\sum_{a=1}^{k-1}\frac{M_a}{(R_a-M_a)R_a}\right)\right)
\end{aligned} \tag{A.7}$$

Ce qui clot le calcul de l'estimateur de la variance l'incidence cumulée en présence de risques compétitifs.

A.2 Résultats des estimations du modèle Quantiles 2.

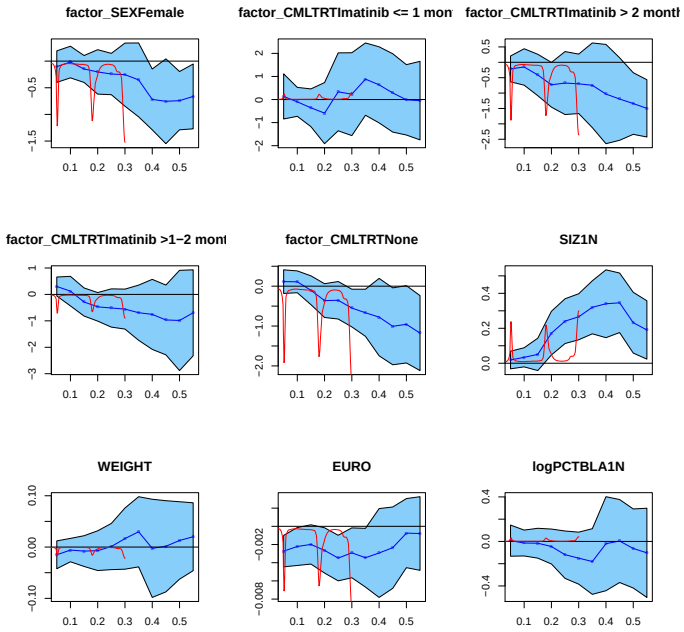

FIGURE A.1 – Estimations des coefficients pour le modèle de regréssion Quantile 2, page 1

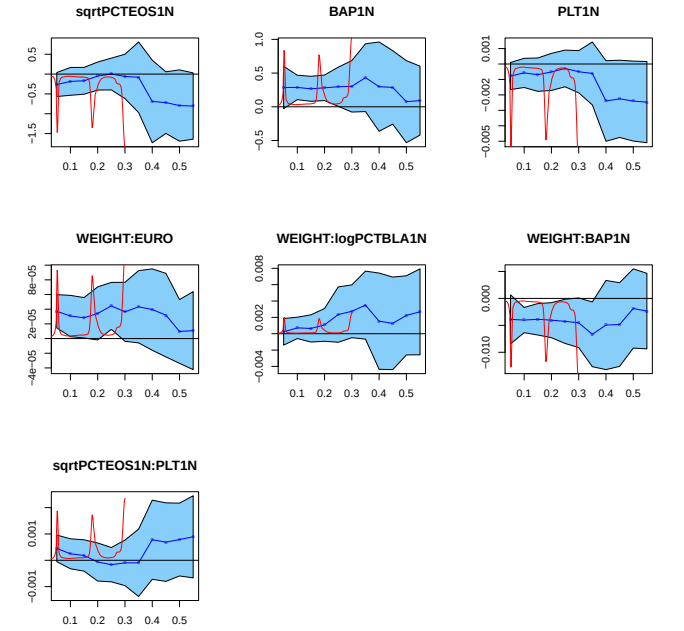

FIGURE A.2 – Estimations des coefficients pour le modèle de regréssion Quantile 2, page 2.

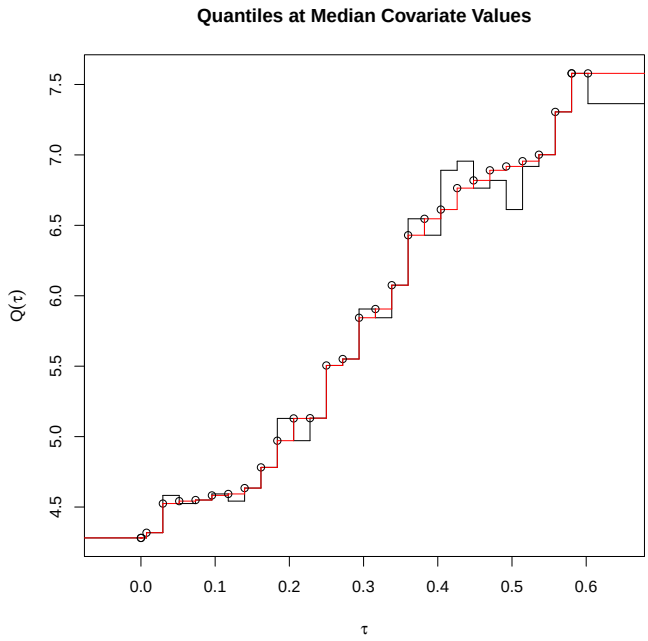

FIGURE A.3 – Estimations de la fonction Quantile du patient médian pour le modèle de regréssion Quantile 1.

Bibliographie

[1] Jean-Christophe Breton, *Processus gaussiens*, Université de La Rochelle (version de décembre 2006).

[2] D.R Cox, *Partial likelihood*, Biometrika **62** (Aug.1975), no. 2, 269–276.

[3] Benjamin Esterni, *Construction d'un modèle de fine et gray par un modèle de cox pondéré.*, Journées des statisticiens des centres (17 juin 2011).

[4] Gooley et al., *Estimation of failure probabilities in the presence of competing risks*, Statist. Med. **18** (1999), 695–706.

[5] Marubini Ettore and Valsecchi Maria Grazia, *Analysing survival data from clinical trials and observational studies*, Statistics un Practice, New York, USA, 1995.

[6] R.J. Gray, *A class of k-sample tests for comparing the cumulative incidence of a competing risk*, Ann Stat. (1988).

[7] W.J Hall and Jon A. Wellner, *Confidence bands for a survival curve from censored data*, Biometrika **67** (Apr.1980), no. 1, 133–143.

[8] Joerg Hasford, *Predicting complete cytogenetic response and subsequaent progression-free survival in 2060 patients with cml on imatinib treatment : the eutos score*, Blood Journal **118** (2011), 686–692.

[9] European Heart Journal, *The logistic euroscore*, European Heart Journal **24** (2003), 1–2.

[10] Roger Koenker, *Censored quantile regression redux*, J. of Statistical Software (2008).

[11] Roger Koenker, *quantreg : Quantile regression*, 2010, R package version 4.53.

[12] D.Y. Lin, L.J. Wei, and Z. Ying, *Checking the cox model with cumlulative sums of martingale-based residuals*, Biometrika **80** (1993), no. 3, 557–572.

[13] Margaret Sullivan Pepe and Motomi Mori, *Kaplan-meier, marginal or conditional probability curves in summarizing competing risks failure time data*, Statistics in Medicine (1993).

[14] R Development Core Team, *R : A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0.

[15] Bohdana Ratitch, Michael O'Kelly, and Sonia Davis, *Missing data in clinical trials : Multiple imputation using sas*, Quintiles - Formation Interne.

[16] D.B. Rubin, *Multiple imputation for nonresponse in surveys*, J. Wiley and Sons, New York (1987).

[17] J.E Sokal and E.B Cox, *Prognostic discrimination in good-risk chronic granulocytic leukemia*, Blood **63** (1983), 789–799.

[18] Anastasios A. Tsiatis, *A large sample study of cox's regression model*, Ann. Statist. (1981).