Mémoire d'initiation à la recherche

LIEN ENTRE LA TAILLE D'UNE TUMEUR ET LE TAUX DE MORT

Alexandre Benharrats, Rosalie Millner, Anouk Ruer

Dauphine | PSL
UNIVERSITÉ PARIS

Encadré par Robin Ryder et Yannick Viossat

M1 Mathématiques et Applications
Majeure Statistiques

2023-2024

Table des matières

1 Introduction et présentation du cadre théorique

Le cancer constitue depuis des années un défi complexe pour la communauté médicale, qui cherche continuellement à trouver des solutions efficaces et durables pour guérir les patients qui en sont atteints. Actuellement, l'approche traditionelle pour traiter le cancer consiste à administrer les traitements anticancéreux à la dose maximale tolérée, c'est-à-dire avant l'apparition d'effets secondaires trop importants, afin de maximiser la probabilité de guérison. On appelera ici MTD ("Maximal Tolerated Dose") ce type de traitement. Cependant, cette approche se heurte à un problème majeur : l'adaptation évolutive des cellules tumorales. En effet, on observe un mécanisme qui va favoriser la survie des cellules les plus résistantes, que l'on peut comparer à la sélection naturelle qui se produit lors de l'utilisation intensive de pesticides contre les espèces nuisibles à l'agriculture. Les traitements tels que la chimiothérapie détruisent uniquement les cellules qui y sont sensibles, laissant ainsi l'opportunité aux cellules résistantes de proliférer. S'il ne reste plus que des cellules résistantes, la tumeur devient incontrôlable et tout traitement ultérieur devient alors inefficace, conduisant inévitablement à la mort du patient. Nous nous plaçons ici dans le cadre simplifié où il n'y a qu'un seul type de traitement disponible. En réalité, lorsqu'un traitement n'est plus performant, on le remplace jusqu'à ce que plus aucun ne soit efficace.

Nous allons donc tenter de voir les cellules cancéreuses comme faisant partie d'un écosystème, bien que celui-ci soit complexe et difficile à comprendre. L'interaction entre les cellules sensibles et résistantes au sein d'une tumeur est un aspect crucial dans la compréhension de la dynamique évolutive du cancer et dans le développement de thérapies et de stratégies adaptatives.

1.1 Thérapie adaptative : une nouvelle approche dans la lutte contre le cancer

La thérapie adaptative propose une approche novatrice visant à contrôler et stabiliser la taille de la tumeur, plutôt que de viser à l'éliminer le plus rapidement possible. Elle repose sur une stratégie de traitement à doses réduites, administrées à différents intervalles de temps. Cette méthode permet d'adapter le traitement au comportement de la tumeur, et d'ajuster les doses ainsi que le moment de leur administration pour maintenir un équilibre délicat entre les cellules sensibles et résistantes.

L'idée générale est que la thérapie adaptative tire parti de la compétition naturelle entre les cellules sensibles et résistantes. L'objectif est d'éviter l'éradication précoce d'un nombre maximal de cellules sensibles afin de retarder l'émergence de la résistance, et donc de maximiser les chances de contrôler sa taille à long terme. Dans les bons cas, les traitements habituels peuvent éliminer massivement les populations sensibles, mais la véritable efficacité réside dans l'attaque

des cellules résistantes en capitalisant sur leur compétition avec les cellules sensibles.

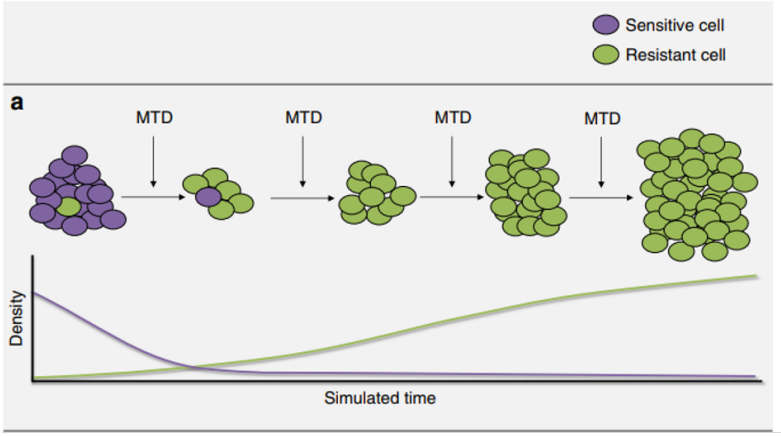


FIGURE 1 – Illustration de la thérapie standard MTD où le traitement est administré de manière continue. [4]

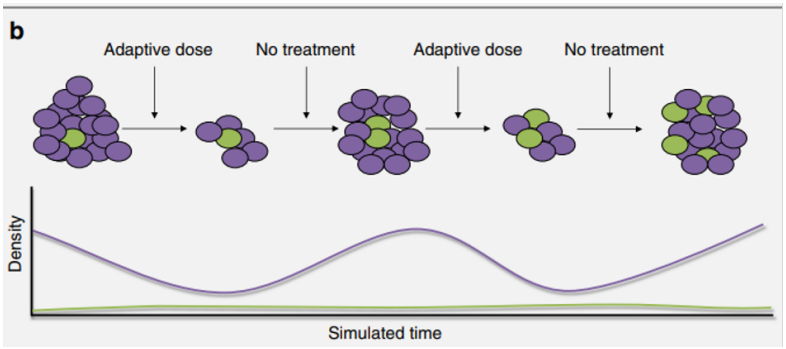


FIGURE 2 – Illustration de la thérapie adaptative où l'administration du traitement s'arrête avant que toutes les cellules sensibles ne soient éliminées.
[4]

Les figures 1 et 2 nous permettent de comparer l'impact théorique des deux types de thérapies sur nos populations de cellules. La thérapie adaptative semble offrir une perspective prometteuse en augmentant les chances de maintenir un contrôle durable sur la tumeur, surpassant ainsi les limites du traitement à dose maximale (MTD). Cette approche met donc bien en lumière l'importance de cette compétition entre les types de cellules, atteignant son efficacité maximale lorsque le nombre de cellules sensibles reste élevé.

Le graphique sur la figure 3 est une représentation simplifiée de ce à quoi pour-

rait ressembler l'évolution de la masse tumorale en fonction du temps selon les différentes thérapies. Dans ce modèle, on suppose que les cellules résistantes sont présentes dès le début, mais en quantités limitées.

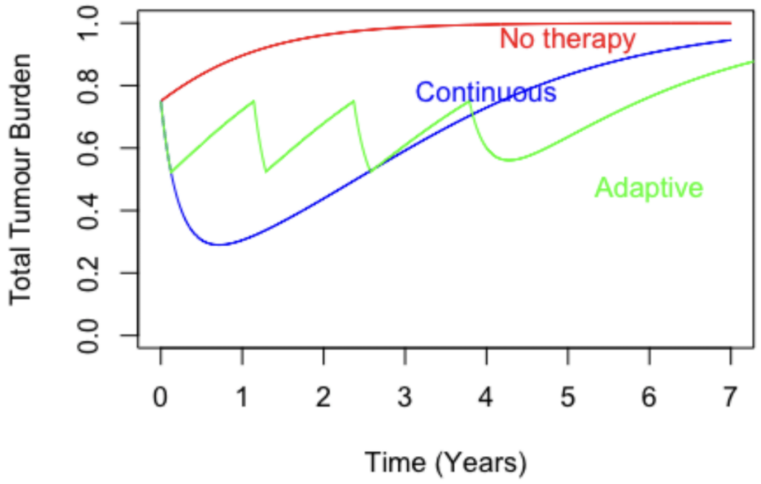


FIGURE 3 – Évolution de la charge tumorale en fonction du temps et du type de thérapie [2]

Sur le graphique (Figure 3) en bleu, nous observons l'effet d'une thérapie MTD : la charge tumorale baisse fortement au début lorsque le traitement intensif est en cours, mais la courbe finit par remonter strictement. Nous pouvons supposer qu'il s'agit du moment où la résistance émerge parmi les cellules tumorales, et que nous ne contrôlons plus la taille de la tumeur par le traitement. En revanche, en vert, nous avons l'effet d'une thérapie adaptative, sous laquelle nous observons une charge tumorale qui augmente et diminue plusieurs fois, fluctuation due à l'administration ou non du traitement. Nous semblons pouvoir mieux contrôler la taille de la tumeur à terme, ou du moins retarder l'émergence de la résistance.

1.2 Une thérapie remise en question : la critique de Mistry

En se référant à la figure 3, il semblerait que la thérapie adaptative devienne avantageuse seulement à partir du moment où la courbe verte passe en dessous de la courbe bleue (environ à l'année 4 en abscisse), autrement dit lorsque la charge tumorale sous une thérapie adaptative devient plus faible que sous une thérapie MTD. Or, cela n'a pas d'intérêt si le patient meurt avant de pouvoir atteindre ce stade.

Plusieurs membres de la communauté scientifique ont ainsi remis l'approche adaptative en question, ce qui est notamment le cas d'un chercheur de l'université de Manchester du nom de Hitesh Mistry, qui en a fait une critique dans une pré-publication datant de 2020 [2].

Selon lui, il y a un manque de considération du fait que les patients doivent survivre jusqu'à un certain temps, tout en portant une charge tumorale plus élevée, avant de pouvoir bénéficier pleinement du traitement. Il qualifie cette notion de *cumulative risk*. La thérapie adaptative ne prendrait pas en compte qu'avoir une plus grande taille de tumeur sur une période prolongée est corrélé à une probabilité de mort plus importante.

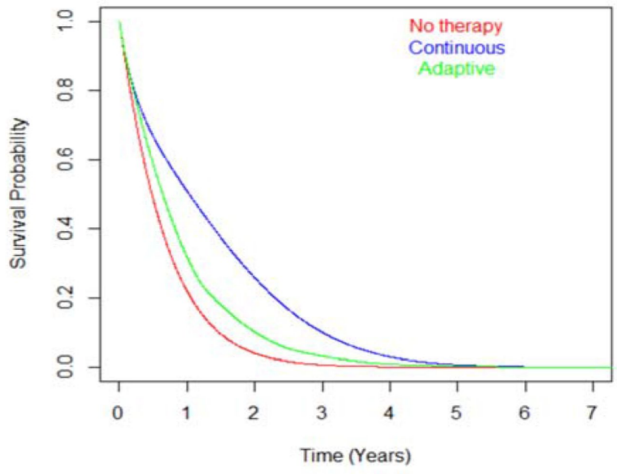


FIGURE 4 – Graphique représentant la probabilité de survie au cours du temps pour les différents types de thérapie d'après l'analyse de Mistry [2].

Selon sa conclusion, le traitement MTD réduit les risques de manière drastique en éradiquant rapidement une proportion maximale de la tumeur, ce qui le rend plus efficace que la thérapie adaptative. Il met en garde contre le fait que les systèmes pré-cliniques actuels ne capturent pas le concept de *cumulative risk* présent dans le système humain, soulignant ainsi la nécessité d'études basées sur de véritables données humaines pour évaluer correctement la supériorité de la thérapie adaptative par rapport au traitement MTD.

Hitesh Mistry met donc le doigt sur un aspect de la thérapie adaptative intéressant et qu'il est pertinent de considérer. Cependant, sa critique repose sur des hypothèses qui sont discutables. En particulier, il fait l'hypothèse que le taux de mort est proportionnel à la taille de la tumeur, sans apporter de justification à ce sujet.

Ainsi, la question du lien entre la taille de la tumeur et le taux de mortalité des patients n'est pas triviale. Les travaux de précédents étudiants ([3]) suggèrent que c'est un point essentiel, et que si le taux de mort n'est pas proportionnel à la taille de la tumeur mais à son carré ou à son cube, les thérapies adaptatives sont théoriquement plus prometteuses. C'est sur cette question que nous porterons nos recherches, en particulier sur la modélisation du taux de mortalité comme une fonction puissance de la taille de la tumeur, et potentiellement d'autres facteurs.

1.3 Schéma d'étude de ce mémoire

L'objectif de ce mémoire est de comprendre comment la taille de la tumeur influe sur le taux de mortalité des individus atteints du cancer. L'analyse de cette corrélation est un point central dans l'étude de l'efficacité d'une thérapie. Dans notre contexte, expliciter la manière de modéliser ce lien permettra de contribuer à une meilleure comparaison entre l'efficacité de la thérapie adaptative par rapport à la thérapie classique.

L'idée est d'explorer différents modèles statistiques qui, en se basant sur des données médicales adéquates, nous permettrons d'estimer au mieux le lien entre la taille de la tumeur et la mortalité.

Dans la suite de ce rapport, nous suivrons le cheminement de recherche que nous avons adopté pour traiter cette question. Dans un premier temps, nous présenterons en détails les données sur lesquelles nous avons basé nos recherches, puis dans un second temps les différentes manières de modéliser cette corrélation tout en précisant les contraintes et limites que chaque modèle peut apporter.

2 Présentation des données

Notre étude s'appuie sur un jeu de données provenant des travaux de Solène Desmée et al.[1], tiré de leur article intitulé "*Nonlinear joint models for individual dynamic prediction of risk of death using Hamiltonian Monte Carlo : application to metastatic prostate cancer*", publié en 2017. Ces données présentent de nombreux arguments en faveur de leur utilisation, comme la présence de covariables pertinentes sous forme d'observations longitudinales, ou encore la restriction de l'étude au périmètre du cancer de la prostate permettant une cohérence de l'analyse des observations entre les différents patients. Il est également à noter que Hitesh Mistry s'est servi de cette étude pour formuler sa critique.

Toutefois, bien que ce jeu de données semble répondre à nos critères de recherche, il est important de souligner qu'il s'agit de données simulées, générées par un modèle mathématique (qui lui, est entraîné à partir de vraies données). Contrairement à une analyse basée sur des données réelles collectées, la pertinence de notre étude dépend ici de la qualité des travaux de Desmée et al. [1] et de la capacité de son modèle à générer des données reflétant la réalité observable.

Dans l'optique de motiver notre choix de jeu de données, nous allons présenter brièvement l'étude de Desmée et al. [1], puis nous exposerons les differentes covariables sur lesquelles nous nous appuierons dans la suite de notre analyse.

2.1 Origine des données et introduction du *PSA*

2.1.1 Contextualisation de la provenance des données

L'article de Desmée et al. [1] vise à prédire le risque de décès chez les patients atteints de cancer de la prostate métastatique résistant à la castration (mCRPC), une forme agressive de cancer de la prostate qui présente une résistance aux traitements standards. Son analyse se base sur un groupe de patients atteints de mCRPC ayant été traités par chimiothérapie de docétaxel en association avec la prednisone, traitement usuel pour ce type de cancer. L'objectif principal de l'étude est de développer un modèle prédictif pour estimer le risque de décès chez les patients atteints de mCRPC en se basant sur les mesures répétées du taux d'antigène prostatique spécifique (PSA), un biomarqueur largement utilisé dans le suivi et le traitement du cancer de la prostate, reflétant l'activité tumorale.

À la suite de ses travaux, une simulation mathématique d'un suivi clinique a été réalisée pour N = 200 patients atteints du mCRPC. Cette simulation inclut une mesure du taux de PSA périodique de 21 jours jusqu'à la date simulée de décès ou de clôture de l'étude, soit sur une période de 1428 jours au maximum.

C'est sur ce jeu de données simulées que nous nous baserons dans la suite de

notre étude. Cependant, il est important de comprendre l'intérêt de l'analyse du PSA dans le cadre de notre travail de recherche sur le lien entre la taille de la tumeur et le taux de mortalité.

2.1.2 Le *PSA* : un indicateur pertinent de la taille de la tumeur

Le PSA, ou antigène prostatique spécifique, est une protéine produite naturellement par les cellules de la prostate. Celle-ci est un biomarqueur souvent utilisé pour évaluer l'état de santé de la prostate. Dans le cadre de patients atteints du cancer de la prostate, l'étude du taux de PSA permet d'avoir des informations sur la taille de la tumeur. En effet, un niveau élevé de PSA dans le sang va indiquer une taille de tumeur plus importante chez un individu, tandis qu'une baisse de ce taux sera un signe du déclin de la tumeur. Cette corrélation permet d'évaluer l'envergure de la tumeur sans en avoir des mesures explicites, qui nécessiteraient des tests d'imagerie plus conséquents.

Cependant, l'étude du PSA se heurte à un problème majeur : le coefficient de corrélation n'est pas uniforme pour tous les patients. En effet, pour une même taille de tumeur, la concentration de PSA dans le sang peut être très différente d'un individu à un autre. L'un des enjeux majeurs de notre étude a été de considérer cette importante variabilité interindividuelle.

2.2 Étude préliminaire : représentation des données et statistiques descriptives

Afin de nous familiariser avec les données et de mieux préparer notre schéma d'étude, nous avons effectué une analyse de chacune de nos covariables. En complément, l'élaboration de statistiques descriptives nous a permis de mieux appréhender la distribution et les caractéristiques de nos variables afin d'affiner notre sélection. De plus, nous avons également créé certaines covariables supplémentaires à partir de nos variables d'intérêt, dans le but de construire le modèle le plus complet possible.

	ID	TIME	r	PSA0	E	Tesc	eta_r	eta_PSA0	eta_E	eta_Tesc	res_error	logPSA	death_day	status
22	1	441	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	0.1795452394	4.809915052	724.44422	1
23	1	462	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	-0.8031730660	3.723311129	724.44422	1
24	1	483	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	0.1344636924	4.757363913	724.44422	1
25	1	504	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	0.4882600067	5.207668646	724.44422	1
26	1	525	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	0.4846065822	5.300607609	724.44422	1
27	1	546	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	0.2302760127	5.142945730	724.44422	1
28	1	567	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	-0.1218694372	4.887538294	724.44422	1
29	1	588	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	-0.4141758414	4.692032878	724.44422	1
30	1	609	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	-0.1123082484	5.090758659	724.44422	1
31	1	630	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	0.5752608239	5.875237869	724.44422	1
32	1	651	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	-0.2685802896	5.128354068	724.44422	1
33	1	672	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	0.0050286935	5.498963200	724.44422	1
34	1	693	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	-0.0093611581	5.581612389	724.44422	1
35	1	714	0.05063908	66.024247	0.5598747	63.40761	-0.06240677	-0.1126908	0.9263099	-0.7776698	-0.5841567734	5.103891123	724.44422	1
36	2	0	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	0.2894926145	4.721716806	312.38614	1
37	2	21	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	-0.5872755481	3.860738740	312.38614	1
38	2	42	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	0.5657161940	5.033564210	312.38614	1
39	2	63	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	-0.6349525146	3.852765498	312.38614	1
40	2	84	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	-0.0402177364	4.467375014	312.38614	1
41	2	105	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	0.2432243613	4.770696255	312.38614	1
42	2	126	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	-0.1517270442	4.395628312	312.38614	1
43	2	147	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	0.1909699476	4.758213001	312.38614	1
44	2	168	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	0.5797451094	5.166880013	312.38614	1
45	2	189	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	-0.0564894299	4.550541394	312.38614	1
46	2	210	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	-0.8634302924	3.763500444	312.38614	1
47	2	231	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	0.7016031360	5.348437699	312.38614	1
48	2	252	0.05806657	83.118304	0.1913234	277.48260	0.07445962	0.1175521	-0.7557771	0.6985045	0.1855399949	4.852282223	312.38614	1

FIGURE 5 – Échantillon des données simulées

2.2.1 Les covariables

Les données extraites de l'étude de Desmée et al. [1] sont des données dites longitudinales : elles présentent l'évolution temporelle des mesures du taux de PSA de patients souffrant d'un cancer de la prostate. Chaque individu, parmi les 200, dispose d'une mesure de son taux de PSA tous les 21 jours pendant 1428 jours, soit la durée de l'étude, sous réserve que celui-ci reste en vie. Cela implique nécessairement que pour les patients restés en vie jusqu'au bout, nous ne disposons pas de données supplémentaires : ceci correspond à la notion de "censure" des données. Celle-ci se produit lorsque le temps de suivi pour un individu se termine sans que l'évènement d'intérêt ne se soit produit. Dans notre cas, l'étude a une durée totale de 1428 jours, donc les patients pour lesquelles la mort n'a pas encore été observée sont dits "censurés" à la fin de l'étude. Nous sommes ainsi face à certaines données qui sont incomplètes, pouvant affecter l'estimation de la distribution de survie et les taux de survie parmi nos individus.

La table 1, présente de manière concise chacune des variables. Parmi celles-ci, plusieurs covariables semblent appropriées pour notre étude, comme l'identifiant du patient, le temps écoulé depuis le début du traitement, ou encore le taux de prolifération des cellules prostatiques. Toutefois, certaines covariables incluses dans le jeu de données telles que eta_r, res_error, etc., servent uniquement à la simulation des données. Nous excluons ces covariables de notre analyse ultérieure, car elles sont spécifiques à la simulation et non pertinentes pour notre étude.

TABLE 1 – Data Dictionary : Description et sélection des covariables

Nom de la covariable	Description	Type	Requis	Source
ID	Identifiant du patient	Factor	YES	dataPSA.txt
TIME	Nombre de jours depuis le début du traitement	Integer	YES	dataPSA.txt
r	Taux de prolifération journalier des cellules prostatiques en l'absence de traitement (constant par individu)	Float	YES	dataPSA.txt
PSA0	Valeur du PSA au début du traitement (mesuré en ng/mL de sang)	Float	YES	dataPSA.txt
E	Effet constant du traitement (constant par individu)	Float	YES	dataPSA.txt
Tesc	Jour où le traitement n'est plus considéré efficace, dans le sens où la taille de la tumeur augmente malgré l'administration du traitement	Integer	YES	dataPSA.txt
eta_r	Variable de simulation	Float	NO	dataPSA.txt
eta_PSA0	Variable de simulation	Float	NO	dataPSA.txt
eta_E	Variable de simulation	Float	NO	dataPSA.txt
eta_Tesc	Variable de simulation	Float	NO	dataPSA.txt
res_error	Variable de simulation	Float	NO	dataPSA.txt
logPSA	Transformation log(PSA)	Float	YES	dataPSA.txt
death_day	Jour de la mort du patient	Integer	NO	dataPSA.txt
status	Survie (0)/décès (1) de l'individu entre t=0 et la fin de l'étude	Logical	YES	dataPSA.txt

2.2.2 Ajustement des données et construction de variables

Il est important de justifier notre choix de conserver la variable *logPSA*, c'est à dire la transformation logarithmique du *PSA*, plutôt que simplement le *PSA*. Comme expliqué dans la section 1.2, nous cherchons à étudier une relation de type puissance entre le taux de mortalité et la taille de la tumeur, exprimée par cette variable, plutôt qu'une relation simplement proportionnelle. Les différents modèles que nous allons utiliser par la suite supposent une relation linéaire des coefficients avec nos covariables d'intérêt. Ainsi, en optant pour cette transformation, l'estimation du coefficient β dans l'expression $\beta \cdot \log(\text{PSA})$ revient à estimer $\log(\text{PSA}^{\beta})$, ce qui correspond à notre objectif de recherche.

Cependant, en vue de la construction d'un modèle pertinent et pour rendre l'étude plus intelligible, nous avons dû modifier la structure de certaines covariables et en créer de nouvelles.

La variable *status*, qui précisait simplement la survie ou non de l'individu sur toute la période d'observation, a été ajustée de manière à ce qu'à chaque mesure, elle puisse indiquer si l'individu décède avant la prochaine observation, soit dans les 21 jours. Elle vaut 1 lorsque ce décès est observé, 0 sinon. Seule la dernière ligne de chaque individu est donc éligible à avoir un *status* à 1.

Le PSA initial (*PSA0*) a également été transformé en *logPSA0* pour une meilleure concordance avec les mesures périodiques du PSA, qui sont déjà transformées en logarithme (*logPSA*).

La variable *Tesc* représentant l'instant où le traitement cesse de faire effet sur la tumeur du patient a également été adaptée pour concorder avec la discrétisation de la variable *TIME*. Une transformation a été appliquée pour arrondir *Tesc* à l'entier le plus proche inférieur, puis multipliée par 21 pour s'aligner avec les intervalles de temps de *TIME*. Cette transformation est exprimée mathématiquement comme suit :

$$\left\lfloor \frac{\text{Tesc}}{21} \right\rfloor \times 21$$

Pour garantir une précision maximale, cette transformation n'a pas été appliquée aux bases de données utilisées pour générer les statistiques descriptives.

Différentes covariables ont également été développées puis ajoutées à nos données. Voici la liste détaillée :

- *vitesse* : représente l'accroissement entre 2 mesures consécutives du *logPSA*, elle est calculée comme le taux d'accroissement du *logPSA* pour chaque période.
- *ecart* : représente l'écart entre le *logPSA* et le *logPSA0* pour chaque mesure. Elle est calculée par la différence "*logPSA* - *logPSA0*".
- *proportion* : représente la proportion de *logPSA* par rapport au *logPSA0* initial. Elle est calculée par le rapport "*logPSA*/*logPSA0*".
- *TimeAbove* : pour chaque mesure, indique depuis combien de temps le taux de PSA est resté supérieur au taux initial. Mesurée en période, elle est calculée comme la somme des périodes consécutives où $logPSA > logPSA0$. Si pour une mesure donnée $logPSA \leq logPSA0$, elle retourne à 0.
- *TimeBelow* : indique la durée pendant laquelle le taux de PSA est resté inférieur au taux initial. Elle est calculée selon le même principe que *TimeAbove*.
- *PSAVariability* : représente la dispersion du taux de PSA pour chaque individu. Il s'agit de l'écart type du *logPSA*, et est constant par individu.
- *PSAStatus* : variable binaire indiquant le statut du taux de PSA selon qu'il est supérieur ou inférieur au taux initial. Elle vaut 0 si $logPSA > logPSA0$, 1 sinon.

2.2.3 Statistiques descriptives

L'idée ici est de comprendre plus en profondeur la structure de nos variables d'intérêt et d'essayer de soulever des informations profitables pour la suite de notre étude.

Nous avons inclus un corrplot (Figure 6), également connu sous le nom de graphique de corrélation, pour visualiser les relations entre les différentes variables de notre ensemble de données. Le corrplot (Figure 6) nous permet d'observer les coefficients de corrélation entre chaque paire de variables, offrant ainsi un aperçu rapide des relations entre les variables. Cette visualisation est essentielle pour comprendre la structure des données et identifier les variables potentiellement corrélées, ce qui peut être crucial dans le processus de sélection des variables à inclure dans notre futur modèle.

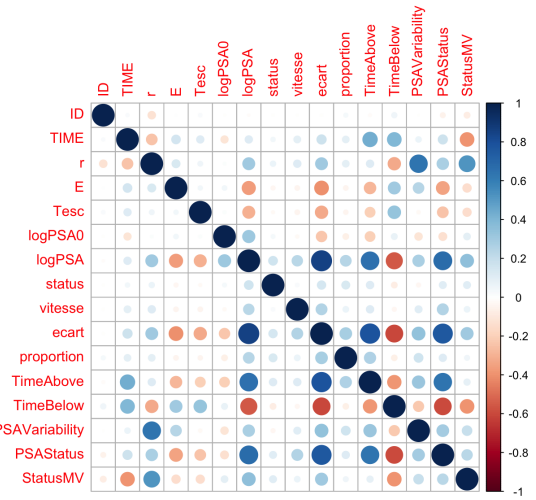


FIGURE 6 – Corrplot de nos variables d'intérêt

Nous observons dans le corrplot (Figure 6) que la corrélation entre la mortalité d'un individu, exprimée par la variable *status*, et son taux de PSA, exprimé par la variable *logPSA*, est nettement plus marquée qu'avec les autres covariables. On remarque également que les autres variables possédant un lien plus ou moins fort avec le *status* sont toutes significativement correlées avec le *logPSA*. Cela vient confirmer notre choix d'étudier le taux de PSA comme variable explicative du taux de mortalité et de l'inclure dans nos futurs modèles.

Dans la table 2, nous pouvons trouver les valeurs extrêmes et quartiles de certaines de nos covariables :

	TIME	*r*	*E*	*Tesc*	*logPSA0*	*logPSA*	*PSAVariability*	*ecart*	*proportion*	*vitesse*
Min.	0.0	0.04219	0.01846	15.82	0.7576	-1.243	0.0000	-7.61997	0.000	-2.1611
1st Qu.	189.0	0.04964	0.18975	108.03	3.0568	1.908	0.8981	-1.63311	0.195	-0.3010
Median	420.0	0.05220	0.39433	157.34	4.1926	3.996	1.3266	-0.23442	0.791	0.0440
Mean	492.6	0.05329	0.41280	181.89	4.1081	4.084	1.5267	-0.02399	59.233	0.0600
3rd Qu.	735.0	0.05687	0.61760	223.94	5.2364	5.833	2.0738	1.39478	4.034	0.4197
Max.	1428.0	0.07134	0.95285	778.16	7.9327	14.066	3.4243	10.16091	25871.930	2.9635

TABLE 2 – Extremums et quartiles

Il est intéressant de noter plusieurs choses sur chaque covariable :
— On observe grâce à la variable *PSAVariability* une hétérogénéité nette des écart de taux de PSA chez les différents patients.
— La valeur maximale de *TIME* est 1428, ce qui indique qu'il y a des patients qui sont en vie à la fin de l'étude : il s'agit de la censure.
— *Tesc* varie entre 16 jours et 778 jours environ, ce qui montre une grande disparité de l'efficacité du traitement chez les individus.

Intéressons-nous maintenant à deux populations distinctes, à savoir les individus ayant survécu jusqu'à la fin de l'étude et ceux décédés durant. L'idée est d'essayer de discerner des singularités sur certaines covariables qui seraient propres à chaque groupe. Dans le table 3 nous pouvons voir comment sont répartis les patients entre ceux décédés et ceux qui ont survécu.

TABLE 3 – Répartition décès/survie

Statut	Nombre
Survivants	31
Décédés	169

Formulé autrement, cela signifie que seul 15,5% de la population initale a survécu au moins jusqu'à la fin de l'étude.

	logPSA		*PSAVariability*		*ecart*		*vitesse*		*Tesc*	
	Survivants	*Décédés*	*Survivants*	*Décédés*	*Survivants*	*Décédés*	*Survivants*	*Décédés*	*Survivants*	*Décédés*
Min	-1.2289	-1.243	0.4082	0.0000	-6.3401	-7.6200	-2.161129	-2.04546	55.55	15.82
1er Quartile	2.1553	2.856	0.7309	0.9238	-3.1651	-1.0746	-0.339305	-0.28333	118.73	98.13
Médiane	2.6279	4.608	1.0890	1.4634	-1.2399	0.1075	0.007579	0.06176	167.18	142.41
Moyenne	4.1843	4.716	1.2880	1.6303	-1.2672	0.5154	0.001430	0.08542	207.38	170.84
3ème Quartile	5.833	6.381	1.8092	2.3045	0.1807	1.9765	0.348048	0.46130	241.06	203.42
Max	10.5150	14.066	2.9605	3.4243	8.0028	10.1609	1.743345	2.96346	624.83	778.16

TABLE 4 – Comparaison des variations du taux de PSA chez les patients "Survivants" et "Décédés"

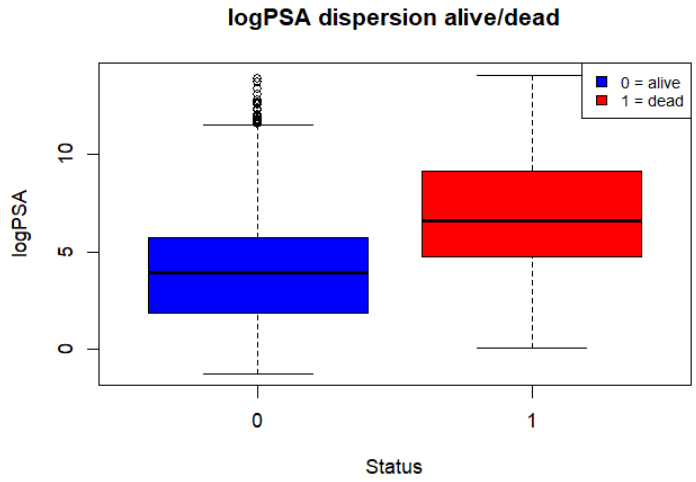


FIGURE 7 – Dispersion du logPSA en fonction du statut du patient

Nous observons dans, la table 4 et sur le boxplot (Figure 7), que le taux de PSA pour les individus "décédés" sont en moyenne plus élévés que chez les patients "survivants". Cependant, il faut garder à l'esprit qu'il existe une grande variabilité inter-individuelle de la relation taux de PSA/taille de la tumeur. Ces observations sont donc à analyser avec précaution.

Néanmoins, lorsque l'on s'intéresse à la variable *ecart*, on observe que la différence entre le log du taux de PSA à chaque mesure et le log du taux de PSA initial (PSA0) est bien plus élévée chez les individus "décédés" que "survivants" : 75% de la population des "survivants" a un écart inferieur à 0.1807 ainsi qu'un écart moyen négatif de -1.2672. En comparaison, la population des "décédés" est à 50% au dessus de 0.1075 dont 25% au dessus de 1.9765. Pour rappel, un écart négatif suggère une amélioration de l'état de la prostate avec un taux de PSA inférieur au taux au moment où le traitement à été commencé, et inversement. De plus, l'étude de la variable *PSAVariability*, qui représente l'écart type par individu, suggère une dispersion des taux de PSA plus élévée pour chaque individu appartenant à la population des "décédés" que chez les individus appartenant à celle des "survivants", indiquant des mouvements plus amples des taux de PSA et donc une plus grande variabilité de la taille de la tumeur. Enfin, les taux de variations exprimés par la variable *vitesse* nous indiquent des mouvements en moyenne extrêmement faibles de l'ordre du millième au sein de la population "survivants", et plus importants chez les "décédés". Cela indique que la population des "décédés" connaît également des variations plus rapides des taux de PSA, en cohérence avec le fait d'avoir des variations plus amples.

En définitive, bien que l'importance absolue du taux de PSA puisse avoir une in-

fluence sur le décès des patients, nos analyses suggèrent que ce sont l'amplitude et la vitesse de la croissance du taux de PSA qui jouent un rôle déterminant dans le décès d'un individu. Il semble également exister une corrélation significative entre un taux de PSA supérieur au taux de PSA initial et le décès de l'individu. Cela paraît cohérent étant donné qu'un taux de PSA supérieur au taux de PSA initial suggère que la tumeur s'est développée.

Cette analyse semble confirmée par l'observation graphique (Figure 8) de l'évolution du *logPSA* dans le temps selon le statut final de l'individu. En bleu sont représentés les individus "survivants" et en vert les individus "décédés". Les points rouges indiquent l'instant où le traitement arrête de faire effet sur la tumeur du patient (covariable *Tesc*).

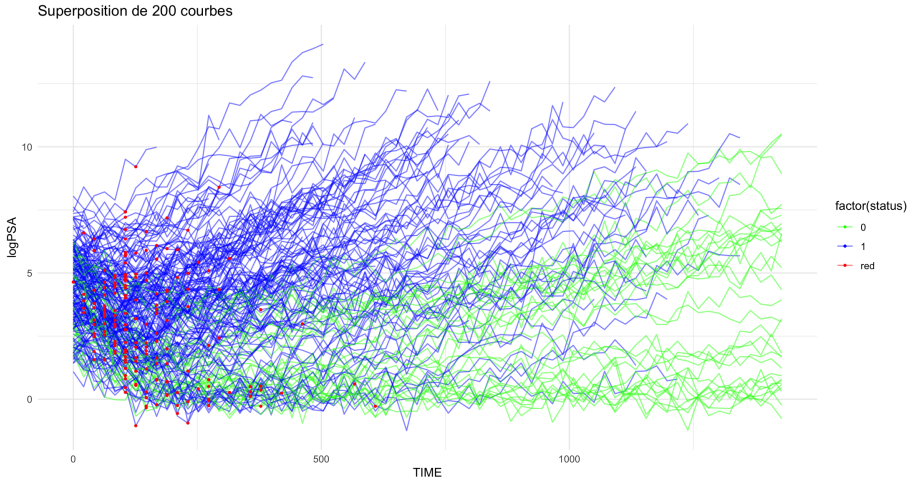


FIGURE 8 – Évolution du *logPSA* dans le temps pour chaque population de patients

3 Vers une modélisation aboutie : comparaison entre modèles de régression et modèles de survie

Après une étude approfondie de nos différentes variables d'intérêt, nous avons entrepris une analyse des différents modèles exploitables pour représenter les évènements survenants dans nos données. Nous avons débuté notre analyse en utilisant des modèles de régression logistique simples pour comprendre les relations entre nos variables d'intérêt et la survie des individus. Cependant, compte tenu de la nature longitudinale de nos données et de l'objectif de prédire le temps jusqu'à l'événement, nous avons progressivement évolué vers des modèles de survie plus sophistiqués. Cette transition nous a permis de capturer de manière plus précise les dynamiques temporelles des événements et d'identifier les facteurs associés à leur occurrence. Dans ce contexte, nous examinons les différences, les avantages et les limitations des modèles de régression par rapport aux modèles de survie, dans le but de parvenir à une modélisation optimale de nos données.

3.1 Modèle de régression linéaire généralisée - la régression logistique

Nous avons amorcé notre analyse en utilisant un modèle de régression linéaire généralisée de type régression logistique, afin d'explorer la relation entre la taille de la tumeur et le taux de mortalité chez les patients présents dans nos données. L'idée est d'estimer la probabilité de mourir d'un individu en fonction de certaines données qui lui sont propres, mais indépendemment du temps.

3.1.1 Fonction de lien logit : un choix approprié

Nous visons ici à prédire le statut d'un patient en fonction de son taux de PSA et de diverses autres covariables. La variable *status* est binaire : elle renvoie 1 si le patient est décédé et 0 sinon. Étant donné la nature de la variable de réponse, la régression logistique est l'approche appropriée pour modéliser cette relation.

Concrètement, le lien logit permet de s'assurer que la sortie du modèle reste dans l'intervalle [0,1], ce qui est nécessaire pour modéliser une probabilité. Lorsque le modèle est appliqué à une situation binaire (par exemple, survie ou décès), la sortie du modèle est la probabilité de survie (ou de décès) pour chaque observation.

Le lien logit est exprimé comme suit :

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

où p est la probabilité d'un événement, ici la probabilité de décès. L'expression $\frac{p}{1-p}$ est le rapport des cotes (odds ratio), représentant le rapport des chances de succès par rapport aux chances d'échec. Dans notre cas, ce rapport représente le rapport des chances de mourir par rapport aux chances de survivre.

3.1.2 Formulation mathématique du modèle

La formulation de notre modèle GLM basé sur la régression logistique est la suivante :

$$P(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ik})$$

Où :
- Y_i représente la variable binaire *status* indiquant le décès du i-ème patient,
- X_{ij} pour $i \in [1, n]$ et $j \in [1, k]$ (n étant le nombre de patients et k le nombre de covariables sélectionnées) est la valeur de la j-ème covariable pour l'individu i,
- β_0 est l'intercept,
- β_j pour $j \in [1, k]$ sont les coefficients de régression associés aux covariables.

3.1.3 Implémentation R du modèle

Nous avons initialement appliqué ce modèle simple à toutes les variables que nous avons jugées pertinentes pour notre étude, puis réduit la dimension de ce modèle en utilisant un processus de sélection de variables basé sur la minimisation du critère d'Akaike (AIC). Ce critère reflète la capacité d'un modèle à atteindre un équilibre entre la pertinence de la prédiction et la complexité du modèle.
Le modèle obtenu pour chaque observation est le suivant :

$$\begin{aligned}\text{logit}(p) = \hat{\beta}_0 + \hat{\beta}_1 \times \mathit{logPSA} + \hat{\beta}_2 \times \mathit{logPSA0} + \hat{\beta}_3 \times \text{r} + \hat{\beta}_4 \times E + \hat{\beta}_5 \times \mathit{Tesc} \\ + \hat{\beta}_6 \times \mathit{vitesse} + \hat{\beta}_7 \times \mathit{TimeBelow} + \hat{\beta}_8 \times \mathit{PSAVariability} \quad (1)\end{aligned}$$

où :
- p représente la probabilité de décès (dans les trois semaines qui suivent, c'est-à-dire avant la prochaine mesure),
- $\hat{\beta}_j$ pour $j \in [1, k]$ sont les coefficients estimés par le modèle que l'on retrouve dans la table 5.

TABLE 5 – Résumé du modèle de régression logistique généralisée (GLM)

	Estimate	Std. Error	z value	Pr(> |z|)	Signif.
(Intercept)	−8.5738011	0.996	-8.603	$< 2 \times 10^{-16}$	***
logPSA	1.3618803	0.096	14.183	$< 2 \times 10^{-16}$	***
logPSA0	−1.0193095	0.094	-10.769	$< 2 \times 10^{-16}$	***
r	87.3909491	19.0556	4.586	4.52×10^{-16}	***
E	4.6973189	0.4885	9.616	$< 2 \times 10^{-16}$	***
Tesc	0.0018978	0.00077	2.457	0.014007	*
vitesse	−0.5914070	0.163	-3.627	0.000286	***
TimeBelow	0.0310261	0.0991	3.128	0.001758	**
PSAVariability	−4.0002815	0.3019	-13.248	$< 2 \times 10^{-16}$	***

3.1.4 Interprétation des résultats

Sous réserve de la validité des hypothèses de ce modèle, on déduit une relation significative entre les variables énoncées précédemment et le statut. En particulier, le coefficient $\hat{\beta}_1$ associé au *logPSA* montre une relation entre la probabilité de mourir et la taux de PSA puissance 1,36.

Il est à noter que contrairement aux hypothèses initiales de Mistry [2], ce modèle a évalué plus pertinent de prendre en compte d'autres facteurs en plus de la taille de la tumeur pour expliquer le taux de mortalité.

Certaines valeurs paraissent contre-intuitives comme la relation négative de *status* avec la variance du PSA (*PSAVariability*) et la *vitesse*. Mais la pertinence de ce modèle est fortement contestable.

3.1.5 Analyse de la validité des hypothèses et limites du modèle

La validité d'un tel modèle est sujette à plusieurs hypothèses, dont certaines ne sont pas vérifiées dans notre contexte, ce qui remet en cause la pertinence de nos résultats.

En premier lieu, les modèles linéaires généralisés supposent l'indépendance entre les observations, considérant que les valeurs observées pour chaque ligne de données ne sont pas influencées par celles des autres. Or, nous disposons de plusieurs observations pour un même individu à different instants ce qui contredit notre hypothèse.

De plus, comme mentionné précédemment, chaque individu présente un taux de PSA relatif à la taille de sa tumeur qui lui est propre : deux patients distincts avec des tumeurs de même taille pourraient avoir des taux de PSA sensiblement différents. Cette variation interindividuelle n'est pas prise en compte par le modèle linéaire généralisé.

En conclusion, bien que ce modèle soit simple à mettre en œuvre, ses limites et le non-respect des hypothèses sous-jacentes nous incitent à envisager l'application d'un modèle plus sophistiqué et mieux adapté à notre situation.

3.2 Modèle linéaire généralisé mixte

Nous cherchons maintenant à estimer le statut d'un patient en fonction de son taux de PSA et d'autres covariables en intégrant dans notre modèle les effets aléatoires liés aux différentes observations pour chaque patient.
Le modèle linéaire généralisé mixte (GLMM) étend le modèle linéaire généralisé (GLM) pour tenir compte du rattachement des observations à des individus distincts.

3.2.1 Modèle mixte et intégration des effets aléatoires

Pour mieux cerner le fonctionnement et l'utilité d'un tel modèle dans notre situation, considérons un exemple simple d'application. Imaginons un échantillon constitué de réponses à un sondage réalisé auprès de membres de différentes communautés. Chaque membre de la communauté répond à plusieurs questions, et ces réponses sont enregistrées comme des observations distinctes dans les données. Dans ce scénario, nous nous attendons à ce que la variation résiduelle des réponses (c'est-à-dire la variation qui n'est pas expliquée par les prédicteurs) ne soit pas indépendante d'une observation à l'autre. En d'autres termes, les réponses des individus au sein d'une même communauté seront probablement plus similaires entre elles que les réponses d'individus provenant de communautés différentes. Cette similitude est due à des facteurs non mesurés qui varient au niveau de la communauté plutôt qu'au niveau de l'individu. Dans cette situation, le modèle mixte permettrait de prendre en compte cette corrélation entre les observations au sein d'une même communauté en permettant aux coefficients du modèle de varier d'une communauté à l'autre, tout en tenant compte des effets fixes des prédicteurs sur la variable réponse.

Les modèles linéaires généralisés à effets mixtes (GLMM pour generalized linear mixed models) fusionnent les caractéristiques des modèles linéaires généralisés et des modèles linéaires mixtes. Les coefficients du prédicteur linéaire varient de manière aléatoire entre les groupes, suivant une distribution normale. En permettant aux coefficients du modèle de varier d'un groupe à l'autre, le GLMM traduit cette situation de variabilité inter-individuelle.

3.2.2 Formulation mathématique du modèle

La formulation du modèle mixte est la suivante :

$$P(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i)$$

Où :

— Y_i représente la variable binaire *status* indiquant le décès du i-ème patient
— X_{ij} pour $i \in [1, n]$ et $j \in [1, k]$ (n étant le nombre de patients et k le nombre de covariables sélectionnées) est la valeur de la j-ème covariable pour l'individu i,
— β_0 est l'intercept,
— β_j pour $j \in [1, k]$ sont les coefficients de régression associés aux covariables,
— ε_i est l'effet aléatoire propre au i-ème individu, représentant la variation individuelle non expliquée par les covariables fixes. Celui-ci est souvent supposé suivre une distribution normale $\mathcal{N}(0, \sigma^2)$ pour un certain σ.

3.2.3 Implémentation R du modèle

Nous avons effectué une régression sur R pour un GLMM à l'aide du package `lme4`.
Nous avons voulu tester un modèle avec toutes les covariables sauf que R nous renvoie un avertissement comme quoi les covariables sont sur des échelles très différentes. Par exemple, la covariable *TIME* s'étend de 0 à 1428 alors que la covariable r est généralement d'ordre 10^{-3}. Pour résoudre ce problème, nous avons choisi de multiplier r et E par 100.
Dans la table 6 et 7 figurent les résultat obtenus pour les effets aléatoires, ainsi que pour les effets fixes du modèle.

Table 6 – Effets aléatoires

Groupes	**Nom**	**Variance**	**Écart-type**
ID	(Intercept)	0	0

Table 7 – Effets fixes

	Estimate	Std. Error	z value	Pr(> |z|)	Signif.
(Intercept)	−8.5738011	0.9575	-8.954	$< 2 \times 10^{-16}$	***
logPSA	1.3618803	0.0960	14.185	$< 2 \times 10^{-16}$	***
logPSA0	−1.0193095	0.0946	-10.770	$< 2 \times 10^{-16}$	***
r	87.3909491	18.1575	4.813	1.49×10^{-6}	***
E	4.6973189	0.4884	9.617	$< 2 \times 10^{-16}$	***
Tesc	0.0018978	0.0008	2.469	0.013535	*
vitesse	−0.5914070	0.1630	-3.629	0.000285	***
TimeBelow	0.0310261	0.0099	3.138	0.001703	**
PSAVariability	−4.0002815	0.3018	-13.256	$< 2 \times 10^{-16}$	***

3.2.4 Interprétation des résultats

On constate que les estimations des coefficients $\hat{\beta}$ sont exactement les mêmes pour le modèle mixte et pour le modèle classique. En effet, le maximum de vraisemblance pour la variance de l'effet aléatoire est atteint pour $\sigma = 0$. Les effets aléatoires étant centrés, ils sont donc nécessairement nuls. Nous obtenons ainsi un modèle mixte qui est équivalent au modèle classique précédent.

3.2.5 Analyse de la validité des hypothèses et limites du modèle

Bien que ce modèle permette d'intégrer les disparités inter-individuelles du *logPSA*, certaines hypothèses ne sont toujours pas vérifiées.

Contrairement au modèle linéaire généralisé, le modèle mixte va établir une indépendance des observations par groupe en rattachant chaque ligne de données à un individu propre, ce qui constitue une amélioration par rapport au modèle précedent. Toutefois, au sein d'un même groupement d'observations, les données vont être considérées par le modèle comme des observations indépendantes les unes des autres, ce qui n'est pas le cas. En effet, de par leur nature longitudinale, chaque observation propre à un individu est directement conditionnée par les observations précédentes. Par exemple, la valeur de la variable *TimeAbove*, qui cumule le nombre de périodes pendant lesquelles le taux de PSA est supérieur au taux initial, dépend directement des lignes d'observations qui la précèdent.

On peut donc affirmer que ce modèle, malgré qu'il soit plus fidèle à la réalité observable, reste peu pertinent et toujours critiquable.

3.3 Modèles de survie : approches statistiques pour l'analyse des événements

3.3.1 Approche heuristique de l'analyse de survie

L'analyse de survie est un ensemble de méthodes statistiques dans lequel on étudie le temps jusqu'à ce qu'un évènement d'intérêt se produise. L'objectif principal est d'étudier la distribution des temps de survie et d'identifier les facteurs qui influent sur ces temps. Dans le cadre de notre mémoire, nous nous intéressons donc au temps avant l'évènement de la mort d'un patient atteint du cancer de la prostate.

Un des éléments principaux d'une analyse de survie est la fonction (ou courbe) de survie. Soit T la variable aléatoire représentant le temps de survie, $F(t) = \mathbb{P}(T \leq t)$ sa fonction de répartition et $f(t) = \frac{dF(t)}{dt}$ sa densité, pour tout réel positif t.
La **fonction de survie** est donnée par $S(t) = \mathbb{P}(T > t) = 1 - F(t)$, elle nous donne la probabilité que le temps de survie soit supérieur à t.

D'autre part, une composante importante de l'analyse de survie est la **fonction de risque instantané**. Celle-ci permet de quantifier le taux de risque instantané de décès auquel un individu est exposé à un temps donné t, conditionnellement au fait qu'il ait survécu jusqu'à ce temps t.

Elle est définie comme :

$$\begin{aligned}\lambda(t) &:= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \cdot \frac{\mathbb{P}(t \leq T < t + \Delta t)}{\mathbb{P}(T \geq t)} \\ &= \frac{f(t)}{S(t)}\end{aligned}$$

Cette fonction de risque instantanée est toujours positive mais peut prendre des valeurs supérieures à 1. On ne peut donc pas, à strictement parler, la voir comme une probabilité. C'est pourquoi on parle plutôt de risque.

Enfin, la notion de **censure** des données, définie précédemment, est un concept important en analyse de survie. Toutes les méthodes d'analyse de survie que nous présenterons dans la suite de notre rapport sont conçues pour tenir compte de la censure afin de produire des estimations précises du temps de survie. Nous allons notamment présenter une courbe de survie de Kaplan-Meier et des modèles de régression de Cox, qui sont capables d'intégrer la censure des données dans le calcul des estimations de survie.

3.3.2 Représentation de la fonction de survie : la courbe de Kaplan-Meier

En général, on ne connaît pas la forme théorique des courbes de survie, mais on peut les approximer. La courbe de survie de Kaplan-Meier (Figure 9) est une courbe échelonnée qui permet de représenter empiriquement la fonction de survie $S(t)$ au fil du temps t. À chaque intervalle de temps observé, la courbe de Kaplan-Meier "chute" en proportion du nombre de décès observé à ce moment-là.

Celle-ci est tracée comme suit : à chaque intervalle de temps, on compte le nombre de décès survenus jusque là, que l'on divise par le nombre de patients

encore en vie. Cela nous donne une proportion que l'on peut alors tracer graphiquement en fonction du temps, nous donnant une représentation de la probabilité de survie empririque des patients depuis le début du traitement. Pour tenir compte de la censure, la courbe de Kaplan-Meier calcule les probabilités en ne considérant que les individus non censurés à ce moment-là. Ainsi, la proportion d'individus encore en vie à chaque instant est ajustée pour refléter le nombre d'individus réellement "à risque" à ce moment précis.

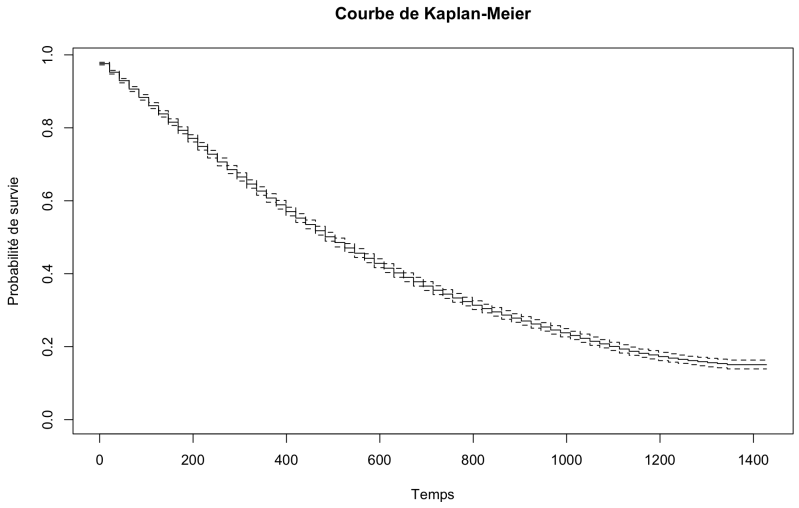


FIGURE 9 – Courbe de Kaplan-Meier appliquée à nos données

Le temps (en jours) est représenté en abscisse et la probabilité de survie en ordonnée. La courbe de Kaplan-Meier (Figure 9) ne prend donc en compte aucune potentielle covariable explicative à part le temps de la survie du patient.

Celle-ci reste tout de même simple à interpréter et peut être un bon moyen de comparer la performance des types de thérapies, en comparant notamment les proportions d'individus survivants sous une thérapie MTD ou adaptative. Par exemple ici, dans nos données sur des patients qui suivent une thérapie MTD, la médiane de survie est à 504, ce qui signifie que la moitié des patients ont survécu au moins 504 jours après le début du traitement. Ce type de courbe est intuitif dans sa compréhension et peut facilement résonner auprès des médecins.

3.4 Un modèle de survie classique : le modèle de régression de Cox

3.4.1 Présentation du modèle

Le modèle de régression de Cox est l'un des modèles les plus couramment utilisés en analyse de survie.
Dans notre contexte, ce modèle permet de mesurer l'effet des covariables sur le risque d'un patient de mourir à un instant t, en intégrant le fait que celui-ci ait survécu jusqu'à cet instant. Ainsi, il se démarque des précedents modèles qui ne prenaient pas en compte ce conditionnement.

Pour ce faire, le modèle de Cox se formalise de la manière suivante :

$$\lambda(t) = \lambda_0(t) \cdot \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$$

où :

- λ est la fonction de risque instantané,
- λ_0 est la fonction de risque de base. Il s'agit du risque instantané lorsque toutes les covariables sont nulles,
- $X_1, ..., X_k$ sont les covariables (k étant le nombre de covariables),
- β_j pour $j \in [1, k]$, sont les coefficients de la régression associés aux covariables. Ils représentent l'effet de chaque covariable sur le risque instantané de décès.

Ainsi, cette formule peut être séparée en 2 parties : celle de gauche dépendant du temps, et celle de droite ne dépendant pas du temps.

Le modèle de Cox est dit semi-paramétrique. En effet, on ne cherche pas à estimer la fonction λ_0 qui est la même pour tous les individus. On s'intéresse plutôt au rapport des risques instantanés de décès pour 2 individus exposés à des facteurs de risque différents (i.e. ayant des valeurs de covariables différentes).

Il est par ailleurs clair que dans ce modèle, lorsque l'on prend le *logPSA* comme variable, on va bien obtenir un risque instantané de mort qui est proportionnel au PSA à une certaine puissance, ce qui concorde avec nos objectifs de recherche.

3.4.2 Hypothèse principale du modèle

L'hypothèse fondamentale du modèle de Cox est celle de **proportionnalité des risques**. En d'autres termes, le rapport des risques instantanés entre deux individus reste constant au fil du temps.

Pour illuster cette hypothèse, prenons deux individus i_0 et i_1 qui ne diffèrent que par une seule covariable, disons la j-ème :

$$\frac{\lambda(t,i_0)}{\lambda(t,i_1)} = \frac{\lambda_0(t)\exp(\beta_1 X_1 + \ldots + \beta_{j-1}X_{j-1} + \beta_j X_j^{i_0} + \beta_{j+1}X_{j+1} + \ldots + \beta_k X_k)}{\lambda_0(t)\exp(\beta_1 X_1 + \ldots + \beta_{j-1}X_{j-1} + \beta_j X_j^{i_1} + \beta_{j+1}X_{j+1} + \ldots + \beta_k X_k)}$$

$$= \exp\{\beta(X_j^{i_0} - X_j^{i_1})\}.$$

Le terme risques proportionnels fait référence à la situation où ce rapport de risque dépend seulement de la différence $(X_j^{i_0} - X_j^{i_1})$ et non pas du temps lui-même. Cela implique donc que l'effet des variables explicatives sur le risque de décès est stable dans le temps.

Ce rapport nous permet d'introduire une nouvelle notion intéressante en analyse de survie : le **hazard ratio**. Dans le même scénario que dans le rapport des risques ci-dessus, si la différence entre la covariable pour les deux individus vaut 1, alors la quantité $exp(\beta)$ permet de quantifier à quel point une augmentation d'**une unité** de la variable en question augmente ou diminue le risque instantané de décès. On a en effet :

$$\frac{\lambda(t,i_0)}{\lambda(t,i_1)} = \exp\{\beta\}$$

$$\Leftrightarrow$$

$$\lambda(t,i_0) = \exp\{\beta\} \cdot \lambda(t,i_1)$$

3.4.3 Un indice de comparaison : la concordance

Pour comparer la précision de différents modèles de Cox, on peut utiliser le coefficient de **concordance**. Fréquemment utilisé en analyse de survie pour évaluer l'exactitude prédictive des modèles, il quantifie de 0 à 1 la capacité d'un modèle à prédire dans le bon ordre le décès des individus dont il connaît les temps de mort.

La concordance est calculée comme suit :
Pour chaque paire d'individus dans le jeu de données, une prédiction individuelle du temps de décès est effectuée par le modèle. Elles sont ensuite classées par ordre de décès, en fonction de quel individu est prédit de décéder en premier.

On procède ensuite à une comparaison entre le classement prédit par le modèle et celui observé dans les données. Si le classement est le même, la paire est dite *concordante*, sinon elle est *discordante*. Cependant, si une paire comporte deux individus tous deux censurés, elle est exclue du calcul de la concordance car leur ordre de survie relatif n'est pas déterminé par les données. En revanche, si seulement l'un des individus de la paire est censuré, il est toujours possible d'évaluer la concordance en fonction de la durée de survie observée. Le coefficient de concordance est alors la proportion suivante :

$$\begin{aligned} C &= \frac{\text{Nombre de paires concordantes}}{\text{Nombre total de paires dont on connaît l'ordre}} \\ &= \frac{2}{N(N-1) - N_c(N_c-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \text{Concordance}(i,j) \end{aligned}$$

où :
— C représente le coefficient de concordance,
— N est le nombre total d'individus dans l'échantillon,
— N_c est le nombre total d'individus censurés dans l'échantillon,
— Concordance(i,j) vaut 1 si la paire (i,j) est concordante et 0 sinon.

La fonction Concordance(i,j) peut être définie comme suit :

$$\text{Concordance}(i,j) = \begin{cases} 1 & \text{si } T_i < T_j \text{ et } O_i < O_j \text{ et } S_i = 0 \\ & \text{ou si } T_i > T_j \text{ et } O_i > O_j \text{ et } S_j = 0 \\ 0 & \text{sinon} \end{cases}$$

où T_i et T_j représentent les temps de survie prédits des individus i et j respectivement, O_i et O_j représentent les temps de survie observés des individus i et j respectivement (en fixant $O_i = 1428$ si l'individu i est censuré), et enfin S_i et S_j représentent leurs statuts de censure (1 pour censure, 0 pour décès observé dans les données).

Ce coefficient varie donc de 0 à 1, où 1 représente une concordance parfaite (toutes les paires sont concordantes) et 0,5 représente une performance aléatoire (autant de paires concordantes que discordantes). Une valeur inférieure à 0,5 indique une performance pire que le hasard. Nous cherchons donc à obtenir un modèle de survie ayant une concordance la plus proche possible de 1.

Cependant, cet indice présente une limite qu'il est important de prendre en compte dans notre étude. Maximiser la concordance revient à pousser le modèle vers un ajustement très fort aux données. Par conséquent, cela peut entraîner un risque de surajustement qui fausserait alors le modèle. Le coefficient de concordance est donc un indicateur pertinent de la qualité du modèle, mais pas suffisant.

3.4.4 Implémentation R du modèle et résultats

Nous avons effectué notre régression de Cox à l'aide du package R `survival`. Pour obtenir le modèle le plus pertinent possible, nous avons procédé en deux étapes :
· Élaboration d'un modèle minimisant le critère d'information d'Akaike.
Pour ce faire, nous avons appliqué un `step` à partir d'un modèle formé par l'ensemble de nos covariables d'intérêt, puis retiré celles désignées comme non-significatives.
`concordance = 0,917  AIC = 1992`
· Élaboration d'un modèle maximisant l'indice de `concordance`.
`concordance = 0,928`
Le modèle minimisant l'AIC ayant une concordance extrêmement proche du modèle construit pour maximiser cet indice, nous avons décider de le sélectionner pour notre étude. Ainsi, nous avons obtenu un modèle avec une concordance maximale tout en prenant en compte le risque de surajustement. Il s'agit du modèle présenté dans la table 8.

Variable	Coefficient	Exp(Coefficient)	Erreur standard	Valeur-z	Pr(>|z|)	Signif
logPSA	8.750e-01	2.399e+00	1.030e-01	8.497	$< 2 \times 10^{-16}$	***
r	2.776e+02	3.529e+120	1.871e+01	14.834	$< 2 \times 10^{-16}$	***
logPSA0	-5.828e-01	5.584e-01	1.158e-01	-5.031	4.89×10^{-07}	***
proportion	7.101e-05	1.000e+00	3.903e-05	1.819	0.068889	.
PSAVariability	-4.665e+00	9.424e-03	2.929e-01	-15.927	$< 2 \times 10^{-16}$	***
TimeAbove	-1.126e-01	8.935e-01	2.405e-02	-4.679	2.88×10^{-06}	***
logPSA :E	3.602e-01	1.434e+00	5.797e-02	6.213	5.18×10^{-10}	***
E :vitesse	-1.586e+00	2.047e-01	3.434e-01	-4.619	3.86×10^{-06}	***
logPSA :TimeAbove	1.074e-02	1.011e+00	3.115e-03	3.447	0.000566	***

TABLE 8 – Coefficients estimés du modèle de Cox retenu

3.4.5 Interprétation des résultats

Ce modèle suggère que de nombreuses covariables sont pertinentes à prendre en compte dans l'étude du risque de décès. En particulier, on observe que le coefficient estimé $\hat{\beta}_1$ associé au *logPSA* vaut 0.875 . Cela suppose donc un lien puissance relativement proche de 1, ce qui est confirmé par l'étude de l'incertitude. Pour un niveau de 95%, l'intervalle de confiance de l'estimation du coefficient associé au *logPSA* vaut environ :

$$[\hat{\beta}_1 \pm se(\hat{\beta}_1) \times q^{\mathcal{N}(0,1)}_{97,5\%}] = [0,673; 1,077]$$

L'intervalle de confiance contenant la valeur 1, il n'apparaît donc pas déconcertant de modéliser le lien entre le taux de PSA et la probabilité de décès de façon proportionnelle, mais à condition d'inclure un grand nombre de variables supplémentaires.

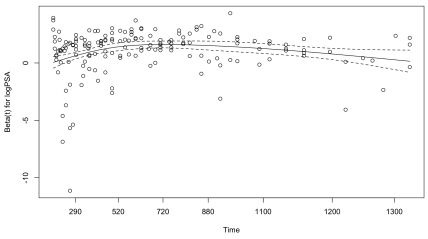
En conclusion, ce modèle suggère que malgré une relation plus ou moins proportionnelle, d'autres facteurs que la taille de la tumeur sont explicatifs du taux de mortalité des patients atteints du cancer de la prostate.

3.4.6 Validité du modèle

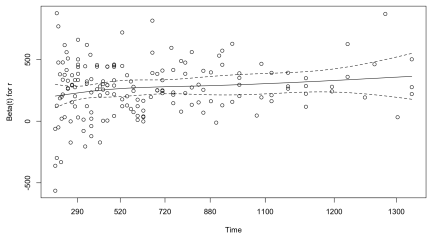
Le but ici est principalement d'analyser si l'hypothèse de proportionnalité de risques est vérifiée.

Pour cela, nous avons utilisé le test `cox.zph` du package `survival`. Ce test permet de calculer les résidus de Schoenfeld pour chaque variable explicative du modèle. Ces derniers mesurent la différence entre les valeurs observées et les valeurs attendues des coefficients de régression.

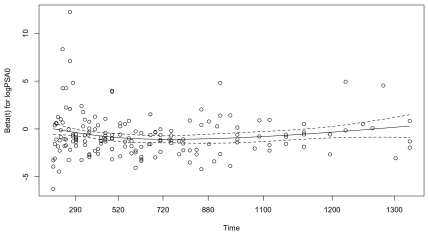
Nous avons visualisé les résultats de cette fonction à l'aide de graphiques (Figure 11). Ceux-ci affichent les résidus en fonction du temps pour chaque variable. Si les lignes de régression sont approximativement horizontales, alors il n'existe pas de tendance claire en fonction du temps et l'on peut considérer que l'hypothèse de proportionnalité des risques est respectée.



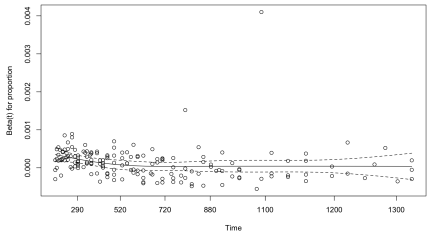
(a) Résidus de Schoenfeld pour logPSA



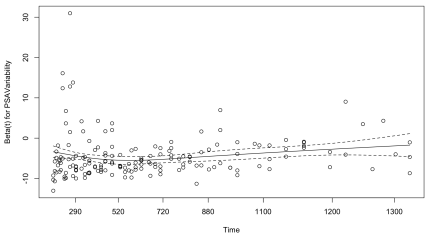
(b) Résidus de Schoenfeld pour r



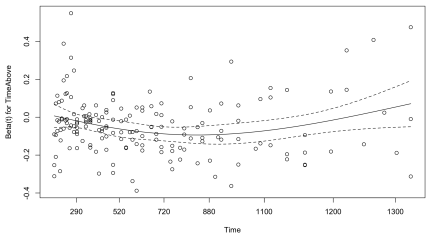
(c) Résidus de Schoenfeld pour logPSA0



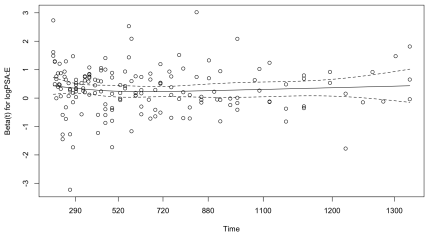
(d) Résidus de Schoenfeld pour proportion



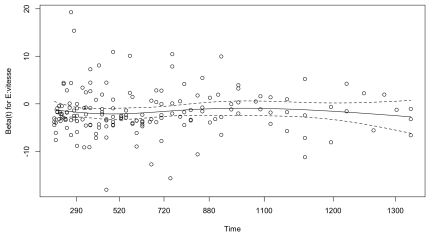
(a) Résidus de Schoenfeld pour PSAVariability



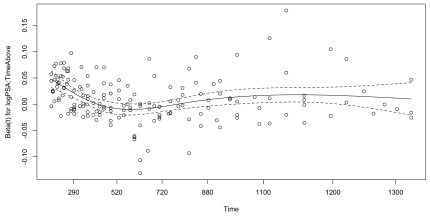
(b) Résidus de Schoenfeld pour TimeAbove



(c) Résidus de Schoenfeld pour logPSA : E



(d) Résidus de Schoenfeld pour vitesse : E



(e) Résidus de Schoenfeld pour logPSA : TimeAbove

FIGURE 11 – Graphiques des résidus pour chaque variable explicative du modèle issus de cox.zph pour tester l'hypothèses de proportionnalité des risques

Ainsi, nous obtenons des courbes qui semblent valider le postulat des risques proportionnels, sauf éventuellement à des échelles très petites.

Cependant, comme pour le modèle linéaire généralisé classique, le modèle de régression de Cox ne prend pas en compte la variabilitié inter-individuelle du taux de PSA relatif aux individus. La fiabilité de notre modèle est donc toujours contestable. De la même manière que précédement, nous allons étudier un modèle qui intègre l'effet mixte au modèle de Cox : le modèle de Cox mixte.

3.5 Vers une modélisation plus réaliste : le modèle de régression Cox mixte

Le modèle de Cox mixte combine les avantages du modèle de Cox avec ceux des modèles mixtes. Il est particulièrement utile dans les situations où les données de survie présentent une corrélation intra-groupe, ce qui est notre cas ici avec la variable *logPSA*, comme expliqué précédemment.

3.5.1 Formulation mathématique du modèle

La fonction de risque instantané pour l'individu i dans le modèle de Cox mixte est donnée par :

$$\lambda_i(t) = \lambda_0(t) \exp(\beta^T X_i + u_i)$$

Où :
- $\lambda_0(t)$ est la fonction de risque de base. On ne cherche pas à l'estimer, et elle est considérée comme identique pour tous les individus de l'échantillon,
- β est le vecteur des coefficients de régression pour les covariables fixes,
- X_i est le vecteur des covariables fixes pour l'individu i,
- u_i est l'effet aléatoire propre au i-ème individu, représentant la variation individuelle non expliquée par les covariables fixes. Celui-ci est souvent supposé suivre une distribution normale $\mathcal{N}(0, \sigma^2)$ pour un certain σ. En d'autres termes, u_i représente les écarts aléatoires par rapport à la moyenne pour chaque individu i.

En ajustant un modèle de Cox mixte, nous estimons à la fois les coefficients de régression β pour les covariables fixes et le paramètre σ^2 pour les effets aléatoires.

3.5.2 Implémentation R du modèle et résultats

Le modèle de Cox mixte peut être implémenté sur R à l'aide du package `coxme`.
Nous avons cherché le modèle nous permettant de minimiser l'AIC au maximum. Nous avons obtenu le modèle présenté dans les tables 9 et 10, ayant un AIC de 1728, qui est donc inférieur à l'AIC du modèle de Cox simple élaboré dans la section précédente :

TABLE 9 – Mixed effects

	sd	variance
Random effects :		
Intercept	3.809027	14.50868

TABLE 10 – Fixed effects

	coef	exp(coef)	se(coef)	p
Fixed effects :				
logPSA	0.683597	1.980991	0.114700	2.52e-09
logPSA : TimeAbove	-0.006410	0.993610	0.002876	0.0258

3.5.3 Interprétation des résultats

On remarque immédiatement que le modèle obtenu est bien moins complexe que ceux mis en place précédemment. Il ne retient comme variable que le *logPSA* et l'interaction *logPSA : TimeAbove*.
Nous avons jugé nécessaire de mener une étude plus approfondie de ce résultat, notamment en raison de sa grande divergence par rapport à notre précédent modèle, le modèle de Cox, qui intégrait un grand nombre de variables explicatives.

Pour tenter de comprendre au mieux ce résultat et de légitimer notre choix de conserver ce modèle, nous avons implémenté un modèle de Cox mixte ayant les mêmes covariables que le modèle de Cox, listées dans la table 8.
Nous obtenons un modèle où :

- le coefficient estimé associé à la variable *logPSA* vaut environ 0,8. Il est proche de celui estimé par notre modèle de Cox mixte présenté en table 10.
- les coefficients de régression estimés associés aux autres variables sont soit très proches de 0, soit jugés non-significatifs par le modèle.

Cela nous pousse donc à privilégier le modèle de Cox mixte qui minimise l'AIC.

Ce modèle nous invite donc à considérer que le risque de décès est une fonction du logPSA élévé à la puissance 2/3 environ, avec $\hat{\beta}_1$ `= 0.683597`. D'autre part, le coefficient $\hat{\beta}_2$ associé à l'interaction entre *TimeAbove* et *logPSA* étant très proche de 0, nous pouvons affirmer que ce facteur ne sera pas fortement explicatif dans le cadre de notre étude. Nous pouvons également ajouter que le modèle nous renvoie un intervalle de confiance de niveau 95% pour le coefficient $\hat{\beta}_1$ associé au *logPSA* qui vaut environ :

$$[\hat{\beta}_1 \pm se(\hat{\beta}_1) \times q^{\mathcal{N}(0,1)}_{97,5\%}] = [0,455; 0,905]$$

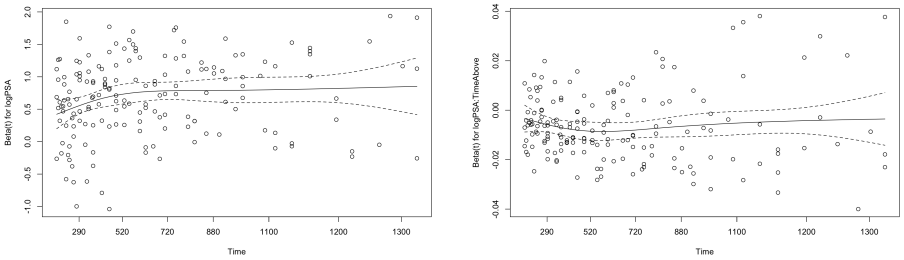
L'étude de l'incertitude nous montre qu'il n'est pas hors de propos de modéliser ce lien avec un β qui se rapproche de 1.

L'étude du hazard ratio est également intéressante. Exprimé par `exp(coef)`, et ici de valeur `1.980991`, ce dernier mesure l'impact de l'écart d'une unité de

logPSA sur le risque instantané de décès. Un tel coefficient indique donc qu'un patient ayant un logPSA supérieur d'1 unité par rapport à un autre verra son risque instantané de décès multiplié par presque 2.

3.5.4 Validité du modèle

De la même manière que pour le modèle de Cox simple, nous devons vérifier si notre modèle satisfait l'hypothèse de proportionnalité des risques. Nous nous basons à nouveau sur les graphiques 12a et 12b issus de la fonction `cox.zph`.



(a) Résidus de Schoenfeld pour logPSA

(b) Résidus de Schoenfeld pour logPSA :TimeAbove

FIGURE 12 – Graphiques des résidus pour chaque variable explicative du modèle de Cox mixte issus de cox.zph

Nous considérons que les courbes obtenues sont suffisamment horizontales pour valider le postulat de proportionnalité des risques.

4 Conclusion

Au cours de notre analyse, nous avons développé différents modèles, chacun présentant des hypothèses plus ou moins conformes à notre contexte d'étude.

Le modèle qui semble être le plus adapté à notre situation est le modèle de Cox mixte. En plus d'être un modèle de survie adéquat à ce genre de données, il a l'avantage de prendre en compte la disparité inter-individuelle du taux de PSA. Nous pouvons donc légitimement considérer que c'est ce modèle qui va décrire les évènements de la manière la plus fidèle à la realité.

Plus généralement, nos différents modèles de Cox (simple et mixte) suggèrent que le risque instantané de décès est une fonction du taux de PSA à une puissance relativement proche de 1. Ce résultat, soutenu par l'étude des niveaux de confiance, semble plutôt tendre vers l'hypothèse de relation proportionnelle d'Hitesh Mistry. Notre analyse ne nous permet en tous cas pas d'affirmer que cette relation se fasse avec une puissance supérieure à 1.

Cependant, les modèles que nous avons mis en place montrent que l'étude du risque de décès ne se réduit pas simplement à l'analyse de la taille de la tumeur. Selon les modèles, il semble plus pertinent d'inclure un plus grand nombre de covariables, ou bien d'intégrer un effet aléatoire individuel. Dans cette mesure, notre étude s'éloigne de l'hypothèse d'Hitesh Mistry.

Il faut malgré tout garder à l'esprit que notre analyse est entièrement basée sur un jeu de données qui non seulement est simulé, mais également limité à 200 patients. De plus, notre étude de la taille de la tumeur se fait par le biais d'un indicateur tiers, le PSA, dont la relation n'est pas explicitement établie, et bien qu'il paraisse pertinent, comporte tout de même ses propres limites. Par conséquent, il est nécessaire d'interpréter ces résultats avec prudence et de reconnaître les limites de notre étude. Une approche plus exhaustive, impliquant un plus grand nombre de patients, pourrait fournir des perspectives supplémentaires sur cette relation complexe.

Références

[1] Solène Desmée, France Mentré, Christine Veyrat-Follet, Bernard Sébastien, and Jérémie Guedj. Nonlinear joint models for individual dynamic prediction of risk of death using hamiltonian monte carlo : application to metastatic prostate cancer. *BMC Medical Research Methodology*, 2017.

[2] Hitesh Mistry. Evolutionary based adaptive dosing algorithms : Beware the cost of cumulative risk. 6 2020.

[3] Anis Toussirt and Alexis Emanuelli. Mémoire. 2022.

[4] Jingsong Zhang, Jessica J. Cunningham, Joel S. Brown, and Robert A. Gatenby. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nature Communications*, 2017.