ANNÉE UNIVERSITAIRE 2022 – 2023

MASTER 2 ÉCONOMIE APPLIQUÉE / PARCOURS INGÉNIERIE DES DONNÉES ET ÉVALUATIONS ÉCONOMÉTRIQUES

ECONOMÉTRIE DES MODÈLES DE DURÉE

UNE ANALYSE SUR LES DONNÉES DE SURVIE DE 137 PATIENTS ATTEINTS D'UN CANCER DU POUMON

LAMON Océane - PEDROT Emma - SEZESTRE Emilien

université angers

Table des matières

1 – Intérêt de l'étude 2
2 – Statistiques descriptives 2
3 – Estimations non-paramétriques. 7
4 – Résultats – estimations non-paramétriques 15
5 – Test de vérification de l'hypothèse de risques proportionnels constants et de lien fonctionnel 16
6 – Estimations semi-paramétriques (stepwise). 22
7 – Résultats – estimations semi-paramétriques 25
8 – Prise en compte du type de cellule cancéreuse 26
9 – Modèle paramétrique 26
10 – Résultats – estimations paramétriques 30
11 – Bibliographie 31

1 – Quel(s) intérêt(s) des données de survie dans le phénomène à l'étude ?

Nous étudions des individus atteints du cancer du poumon. Les données de survie permettent de modéliser pour chaque patient le temps passé jusqu'à la survenue de l'événement (i.e le décès dans notre cas, mais l'on pourrait aussi bien étudier le temps passé jusqu'à une rechute) ou la sortie de l'étude (censure), tout en recherchant quelles variables ont un impact significatif positif ou négatif sur le temps de survie : inhérentes aux caractéristiques de l'individu, à savoir l'âge, l'impact de la maladie sur leur mode de vie actuel ou encore le type de traitement, ou aux caractéristiques de la maladie en elle-même, autrement dit le type de cellule cancéreuse; petite, large, adeno ou squameux.

L'utilisation des données de survie est d'autant plus primordiale que les patients ont des trajectoires de traitement très différentes (certains ayant déjà reçu d'autres traitements en amont par exemple) et peuvent décéder à des moments très différents.

De ce fait, ces analyses peuvent aiguiller les professionnels de la santé dans une meilleure compréhension des facteurs qui peuvent influencer la durée de vie des patients atteints d'un cancer du poumon, voire de déterminer les traitements qui sont les plus efficaces. L'efficacité du traitement pouvant fortement varier en fonction de l'étendue de la maladie et de la réaction individuelle au traitement, il est donc d'autant plus important de prendre en considération toutes les caractéristiques des patients et de la maladie, et de mener des comparaisons entre groupes de patients dans la recherche du traitement le plus efficace.

2 – Décrire les individus à l'étude à l'aide des variables AGE, DD, et PERF

Statistiques simples

Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum	Libellé
DD	137	8.77372	10.61214	1202	1.00000	87.00000	DD
AGE	137	58.30657	10.54163	7988	34.00000	81.00000	AGE
PERF	137	58.56934	20.03959	8024	10.00000	99.00000	PERF

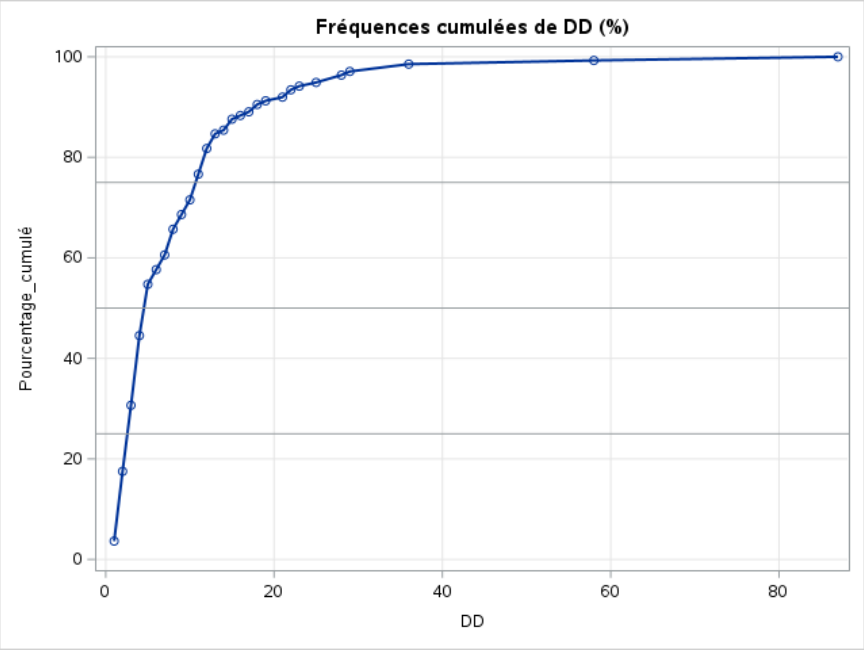
Les individus à l'étude sont au nombre de 137. Ils ont entre 34 et 81 ans, avec une moyenne d'âge de 58 ans. Parmi eux, plus de 35% ont entre 55 et 65 ans et 32,12% ont plus de 65 ans, quand seuls 15,3% (21 individus) ont moins de 45 ans.

Classe_age

Classe_age	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
45-55 ans	23	16.79	23	16.79
55-65 ans	49	35.77	72	52.55
Moins de 45 ans	21	15.33	93	67.88
Plus de 65 ans	44	32.12	137	100.00

En complément à l'âge, les individus peuvent être caractérisés en fonction d'autres variables : DD et PERF.

La variable DD correspond à la durée de la maladie en nombre de mois observée pour chaque individu durant l'étude, jusqu'au décès ou jusqu'à la censure en fonction du statut de l'individu. En moyenne, la durée de la maladie est de 8,8 mois, avec des observations allant d'un mois minimum à 87 mois maximum (soit plus de 7 ans). Plus spécifiquement, 30,66% des individus ont une durée de maladie de trois mois ou moins, 54,7% de 5 mois ou moins, 76,6% de 11 mois ou moins, et enfin 90,5% de 18 mois ou moins. Seuls trois individus sur les 137 ont une durée de maladie supérieure à 29 mois.



La variable PERF est un score en pourcentage pouvant aller de 0% (mort) à 100% (pleine santé). Cette variable est construite à partir de l'indice de Karnofsky, échelle permettant d'évaluer l'autonomie et la dépendance d'un patient à travers différents états, notamment utilisée pour évaluer ses capacités à survivre à une chimiothérapie contre le cancer. Ainsi, plus le score est élevé, et moins les patients sont fragilisés et impactés dans leur quotidien du fait de la maladie : en présence de cette dernière, certains sujets peuvent être incapables de travailler voire totalement invalides et requérir une assistance constante pour subvenir à leurs besoins, tandis que d'autres n'en ressentiront aucun symptôme et ne seront donc pas affectés dans leurs activités quotidiennes.

Le tableau ci-dessous présente les différentes états intermédiaires de l'échelle de Karnofsky, divisés en trois groupes principaux (0-40% pour les individus incapables de subvenir à leurs besoins, 50-70% pour les individus incapables de travailler mais pouvant subvenir à leurs besoins, et 80-100% pour les individus capables de travailler) et en dix sous-groupes pour illustrer les spécificités à l'intérieur d'un même groupe (par tranche de 10%).

Définition	%	Critères
Capable de mener une activité normale et de travailler ; pas besoin de soins particuliers	100	Normal ; pas de plaintes ; pas d'évidence de maladie
	90	Capable d'une activité normale ; signes ou symptômes mineurs en relation avec la maladie
	80	Activité normale avec effort ; signes ou symptômes de la maladie
Incapable de travailler ; capable de vivre à domicile et de subvenir à la plupart de ses besoins	70	Capacité de subvenir à ses besoins ; incapable d'avoir une activité normale et professionnelle active
	60	Requiert une assistance occasionnelle mais est capable de subvenir à la plupart de ses besoins
	50	Requiert une assistance et des soins médicaux fréquents
Incapable de subvenir à ses besoins ; requiert un équivalent de soins institutionnels ou hospitaliers	40	Invalide ; requiert des soins et une assistance importants
	30	Sévèrement invalide ; hospitalisation indiquée bien que le décès ne soit pas imminent
	20	Extrêmement malade ; hospitalisation nécessaire ; traitement actif de soutien nécessaire
	10	Mourant ; mort imminente
	0	Décédé

Karnofsky D. The clinical evaluation of chemotherapeutic agents in cancer. Columbia University Press, New-York 1949 :191-205

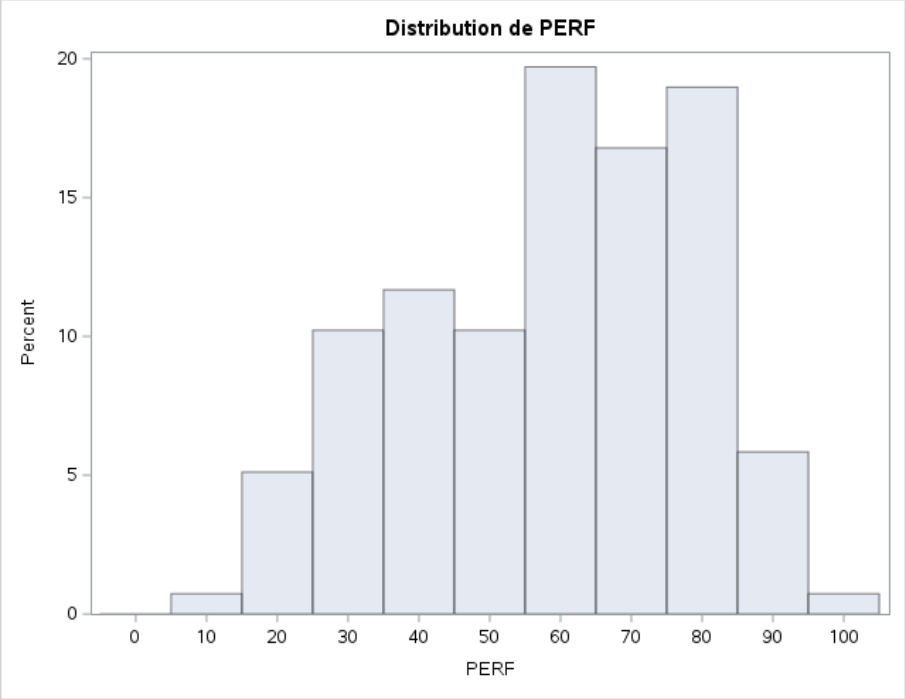
Concernant nos individus, nous pouvons dans un premier temps constater que seuls huit d'entre eux soit 6% de l'échantillon ont un score de 10 ou 20% sur l'échelle de Karnofsky et sont donc extrêmement malades voire mourants.

27,74% de nos individus font partie du groupe caractérisé par l'incapacité totale de subvenir à leurs besoins. 48,18% font partie du groupe caractérisé par l'incapacité de travailler mais la capacité de subvenir à leurs besoins, tandis que les 24,09% restants font partie du dernier groupe et ne sont donc que très peu voire pas influencés par la maladie au quotidien. L'indice moyen relevé s'élève à 58,3%, ce qui caractérise l'individu moyen par une incapacité de travailler et des besoins ponctuels d'assistance mais une certaine autonomie dans leur vie de tous les jours.

PERF				
PERF	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
10	1	0.73	1	0.73
20	7	5.11	8	5.84
30	14	10.22	22	16.06
40	16	11.68	38	27.74
50	14	10.22	52	37.96
60	27	19.71	79	57.66
70	23	16.79	102	74.45
75	2	1.46	104	75.91
80	24	17.52	128	93.43
85	1	0.73	129	94.16
90	7	5.11	136	99.27
99	1	0.73	137	100.00

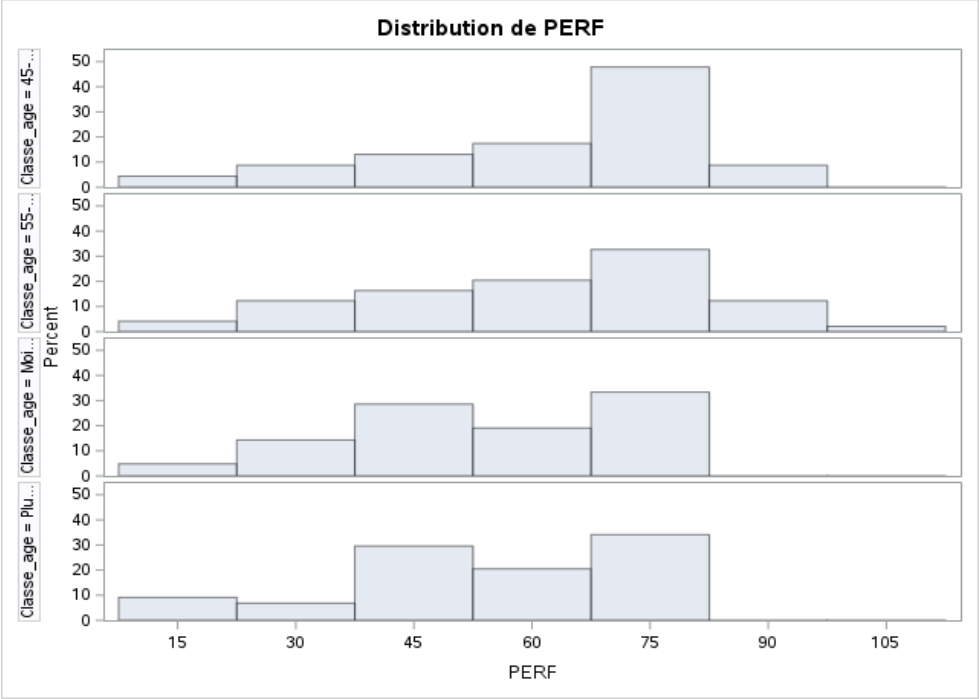
Quant à la répartition de nos individus dans les différents sous-groupes, nous pouvons constater une forte concentration des effectifs (>15% de l'échantillon) dans les catégories 60-70-80%, et une autre concentration relativement moindre (>10% de l'échantillon) dans les catégories 30-40-50%. Autrement dit, une grande partie de nos individus sont affectés par la maladie sans pour autant être invalides, certains peuvent mener une activité professionnelle avec effort tandis que d'autres n'en sont pas capables, mais ils sont tous en mesure de subvenir à leurs besoins par eux-mêmes.

D'autre part, une autre grande partie des individus ont un état de santé plus faible, allant du besoin fréquent d'assistance médicale à l'invalidité, étant donc bien plus affectés par la maladie au quotidien : ils ne sont plus en mesure de répondre à leurs besoins sans assistance.



Par la suite, bien que les individus soient affectés par la maladie de différentes façons, nous nous sommes demandé si l'âge pouvait avoir un impact sur leur autonomie et leur indépendance, les individus les plus touchés pouvant être ceux déjà fragilisés par l'âge. Nous avons alors choisi de diviser notre échantillon en quatre classes d'âge : moins de 45 ans (20 individus), 45-55 ans (22 individus), 55-65 ans (48 individus) et plus de 65 ans (43 individus).

Avec le graphique suivant présentant les distributions de la variable PERF en fonction des classes d'âge, nous ne constatons pas de réelles différences entre classes d'âge, si ce n'est que les distributions des classes 45-55 et 55-65 sont relativement similaires, de même que les classes 45- et 65+. Ainsi, l'âge ne semble pas avoir d'influence sur l'indice de Karnofsky.



Enfin, pour conclure notre description de la population étudiée, nous pouvons analyser les corrélations entre nos variables d'intérêt pour nous renseigner sur les impacts croisés sur notre population. En effet, nous pouvons remarquer que les individus ayant une durée de maladie longue ont généralement un score sur l'échelle de Karnofsky plus faible, et inversement, ce qui signifie que durée de la maladie et dégradation de l'autonomie et de l'indépendance s'influencent mutuellement de façon significative.

Les autres coefficients n'étant pas significatifs, nous pouvons en conclure que l'âge et l'indice de Karnofsky ne s'influencent pas mutuellement (ce qui est en adéquation avec nos observations précédentes), de même que l'âge et la durée de la maladie

Coefficients de corrélation de Pearson, N = 137 Proba > |r| sous H0: Rho=0			
	DD	**AGE**	**PERF**
DD **DD**	1.00000	-0.03342 0.6982	-0.18371 0.0316
AGE **AGE**	-0.03342 0.6982	1.00000	-0.09498 0.2696
PERF **PERF**	-0.18371 0.0316	-0.09498 0.2696	1.00000

3 – Proposer un diagnostic sur le phénomène à l'étude en réalisant une estimation non paramétrique sur l'ensemble des individus, puis à l'aide des variables CT, PERF, TX et PRIORTX dont les modalités sont à recoder en 0 et 1. Commenter tous vos résultats en donnant un exemple d'interprétation des résultats des estimations non paramétriques de la fonction de densité de probabilité, de survie, de risque instantané, et de risque cumulé.

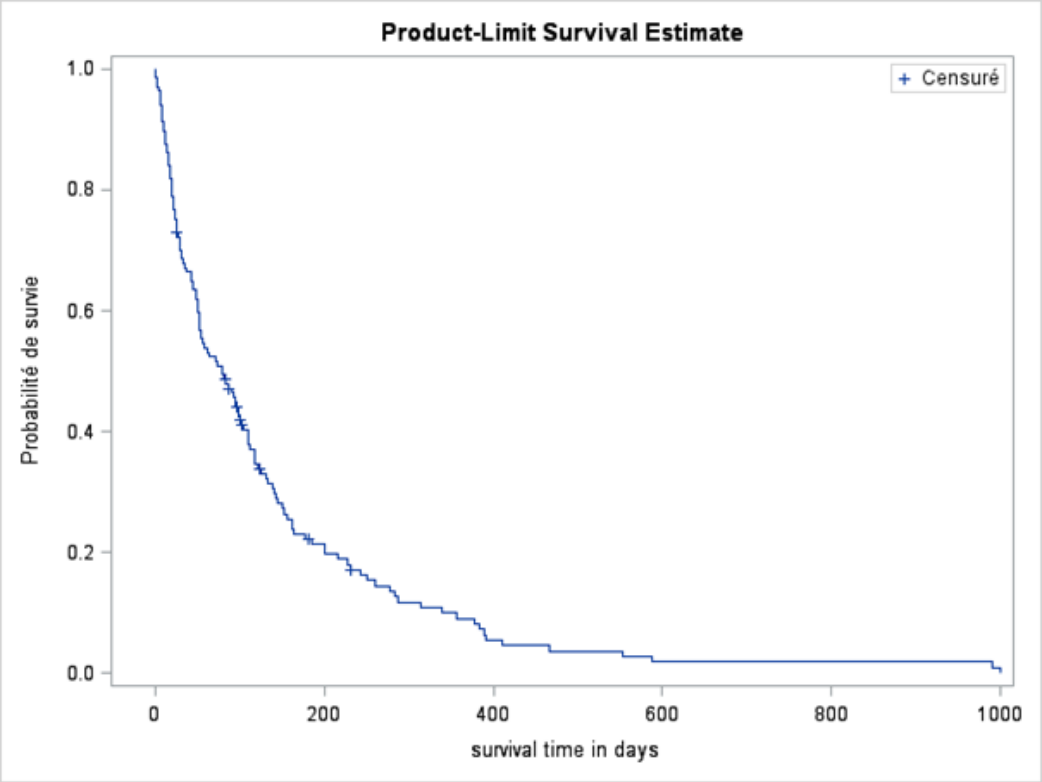
Analyse de l'échantillon total

Dans un premier temps, nous réaliserons une analyse non-paramétrique sur l'ensemble de notre échantillon afin de comprendre comment évolue le temps de survie de nos individus.

Statistiques descriptives pour variable temps SURVT

Estimations du quartile				
		95% Intervalle de confiance		
Pourcentage	Valeur estimée du point	Transformation	[Inférieur	Supérieur)
75	162.000	LINEAR	126.000	228.000
50	80.000	LINEAR	52.000	103.000
25	25.000	LINEAR	19.000	35.000

Tout d'abord, on remarque que sur nos 137 individus 128 d'entre eux sont morts soit 93.4 % de notre échantillon les autres ont été sortis de l'étude. La probabilité qu'un individu meurt avant 30 jours est de 0.3459 (1-0.6541). 50 % des individus ont un temps de survie supérieur ou égal à 80 jours.



Le graphique ci-dessus permet de confirmer nos résultats : on constate qu'à mesure que le temps augmente, la probabilité de survivre au cancer du poumon diminue.

Analyse selon le type de cellule cancéreuse (variables CT)

CT1

Nous allons nous intéresser à la variable CT1 qui renseigne sur le fait que la cellule cancéreuse est large ou non. Nous allons vérifier si ce type de cellule impact la durée de survie de nos patients.

Tout d'abord, nous observons que la proportion de patients décédés selon les groupes est au-dessus des 92 % (96 % des patients avec une large cellule sont décédés et 92.3 % des patients qui n'ont pas ce type de cellule sont décédés également). Le tableau ci-dessous permet de visualiser ces résultats.

Récapitulatif du nombre de valeurs censurées et non censurées

Niveau de discrétisation	CT1	Total	Echec	Censuré	Pourcentage censuré
1	0	110	102	8	7.27
2	1	27	26	1	3.70
Total		137	128	9	6.57

Pour comprendre les différences entre ces deux groupes, nous pouvons nous intéresser à la proportion de patients qui survivent au-delà de 162 jours. On constate alors que 88 % des patients ayant une cellule de type CT1 survivent au-delà de 162 contre 18 % des patients qui n'ont pas une cellule large. Il semblerait alors que le fait de ne pas avoir une cellule large diminue les chances de survie. Cela se confirme également par le fait que la survie moyenne des patients CT1 soit de 170.5 jours tandis que les patients n'ayant pas une cellule large ont une survie moyenne de 124.25 jours.

Afin de vérifier si ce type de cellule influe sur le temps de survie, il convient de prendre en compte les résultats des tests log-rank, log(R) et Wilcoxon. L'on constate que pour deux tests sur trois la p-value est supérieure à 0.05 : on accepte alors h0, l'hypothèse selon laquelle il y a une absence de différence entre les 2 groupes.

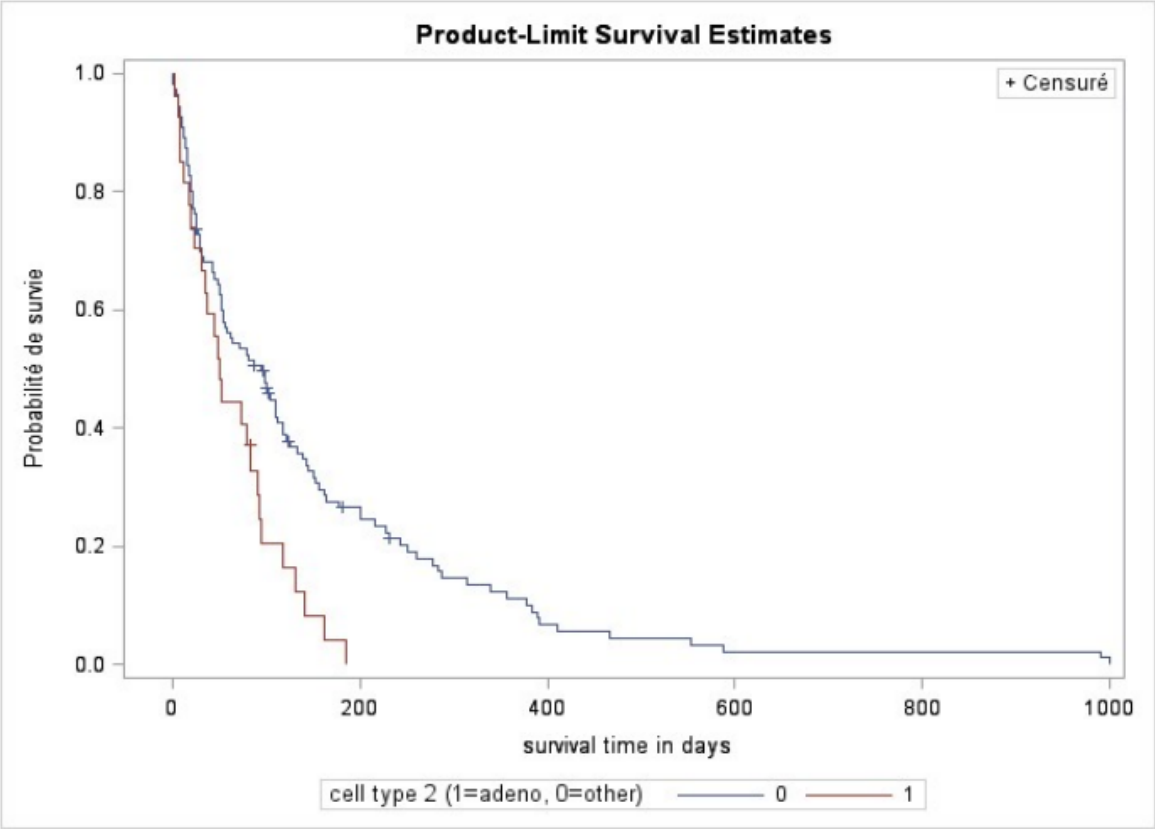
Test d'égalité sur niveaux de discrétisation

Test	Khi-2	DDL	Pr > khi-2
Log-rang	3.0205	1	0.0822
Wilcoxon	7.7837	1	0.0053
-2Log(LR)	3.0094	1	0.0828

CT2

Maintenant que nous savons que les cellules larges n'influent pas significativement sur le temps de survie, qu'en est-il des cellules adenos ? C'est ce que la variable CT2 nous permet de comprendre.

La proportion de patients qui ne survivent pas à leur cancer est la même quelle que soit le groupe que pour la variable CT1. Néanmoins, on constate une grande différence entre les deux groupes dans la proportion de survie au-delà de 162 jours. Effectivement, 28 % des individus n'ayant pas une cellule adeno survivent au-delà de 162 jours tandis que seulement 4 % des individus ayant une cellule adeno survivent au-delà de cette durée. Le graphique ci-dessous permet de visualiser cette différence :



La courbe rouge représente la courbe de survie des patients ayant une cellule de type CT2 et la bleu nous permet de voir l'évolution de la survie des autres patients. On constate qu'aucun patient avec une cellule adeno ne survit au-delà de 200 jours. Il semble alors que le fait d'avoir ce type de cellule diminue le temps de survie des patients. On remarque que la p-value de l'ensemble des tests est inférieure à 0.05 : il existe donc une différence significative entre nos deux groupes de patients.

Test d'égalité sur niveaux de discrétisation			
Test	**Khi-2**	**DDL**	**Pr > khi-2**
Log-rang	8.1920	1	0.0042
Wilcoxon	4.0087	1	0.0453
-2Log(LR)	10.9248	1	0.0009

CT3

La variable CT3 nous permet de savoir si l'individu a une petite cellule ou non. Cela nous permet de comprendre si une petite cellule peut influer sur le temps de survie des patients atteints du cancer du poumon.

Tout d'abord, 48 patients sur 137 avaient une cellule de type CT3, soit 35 % de notre population. La probabilité de survivre au-delà des 162 jours pour eux était de 8 % tandis que les patients qui n'avaient pas une petite cellule avaient une probabilité de 17 %. On constate que le fait d'avoir une petite cellule semble impacter négativement le temps de survie des individus. De plus, 50 % des patients ayant une cellule de type CT3 avaient un temps de survie inférieur ou égal à 51 jours contre 105 jours pour les autres patients. Ces résultats sont confirmés par les tests qui affirment qu'il existe une différence significative entre les deux groupes.

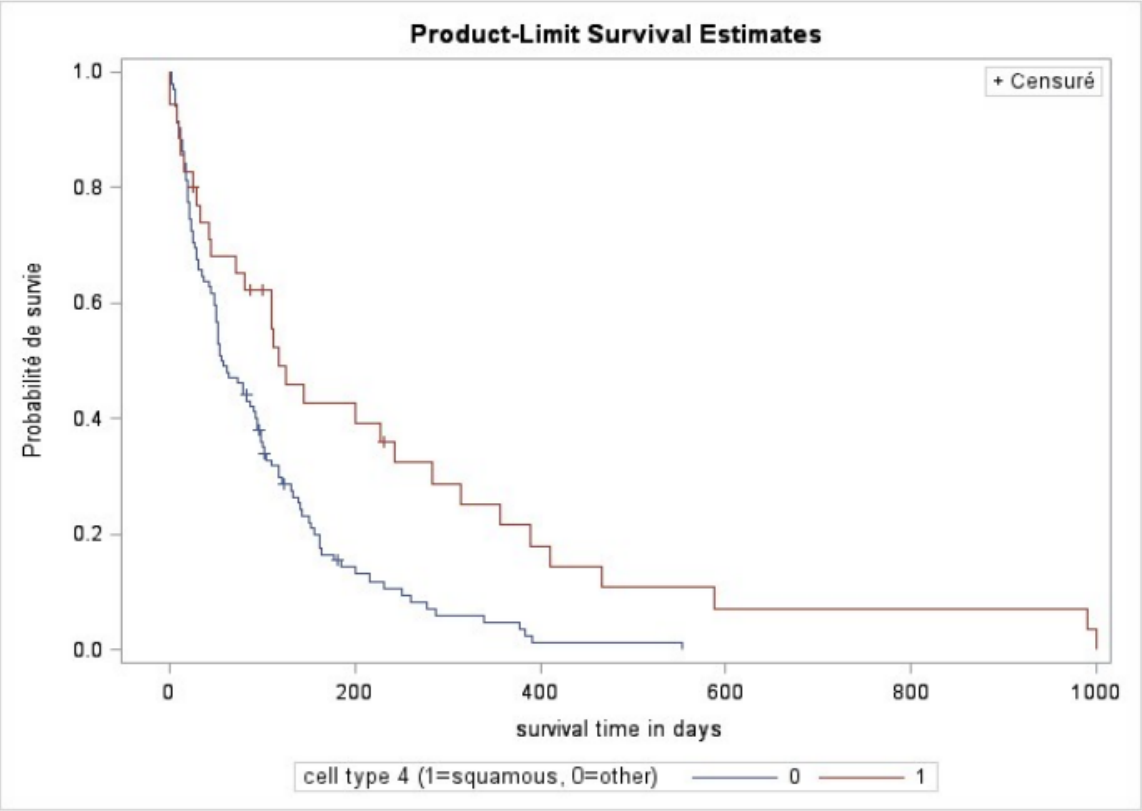
Test d'égalité sur niveaux de discrétisation			
Test	Khi-2	DDL	Pr > khi-2
Log-rang	10.2025	1	0.0014
Wilcoxon	9.3017	1	0.0023
-2Log(LR)	14.3868	1	0.0001

CT4

La variable CT4 permet de savoir si le patient à une cellule squameuse ou non. Dans notre étude, 88 % des patients ayant une cellule squameuse n'ont pas survécu, ils étaient 95 % dans le groupe des patients n'ayant pas une cellule squameuse. De plus, on constate que seulement 17 % des patients survivent au-delà des 162 jours, ils étaient 42 % dans le groupe des CT4. Il semblerait que le fait d'avoir une cellule squameuse favoriserait un temps de survie plus long. En effet, le temps de survie moyen des patients avec cellule squameuse est de 230 jours tandis qu'il est de 100 jours pour les autres patients. Les tests de Wilcoxon, log-rank et log(R) permettent de confirmer l'existence d'une différence significative.

Test d'égalité sur niveaux de discrétisation			
Test	Khi-2	DDL	Pr > khi-2
Log-rang	10.5313	1	0.0012
Wilcoxon	4.7887	1	0.0286
-2Log(LR)	17.8394	1	<.0001

On remarque également que cette différence est clairement visible sur le graphique des temps de survie ci-dessous :



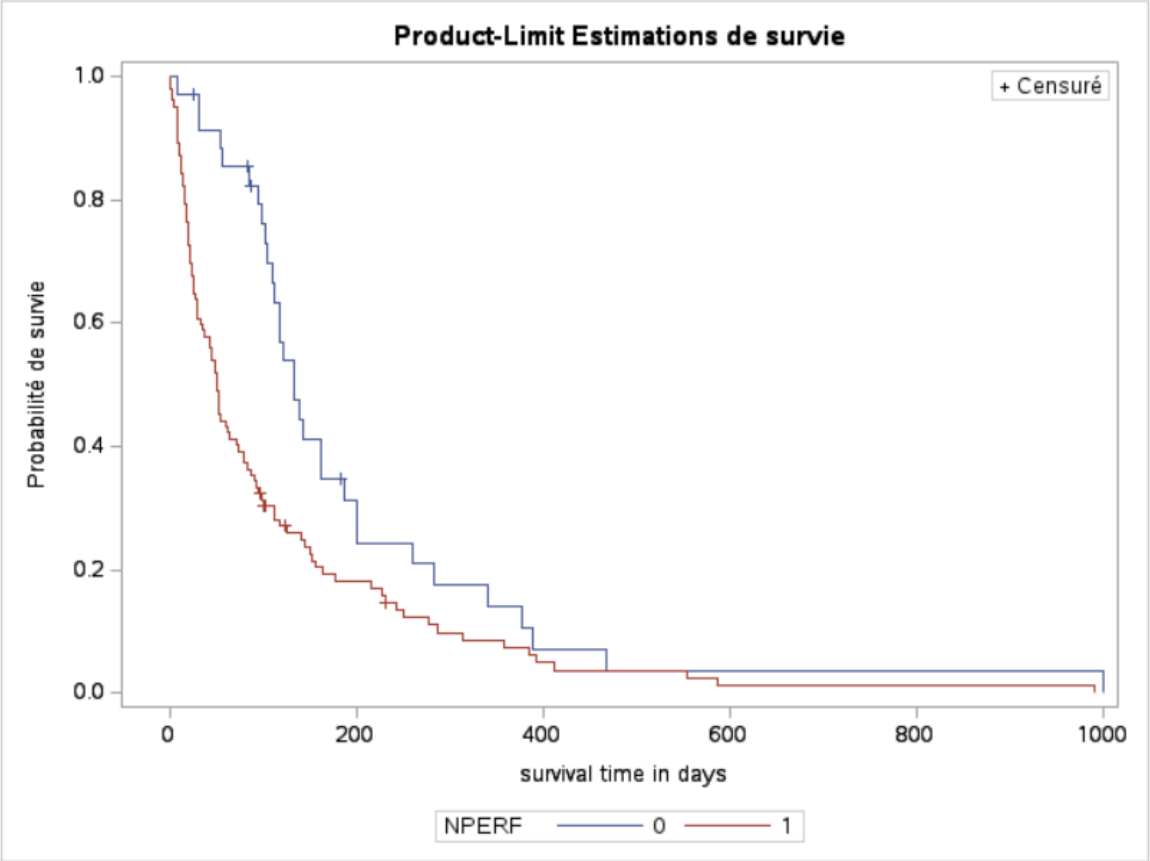
NPERF

La variable PERF nous renseigne sur la capacité des patients à être autonome et à mener une activité normale. Elle se base sur le score de Karnofsky, situé sur une échelle de 0 à 100. Pour notre analyse, nous avons dû recoder cette variable en binaire. Nous avons donc choisi de séparer les patients en deux groupes : ceux qui peuvent réaliser une activité normale (NPERF=0) et ceux qui ne le peuvent pas ou alors avec difficulté (NPERF=1). Pour cela, nous avons étudié les caractéristiques des étapes intermédiaires de l'échelle de performance de Karnofsky afin de déterminer une valeur seuil qui séparerait les deux groupes.

Nous avons ainsi choisi de fixer la valeur seuil à 75 pour construire nos deux groupes, pour avoir d'une part les individus à partir de cette valeur considérés comme assez autonomes avec peu ou pas d'impact de la maladie sur leurs activités quotidiennes, contre d'autre part les individus strictement en dessous de ce seuil considérés comme au minimum incapables d'être actifs professionnellement parlant, voire incapables de subvenir à leurs besoins par eux-mêmes pour les scores les plus bas.

A l'issue de cette modification, nous avons constaté que seuls 25 % des patients étaient en capacité de réaliser une activité normale, parmi lesquels 89% des individus sont morts (contre 95 % dans le groupe des individus peu/pas autonomes). Les individus considérés comme peu autonomes avaient une probabilité de survie après 162 jours de 34 % tandis que les individus considérés comme autonomes avaient une probabilité de 30 %. De plus, la durée de survie moyenne des individus peu autonomes était de 111 jours contre 195 pour les autonomes. Il semblerait que le fait d'être autonomes ou non influe sur la durée de survie, c'est ce que confirment les tests suivants :

Test d'égalité sur niveaux de discrétisation			
Test	Khi-2	DDL	Pr > khi-2
Log-rang	8.4369	1	0.0037
Wilcoxon	16.0023	1	<.0001
-2Log(LR)	9.8115	1	0.0017

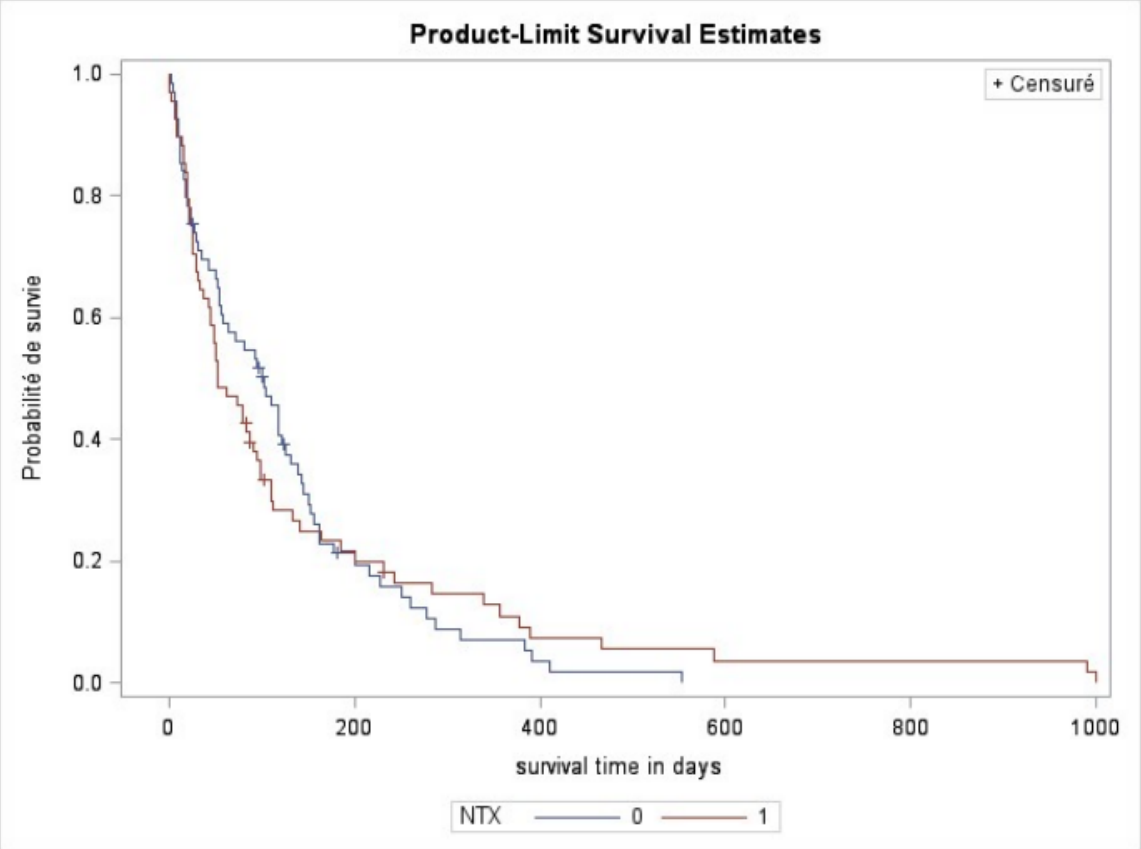


L'étude graphique des courbes des temps de survie nous permettent de confirmer la différence de temps de survie entre les personnes dites "dépendantes" et les personnes indépendantes.

NTX

La variable TX permet de savoir si le patient a reçu le nouveau traitement ou s'il a reçu un traitement standard. Nous avons recodé cette variable en binaire : 1 si le patient a le nouveau traitement et 0 si non. 49 % des patients de notre étude ont reçu le nouveau traitement. La proportion de personnes décédées durant l'étude selon les groupes est similaire : 94 % pour ceux ayant reçu le nouveau traitement et 93 % pour les autres. La probabilité de survivre au-delà de 162 jours pour les patients n'ayant pas reçu le nouveau traitement est de 22 %, elle est de 24 % pour les autres patients. On constate néanmoins une grande différence entre les

groupes avec la médiane. En effet, 50 % des patients n'ayant pas reçu le nouveau traitement ont un temps de survie supérieur ou égal à 52 jours contre 123 jours pour les patients avec le nouveau traitement. Cependant, cette différence s'explique par le fait qu'un seul patient dans le groupe du nouveau traitement a survécu au-delà des 900 jours comme on peut l'observer ci-dessous :



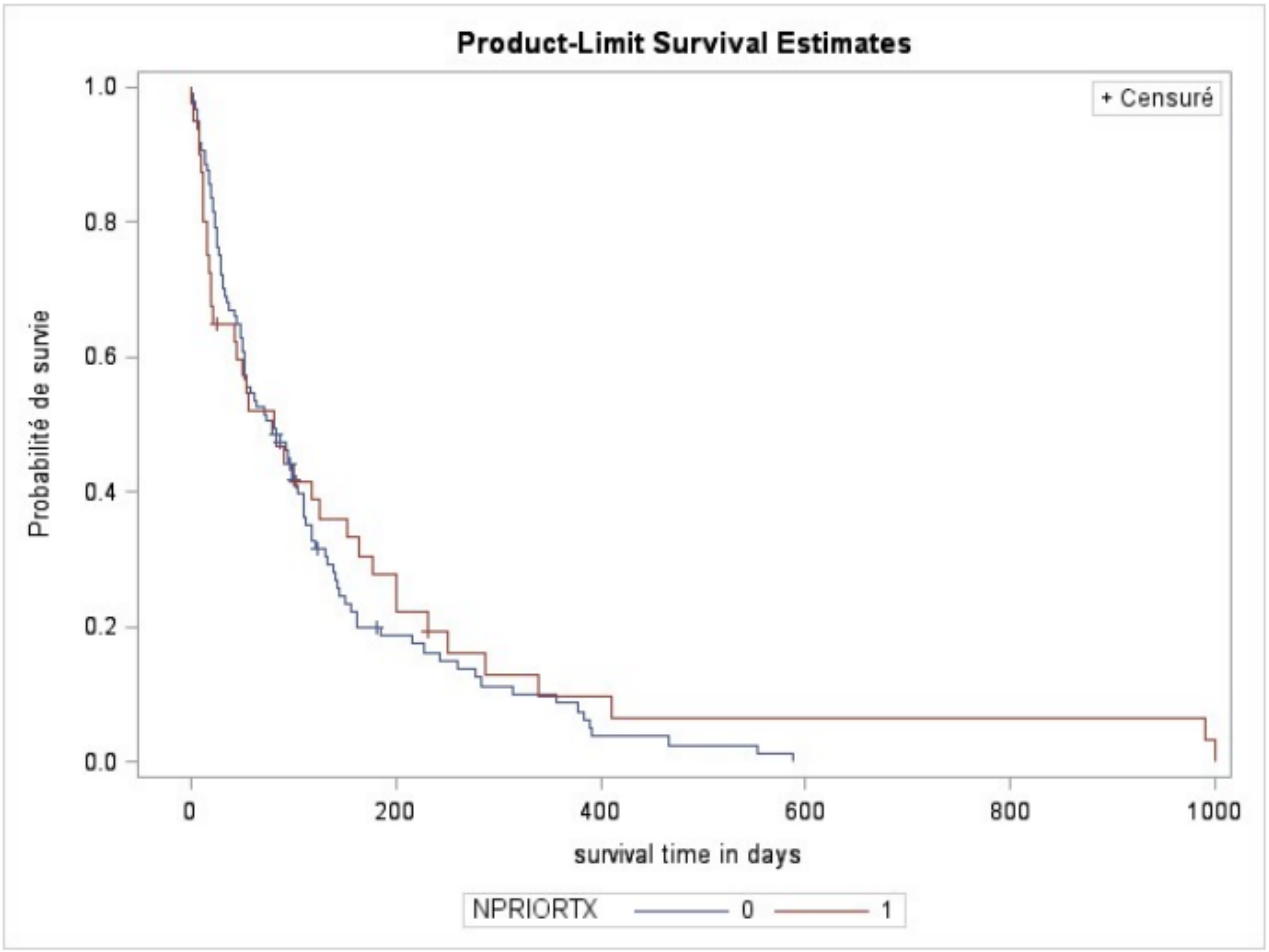
De plus, les tests de Wilcoxon, log-rank et log(R) permettent de valider le fait que le nouveau traitement n'a pas plus d'impact sur le temps de survie des patients que l'ancien traitement.

NPRIORTX

La variable PRIORTX permet de savoir si le patient a eu accès à un traitement antérieur ou non. Nous l'avons codé en binaire avec PRIORTX = 1 si le patient a reçu un traitement antérieur et PRIORTX = 0 si non. Seulement 29 % des patients ont eu accès à un traitement antérieur, parmi eux 92 % n'ont pas survécu au cancer, la proportion est similaire pour les patients qui n'ont pas reçu de traitement (93 %).

Récapitulatif du nombre de valeurs censurées et non censurées					
Niveau de discrétisation	NPRIORTX	Total	Echec	Censuré	Pourcentage censuré
1	0	97	91	6	6.19
2	1	40	37	3	7.50
Total		137	128	9	6.57

Pour comprendre si le fait d'avoir reçu un traitement antérieurement influence la durée de survie, nous pouvons comparer la probabilité de survie après 162 jours pour nos deux groupes. Tout d'abord, on constate que pour le groupe PRIORTX = 1 la probabilité de survivre au-delà des 162 jours est de 33 % tandis que pour l'autre groupe elle est de 19 %. Il semble alors que le fait d'avoir eu un traitement antérieur favorise une durée de survie plus élevée. C'est ce que l'on peut observer sur le graphique ci-dessous :



De plus nous avons réalisé les tests suivant afin de valider si la différence entre les groupes est significative ou non :

Test d'égalité sur niveaux de discrétisation			
Test	Khi-2	DDL	Pr > khi-2
Log-rang	0.5014	1	0.4789
Wilcoxon	0.0574	1	0.8107
-2Log(LR)	1.9605	1	0.1615

Effectivement, on constate que les p-values des tests de wilcoxon, log-rank et log(R) sont supérieures à 0.05. Cela signifie que la différence entre les deux groupes n'est pas significative et que le fait d'avoir eu un traitement antérieur n'influence pas de manière significative le temps de survie des patients.

4 – Quel(s) enseignement(s) à tirer de ces résultats ?

Variables	Différence significative entre les deux groupes	Différence non significative entre les deux groupes
CT1 *(cellule large)*		X
CT2 *(cellule adeno)*	X *influence - sur le temps de survie*	
CT3 *(cellule petite)*	X *influence - sur le temps de survie*	
CT4 *(cellule squameux)*	X *influence + sur le temps de survie*	
NPERF *(1 si peu autonome et 0 sinon)*	X *influence - sur le temps de survie*	
NTX *(1 si nouveau traitement et 0 sinon)*		X
NPRIORTX *(1 si traitement antérieur et 0 sinon)*		X

Grâce aux différentes estimations non-paramétriques vues à la question précédente, nous avons pu en apprendre davantage sur la survie de nos patients et sur l'influence de certains facteurs tels que les types de cellule mais encore les traitements reçus.

Nous avons constaté que le temps de survie médian de nos patients était de 80 jours. Pour comprendre ce qui pouvaient influer sur ce temps, nous nous sommes tout d'abord intéressés aux types de cellules (large, adéno, petite, squameuse). Il s'est avéré que seule la variable CT1 et donc le fait d'avoir une cellule large ou non n'avait pas d'impact significatif sur le temps de survie. Le fait d'avoir soit une petite cellule ou adeno exerce une influence négative sur le temps de survie des patients. En effet, il s'est avéré que les patients ayant une petite cellule avaient un temps de survie moyen largement inférieur aux autres patients (65 jours contre 149 jours), il en va de même pour les patients avec une cellule adeno (79 jours contre 170 jours). A l'inverse, le fait d'avoir une cellule de type squameuse influence positivement le temps de survie, en effet les patients ayant ce type de cellule ont un temps de survie moyen de 230 jours tandis que celui des autres patients est de 100 jours.

Ces trois types de cellules ne sont pas les seules variables à impacter la durée de survie des patients : le fait de ne pas être assez autonomes (avec peu ou pas d'impact de la maladie sur leurs activités quotidiennes) impact négativement le temps de survie des malades. Il semblerait que ces patients ont un temps de survie moyen de 110 jours contre 195 jours pour les patients plus autonomes. Par conséquent, le fait d'être autonome (avec peu ou pas d'impact de la maladie sur leurs activités quotidiennes), permettrait d'augmenter la durée de survie.

Malheureusement, nous avons remarqué que le temps de survie n'était ni impacté par le fait d'avoir eu un traitement antérieur ou par le fait d'avoir bénéficié du nouveau traitement. Il semble donc que le nouveau traitement n'ait pas de réel impact sur le cancer du poumon.

5 – Réaliser un test de vérification de l'hypothèse de risques proportionnels constants et de lien fonctionnel des facteurs explicatifs d'intérêt (de votre choix).

Vérification de l'hypothèse de risque constant :

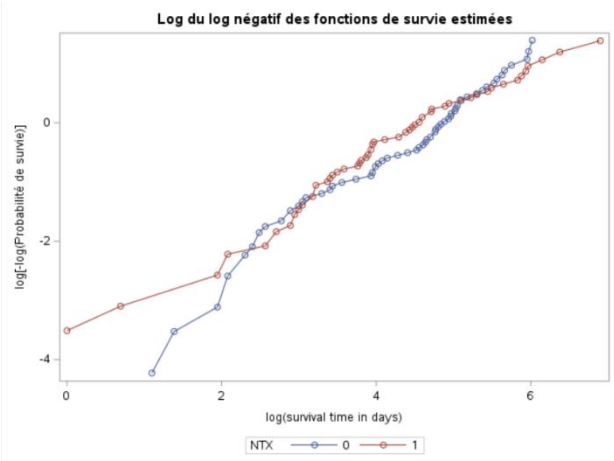
Le test de vérification de l'hypothèse de risques proportionnels constant est important car nous utilisons dans notre étude un modèle de Cox qui est un modèle semi-paramétrique à risques proportionnels : le risque de l'événement dans un groupe donné est un multiple constant du risque dans tous les autres groupes. Cette hypothèse implique que, comme mentionné ci-dessus, les courbes de risque pour les groupes doivent être proportionnelles et ne peuvent pas se croiser.

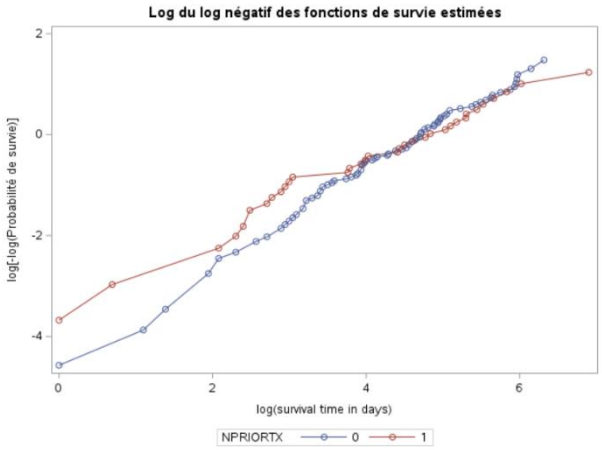
Dans notre étude, nous allons vérifier l'hypothèse de risques proportionnels des facteurs explicatifs suivants: CT1, CT2, CT3, CT4, NTX, NPRIORTX, NPERF

Nous avons donc compilés ci-dessous les courbes log-log négatif des fonctions de survie pour chaque variables, ainsi que les résultats aux tests associés :

Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	0.5014	1	0.4789
Wilcoxon	0.0574	1	0.8107
-2Log(LR)	1.9605	1	0.1615

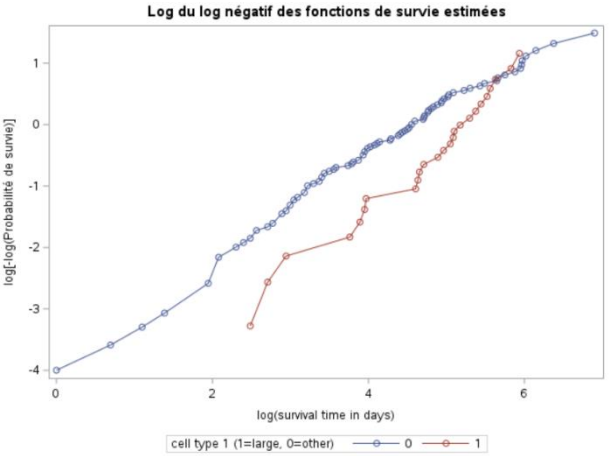
Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	0.0082	1	0.9277
Wilcoxon	0.9608	1	0.3270
-2Log(LR)	0.2758	1	0.5995

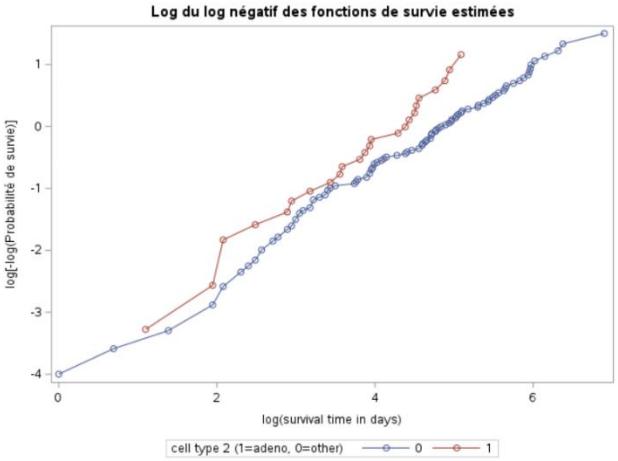




Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	3.0205	1	0.0822
Wilcoxon	7.7837	1	0.0053
-2Log(LR)	3.0094	1	0.0828

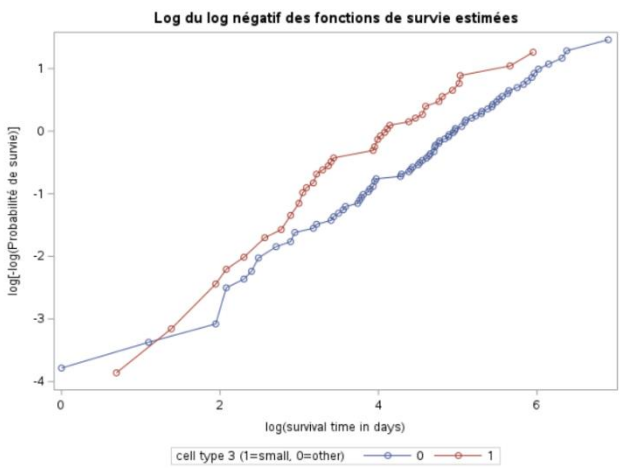
Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	8.1920	1	0.0042
Wilcoxon	4.0087	1	0.0453
-2Log(LR)	10.9248	1	0.0009

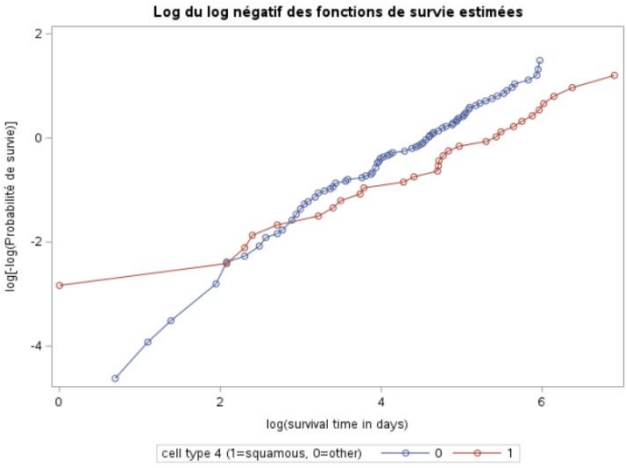




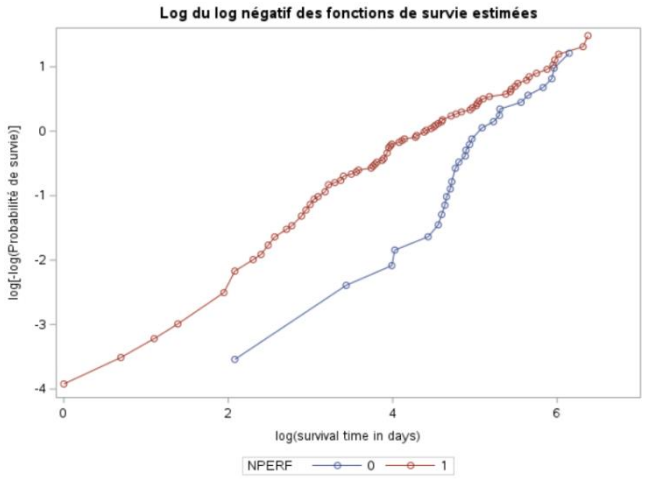
Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	10.2025	1	0.0014
Wilcoxon	9.3017	1	0.0023
-2Log(LR)	14.3868	1	0.0001

Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	10.5313	1	0.0012
Wilcoxon	4.7887	1	0.0286
-2Log(LR)	17.8394	1	<.0001





Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	8.4369	1	0.0037
Wilcoxon	16.0023	1	<.0001
-2Log(LR)	9.8115	1	0.0017



Les graphiques s'interprètent de la manière suivante : il s'agit de l'écart entre les différents groupes, par exemple l'écart des fonctions de survie entre ceux qui ont reçu le nouveau traitement et ceux qui n'ont pas reçu de nouveau traitement. Pour vérifier l'hypothèse de risques proportionnels constants, on peut utiliser 2 indicateurs : les tests de risques proportionnelles (Log-rank, Wilcoxon, 2log(LR)) et comparer les écarts et la forme des fonctions de survie. Si les courbes se croisent ou se séparent de manière significative, cela veut dire que l'hypothèse de risques proportionnels constants n'est pas respectée. Au contraire, si les courbes sont parallèles, l'hypothèse de risques proportionnels constants est respectée.

Nous avons réalisé un tableau qui récapitule l'ensemble des résultats aux tests :

Tests	NTX	NPRIORTX	CT1	CT2	CT3	CT4	NPERF
Log-rang	Non	Non	10%	1%	1%	1%	1%
Wilcoxon	Non	Non	1%	5%	1%	5%	1%
-2Log(LR)	Non	Non	10%	1%	1%	1%	1%

De plus, via l'observation graphique on remarque que l'ensemble des courbes se croisent, sauf celle de la variable CT2, et l'ensemble des courbes se rapprochent énormément, sans les tests, nous pourrions rejeter les hypothèses de risques proportionnels constant pour l'ensemble des variables.

Cependant, nous décidons de conserver CT1 CT2 CT3 CT4 et NPERF pour réaliser une estimation semi paramétrique, afin d'estimer les ratio de risque associés à chaque caractéristique.

Nous obtenons les résultats suivants :

Analyse des valeurs estimées du maximum de vraisemblance									
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%		Libellé
CT1	1	0.26444	0.27639	0.9154	0.3387	1.303	0.758	2.239	cell type 1 (1=large, 0=other)
CT2	1	1.22564	0.29425	17.3492	<.0001	3.406	1.913	6.064	cell type 2 (1=adeno, 0=other)
CT3	1	0.90824	0.25295	12.8922	0.0003	2.480	1.511	4.072	cell type 3 (1=small, 0=other)
CT4	0	0	.	.	.	.	.	.	cell type 4 (1=squamous, 0=other)
NPERF	1	0.58711	0.21693	7.3249	0.0068	1.799	1.176	2.752	

On remarque que CT1 n'est pas significatif et que l'on ne peut pas estimer le ratio de risque de CT4 car il s'agit d'une variable liée à CT1, CT2, CT3 qui prend 1 si CT1, CT2 et CT3 sont = à 0, nous allons donc retirer CT1. CT4 n'étant pas significatif non plus, on la retire de l'estimation.

On a donc le modèle final suivant :

Analyse des valeurs estimées du maximum de vraisemblance									
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%		Libellé
CT2	1	1.09782	0.25771	18.1475	<.0001	2.998	1.809	4.967	cell type 2 (1=adeno, 0=other)
CT3	1	0.78651	0.21324	13.6041	0.0002	2.196	1.446	3.335	cell type 3 (1=small, 0=other)
NPERF	1	0.57812	0.21662	7.1223	0.0076	1.783	1.166	2.726	

Avec les résultats suivants aux tests d'hypothèses des risques proportionnels :

Test de la borne supérieure pour l'hypothèse des risques proportionnels				
Variable	Valeur absolue maximale	Réplications	Valeur initiale	Pr > ValAbsMax
CT2	1.2514	50	25000	0.1000
CT3	1.1742	50	25000	0.1200
NPERF	2.2110	50	25000	<.0001

L'interprétation de ce test est similaire à celui de Kolmogorov, c'est à dire que l'on pourras considérer les ratio de risque (ici nommés "Rapport de risque") seulement si la p-value est supérieur à 0,05. Cela nous amène à conclure que la variable NPERF, à un impact non significatif sur l'estimation. Ainsi on pourras estimer les variables CT2 et CT3.

Donc pour conclure, nous pouvons dire qu'avoir une cellule cancéreuse de type "adeno" augmente de 2,998 fois la probabilité de connaître l'événement. De plus, avoir une petite cellule cancéreuse fait augmenter de 2,196 fois la probabilité de connaître l'événement.

Vérification du lien fonctionnel :

Outre cela, il est important de vérifier que chaque variable explicative du modèle suit la forme fonctionnelle. En effet, lorsque l'on étudie la forme fonctionnelle, on étudie en réalité le lien entre la variable dépendante (c'est-à-dire le temps de survie) et les variables explicatives

Si la forme fonctionnelle d'une variable n'est pas vérifiée, cela peut conduire à des erreurs d'estimation des coefficients.

Dans ce cadre, nous utiliserons le test de Kolmogorov qui permet de comparer 2 distributions (ici la fonction de répartition), l'une théorique et l'autre observée. Ce test permettra de calculer les différences maximales entre ces 2 distributions, appelées « Distance de Kolmogorov ». Pour finir, en fonction de cette distance, on rejettera ou non l'hypothèse nulle H0 : l'échantillon de données suit une distribution de probabilité spécifique (comme la Loi Normal par exemple).

Concrètement, si l'hypothèse nulle est rejetée, on ne pourrait pas interpréter le ratio de risque, calculé avant le test, car il serait biaisé.

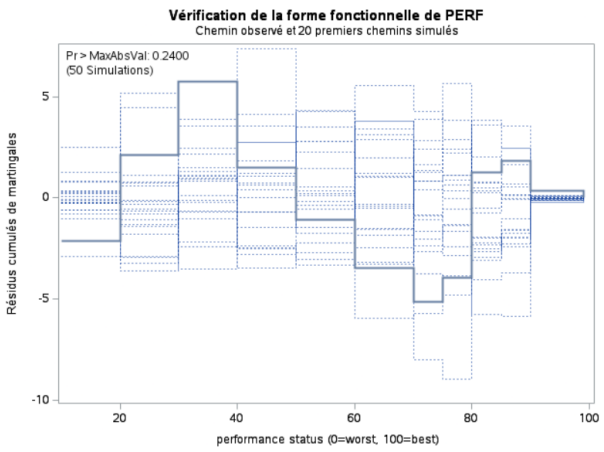
Dans le cadre de notre dossier, nous allons vérifier la forme fonctionnelle de la relation entre le risque et le score de l'échelle de Karnofsky ainsi que l'âge qui sont les 2 seules variables continues de notre BDD. Nous utiliserons également les variables explicatives précédemment retenues.

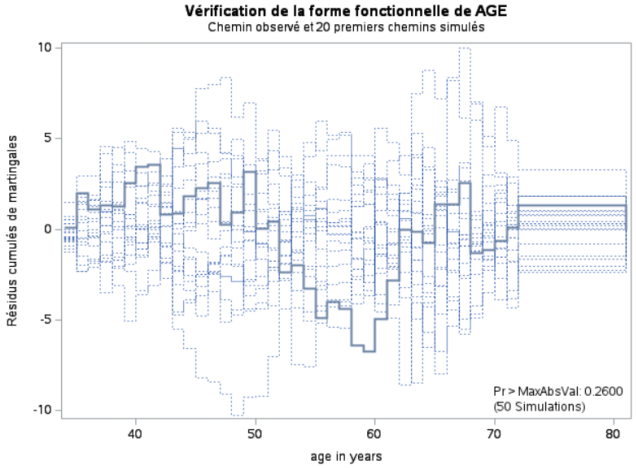
Ainsi, nous obtenons le modèle de Cox suivant, avec Age et Perf en covariable :

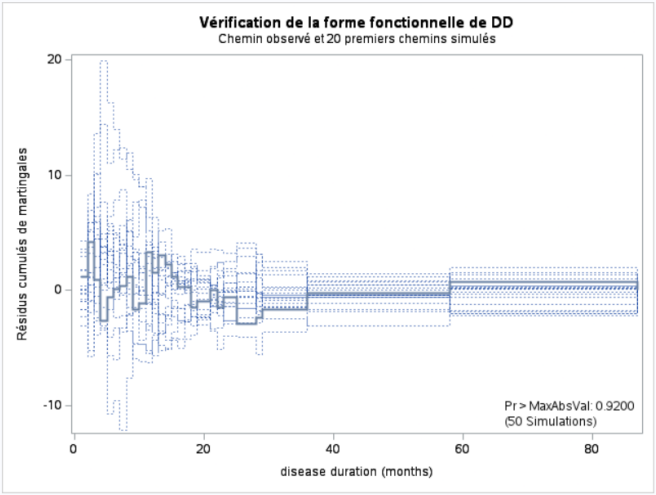
Analyse des valeurs estimées du maximum de vraisemblance									
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%		Libellé
PERF	1	-0.03166	0.00543	33.9279	<.0001	0.969	0.959	0.979	performance status (0=worst, 100=best)
AGE	1	-0.00561	0.00911	0.3788	0.5382	0.994	0.977	1.012	age in years
DD	1	0.00280	0.00818	0.1171	0.7322	1.003	0.987	1.019	disease duration (months)
CT2	1	0.85358	0.29553	8.3425	0.0039	2.348	1.316	4.190	cell type 2 (1=adeno, 0=other)
CT3	1	0.39432	0.26182	2.2683	0.1320	1.483	0.888	2.478	cell type 3 (1=small, 0=other)
CT4	1	-0.32138	0.27662	1.3498	0.2453	0.725	0.422	1.247	cell type 4 (1=squamous, 0=other)

La première chose remarquable c'est que l'âge, DD, CT3 et CT4 ne sont pas significatif, nous essayerons donc de faire une estimation en retirant l'âge.

Ensuite, nous avons la vérification des formes fonctionnelles de PERF, AGE et DD :







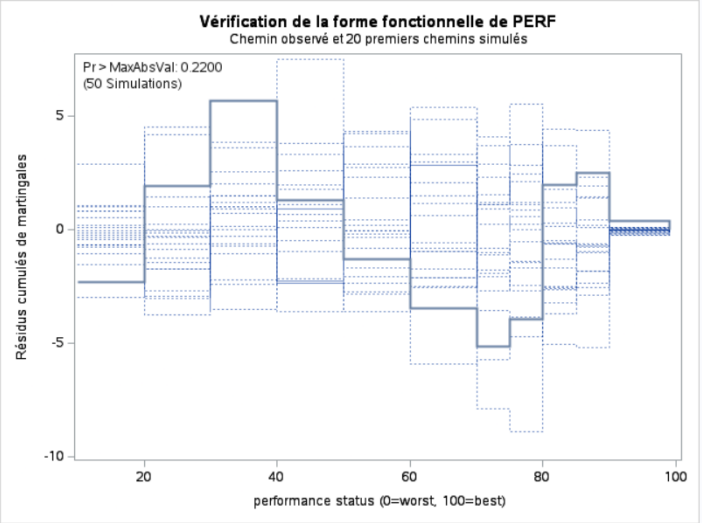
Avec les résultats du test de Kolmogorov associés :

Test de la borne supérieure pour la forme fonctionnelle				
Variable	Valeur absolue maximale	Réplications	Valeur initiale	Pr > ValAbsMax
PERF	5.6169	50	25000	0.2400
AGE	6.9553	50	25000	0.2400
DD	4.2090	50	25000	0.9200

u vu des résultats, on accepte, l'hypothèse nulle, nos 3 variables suivent une distribution adéquate, on pourra donc interpréter le ratio de risque de PERF (AGE n'étant pas significatif, on ne peut quand même pas interpréter son ratio de risque). Nous allons donc réaliser une seconde estimation du modèle de COX avec uniquement le score de l'échelle de Karnofsky en covariable (en retirant CT4 et DD du modèle, l'ensemble de nos variables sont significatives). Nous obtenons donc les résultats suivants :

Analyse des valeurs estimées du maximum de vraisemblance								
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%	Libellé
PERF	1	-0.03032	0.00512	35.0778	<.0001	0.970	0.960 0.980	performance status (0=worst, 100=best)
CT2	1	1.00142	0.25821	15.0416	0.0001	2.722	1.641 4.515	cell type 2 (1=adeno, 0=other)
CT3	1	0.57101	0.21568	7.0092	0.0081	1.770	1.160 2.701	cell type 3 (1=small, 0=other)

Avec la vérification de la forme fonctionnelle (p-value = 0,22) :



On accepte donc H0, et l'on peut interpréter le ratio de risque de la variable PERF. Ainsi, nous pouvons dire que l'augmentation d'un point au score de perf correspond à une augmentation de la probabilité de connaître l'événement (ici la mort) de 0,97 fois (soit 1,0308 fois moins de chance de connaître l'événement).

Calcul de 1,0308 : e-0,03032(-1)= 1,0308*

6 – Proposer un diagnostic sur le phénomène à l'étude en réalisant des estimations semi-paramétriques stepwise avec d'une part les variables TX (recodée), PRIORTX (recodée), âge et d'autre part avec tous les facteurs explicatifs. Écrire la forme théorique de chaque modèle et commenter les résultats obtenus dans chaque modèle.

Nous allons mettre en place deux estimations semi-paramétriques stepwise, cela va nous permettre de prendre en compte le temps à l'aide du modèle de Cox. D'une manière générale, le modèle de Cox peut s'écrire sous la forme suivante :

$$h(t, X(t)) = h_0(t) exp(\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \delta_i X_i \times g_i(t))$$

où t représente le temps, X les variables explicatives et $g_i(t)$ la fonction du temps pour le i-ème facteur explicatif.

Tout d'abord, nous allons réaliser une estimation semi-paramétrique stepwise avec les variables NTX, NPRIORTX et âge. La forme théorique de ce modèle s'écrit comme suit :

$$h(t) = h_0(t)\, exp(\beta_1 \times NTX + \beta_2 \times NPRIORTX + \beta_3 \times \hat{a}ge\,)$$

Sous SAS, nous utilisons la méthode EFRON qui permet de calculer un risque moyen pour tous les événements simultanés. Nous obtenons alors les résultats suivants :

La procédure PHREG

Informations sur le modèle		
Table	TP.VETS	
Variable dépendante	SURVT	SURVT
Variable de censure	STATUS	STATUS
Valeur(s) de censure	0	
Ties Handling	EFRON	

Nombre d'observations lues	137
Nombre d'observations utilisées	137

Récapitulatif du nombre d'événements et de valeurs censurées			
Total	Evénement	Censuré	Pourcentage censuré
137	128	9	6.57

Note: No effects met the 0.05 level for entry into the model. There are no explanatory variables in the model.

-2 LOG L = 1010.898

On remarque qu'aucune des variables de notre modèle n'est significative : elles ne sont donc pas conservées.

A présent, nous allons constituer un second modèle qui regroupe l'intégralité des variables. Ce nouveau modèle peut s'écrire de la forme suivante :

$$h(t) = h_0(t)\, exp(\beta_1 \times CT1 + \beta_2 \times CT2 + \beta_3 \times CT3 + \beta_4 \times CT4 + \beta_5 \times NPERF + \beta_6 \times DD + \beta_7 \times \hat{a}ge + \beta_8 \times NTXT + \beta_4 \times NPRIORTX)$$

Ce nouveau modèle stepwise comporte trois étapes. La première nous permet de remarquer que la variable CT4, qui nous indique si le patient a une cellule squameux ou non, est sélectionnée par le modèle. Elle est effectivement significative car la p-value de ses tests sont inférieures à 0.05.

Etape 1. Effet CT4 saisi. Le modèle contient les effets suivants :

CT4

Etat de convergence
Critère de convergence (GCONV=1E-8) respecté.

Statistique d'ajustement du modèle		
Critère	Sans covariables	Avec covariables
-2 LOG L	1010.898	999.524
AIC	1010.898	1001.524
SBC	1010.898	1004.376

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	11.3737	1	0.0007
Score	10.5453	1	0.0012
Wald	10.1976	1	0.0014

La seconde étape permet de sélectionner la variable concernant les caractéristiques des cellules larges (CT1). Comme pour la variable précédente ses p-value pour les tests sont inférieures à 0.05.

Etape 2. Effet CT1 saisi. Le modèle contient les effets suivants :

CT1 CT4

Etat de convergence
Critère de convergence (GCONV=1E-8) respecté.

Statistique d'ajustement du modèle		
Critère	Sans covariables	Avec covariables
-2 LOG L	1010.898	986.390
AIC	1010.898	990.390
SBC	1010.898	996.094

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	24.5079	2	<.0001
Score	25.1115	2	<.0001
Wald	23.7553	2	<.0001

La troisième et dernière étape permet de sélectionner la variable NPERF qui concerne l'autonomie des individus.

Etape 3. Effet NPERF saisi. Le modèle contient les effets suivants :

CT1 CT4 NPERF

Etat de convergence
Critère de convergence (GCONV=1E-8) respecté.

Statistique d'ajustement du modèle		
Critère	Sans covariables	Avec covariables
-2 LOG L	1010.898	979.585
AIC	1010.898	985.585
SBC	1010.898	994.141

Test de l'hypothèse nulle globale : BETA=0			
Test	khi-2	DDL	Pr > khi-2
Rapport de vrais	31.3128	3	<.0001
Score	31.5665	3	<.0001
Wald	29.9867	3	<.0001

Pour conclure l'analyse de ce modèle on remarque alors que sur l'ensemble de nos facteurs explicatifs, seuls trois sont sélectionnés. Les variables CT1, CT4 et NPERF sont les plus pertinentes pour analyser la durée de survie.

Analyse des valeurs estimées du maximum de vraisemblance

Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%		Libellé
CT1	1	-0.75253	0.23829	9.9731	0.0016	0.471	0.295	0.752	CT1
CT4	1	-1.00878	0.23773	18.0064	<.0001	0.365	0.229	0.581	CT4
NPERF	1	0.53058	0.21179	6.2759	0.0122	1.700	1.122	2.575	

Récapitulatif sur la sélection Stepwise

Etape	Effet		DDL	Nombre dans	Khi-2 du score	Khi-2 de Wald	Pr > khi-2	Libellé de l'effet
	Saisi	Supprimé						
1	CT4		1	1	10.5453		0.0012	CT4
2	CT1		1	2	12.4872		0.0004	CT1
3	NPERF		1	3	6.4143		0.0113	

Contrairement à ce que nous avons vu dans le cadre de l'analyse non-paramétrique, il semblerait que le fait d'avoir une cellule de type CT1 ou CT4 impacterait de façon négative la durée de survie. La variable NPERF, quant à elle, démontre le fait qu'un individu moins autonome a plus de chance de survivre longtemps qu'un autre patient.

7 – Quel(s) enseignement(s) à tirer de ces modèles ?

Ces deux estimations semi-paramétriques avaient pour objectif de déterminer un possible effet significatif de nos variables sur la durée de survie des patients à l'aide d'une procédure stepwise. Cela signifie que le modèle ne retient que les variables ayant un effet significatif sur le temps de survie.

Dans un premier temps, nous étions contraints dans notre choix de variable et avons créé notre modèle à partir des variables NTX, NPRIORTX et âge. A l'issue de la procédure stepwise, aucune de ces variables n'a été retenue. Nous pouvons donc tirer de ce premier modèle une première conclusion, à savoir que ni le fait d'être traité avec le nouveau traitement, ni le fait d'avoir déjà eu un traitement antérieur, ni l'âge, ont un effet significatif sur le temps de survie.

Dans un second temps, nous avons réalisé une deuxième estimation stepwise en intégrant cette fois-ci l'ensemble de nos variables explicatives. Au terme de la procédure stepwise, le modèle obtenu est cette fois-ci composé de trois variables explicatives dont nous savons d'ores et déjà qu'elles ont un effet significatif sur le temps de survie. A l'inverse, nous savons également que toutes les variables non-retenues n'ont pas d'effet significatif sur le temps de survie.

Ainsi, nous pouvons tirer de ce deuxième modèle une seconde conclusion, à savoir que les variables CT1 (cellule cancéreuse de type large), CT4 (cellule cancéreuse de type squameuse) et NPERF (patients dont l'autonomie est fragilisée par la maladie) ont respectivement un impact significatif négatif, négatif et positif

sur la durée de survie des patients.

Cela entre néanmoins en contradiction avec nos résultats précédents, dans la mesure où nous avions déterminé suite à l'estimation de modèles non-paramétriques que CT1 n'avait aucun impact significatif sur le temps de survie, que CT4 avait certes un effet significatif mais positif sur le temps de survie, de même que NPERF qui avait un effet significatif mais négatif sur le temps de survie.

8 – Comment prendre en compte le type de cellule cancéreuse dans les modèles précédents ?

Pour prendre en compte le type de cellule cancéreuse dans les modèles précédents, nous aurions pu créer un profil unique inhérent à chaque cellule compte tenu des quatre caractéristiques qu'elle peut présenter ou non : large et/ou petite et/ou adeno et/ou squameuse, pour générer à l'issue seize profils type différents. Néanmoins, après une observation plus poussée de notre jeu de données, nous nous sommes rendu compte que chaque individu ne présentait qu'une unique caractéristique parmi les quatre, excluant les trois autres de facto.

De ce fait, nous avons procédé de la bonne façon dans notre manière de prendre en compte le type de cellule cancéreuse dans les modèles précédents. En effet, pour le modèle non-paramétrique, nous avons effectué des analyses séparées pour chaque type de cellule cancéreuse en stratifiant sur ces variables : nous avons estimé les courbes de survie et effectué des tests de comparaison de survie pour chaque type de cellule cancéreuse séparément, pour ainsi déterminer quelle caractéristique avait ou non un effet significatif sur la durée de vie des patients.

Par la suite, pour le modèle semi-paramétrique, en l'absence de la procédure stepwise imposée par la question concernée, qui nous a contraint dans notre choix de variables en ne gardant automatiquement que celles qui sont significatives, nous aurions inclus les variables binaires du type de cellule cancéreuse comme des covariables dans le modèle. Cela nous aurait ainsi permis de contrôler l'effet du type de cellule cancéreuse sur la durée de survie des individus tout en ajustant pour d'autres facteurs tels que l'âge, le type de traitement, l'indice de Karnofsky, etc. Nous aurions également pu explorer les interactions entre le type de cellule cancéreuse et les autres covariables, pour déterminer si l'effet de la cellule cancéreuse diffère selon le niveau des autres covariables.

9 – Modèle paramétrique

Pour finir, nous allons proposer un prolongement appliquant un modèle paramétrique à nos données.

Les modèles paramétriques se distinguent des autres par le fait que la loi de survie est connue à l'avance. Ainsi, ce modèle suppose une distribution particulière des données. Dans le cadre de l'économétrie des modèles de durée, cette distribution peut prendre trois formes : loi exponentielle (fonction de risque constant), loi de Weibull (croissant ou décroissant) et loi logistique.

De plus il existe deux familles de modèles, les modèles à risque proportionnel et les modèles à sorties

accélérées (AFT). Le modèle à sorties accélérées nous donne le facteur d'accélération de la survie en fonction des caractéristiques xi.

Dans ce prolongement, nous allons particulièrement nous intéresser à la variable "NTX" représentant le nouveau traitement, qui, comme nous l'avons vu, n'a eu jusqu'à présent aucun impact significatif sur le temps de survie dans nos estimations non-paramétriques et semi-paramétriques.

La première hypothèse que nous cherchons à vérifier est celle impliquant que le nouveau traitement serait meilleur que l'ancien sur certains types de cellules cancéreuses. Nous avons donc réalisé un modèle de Cox stratifié par type de cellule cancéreuse avec comme variable explicative la variable NTX. Ces quatre modèles ne sont pas concluants, car NTX n'est en effet pas significative dans chacun des modèles.

La seconde hypothèse à vérifier implique que le nouveau traitement est plus efficace lorsqu'il est administré plus tôt. Pour cela, nous allons utiliser la variable DD, qui correspond à la durée de la maladie en mois au moment où l'individu rentre dans l'étude. Comme nous l'avons vu dans nos statistiques descriptives, les individus ont en moyenne une maladie qui a duré 8,77 mois, cependant du fait de la présence de quelques valeurs atypiques (maximum : 87 mois), nous avons choisi la médiane qui est d'une valeur de 5 mois.

Nous allons donc créer un groupe NDD = 1 pour les personnes dont la maladie dure depuis 5 mois ou moins. Ensuite, nous estimerons avec un modèle de Cox stratifié par la variable NDD. Puis nous ajusterons progressivement en fonction de nos résultats.

Cependant, encore une fois, l'ensemble de nos résultats ne sont pas significatifs, même en diminuant progressivement (avec un pas de -1) les mois de 5 à 2, et en augmentant le seuil à 8. Nous avons également réalisé une estimation sur les individus qui ont la maladie depuis le plus longtemps (DD >= 15). Ainsi, cela montre que le nouveau traitement n'a aucun impact significatif sur le temps de survie, même en considérant les différences d'avancement (tôt comme tardif) de la maladie au moment du début du traitement entre les individus.

Notre troisième hypothèse est l'addition des deux précédentes, nous faisons l'hypothèse que le nouveau traitement traite mieux un patient dont le cancer est détecté plus tôt et certains types de cellules cancéreuses. Ainsi nous avons simplement croisé la variable NDD (= 1 si DD <= 5, 0 sinon) avec CT1, CT2 etc… Ensuite nous avons réalisé la même chose, c'est-à-dire quatre modèles de Cox stratifié par les variables croisées.

Pour la première fois, nous avons un résultat significatif pour la variable croisées NDDCT1 :

Analyse des valeurs estimées du maximum de vraisemblance								
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%	
NTX	1	1.16078	0.64873	3.2016	0.0736	3.192	0.895	11.385

Ce résultat est à relativiser car il n'est que significatif au seuil de 10 %. De plus, il est contradictoire car il semblerait que bénéficier du nouveau traitement, lorsque que l'on a un cancer depuis peu et de type large augmente de 3,192 fois la probabilité de connaître l'événement.

Nous avons également réalisé une cinquième étude, cette fois-ci en stratifiant par NTX=1 et en exprimant CT1, CT2, CT3, CT4 et NDD. Nous avons obtenu les résultats suivants :

Analyse des valeurs estimées du maximum de vraisemblance									
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%		Libellé
CT1	1	0.73959	0.40698	3.3024	0.0692	2.095	0.944	4.652	cell type 1 (1=large, 0=other)
CT2	1	1.43686	0.41680	11.8845	0.0006	4.207	1.859	9.524	cell type 2 (1=adeno, 0=other)
CT3	1	1.60758	0.42987	13.9852	0.0002	4.991	2.149	11.590	cell type 3 (1=small, 0=other)
DD	1	0.00389	0.00965	0.1630	0.6864	1.004	0.985	1.023	disease duration (months)

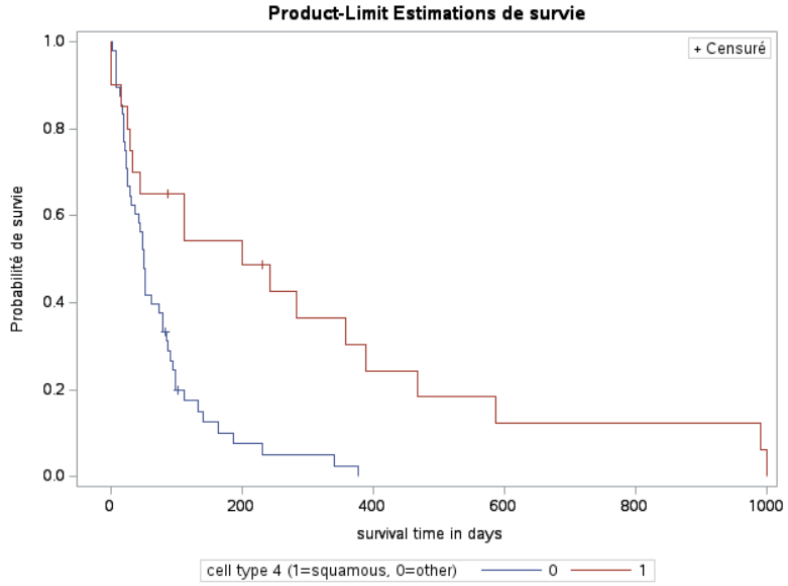
Analyse des valeurs estimées du maximum de vraisemblance									
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	khi-2	Pr > khi-2	Rapport de risque	Intervalle de conf. du rapport de hasard à95%		Libellé
CT4	1	-0.73959	0.40698	3.3024	0.0692	0.477	0.215	1.060	cell type 4 (1=squamous, 0=other)
CT2	1	0.69727	0.40059	3.0297	0.0818	2.008	0.916	4.404	cell type 2 (1=adeno, 0=other)
CT3	1	0.86799	0.41082	4.4640	0.0346	2.382	1.065	5.329	cell type 3 (1=small, 0=other)
DD	1	0.00389	0.00965	0.1630	0.6864	1.004	0.985	1.023	disease duration (months)

Ainsi, nous pouvons constater que l'ensemble de nos variables sont significatives (sauf DD) au seuil de 1% pour CT2 et CT3, 10 % pour CT1 et CT4.

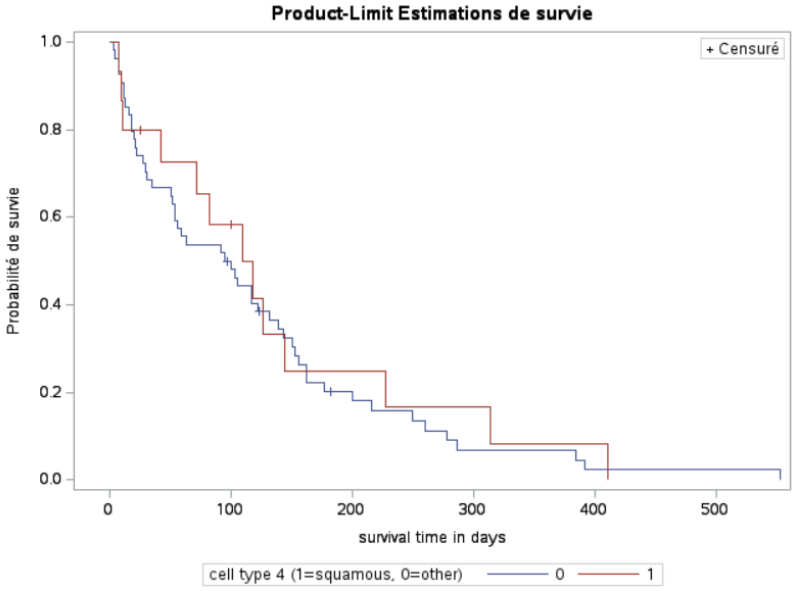
De plus, il est intéressant de constater que parmi les individus ayant bénéficié du nouveau traitement, avoir une cellule cancéreuse de type large, adeno et petite semble faire augmenter la probabilité de connaître l'événement. Cependant, on retrouve le phénomène inverse pour CT4. Ainsi, nous pouvons dire que parmi les personne bénéficiant du nouveau traitement, avoir une cellule cancéreuse squameuse fait diminuer la probabilité de connaître l'événement de 2,095 fois (e-0,73959*-1).

Via une estimation non-paramétrique, on trouve les résultats suivants :

Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	12.6895	1	0.0004
Wilcoxon	5.6064	1	0.0179
-2Log(LR)	26.1590	1	<.0001



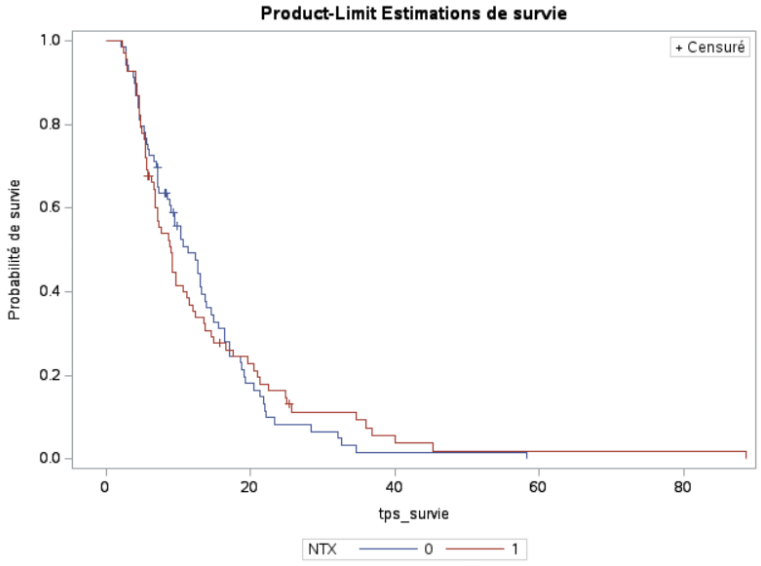
Si l'on compare avec le temps de survie lorsque NTX = 0, on trouve le résultat suivant :



Il y a donc clairement un impact du nouveau traitement sur le temps de survie.

Notre quatrième hypothèse concerne le temps de survie, en effet, l'étude dure longtemps (le maximum de la variable SURVT étant de 999 jours). Nous avons ainsi additionné la variable SURVT et la variable DD, nous avons alors le temps de survie complet à la maladie. Nous nous sommes servi de cette variable afin de remplacer SURVT puis nous avons réalisé une estimation non paramétrique en stratifiant avec NTX. Nous obtenons les résultats ci-dessous :

Test d'égalité sur les niveaux de discrétisation			
Test	khi-2	DDL	Pr > khi-2
Log-rang	0.0586	1	0.8087
Wilcoxon	0.4236	1	0.5151
-2Log(LR)	0.0473	1	0.8279



Cependant, encore une fois, les résultats ne sont pas significatifs, comme le montrent les tests.

Pour finir, nous avons donc réalisé un modèle paramétrique. Nous avons testé avec les trois types de modèles (exponentiel, Weibull et logistique), avec le modèle suivant : CT2 CT3 CT4 PERF NPERF NTX ntx_ct4 (ntx_ct4 étant une variable croisée). L'approche graphique nous a amené à choisir la loi exponentielle. Ensuite nous avons tâtonné jusqu'à trouver le modèle qui minimise les AIC et BIC.

Le modèle qui minime l'AIC, le BIC et l'exponential BIC est le suivant :

Analyse des paramètres estimés du maximum de vraisemblance							
Paramètre	DDL	Estimation	Erreur type	Intervalle de confiance à95%		Khi-2	Pr > khi-2
Intercept	1	3.3564	0.3497	2.6709	4.0418	92.10	<.0001
CT2	1	-0.6613	0.2740	-1.1984	-0.1242	5.82	0.0158
CT3	1	-0.4585	0.2251	-0.8996	-0.0174	4.15	0.0416
PERF	1	0.0281	0.0048	0.0186	0.0376	33.56	<.0001
NTX	1	-0.3726	0.2115	-0.7871	0.0418	3.11	0.0780
ntx_ct4	1	0.6590	0.3233	0.0254	1.2927	4.16	0.0415
Echelle	0	1.0000	0.0000	1.0000	1.0000		
Forme de Weibull	0	1.0000	0.0000	1.0000	1.0000		

Ainsi, l'ensemble de nos variables est significatif au seuil de 10 %. Nous pouvons donc conclure avec les interprétations suivantes. Le fait d'avoir bénéficié du nouveau traitement diminue le risque de 31 % : (e-0,32726 =06889). De plus l'augmentation d'un point de PERF a un coefficient de facteur d'accélération de 1,02.

10 – Quel(s) enseignement(s) à tirer de ces modèles ?

Ces différentes estimations nous amènent à mettre en lumière plusieurs conclusions. Dans un premier temps, nous avons d'abord pas réussi à isoler un effet significatif du nouveau traitement en fonction des différents types de cellules cancéreuses, ni en fonction de l'avancement de la maladie au moment de l'administration du traitement. Notre seul effet significatif (au seuil de 10%) relevé se révèle par ailleurs contradictoire, dans la mesure où bénéficier du nouveau traitement lorsque que l'on a un cancer depuis peu de temps et une cellule cancéreuse de type large simultanément augmenterait de 3,192 fois la probabilité de connaître l'événement.

Par la suite, nous avons pu déterminer qu'en stratifiant en fonction de la présence du nouveau traitement, avoir une cellule cancéreuse de type large, adeno ou petite semble augmenter la probabilité de connaître l'événement, tandis qu'avoir une cellule cancéreuse de type squameux semble diminuer cette probabilité, ce qui implique l'existence d'un impact du nouveau traitement sur le temps de survie.

Enfin, à l'issue de l'estimation de notre modèle paramétrique, nous avons été en mesure de déterminer que le fait d'avoir bénéficié du nouveau traitement diminue le risque de connaître l'évènement de 31 %.

11 – Bibliographie

Cox Proportional-Hazards Model - Easy Guides - Wiki - STHDA. http://www.sthda.com/english/wiki/cox-proportional-hazards-model. Consulté le 21 mars 2023.

Définition indice de Karnofsky. https://www.e-cancer.fr/Dictionnaire/I/indice-de-Karnofsky. Consulté le 21 mars 2023.

« Régression de Cox ». *Wikipédia*, 14 juillet 2022. *Wikipedia*, https://fr.wikipedia.org/w/index.php?title=R%C3%A9gression_de_Cox&oldid=195304215.

SAS Help Center: Example 89.12 Model Assessment Using Cumulative Sums of Martingale Residuals. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_phreg_examples12.htm#statug.phreg.phr10be. Consulté le 21 mars 2023.

Stensrud, Mats J., et Miguel A. Hernán. « Why Test for Proportional Hazards? » *JAMA*, vol. 323, n^{o} 14, avril 2020, p. 1401. *DOI.org (Crossref)*, https://doi.org/10.1001/jama.2020.1267.

Test de Kolmogorov. http://www.jybaudot.fr/Inferentielle/kolmogorov.html. Consulté le 21 mars 2023.

Test de Kolmogorov, test de Kolmogorov-Smirnov. https://www.bibmath.net/dico/index.php?action=affiche&quoi=./k/kolmogorovtest.html. Consulté le 21 mars 2023.