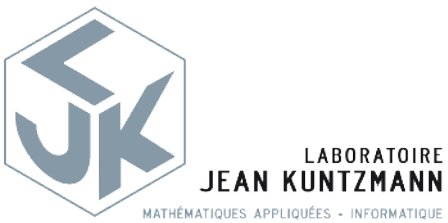
Modélisation mutli-états pour l'analyse de données de survie de patients atteints de cancer

Rapport de projet tutoré M2

Diagne Marame, McKenna Kevin, Pittion Florence, Silvestre Théo, Wang Shuyu

Sous la direction d'Adeline Leclercq Samson et Florent Chuffart



Octobre 2021 - Février 2021

Table des matières

1 Introduction

Pendant une décennie, le programme Cancer Genome Atlas (TCGA) a collecté des données d'annotation clinico-pathologiques ainsi que des profils moléculaires multiplateformes de plus de 11 000 tumeurs humaines dans 33 types de cancers différents. Pour garantir une utilisation appropriée de ce vaste ensemble de données cliniques, un article récent [1] propose un ensemble de données standardisé nommé TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR). Cette base de données intégre notamment les résultats cliniques autour de la survie (DSS *Disease Specific Survival*, PFI *Progression Free Interval*). Ce jeu de données offre une opportunité sans précédent de construire et d'étudier des modèles de survie multi-états à large échelle et sur des données réelles.

L'objectif est i) d'identifier des modèles de survie intégrant les évènements relatifs à la progression de la maladie, ii) de faire un état des lieux des outils disponibles implémentant ces modèles et iii) de mettre en oeuvre ces modèles (packages R) avec les outils identifiés dans le cadre du jeu de données TCGA-CDR.

Ce travail demandé par Florent s'inscrit dans le cadre de son activité d'analyse de données multi-omiques au sein de l'équipe EpiMed (IAB - INSERM (U1209) - CNRS (UMR5309)). Ce travail est également encadré par Adeline qui nous a fait profiter de son expérience autour des modèles de survies, des processus markoviens et plus généralement de la modélisation statistique.

1.1 Equipe EpiMed IAB

L'IAB (Institute for Advanced Biosciences) développe un ensemble de programmes de recherche fondamentale et translationnelle dans les domaines de l'épigénétique, de l'environnement, de la plasticité cellulaire, appliquées aux cancers, aux infections parasitaires, à l'infertilité et aux effets de l'environnement durant les premières étapes de la vie. L'équipe Epimed assure une activité de recherche translationnelle dans le domaine de l'épigénétique et notamment le développement et la mise en œuvre des outils bio-informatiques pour l'analyse des données "omiques" (génomiques et épigénétiques) pour l'ensemble des activités de l'IAB. L'activité EpiMed repose sur sa capacité d'analyse des données à grande échelle, essentielle à la compréhension de l'épigénome. Cette activité est renforcée par la constitution d'une cellule d'analyses bio-informatiques et l'organisation d'une base de données interactive associant les données "omics" et cliniques et/ou biologiques, utilisable par les scientifiques et les médecins impliqués dans les projets translationnels EpiMed. L'équipe est dirigée par Sophie Rousseaux et est constituée des ingénieurs de recherche en informatique Ekaterina Flin et Florent Chuffart et d'une ingénieure d'étude en biologie Anne-Laure Vitte.

Notre point de départ de travail est l'article "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics." [1] qui utilise les données du projet TCGA, nous allons donc dans un premier temps expliquer le but et les données produites par ce projet puis nous ferons une rapide synthèse du travail de l'équipe de l'article.

1.2 Projet TCGA (The Cancer Genoma Atlas)

Ce programme de génomique du cancer, a permis de caractériser plus de 20 000 échantillons primaires de cancer, et d'échantillons normaux correspondant, au niveau moléculaire. 33 types de cancer ont été documentés. Ce projet a débuté en 2006, rassemblant des chercheurs de diverses disciplines et de multiples institutions. Durant plus d'une décennie le projet TCGA a généré plus de 2,5 pétaoctets de données génomiques, épigénomiques, transcriptomiques et protéomiques. Ces données, qui ont déjà permis d'améliorer les capacités à diagnostiquer, traiter et prévenir le cancer, resteront accessibles au public et pourront être utilisées par tous les membres de la communauté des chercheurs. [2]

1.3 "An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics."

L'objectif des auteurs de cet article est de proposer des données adaptées depuis le projet TCGA pour l'analyse de survie des patients atteints de cancer. L'objectif est de travailler autour de quatre critères de survie et de les recommander différemment selon le type de données que l'on souhaite traiter. L'OS (Overall Survival), le DSS (Disease Specific Survival) qui affine la variable précédente en ne concernant que les évènements de décès lié à la maladie, le PFI (Progression Free Interval) dont l'évènement est la rechute de la maladie (cependant une mort compte aussi comme une rechute), et DFI (Disease Free Interval) dont l'évènement est la découverte d'une nouvelle tumeur.

Concernant le PFI et le DFI, la durée minimale de suivi pour ces critères est plus courte car les patients développent généralement une récidive ou une progression de la maladie avant de mourir de leur maladie. La sélection d'un critère de survie spécifique dépend également de l'objectif de l'étude. Pour chacun de ces critères les auteurs ont réalisé des courbes de survie pour comprendre leur impact sur la différence de modélisation de la survie selon les différents types de cancer, voir figure 1.

Pour la suite de notre projet nous n'allons utiliser que les critères DSS et PFI. Le critère DSS prend en compte le décès causé par le cancer, c'est ce que nous voulons modéliser et le PFI parait être un meilleur indicateur intermédiaire de la survie car nous observons qu'il est plus souvent renseigné que le DFI.

Nous n'irons pas plus loin sur les résultats de cet article, l'objectif pour nous étant d'utiliser la base de données de survie que les auteurs ont produit et donc les critères évoqués ci dessus. Nous utiliserons aussi certaines covariables cliniques disponibles que nous allons détailler plus loin.

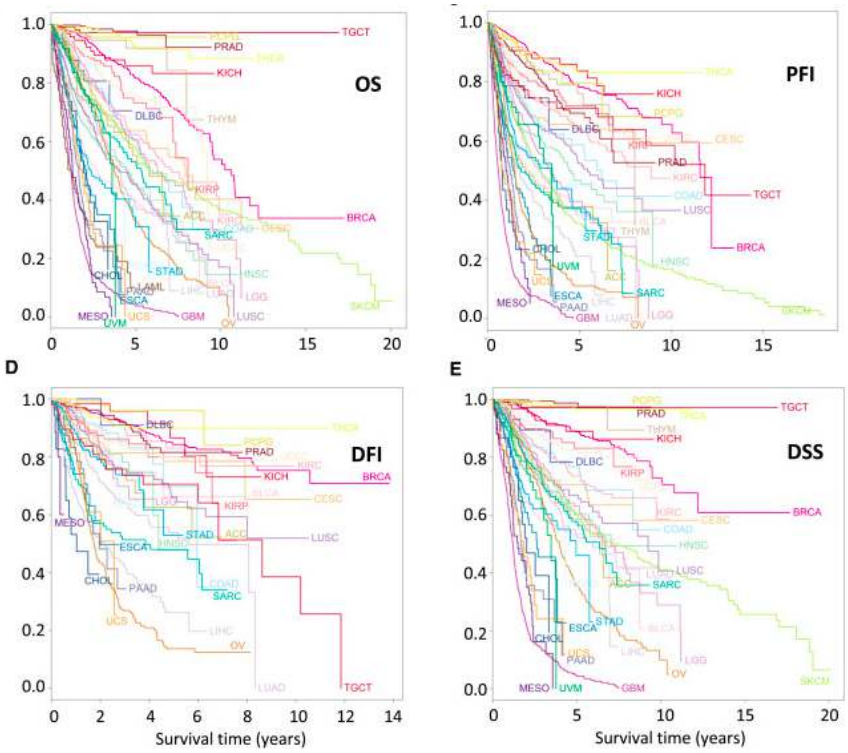


FIGURE 1 – Courbe de survie des cancers en fonction des quatre critères OS, DSS, PFI, DFI [1]

2 Exploration de données

2.1 Statistiques descriptives

Le jeu de données enregistre des informations sur plus de 11.000 patients atteints de cancer. Il est composé de 34 variables de données cliniques et de survie des patients dont principalement :

- le type de cancer (BRCA pour le cancer du sein, OV pour le cancer de l'ovaire, BLCA pour le cancer de la vessie, etc) ;
- l'âge du patient au diagnostic initial du cancer ;
- le genre du patient ;
- le stade du cancer (*I*, *II*, *III*, *IV*) ;
- le dernier état vital (mort ou vivant) ;
- le dernier statut tumoral ;
- les différentes données de survie OS, DSS, PFI et DFI évoquées plus haut.

Des analyses descriptives basiques peuvent donner une première idée de la structure du jeu de données et des cancers :

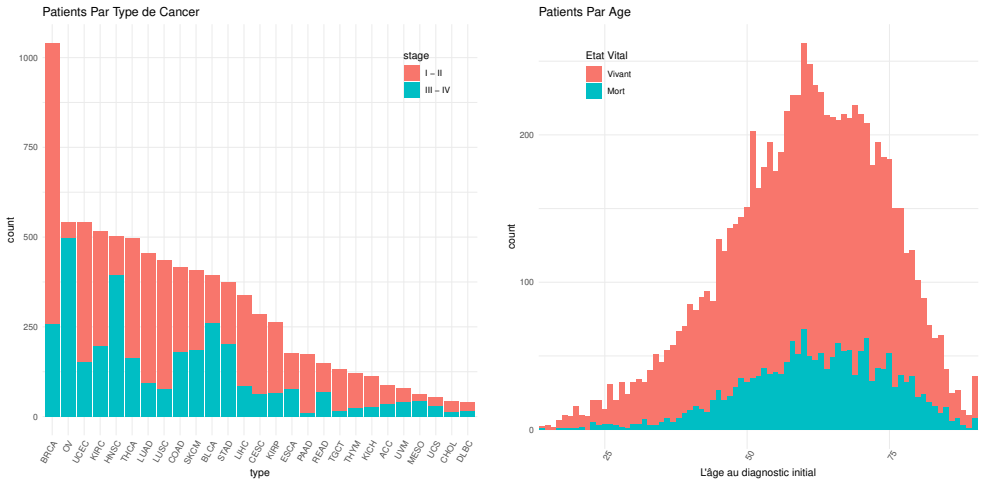


FIGURE 2 – Nombre de patients par type de cancer (avec la répartition dans stade de cancer) et par âge au diagnostic initial (avec le statut vital)

Comme le graphique ci-dessus le laisse voir, les effectifs de patients par type de cancer sont très variés. Ainsi certains cancers comme le *CHOL* concernent très peu d'individus, ce qui peut impacter les résultats numériques de nos modèles. A l'inverse, le *BRCA* et l'*OV* sont ceux concernant le plus d'individus et donc seront naturellement nos premiers choix de modélisation. La covariable sur laquelle nous allons focaliser notre attention est celle du stade du cancer, on attend biologiquement un effet sur la survie. Le nettoyage de cette variable qui est composée en partie de celle du stade tumoral et du stade clinique, va occasionner une réduction supplémentaire du nombre d'individus.

Les deux événements DSS et PFI sont explicités dans le schéma de la Figure 3. Les données de survie sont une mesure de temps, ici à partir du diagnostic de la maladie, en pratique une opération pour retirer une tumeur, jusqu'à l'observation de l'événement observé. Pour rappel le DSS est le temps de la mort suivant la maladie (en pratique présence d'une tumeur ou diagnostic clinique) et le PFI est le temps de l'apparition de la tumeur (ou la mort), il peut coïncider avec DSS. Ces données sont entachées de censure dites à droite, quand pour un individu l'événement

ne se produit pas avant la fin du suivi ou que tout simplement il quitte l'étude en cours de route.

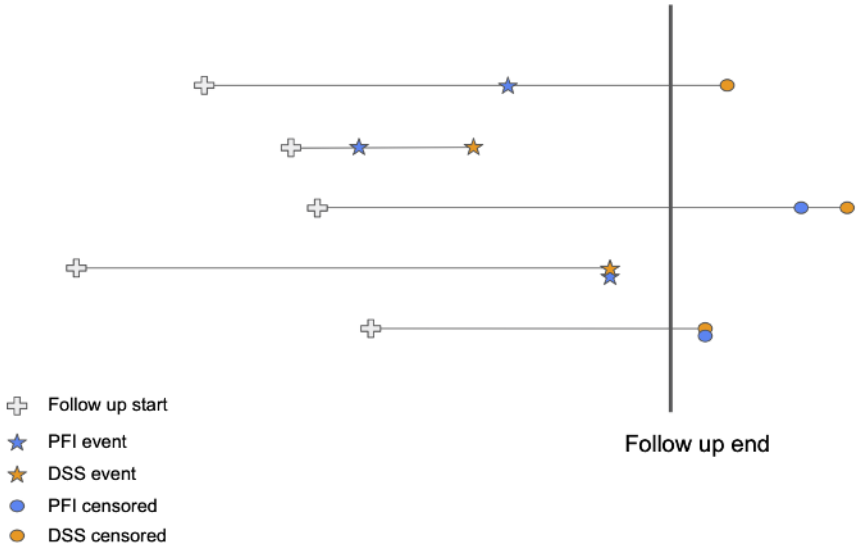


FIGURE 3 – Schéma de la censure sur les événements DSS et PFI.

La Figure 4 permet d'apprécier les différences de survie selon l'événement DSS et PFI :

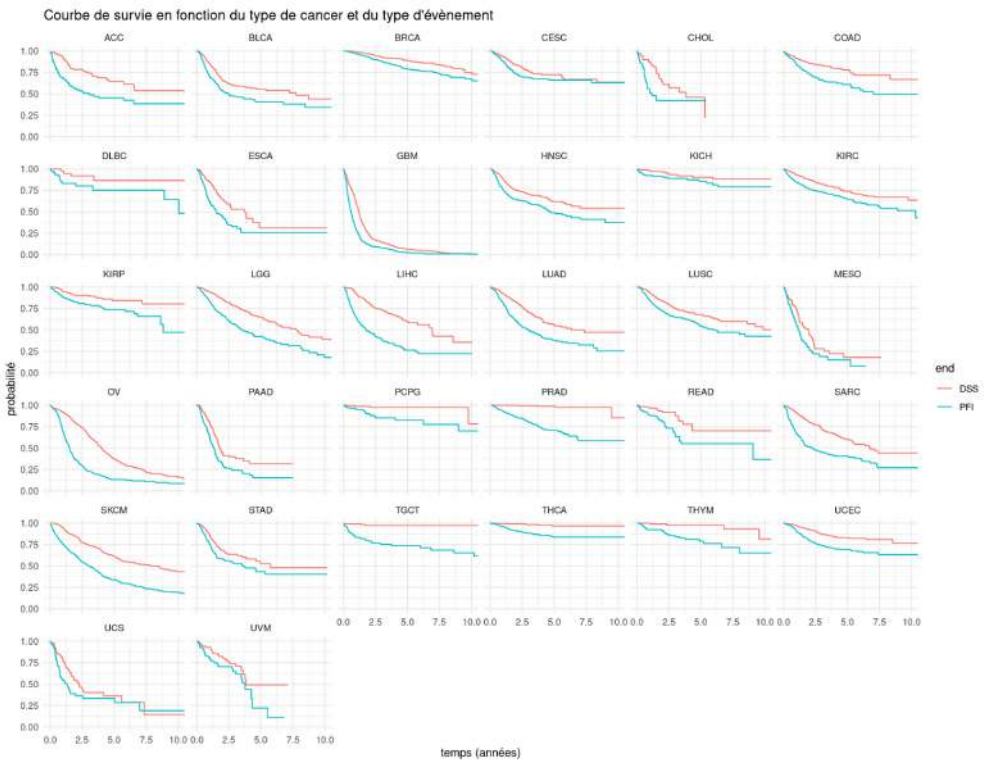


FIGURE 4 – Courbes de survie des différents cancers en fonction des types de survie DSS et PFI.

Cette représentation permet, en première analyse, de détecter les cancers les plus agressifs, c'est-à-dire ceux pour lesquels les patients meurent rapidement, ceux qui meurent sans rechute, ou au contraire qui passent par une rechute. En cela les deux cancers *BRCA* et *OV* semblent être de bons candidats à une modélisation multi-états, c'est-à-dire pouvant passer par une rechute avant la mort, tout en proposant deux profils assez différents. Le cancer *BRCA* est peu mortel à court terme quand l'*OV* montre une forte rechute dans les premières années suivant la mort, dans les 10 ans après diagnostic. Nous allons maintenant rentrer plus dans le détail en ce qui concerne le traitement des données.

2.2 Traitement des données

Les données sont assez hétérogènes du fait qu'elles proviennent de nombreuses sources. Un premier nettoyage a été effectué par l'équipe du *TCGA* et c'est à celui-ci que nous allons nous fier. La variable DSS est théoriquement dérivée du statut vital et tumoral du patient, ainsi que d'observations cliniques (qualitatives). Cependant en pratique, il y a des écarts à cette norme. Le PFI n'en est pas exempt. Une passe de nettoyage est effectuée pour retirer les valeurs manquantes relatives au DSS et au PFI, ainsi qu'aux covariables d'intérêt, ici le stade de la maladie dérivé de la variable de stade tumoral et de stade clinique. Certains artefacts restant (en petit nombre sur le total) sont retirés comme des individus observant un PFI après un DSS ou ayant ces deux événements au temps 0. Enfin, les stades sont regroupés en *I - II* et *III - IV* pour simplifier les résultats, le premier et deuxième stade donnant des pathologies similaires, de même pour le troisième et quatrième.

Pour utiliser un modèle multi-état, il faut définir quels sont les états, comme représenté dans la Figure 5 :

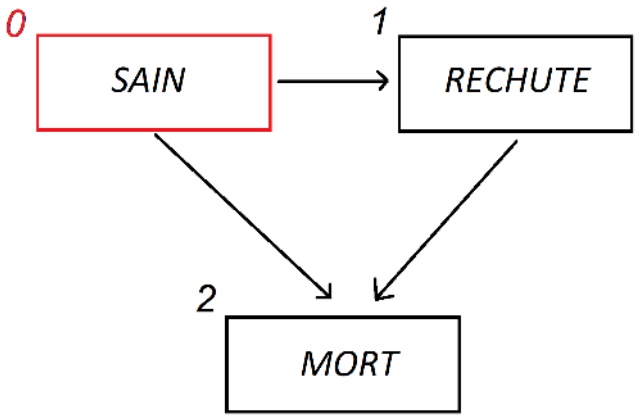


FIGURE 5 – Schéma explicitant les états et transitions possibles.

Selon le schéma 5, les individus peuvent expérimenter trois états : sain (à la suite de l'opération retirant la tumeur), malade (rechute) et mort. Ils peuvent passer de l'état sain à malade puis de malade à mort ou directement à la mort. Les données peuvent être censurées, un individu n'est pas toujours mort lors de sa dernière trace dans le suivi.

Dans le jeu nettoyé, tous les individus commencent le suivi en étant sain, dans l'état *0*. Des variables de transitions sont créées à partir de *PFI* et *DSS*. La transition de *0* à *1* s'effectue au temps de *PFI* quand celui-ci n'est pas égal à celui de *DSS*. La transition de *0* à *2* s'effectue au

temps de *DSS* quand l'individu n'a pas expérimenté de rechute, c'est-à-dire quand le temps du *PFI* est égal à celui du *DSS*. Le temps de la transition de *1* à *2* est pris à partir de l'arrivée à l'état *1* et est la différence du temps de *PFI* et *DSS* quand le patient observe une rechute. Quand le patient observe une transition de *0* à *1*, la transition de *0* à *2* est censurée au temps final et inversement, avec comme spécificité que la transition de *1* à *2* est censurée au temps *0* puisque l'individu n'est jamais arrivé à l'état *1*. La Figure 6 résume les trois transitions possibles :

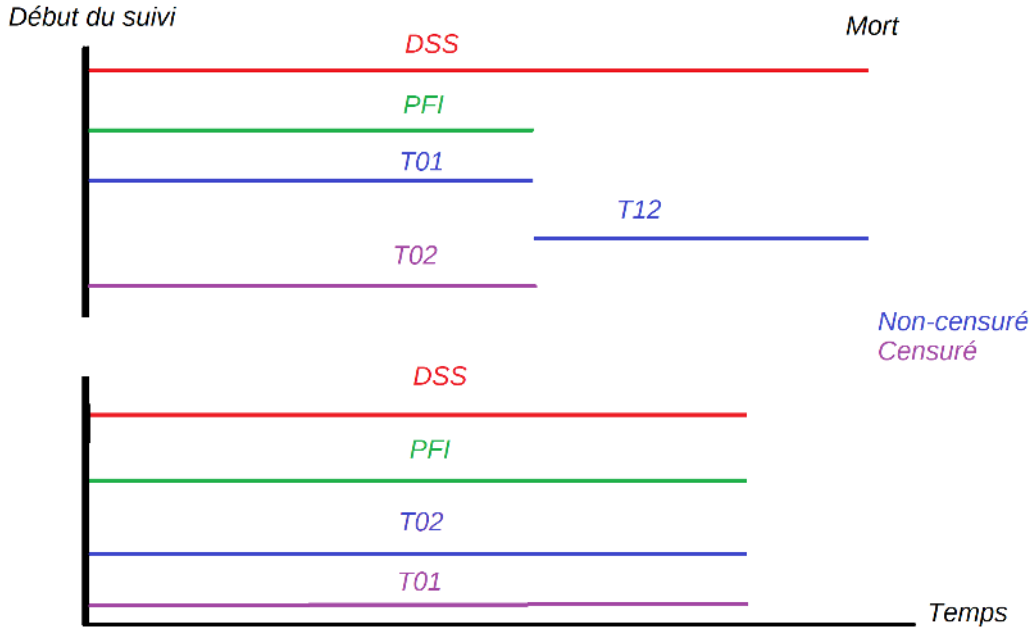


FIGURE 6 – Les deux scénarios principaux. Avec *T01* la transition de *0* à *1*, etc. Dans le deuxième cas, *T12* est censuré au temps 0 puisque l'individu n'atteint jamais l'état 1.

En réalité, il est possible qu'un individu passe de l'état malade à sain, mais ce cas n'est pas pris en compte dans le jeu de données.

Les modèles de survie à plusieurs états en compétition ne sont pas retenus dans notre analyse puisqu'ils ne peuvent pas dans leur formulation basique intégrer d'état intermédiaire. En un sens, un modèle de survie multi-état est plus raffiné, ce qui amène à une autre formulation. Ce passage ne se fait pas sans hypothèse, que nous verrons dans la partie suivante.

3 Les différents modèles

3.1 Modèle de survie

Les développements suivant sont extraits en partie du mémoire de master de Torunn Heggland de l'université d'Oslo [3].

Soit T une variable aléatoire représentant le temps auquel survient un événement, dans notre cas le décès d'un patient. On définit la fonction de survie représentant la probabilité pour un patient de survivre au moins jusqu'à un temps $t > 0$:

$$S(t) = \mathbb{P}(T > t) \tag{1}$$

La fonction de risque instantanée λ représente le risque instantané d'un patient de décéder sachant qu'il a survécu au moins jusqu'au temps t, elle est directement liée à la fonction de survie en utilisant la formule de Bayes :

$$\lambda(t) = \lim_{h\to 0} \frac{\mathbb{P}(t \leq T \leq t+h \mid T > t)}{h} = \lim_{h\to 0} \frac{\mathbb{P}(t \leq T \leq t+h)}{h} \frac{1}{S(t)} = \frac{f(t)}{S(t)} \tag{2}$$

où $f(t)$ est la densité de probabilité de T. Ceci nous mène directement à la relation suivante, en utilisant le fait que $f(t) = -S'(t)$, S' étant la dérivée de S :

$$\frac{S'(t)}{S(t)} = -\lambda(t)$$

Donnant ainsi en intégrant sur $[0, t]$:

$$S(t) = \exp\Big(-\int_0^t \lambda(u)du \Big) \tag{3}$$

La fonction de survie peut aussi être obtenue d'une autre façon, en utilisant des probabilités conditionnelles sur une discrétisation $0 = t_0 < \ldots < t_n = t$ du temps :

$$S(t_n) = \mathbb{P}(T > t_n, T > t_{n-1}) = \mathbb{P}(T > t_n \mid T > t_{n-1})\mathbb{P}(T > t_{n-1})$$

En répétant le processus pour $\mathbb{P}(T > t_{n-1})$ on obtient :

$$S(t_n) = \prod_{i=1}^{n} \mathbb{P}(T > t_i \mid T > t_{i-1}) = \prod_{i=1}^{n} \Big(1 - \mathbb{P}(T \leq t_i \mid T > t_{i-1})\Big)$$

En faisant tendre n vers l'infini on obtient via le produit intégrale (continue) :

$$S(t) = \prod_{u\in[0,t]} 1 - \lambda(u)du \tag{4}$$

En intégrant (3) à (4) on obtient l'équation :

$$\prod_{u\in[0,t]} \Big(1-\lambda(u)\Big)du = \exp(\ -\int_0^t\ \lambda(u)du\ ) \tag{5}$$

Avec des données collectées, la fonction de survie peut être estimée via l'estimateur de Kaplan-Meier, qui est une version empirique de (4). Sa définition avec un échantillon $(T_i)_{i\in\{1,\dots,n\}}$ est :

$$\widehat{S}(t) = \prod_{0<T_i\leq t} 1-\frac{d_i}{N_i} \tag{6}$$

Avec d_i le nombre d'individus observant l'événement au temps T_i et N_i le nombre d'individus encore à risque entre t_{i-1} et T_i.

Ce modèle peut être agrémenté de covariable dans un modèle de survie à risque proportionnel, ici modèle de Cox. Si l'on note $Z_1, Z_2, \dots, Z_n$, $n \in \mathbb{N}$ les covariables, la fonction de risque instantané se formule de la sorte :

$$\lambda(t, Z) = \lambda_0(t)\ \exp\Big(\sum_{k=1}^{n}\beta_k Z_k\Big) \tag{7}$$

avec λ_0 le risque instantané du modèle nul et les β_k des coefficients obtenus par régression avec comme critère la maximisation de la vraisemblance.

Une fois tout ceci défini dans le cas d'un seul type d'événement, la transition peut-être faite avec un cas où peuvent s'enchaîner plusieurs événements. Il n'est plus seulement question du décès d'un individu, mais de trajectoires comme tomber malade puis décéder ou bien directement décéder. Une des classes de modèles permettant cela est celle des modèles de survie multi-états.

3.2 Modèle de survie multi-états

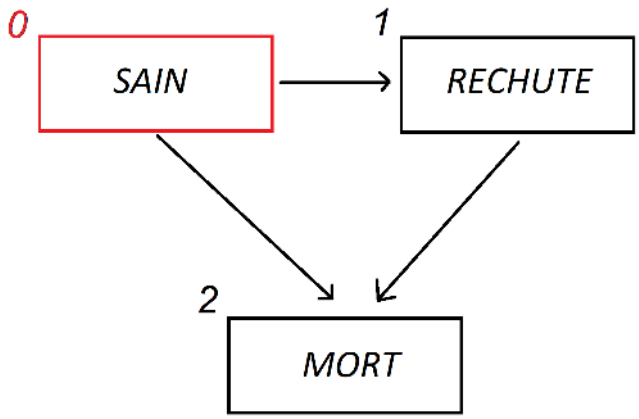
Dans le cadre des modèles multi-états, on utilise une suite de variables aléatoires $(X_t)_{t\in T}$ avec $T \subset \mathbb{R}^+$ et $X_t \in S$, $S = \{1, \dots, m\}$, $m \in \mathbb{N}$. Ainsi dans notre cas, à chaque temps t entre 0 et la fin du suivi, l'individu est dans un état 0, 1 ou 2. Les probabilités de passage d'un état à l'autre sont dépendantes du chemin suivi depuis le temps de départ :

$$\forall i, j \in S,\ 0 < s < t,\ p_{ij}(s,t) = \mathbb{P}(X(t) = j | X(s) = i, \mathcal{F}_{s-}) \tag{8}$$

L'équation se lit de la sorte : la probabilité pour un individu d'être au temps t dans l'état j sachant qu'il est dans l'état i au temps s. La filtration $\mathcal{F}_{s-}$ représente les états aux temps précédents s. Les intensités de transition inter-états sont définies à partir de (7) :

$$\alpha_{ij}(t) = \lim_{h\to+\infty} \frac{p_{ij}(t, t+h)}{h} \tag{9}$$

Compte tenu des restrictions de notre modèle, tout état ne peut-être atteint à partir de n'importe quel état. Plus précisément un individu peut suivre le schéma suivant :



La trajectoire possible d'un individu peut être, avec deux temps $s < t$, $X_0 = 0$, $X_s = 1$ et $X_t = 2$. Les intensités de transition forment ainsi une matrice $A(t)$:

$$A(t) = \begin{pmatrix} -(\alpha_{01}(t) + \alpha_{02}(t)) & \alpha_{01}(t) & \alpha_{02}(t) \\ 0 & -\alpha_{12}(t) & \alpha_{12}(t) \\ 0 & 0 & 0 \end{pmatrix}$$

Cette matrice se lit avec les états de départ en ligne et les états d'arrivée en colonnes, la dernière colonne est nulle puisque le décès ne mène à aucun autre état.

En l'état, ces équations ne donnent pas de cadre théorique satisfaisant, c'est pour cela qu'une hypothèse fréquente est faite ici, l'hypothèse markovienne :

$$p_{ij}(s,t) = \mathbb{P}(X(t) = j | X(s) = i), \ s < t \tag{10}$$

L'équation (10) simplifie la (8). Celle-ci statut que la probabilité qu'un individu soit dans un temps ne dépend que de son état au temps précédent, oubliant ainsi les états passés. Ce cadre théorique permet l'utilisation de l'équation différentielle de Kolmogorov permettant la détermination des probabilités de transition en fonction des intensités, en notant P la matrice de celles-ci :

$$\frac{\partial}{\partial t} P(s,t) = P(s,t) \ A(t) \tag{11}$$

Les probabilités nous intéressant étant p_{00}, p_{01}, p_{02}, p_{11} et p_{12} et sachant que $p_{02} = 1 - p_{00} - p_{01}$ et $p_{12} = 1 - p_{11}$, on se concentre sur les 3 premières probabilités :

$$\begin{aligned} \frac{\partial}{\partial t} p_{00}(s,t) &= -(\alpha_{01}(t) + \alpha_{02}(t)) \ p_{00}(s,t) \\ \frac{\partial}{\partial t} p_{01}(s,t) &= \alpha_{01}(t) p_{00} - \alpha_{12}(t) \ p_{01}(s,t) \\ \frac{\partial}{\partial t} p_{11}(s,t) &= -\alpha_{12}(t) \ p_{11}(s,t) \end{aligned}$$

En résolvant les équations différentielles (en annexe A1), on obtient les résultats suivant :

$$\begin{aligned} p_{00}(s,t) &= exp\left(- \int_s^t \alpha_{01}(u) + \alpha_{02}(u) \ du \right) \\ p_{01}(s,t) &= \int_s^t p_{00}(s,u) \alpha_{01}(u) p_{11}(u,t) \ du \\ p_{11}(s,t) &= exp\left(- \int_s^t \alpha_{12}(u) \ du \right) \end{aligned} \tag{12}$$

Les fonctions $p_{00}(s,t)$ et $p_{11}(s,t)$ sont similaires au cas de survie à deux états.

Pour ce qui est de l'estimation de ces fonctions, il faut utiliser l'équation de Chapman-Kolmogorov :

$$p_{ij}(s,t) = \sum_{k \in S} p_{ik}(s,u)\ p_{kj}(u,t) \tag{13}$$

avec $u \in [s,t]$. En étendant à toutes les transitions et en discrétisant l'intervalle temporel en une suite $(t_i)_{i \in \{1,\dots,n\}}$ telle que $t_0 = s$ et $t_n = t$, on obtient :

$$P(t_0 = s, t_n = t) = \prod_{i=0}^{n-1} P(t_i, t_{i+1})$$

De plus en utilisant l'équation (10) sur les intervalles en supposant que leur longueur tende vers 0 on obtient :

$$\begin{array}{lll} & P(t_i,t_{i+1}) - P(t_i,t_i) & \underset{n\to 0}{\sim} \quad P(t_i,t_i)A(t_{i+1})(t_{i+1}-t_i) \\ \iff & P(t_i,t_{i+1}) & \underset{n\to 0}{\sim} \quad I + A(t_{i+1})\ (t_{i+1}-t_i) \end{array}$$

En injectant ce résultat dans le précédent on obtient :

$$P(s,t) = \prod_{i=0}^{n-1} \Big(I + A(t_{i+1})\ (t_{i+1} - t_i)\Big)$$

en suivant le même schéma que pour l'équation (4) on obtient son équivalent multi-état :

$$P(s,t) = \prod_{u \in [s,t]} \Big(I + A(u)\Big)\ du \tag{14}$$

Pour ce qui est des estimations des fonctions de probabilités (11), les estimateurs sont les suivants :

$$\begin{array}{rcl} \widehat{p}_{00}(s,t) & = & \displaystyle\prod_{s<T_j\le t} \left(1 - \frac{d_0(T_j)}{N_0(T_j)}\right) \\ \widehat{p}_{01}(s,t) & = & \displaystyle\sum_{s<T_j\le t} \widehat{p}_{00}(s,T_{j-1})\ \frac{d_{01}(T_j)}{N_{01}(T_j)}\ \widehat{p}_{11}(T_j,t)\ du \\ \widehat{p}_{11}(s,t) & = & \displaystyle\prod_{s<T_j\le t} \left(1 - \frac{d_1(T_j)}{N_1(T_j)})\right) \end{array} \tag{15}$$

Avec $d_0(T_j)$ le nombre de départs de l'état 0 vers 1 ou 2 au temps T_j et $N_0(T_j)$ le nombre d'individus encore à l'état 0 juste avant T_j. La probabilité de transition de 0 à 0 est simplement un estimateur de Kaplan-Meier de l'événement quitter l'état 0. $d_{01}(T_j)$ et $N_{01}(T_j)$ sont les quantités analogues à d_0 et N_0 mais pour le passage de l'état 0 à 1, de même que $d_{12}(T_j)$ et $N_{12}(T_j)$ pour le passage de 1 à 2. Cependant, dans le cas où le temps de passage de l'état 1 à 2 est pris à partir de l'arrivée à l'état 1 et non depuis le début de l'étude, (14) change quelque peu. En notant Θ_i cette nouvelle variable aléatoire, les formules deviennent :

$$\begin{array}{rcl} \widehat{p}_{01}(s,t) & = & \displaystyle\sum_{s<T_j\le t} \widehat{p}_{00}(s,T_{j-1})\ \frac{d_{01}(T_j)}{N_{01}(T_j)}\ \widehat{p}_{11}(0,t-T_j)\ du \\ \widehat{p}_{11}(0,\theta) & = & \displaystyle\prod_{\Theta_j\le\theta} \left(1 - \frac{d_1(\Theta_j)}{N_1(\Theta_j)})\right) \end{array} \tag{16}$$

3.3 Ajout de covariables

Les modèles mutli-états peuvent intégrer des covariables au moyen d'une réécriture des intensités de transition en fonction de ces dernières. Si l'on note $Z_1, Z_2, \ldots, Z_n$, $n \in \mathbb{N}$ les covariables, les intensités de transition se formulent de la sorte :

$$\alpha^{ij}(t, Z) = \alpha_0^{ij}(t) \exp\Big(\sum_{k=1}^{n} \beta_k^{ij} Z_k\Big) \tag{17}$$

Ici $\alpha_0^{ij}(t)$ représente l'intensité de transition du modèle nul et les β_k^{ij} sont des coefficients obtenus par régression, correspondant à la transition de i à j et à la covariable Z_k. La formule est semblable à celle d'un modèle de Cox, la régression s'effectue aussi par maximum de vraisemblance.

Dans cette approche, il est donc question d'obtenir des coefficients par maximisation de la vraisemblance afin de principalement obtenir des taux de risque associés aux covariables, à l'instar d'un modèle de Cox. Cependant, un modèle multi-état ajoute la dépendance des processus représentant les trajectoires en terme d'état suivies par les patients. La comparaison d'un modèle de Cox simple sur les transitions et de son équivalent multi-état va être effectué dans la partie suivante afin de déterminer la valeur ajouté de ce dernier modèle.

3.4 Simulations

La simulation d'un jeu de données respectant les hypothèses de départ peut être faite aisément. L'algorithme suivant permet de générer des temps de passage des états 0 à 1, 0 à 2 et 1 à 2 :

- Définir des intensités de transition constantes que l'on notera α_{01}, α_{02} et α_{12}.

1. Tirer une réalisation T_1 selon une loi exponentielle de paramètre $\alpha_{01} + \alpha_{02}$, c'est-à-dire de loi le minimum de deux lois exponentielles de paramètre respectif α_{01} et α_{02}.

2. Tirer une réalisation B selon une loi de Bernouilli de paramètre $\dfrac{\alpha_{02}}{\alpha_{01} + \alpha_{02}}$.

3.1 Si $B = 0$, une transition de 0 à 1 s'effectue au temps T_1 et on passe au point suivant.

   (a) Tirer une réalisation T_2 de loi exponentielle de paramètre α_{12}, qui est le temps de transition de 1 à 2.

3.2 Si $B = 1$, une transition de 0 à 2 s'effectue au temps T_1.

4. Réitérer l'algorithme suivant le nombre d'individus souhaité.

Avec cet algorithme, tous les individus atteignent la mort. Cette version produit un processus markovien, homogène en temps puisque les coefficients de transitions ne varient pas dans le temps. Le résultat est présenté sous forme de tableau dans la figure 7, on peut y voir que tous les individus décèdent bien.

	DSS	DSS.time	PFI	PFI.time	T01	T01.time	T02	T02.time	T12	T12.time
1	1	1.7686310	1	0.6429939	1	0.6429939	0	1.7686310	1	1.125637139
2	1	5.3261861	1	0.9507761	1	0.9507761	0	5.3261861	1	4.375409940
3	1	22.2710860	1	6.3763481	1	6.3763481	0	22.2710860	1	15.894737901
4	1	0.2410896	1	0.2410896	0	0.2410896	1	0.2410896	0	0.000000000
5	1	1.8451463	1	1.8451463	0	1.8451463	1	1.8451463	0	0.000000000
6	1	6.0559799	1	2.7225932	1	2.7225932	0	6.0559799	1	3.333386732
7	1	0.3037968	1	0.3037968	0	0.3037968	1	0.3037968	0	0.000000000
8	1	4.7396674	1	4.7396674	0	4.7396674	1	4.7396674	0	0.000000000
9	1	4.0423726	1	0.3766025	1	0.3766025	0	4.0423726	1	3.665770103
10	1	2.8292136	1	2.3058748	1	2.3058748	0	2.8292136	1	0.523338805
11	1	3.6408125	1	0.1344485	1	0.1344485	0	3.6408125	1	3.506363972
12	1	9.5338752	1	4.4028904	1	4.4028904	0	9.5338752	1	5.130984876
13	1	0.5958772	1	0.5958772	0	0.5958772	1	0.5958772	0	0.000000000
14	1	1.8152519	1	0.4579470	1	0.4579470	0	1.8152519	1	1.357304812
15	1	1.7769958	1	0.6505991	1	0.6505991	0	1.7769958	1	1.126396696
16	1	1.6685728	1	0.2307638	1	0.2307638	0	1.6685728	1	1.437809011
17	1	3.7629590	1	3.6316472	1	3.6316472	0	3.7629590	1	0.131311848
18	1	14.0715857	1	3.9733483	1	3.9733483	0	14.0715857	1	10.098237494
19	1	6.7932889	1	4.8332937	1	4.8332937	0	6.7932889	1	1.959995135
20	1	0.9197312	1	0.9180213	1	0.9180213	0	0.9197312	1	0.001709821

FIGURE 7 – Jeu de données simulé avec les paramètres d'intensité de transition égaux à 0.3 pour les transitions de 0 à 1 et de 1 à 2 et 0.1 pour la transition de 0 à 2, le tout sans censure.

Une version censurée peut-être créée en la définissant comme un état en concurrence avec 1 et 2, c'est-à-dire en lui assignant une intensité de transition de sorte qu'elle suive une loi exponentielle, voir annexe A2. La figure 8 représente un tel jeu, où peut être vu que des individus ne décède pas avant la fin du suivi.

	DSS	DSS.time	PFI	PFI.time	T01	T01.time	T02	T02.time	T12	T12.time
1	1	0.4271754	1	0.4271754	0	0.4271754	1	0.4271754	0	0.0000000
2	0	1.7187908	1	0.9139363	1	0.9139363	0	1.7187908	0	0.8048545
3	1	1.3348168	1	1.3348168	0	1.3348168	1	1.3348168	0	0.0000000
4	1	0.8936851	1	0.4867607	1	0.4867607	0	0.8936851	1	0.4069244
5	0	0.4080925	0	0.4080925	0	0.4080925	0	0.4080925	0	0.0000000
6	0	3.4613812	1	0.9705966	1	0.9705966	0	3.4613812	0	2.4907846
7	0	0.7918394	0	0.7918394	0	0.7918394	0	0.7918394	0	0.0000000
8	0	0.6336928	0	0.6336928	0	0.6336928	0	0.6336928	0	0.0000000
9	0	0.5749536	0	0.5749536	0	0.5749536	0	0.5749536	0	0.0000000
10	1	1.0798564	1	1.0798564	0	1.0798564	1	1.0798564	0	0.0000000
11	0	0.7386786	0	0.7386786	0	0.7386786	0	0.7386786	0	0.0000000
12	1	0.4854833	1	0.1274652	1	0.1274652	0	0.4854833	1	0.3580182
13	1	1.2642170	1	1.2642170	0	1.2642170	1	1.2642170	0	0.0000000
14	0	0.4989707	0	0.4989707	0	0.4989707	0	0.4989707	0	0.0000000
15	0	0.3783810	0	0.3783810	0	0.3783810	0	0.3783810	0	0.0000000
16	1	0.5350322	1	0.2499030	1	0.2499030	0	0.5350322	1	0.2851292
17	0	0.4665753	0	0.4665753	0	0.4665753	0	0.4665753	0	0.0000000
18	0	0.8185155	1	0.2092639	1	0.2092639	0	0.8185155	0	0.6092515
19	0	1.2519255	0	1.2519255	0	1.2519255	0	1.2519255	0	0.0000000
20	0	1.0740638	0	1.0740638	0	1.0740638	0	1.0740638	0	0.0000000

FIGURE 8 – Jeu de données simulé avec les paramètres d'intensité de transition égaux à 0.3 pour les transitions de 0 à 1 et de 1 à 2 et 0.1 pour la transition de 0 à 2, le tout avec censure (d'intensité de transition égale à 0.9).

En pratique, les simulations nous permettent de vérifier le comportement des formules et packages de survie *R* utilisés. Nous allons utiliser des jeux de données simulées par la suite pour se faire une idée de la robustesse du modèle avec différents packages de *R* afin notamment d'obtenir des taux de risques au sein de modèles de Cox.

4 Packages R

Nous avons exploré plusieurs packages sous R pour modéliser nos données avec un modèle de survie multi-états et avoir des résultats reproductibles.

4.1 `msm`

Le package `msm` sous R est un package conçu pour créer les modèles markovien multi-états sur des données de survie. Son utilisation à partir d'un jeu de données se fait comme suit :

Pour chaque patient,

— identifier l'état de départ au temps 0 (en l'occurrence toujours 0) et créer une ligne du jeu de données contenant l'identifiant du patient, son état au temps donné,

— répéter l'opération pour les temps suivant,

— si l'individu a un suivi censuré, identifier le dernière état et insérer une ligne avec ce dernier au temps de censure.

Le résultat sur un jeu de données simulé est présenté dans la figure 9.

	indiv	state	time
1	1	1	0.00000000
2	1	2	1.88795458
3	1	2	1.92291841
4	2	1	0.00000000
5	2	1	0.68336978
6	3	1	0.00000000
7	3	2	0.36426682
8	3	2	1.36990732
9	4	1	0.00000000
10	4	2	0.34948815
11	4	2	1.11564939
12	5	1	0.00000000
13	5	1	0.22689039
14	6	1	0.00000000
15	6	1	0.23198679
16	7	1	0.00000000
17	7	1	2.53738149
18	8	1	0.00000000
19	8	1	0.01116425
20	9	1	0.00000000
21	9	1	0.07563412
22	10	1	0.00000000
23	10	1	0.12722097

FIGURE 9 – Jeu de données simulées, formaté pour son utilisation avec le package `msm`.

Ensuite, la fonction `msm` maximise le maximum de vraisemblance des intensités de transitions telle qu'elles sont définies dans l'équation 17, via une descente de gradient. La fonction nécessite au préalable de lui donner en entrée une matrice indiquant quelles transitions inter-états sont possibles. Enfin, elle permet d'ajouter des covariables au modèle.
Les résultats sont les intensités de transition avec intervalle de confiance à 95% pour chacune d'elles, et si le modèle comporte des covariables, les hazard ratios leur correspondant avec intervalle de confiance à 95%, en prenant la première valeur de la covariable comme référence.

Le package `msm` contient également une fonction pour faire du bootstrapping sur les intensités de transition (`boot.msm`). Cette fonction permet dans certains cas d'obtenir des intervalles

de confiance plus raisonnable que selon la méthode classique.

Le package `msm` n'est pas très simple d'utilisation, il nécessite des pré-traitements sur le jeu de données, chose qui est plus aisé avec le package `mstate` dont il est question dans la partie suivante.

4.2 `mstate`

Le package `mstate` est conçu pour faire tourner des modèles de survie à risques compétitifs et des modèles de survie multi-états. Cette partie vise à présenter de manière synthétique la procédure dans le cas des modèles multi-états seulement [4] :

— Établir la matrice de transition avec la fonction `transMat` : elle spécifie quelles transitions directes sont possibles et attribue des numéros ordonnés aux transitions pour référence future,

— Préparer les données en format long (i.e. avec des colonnes individu, état et temps) avec la fonction `msprep` : chaque ligne correspond à une transition pour laquelle un patient est à risque contrairement au données initiales qui ont une colonne par type d'événement,

— Ajout des covariables avec la fonction `expand.covs`,

— Estimer des effets de covariables : à l'aide de la régression de Cox sur le modèle du package `Survival`, classique en survie.

Le jeu de données obtenue à la suite de la fonction `msprep` est représenté dans la figure 10

```
Data:
   id from to trans Tstart      Tstop       time status  stage stage.1 stage.2 stage.3
1   1    1  2     1      0 10.9650924 10.9650924      0   I-II       0       0       0
2   1    1  3     2      0 10.9650924 10.9650924      0   I-II       0       0       0
3   2    1  2     1      0  4.0355921  4.0355921      0   I-II       0       0       0
4   2    1  3     2      0  4.0355921  4.0355921      0   I-II       0       0       0
5   3    1  2     1      0  3.9644079  3.9644079      0   I-II       0       0       0
6   3    1  3     2      0  3.9644079  3.9644079      0   I-II       0       0       0
7   4    1  2     1      0  0.9527721  0.9527721      0 III-IV       1       0       0
8   4    1  3     2      0  0.9527721  0.9527721      0 III-IV       0       1       0
9   5    1  2     1      0  4.0438056  4.0438056      0   I-II       0       0       0
10  5    1  3     2      0  4.0438056  4.0438056      0   I-II       0       0       0
11  6    1  2     1      0  4.0273785  4.0273785      0 III-IV       1       0       0
12  6    1  3     2      0  4.0273785  4.0273785      0 III-IV       0       1       0
13  7    1  2     1      0  0.8295688  0.8295688      0   I-II       0       0       0
14  7    1  3     2      0  0.8295688  0.8295688      0   I-II       0       0       0
15  8    1  2     1      0  0.7091034  0.7091034      0   I-II       0       0       0
16  8    1  3     2      0  0.7091034  0.7091034      0   I-II       0       0       0
17  9    1  2     1      0  1.1964408  1.1964408      0   I-II       0       0       0
18  9    1  3     2      0  1.1964408  1.1964408      0   I-II       0       0       0
19 10    1  2     1      0  3.6167009  3.6167009      0   I-II       0       0       0
20 10    1  3     2      0  3.6167009  3.6167009      0   I-II       0       0       0
```

FIGURE 10 – Jeu de données sur le cancer du sein, formaté avec la fonction `msprep`.

On remarque la présence de colonne indiquant le numéro de la transition, spécifié dans la matrice crée avec la fonction `transMat`. Une colonne statut indique si la transition est censurée ou non (si elle a lieu). Les *stage.1*, *stage.2* et *stage.3* indiquent pour chaque transition quel est l'instance de la covariable, ici binaire.

Ce package permet ensuite via la fonction `coxph` de créer un modèle mutli-état avec des

covariable. Un exemple de syntaxe est le suivant, ressemblant fortement (même commande) avec un modèle de Cox simple du package `survival` :

```
1         coxph(Surv(Tstart, Tstop, status) ~ strata(trans) +
2         stage.1 + stage.3 + stage.3, data = sortie_msprep, method
              = "breslow")
```

Les résultats des hazard ratios avec intervalles de confiance. Ce package est plus simple d'utilisation dû aux fonction faisant elles-mêmes le pré-traitement.

5 Application sur les données

5.1 Estimation d'un modèle multi-état sur les données

Un équivalent des courbes de survie dans un modèle multi-état peut être les probabilités de transitions. Celles-ci s'estiment via les estimateurs définis dans la section 3. Le graphe de la figure 11 est tracé à partir d'un jeu de données simulé avec des paramètres d'intensité de transition égaux tous à 0.3 et sans censure.

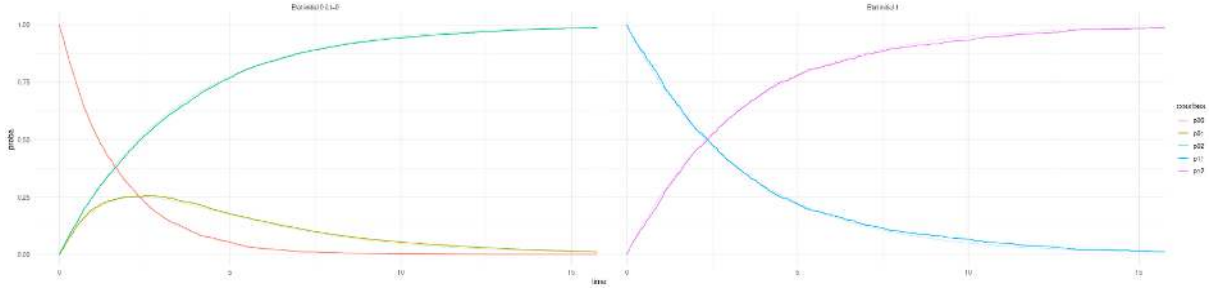


FIGURE 11 – Probabilités de transition depuis l'état 0 à gauche et 1 à droite au temps 0 (en années) pour un jeu de données simulées.

La courbe verte donne la probabilité d'être dans l'état 2 au temps t sachant que l'individu était à l'état 0 au temps 0. Ce graphe donne l'allure général des probabilités de transitions, la transition de 0 à 0 ressemble à une courbe de survie classique, celle de 0 à 2 à une courbe de survie inversée tandis que celle de la transition 0 à 1 est en cloche dans les premières années. Pour les transition à partir de l'état 1, la courbe à destination de de l'état 1 ressemble à une courbe de survie quand celle à destination de l'état 2 ressemble à l'inverse. Dans les cas avec censure, les courbes ont des plateaux et des écarts verticaux dus à la censure, comme il peut être vu dans les figures 12 et 13.

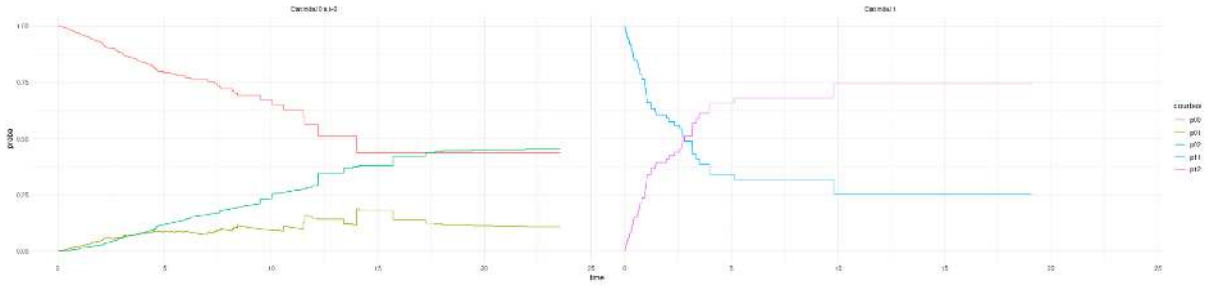


FIGURE 12 – Probabilités de transition depuis l'état 0 à gauche et 1 à droite au temps 0 (en années) pour un jeu BRCA.

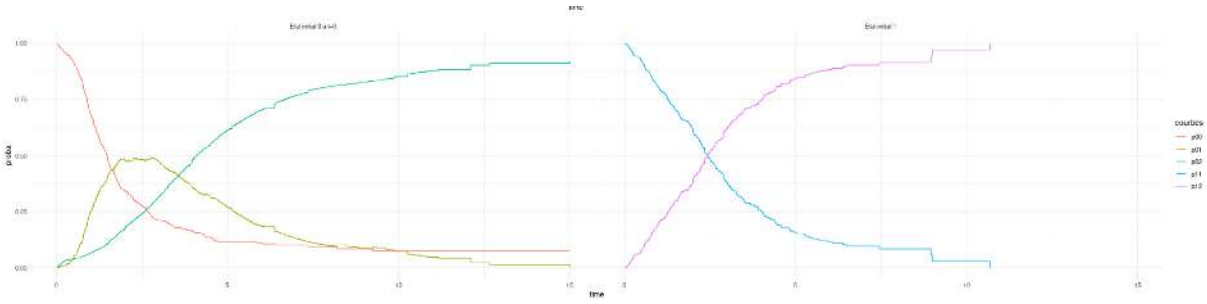


FIGURE 13 – Probabilités de transition depuis l'état 0 à gauche et 1 à droite au temps 0 (en années) pour le jeu OV.

Le cancer OV ressemble plus au comportement du jeu simulé que le BRCA, c'est parce qu'il à beaucoup moins de censure. Par ailleurs il peut être clairement vu qu'il expérimente dans les 3 premières années majoritairement un stagnation dans l'état sain ou une rechute, à l'horizon 10 ans la mort direct est plus probable que les deux autres scénarios. En ce qui concerne le BRCA, il expérimente beaucoup moins de mort, stagne à 50% d'individus encore sain à l'horizon 20 ans partagé avec la mort directe, et peu de probabilité de rechute. En somme le BRCA est beaucoup moins agressif que l'OV. De manière à intégrer des covariables, nous allons utiliser par la suite des packages dédiés sous *R* aux modèles de survie multi-états.

5.2 Comparaison des méthodes

L'objectif sur les données est de comparer une approche basique qui est d'appliquer des modèles de Cox, avec l'approche de modèle de survie multi-état. Pour ce faire nous allons prendre le cancer du sein (BRCA) et le cancer des ovaires (OV), sur lesquelles nous appliquerons des modèles de Cox sur les variables *DSS* et *PFI* avec comme covariable le stade de la tumeur. La mise en oeuvre sous *R* se fait via le package *Survival* avec la fonction *coxph*. Les données d'entrée doivent être décomposées en deux vecteurs, un avec les temps et un autre indiquant la censure avec un 0 et la non censure avec un 1.

Puis nous utiliserons les deux packages `msm` et `mstate` pour construire des modèles multi-états avec la même covariable, puis comparerons les hazard ratios obtenus.

Nous enchaînerons avec des jeux de données simulées avec des intensités de transitions similaires pour comparer les résultats. Le jeu de données simulées est composé de deux simulations de taille les effectifs des individus dans le stade $I-II$ et des individus dans le stade $III-IV$, le jeu final étant la concaténation de ces deux jeux la. Les intensités de transitions s'estiment via la fonction `crudeinits.msm` du package `msm`. L'intensité de censure s'estime via le maximum de vraisemblance d'une loi exponentielle, le temps de la variable *PFI* simulée étant le minimum du temps de censure, de passage de *0* à *1* et de *0* à *2*, soit une loi exponentielle de paramètre la somme des intensités (voir Annexe A.2).

5.2.1 Résultats

Nous nous intéressons particulièrement aux résultats de hazard ratios, de voir si nous trouvons des différences entre chaque changement d'étape. Avec un hazard ratio égale à *1*, il n'y a pas d'effet et les changements d'étapes se réalisent avec le même risque. Si le hazard ratio est significativement au dessus de *1*, le risque de diminution du temps de survie est augmenté, si le HR est significativement en dessous de *1*, le risque du diminution du temps de survie est diminué. Ce qui est biologiquement attendu est qu'un stade de cancer avancé (stade $III-IV$)

devrait augmenter le risque de diminution de la survie.

Un premier résultat numérique visible dans la table 1 montre des HR significativement supérieurs à 1 pour les deux cancers (BRCA et OV) et les deux évènements *DSS* et *PFI*.

	BRCA				OV			
	HR	lower	upper	p-val	HR	lower	upper	p-val
PFI	3.06	2.17	4.32	0	2.32	1.45	3.73	0
DSS	3.99	2.57	6.21	0	2.22	1.18	4.17	0.01

Table 1 – HR et IC des Cox marginaux

D'autre part nous pouvons observer la répartition des censures sur la table 2. Les censures dans BRCA sont assez déséquilibrées puisque c'est un cancer peu agressif, peu d'évènements sont observés donc il y a beaucoup de censure. Pour le cancer *OV*, qui est un cancer agressif, les effectifs de censure sont plutôt équilibrés.

		BRCA		OV	
		réelles	simus	réelles	simus
PFI	Censurées	909	920	140	148
	Non Censurées	132	121	401	393
DSS	Censurées	962	942	241	273
	Non Censurées	67	98	300	268
T01	Censurées	943	956	192	195
	Non Censurées	98	85	349	346
T02	Censurées	1008	1005	490	494
	Non Censurées	33	36	51	47
T12	Censurées	995	1010	292	320
	Non Censurées	46	31	249	221

Table 2 – Effectifs des censures

Le jeu simulé pour *OV* respecte les proportions des données observées tandis que pour *BRCA*, la transition de *1* à *2* observe un effectif moindre, en-dessous de celui de la transition de 0 à 2 (il ne semble pas possible d'inverser cette tendance après plusieurs simulations). Ceci est un signe que le jeu de données observé pourrait ne pas respecter totalement les hypothèses du modèle, potentiellement une censure pas totalement aléatoire.

Sur les graphiques suivants (figure 14 et figure 15), les positions des HR et de leurs intervalles de confiance peuvent être visuellement appréciées.

Nous retrouvons les HR des modèles de Cox marginaux qui vont nous servir de références pour éprouver les modèles multi états.

Pour le cancer BRCA (figure 14), tous les HR sont significativement supérieurs à 1 pour les modèles multi-états à l'exception de la transition de 0 à 2 du package `msm` où le modèle n'a pas convergé. Si nous regardons la partie simulée, les résultats sont très similaires mais les HR pour la transitions de 0 à 1 deviennent non significatifs (ils étaient cependant à la limite pour les

données réelles) pour les deux packages et nous retrouvons une forte instabilité de la transition de 0 à 2 pour le package `msm`.

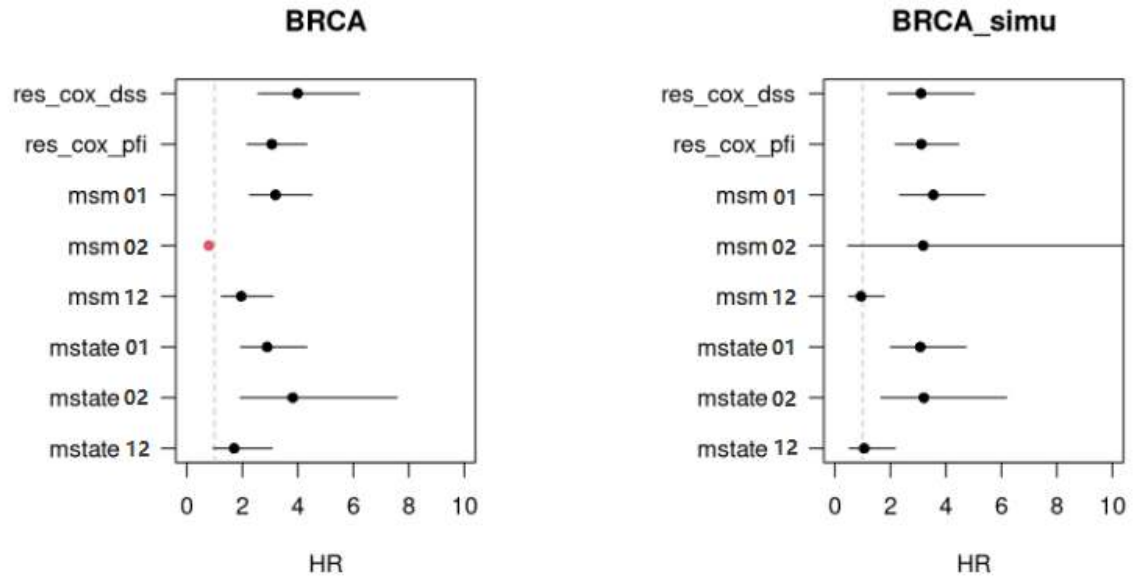


FIGURE 14 – Résultats

Pour le cancer OV (figure 15), tous les HR sont significativement supérieurs à 1 pour les transition de 0 à 1 des modèles, pour les deux packages, et ils ne sont pas significatifs pour le reste. Si nous regardons la partie simulée les résultats sont très similaires. Cette dernière figure montre que les simulations se comportent de manière similaire aux données.

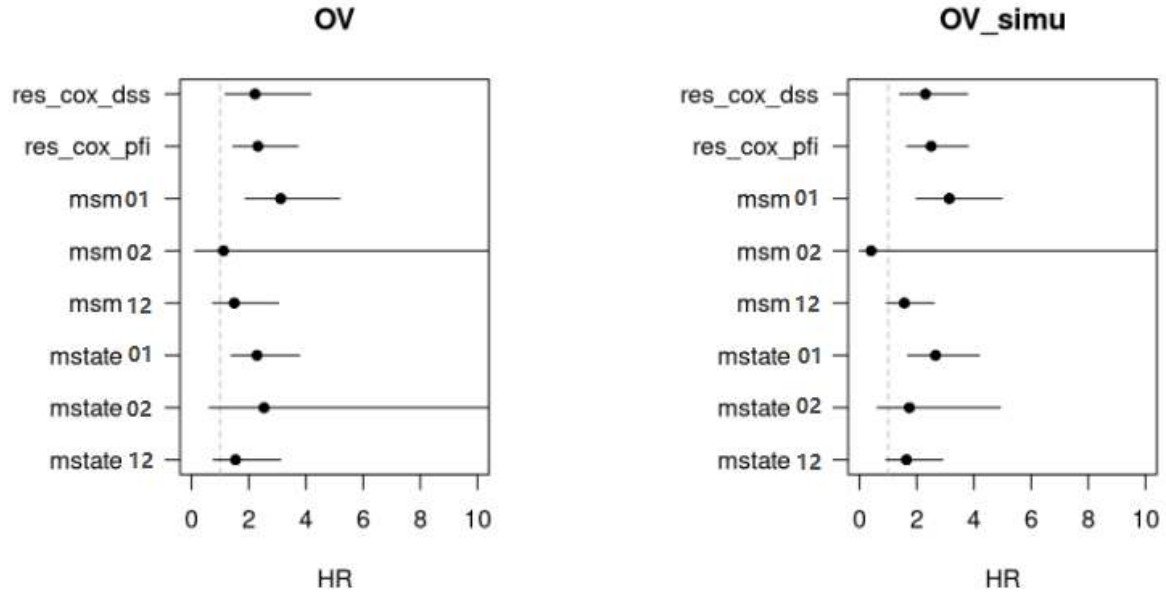


FIGURE 15 – Résultats

Les mêmes graphes pour tous les cancers sont disponibles en annexe A.3. Les données simulées donnent des résultats très similaires aux données observées, nous encourageant à valider les hypothèses du modèle pour ces dernières, avec néanmoins les réserves énoncées plus haut pour le cancer *BRCA*.

5.2.2 Travail en cours

Un premier travail sur la variation des paramètres dans les données simulées nous permet de comprendre la décomposition des valeurs de HR dans le modèle multi-états. Dans la figure 16, les valeurs des paramètres sont dans l'ordre suivant : l'effectif du groupe de stade de cancer $I - II$, les intensités de transition de la transition de l'état 0 à 1, 0 à 2 et 1 à 2 du groupe de stade de cancer $I - II$, l'effectif du groupe de stade de cancer $III - IV$, les intensités de transition de la transition de l'état 0 à 1, 0 à 2 et 1 à 2 du groupe de stade de cancer $III - IV$, le paramètre de censure.

Dans le premier graphique seule l'intensité de transition de la transition de 0 à 1 varie (de 0.02 à 0.06). Les HR des modèles de Cox pour les évènements *DSS* et *PFI* sont significativement supérieurs à 1 alors que si l'on regarde les HR des différentes transitions de la modélisation multi état seuls les HR pour la transition de 0 à 1 sont significativement supérieurs à 1. De la même manière pour le second graphique ou seule l'intensité de transition de la transition de l'état 0 à 2 varie, on voit que seul l'HR de la transition 0 à 2 est significativement supérieur à 1.

Ce résultat, même s'il parait intuitif, permet de bien comprendre les différences de résultats obtenus, les HR en particulier, pour un simple modèle de Cox ou une modélisation de multi état.

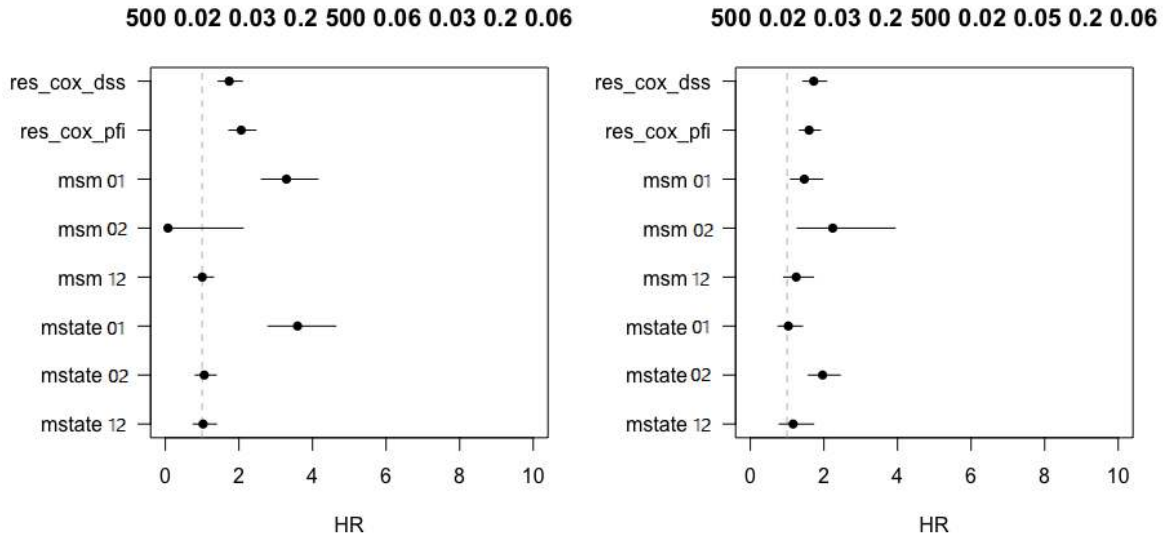


FIGURE 16 – Effet de la variation des intensités de transition pour les stade de cancer pour transition 0 à 1 et 0 à 2 sur la significativité des HR

6 Conclusion

Les modèles de survie multi-états donnent plus d'information que les modèles de Cox en cela qu'ils peuvent détailler les transitions. Cependant ils supposent des hypothèse plus fortes comme le fait que les intensités de transitions ne dépendent pas du temps (ou bien il faut trouver un moyen d'évacuer la dépendance dans une covariable) et l'aspect markovien du processus.

6.1 Les difficultés rencontrées

Plusieurs difficultés des côtés programmation et modélisations se sont présentés à nous :

— La création de l'algorithme de simulations de données avec censures, qui a nécessité plusieurs aménagements.
— Les données comportaient des écarts à la norme indiquée dans la notice de construction de celles-ci, il a fallu faire des choix sur quoi garder et quoi retirer.

— Il a fallu mettre en place des tests pour les morceaux de codes implémentés de sorte à en garantir l'intégrité du résultat et traquer les potentiels erreurs de code quand ils étaient modifiés sur le github.

— L'évaluation la robustesse des hypothèses du modèle sur les données, notamment en comparant avec des données simulées.

— L'utilisation du package `msm` non intuitive pouvant mener à des erreurs conceptuelles.

— La visualisation et la comparaison des courbes de survie non aisée avec le nombre de cancers différents, la création d'une application `shiny` dédiée à cela nous a bien aidée.

6.2 Les améliorations possibles

La piste d'amélioration qui irait dans le sens de la compréhension des hazard ratios dans la modélisation des modèles mutli-états serait de réaliser une analyse de sensibilité sur les données simulées (comment entrevue dans la section travail en cours). L'idée serait de rechercher les paramètres d'intensité de transition et de censure qui fourniraient des configurations types, de sorte à avoir une cartographie des différents phénomènes. Par exemple, avoir dans un modèle de Cox des HR pour une covariable (pour DSS et PFI) qui ne serait pas significatif et lorsque que l'on passerait à une modélisation multi-états, les HR des transitions seraient pour certains significatifs. Cette analyse consisterait donc à faire varier une grille de paramètres et de répliquer l'opération avec des graines de génération aléatoire différentes. Une autre possibilité peut être de générer les populations des stade $I - II$ et $III - IV$ selon des processus différents.

6.3 Bilan et remerciements

Ce projet nous a permis de bien nous immerger dans les difficultés de modélisation de la survie. La partie théorique n'a pas été facile à appréhender pour nous tous, cependant l'application avec les packages nous a permis de mieux nous approprier la modélisation. Nous remercions Adeline pour nous avoir soutenu et aiguillé grâce à son expérience dans ce type de modèles. Nous remercions Florent pour le challenge qu'il nous a proposé et pour son aide dans la compréhension du contexte clinique et pour nous avoir assisté dans la conceptualisation de notre code autour de ce projet.

Références

[1] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, et al. An integrated TCGA Pan-Cancer clinical data resource to drive High-Quality survival outcome analytics. *Cell*, 173(2) :400–416.e11, April 2018.

[2] https ://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[3] Torunn Heggland. Estimating transition probabilities for the illness-death model. *Master's Thesis for the degree Modelling and Data Analysis at the Faculty of Mathematics and Natural Sciences in the University of Oslo*, November 2015.

[4] Hein Putter. Tutorial in biostatistics : Competing risks and multi-state models analyses using the mstate package. November 2021.

A Annexes

A.1 Obtention des équations de (11)

Les équations à résoudre sont :

$$\begin{aligned}
\frac{\partial}{\partial t}p_{00}(s,t) &= -\big(\alpha_{01}(t)+\alpha_{02}(t)\big)\ p_{00}(s,t)\\
\frac{\partial}{\partial t}p_{01}(s,t) &= \alpha_{01}(t)p_{00}(s,t)-\alpha_{12}(t)\ p_{01}(s,t)\\
\frac{\partial}{\partial t}p_{11}(s,t) &= -\alpha_{12}(t)\ p_{11}(s,t)
\end{aligned}$$

On commence par résoudre la première équation qui est une équation différentielle du linéaire première ordre à coefficients non constants sans second membre, le schéma de résolution est le suivant en supposant que $\forall t \in [t,+\infty[,\ p_{00}(s,t) \neq 0$ et $p_{00}(s,s)=1$:

$$\begin{aligned}
& & \frac{\partial}{\partial t}p_{00}(s,t) &= -\big(\alpha_{01}(t)+\alpha_{02}(t)\big)\ p_{00}(s,t)\\
\Longrightarrow & & \frac{\partial}{\partial t}p_{00}(s,t)\ \frac{1}{p_{00}(s,t)} &= -\big(\alpha_{01}(t)+\alpha_{02}(t)\big)\\
\Longrightarrow & \int_s^t \frac{\partial}{\partial u}p_{00}(s,u)\ \frac{1}{p_{00}(s,u)}\ du &= -\int_s^t \Big(\alpha_{01}(u)+\alpha_{02}(u)\Big)\ du\\
\Longrightarrow & ln(p_{00}(s,t))-ln(p_{00}(s,s)) &= -\int_s^t \Big(\alpha_{01}(u)+\alpha_{02}(u)\Big)\ du\\
\Longrightarrow & & p_{00}(s,t) &= exp\Big(-\int_s^t \Big(\alpha_{01}(u)+\alpha_{02}(u)\Big)\ du\Big)
\end{aligned}$$

De plus en utilisant la même méthode la formule analytique de p_{11} peut être obtenue. La dernière équation à résoudre est donc celle de p_{01} qui est une équation différentielle du linéaire première ordre à coefficients non constants avec second membre, le schéma de résolution est la résolution de l'équation homogène avec $p_{01}(s,s)=0$ puis l'utilisation de la méthode de variation de la constante pour trouver une solution particulière. Pour l'équation homogène :

$$\frac{\partial}{\partial t}p_{00}(s,t)+\alpha_{12}(t)\ p_{01}(s,t)=0$$

La fonction $p_{01}(s,t)=0$ est solution et vérifie bien la condition initiale, elle est donc la seule (théorème de Cauchy-Lipschitz). Pour trouver une solution particulière, on suppose que $\forall t \in [t,+\infty[,\ p_{01}(s,t) \neq 0$:

$$
\begin{aligned}
\frac{\partial}{\partial t} p_{01}(s,t) &= -\alpha_{12}(t)\ p_{01}(s,t) \\
\Longrightarrow \quad \int_s^t \frac{\partial}{\partial u} p_{01}(s,u)\ \frac{1}{p_{01}(s,u)}\ du &= -\int_s^t \alpha_{12}(u)\ du \\
\Longrightarrow \quad ln(p_{01}(s,t)) - A &= -\int_s^t \alpha_{12}(u)\ du \\
\Longrightarrow \quad p_{01}(s,t) &= C\ exp\Big(-\int_s^t \alpha_{12}(u)\ du\Big)
\end{aligned}
$$

Avec $A, C \in \mathbb{R}$. La méthode de la variation de la constante consiste à poser la constante C comme fonction de t, remplacer dans l'équation avec second membre p_{01} par l'expression obtenue. Ainsi on obtient :

$$
\begin{aligned}
\Big(C'(t) + \alpha_{12}C(t) - \alpha_{12}C(t)\Big)\ exp\Big(-\int_s^t \alpha_{12}(u)\ du\Big) &= \alpha_{01}(t)p_{00}(s,t) \\
\Longrightarrow \quad C'(t) &= \alpha_{01}(t)\ p_{00}(s,t)\ exp\Big(\int_s^t \alpha_{12}(u)\ du\Big) \\
\Longrightarrow \quad C(x) &= \int_s^t \alpha_{01}(x)p_{00}(s,x)exp\Big(-\int_x^s \alpha_{12}(u)du\Big)du
\end{aligned}
$$

Ainsi $\int_s^t \alpha_{01}(x)\ p_{00}(s,x)\ exp\Big(-\int_x^s \alpha_{12}(u)du\Big)du\ exp\Big(-\int_s^t \alpha_{12}(u)\ du\Big) = \int_s^t \alpha_{01}(x)\ p_{00}(s,x)\ exp\Big(-\int_x^t \alpha_{12}(u)du\Big)\ du$ est une solution particulière, et puisque la solution homogène est la fonction constante nulle, c'est aussi la solution générale. Pour finir les solution obtenues sont :

$$
\begin{aligned}
p_{00}(s,t) &= exp\left(-\int_s^t\ \alpha_{01}(u) + \alpha_{02}(u)\ du\right) \\
p_{01}(s,t) &= \int_s^t\ p_{00}(s,u)\alpha_{01}(u)p_{11}(u,t)\ du \\
p_{11}(s,t) &= exp\left(-\int_s^t\ \alpha_{12}(u)\ du\right)
\end{aligned}
$$

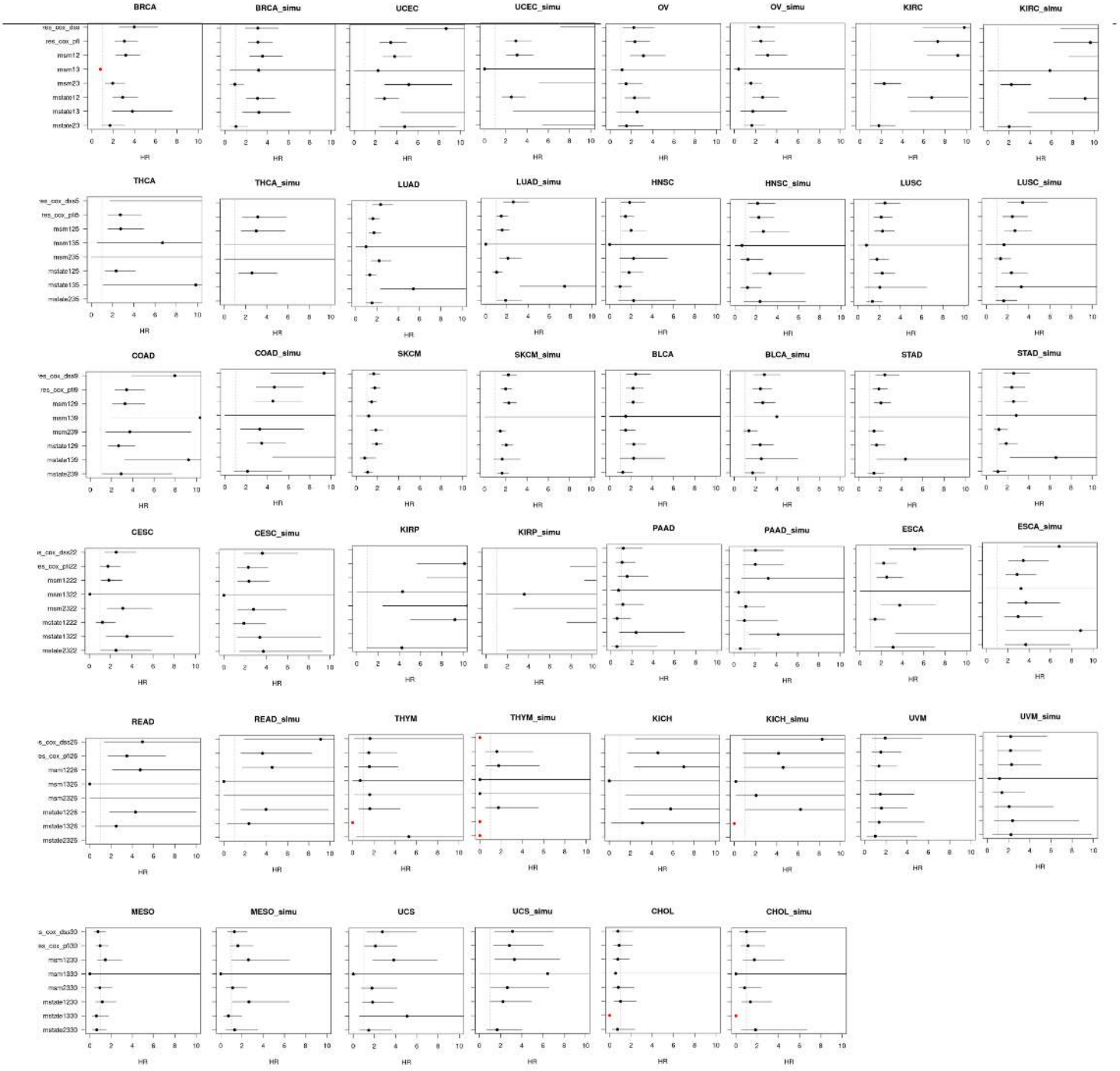
A.2 Simulation d'un jeu de données censuré

Le code ci-dessous est rédigé en R de manière vectorielle, c'est-à-dire que les temps et les censures sont générés pour n individus puis les transitions sont décidées par après. Les données de censure sous R sont constituées d'un vecteur de temps et d'un vecteur indiquant l'éventuel censure (0 si censuré et 1 sinon).

```r
process <- function(n, i01, i02, i12, i00=NULL, seed=1){
  set.seed(seed) #fixation de la graine aléatoire
  tcensure <- rexp(n,i00)     #temps de censure pour chaque individu
  tmin0102 <- rexp(n,i01+i02) #minimum du temps de passage de 0 à 1 et de 0 à 2
  t1 <- apply(matrix(c(tcensure,tmin0102),byrow = FALSE,ncol=2),1,min) #minimum du temps de passage de 0 à 1 et de 0 à 2 et du temps de censure
  s1 <- ifelse(tmin0102 < tcensure,rbinom(n, 1, i02/(i01+i02))+1,0) #détermination de l'état atteint (1 ou 2 ou censure à 0)
  tmin12 <- rexp(n,i12)
  t2 <- apply(matrix(c(tmin12,tcensure-t1),byrow = FALSE,ncol=2),1,min) #minimum temps de passage de 1 à 2 et du temps de censure
  s2 <- ifelse(tmin12 < tcensure,1,0) #détermination de la censure de la transition 1 -> 2
  df <- data.frame("DSS"=ifelse(s1 == 2 | (s1 == 1 & s2 == 1),1,0),"DSS.time"=ifelse(s1==1,t1+t2,t1),  #temps du décès simple
                   "PFI"=ifelse(s1 > 0,1,0),"PFI.time"=t1,                                            #temps de la rechute ou de mort
                   "T01"=ifelse(s1==1,1,0),"T01.time"=t1,                                             #temps de rechute
                   "T02"=ifelse(s1-1 > 0,1,0),"T02.time"=ifelse(s1==2 | s1==0,t1,t1+t2),              #temps de mort sans rechute
                   "T12"=ifelse(s1==1 & s2==1,1,0),"T12.time"=ifelse(s1==1,t2,0))                     #temps de mort avec rechute
  return(df)
}
```

FIGURE 17 – Code en langage R permettant de simuler un jeu de données censuré.

A.3 Résultats sur tous les cancers



Hazard Ratio résultats pour chaque cancer et simulation