RAPPORT DU PROJET ANALYSE DE SURVIE

2024 - 2025

Rédiger à
L'Université de Paris Saclay

par
ATHOUMANI Ibroihima

Survival analysis of mortality of HIV patients

Table des matières

1 Introduction

Ce projet vise à analyser la survie des données longitudinales dans une étude qui déterminera la mortalité des patients atteints du VIH. Le VIH, ou virus de l'immunodéficience humaine, affaiblit le système immunitaire en détruisant les cellules T-lymphocytes, en particulier les cellules CD4. L'évolution de la maladie peut être surveillée par le comptage des cellules CD4, un biomarqueur essentiel indiquant le déclin de la santé d'un individu, sensiblement jusqu'à l'apparition du SIDA. Dans cette étude longitudinale, des patients VIH ayant échoué ou ne tolérant pas la thérapie AZT (zidovudine) ont été répartis aléatoirement en deux groupes de traitement, ddI (didanosine) et ddC (zalcitabine), dans le but de comparer l'efficacité et la sécurité de ces antirétroviraux.

L'objectif du projet est de répondre à des questions de recherche précises, notamment : Quels sont les facteurs influençant la survie des patients ? Existe-t-il une différence entre l'efficacité des traitements ? Quelle est l'association entre le taux de cellules CD4 et le risque de mortalité ?

2 Modélisation et méthodologie

Dans cette partie, nous allons expliciter les méthodes utilisées pour atteindre les objectifs fixés.

2.1 Préparation des données

Cette première étape consiste à charger les données et à vérifier que chaque variable est bien définie et correctement formatée. Cela permet d'identifier immédiatement les éventuelles incohérences. Les valeurs manquantes ou les doublons peuvent fausser les résultats de l'analyse. Il est donc essentiel de les détecter et de les traiter en générant une méthode appropriée, comme l'imputation.

Dans notre cas on a constaté que il n'y a pas des valeurs manquantes ni des doublons.
Ce pendant, pour mieux faire les statistiques sur les patients, je juge nécessaire de créer un dataset que je nommerai 'unique-data', en supprimant les colonnes CD4 et time-obs. Cela nous évite de compter les répétitions des patients en fonction des observations de la quantité des CD4 chaque deuxième mois.

2.2 Statistiques descriptives et visualisation

cette étape de statistiques descriptives et de visualisation fournit une vue d'ensemble des données et des tendances, permettant d'identifier les motifs et les anomalies avant d'approfondir avec des analyses statistiques ou des modèles plus complexes. Elle est essentielle pour bien comprendre la structure des données et orienter les analyses suivantes.

2.2.1 Statistiques descriptives

subject	time	death	treatment	sex	prev_infection	azt
Min. : 1.0	Min. : 0.47	0:279	1:237	0: 45	0:307	0:292
1st Qu.:117.5	1st Qu.:10.23	1:188	2:230	1:422	1:160	1:175
Median :234.0	Median :13.20					
Mean :234.0	Mean :12.63					
3rd Qu.:350.5	3rd Qu.:16.23					
Max. :467.0	Max. :21.40					

FIGURE 1 – Statistique descriptive

Ce tableau nous donne une vision globales sur nos données en décrivant chaque colonne. En se basant sur ce dernier on sait que on a 467 patients sur le fichier dont 422 homes et 45 femmes. On sait aussi que 237 patients ont été traité avec ddc et 230 ont été traité avec ddi. On a 307 patients avec AIDS et 160 sans AIDS. On a aussi 292 patients avec AZT intolerance et 175 patients avec AZT failure. Et on a aussi des informations sur la duree, on sait que la duree minimum est de 0,47, la duree maximum est de 21,40 et la duree moyenne est de 12,63.

2.2.2 visualisation

La visualisation rend les données plus compréhensibles et permet de repérer les tendances, les corrélations ou les anomalies qui pourraient être difficiles à identifier avec des chiffres seuls.[1]

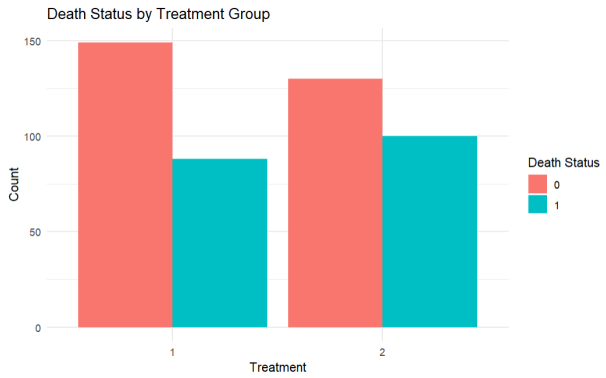


FIGURE 2 – death by group traitement

L'analyse de ce diagramme montre que le nombre de décès parmi les patients traités par DDC est inférieur à celui des patients traités par DDI. Ce constat nous permet d'émettre l'hypothèse que le traitement par DDC est plus efficace que celui par DDI. Toutefois, cette hypothèse ne peut être confirmée immédiatement ; des informations supplémentaires sont nécessaires pour en valider la véracité.

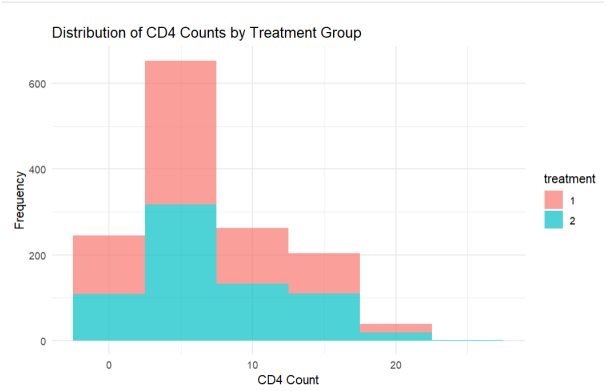


FIGURE 3 – CD4 by group traitement

L'analyse de cet histogramme fournit des informations sur la distribution des CD4 dans chaque groupe de traitement. On observe que le nombre de CD4 est plus élevé dans le groupe de patients traités par DDC par rapport au groupe de patients traités par DDI. Cette observation soutient l'hypothèse avancée dans le diagramme précédent : si le nombre de CD4 augmente pendant le traitement, cela implique une efficacité du traitement.

Les analyses des graphiques précédents soutiennent l'hypothèse selon laquelle le traitement par DDC est plus efficace que le traitement par DDI. Il nous reste à confirmer cette hypothèse dans les paragraphes suivants, en analysant les courbes de survie et les tests de corrélation.

Avant de commencer l'analyse de corrélation, examinons la distribution des CD4 par rapport au facteur pré-infection. Cette distribution nous fournira des informations sur la quantité de CD4 selon le groupe de patients dans ce facteur.

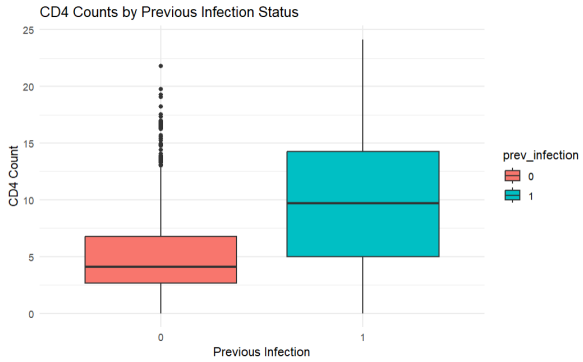


FIGURE 4 – CD4 by group prev-infection

L'analyse de ces boites nous montre évidement que les patients qui ont un AIDS ont une faibles quantités de CD4 par rapport aux qui n'ont pas. On peut conclure que les patients qui ont un AIDS sont plus exposer et plus fragile à des maladies.

L'observation et l'analyse des visualisations précédentes nous fournissent des informations approfondies sur la distribution des données et permettent de formuler des hypothèses. Pour renforcer ces hypothèses, nous allons étudier les corrélations entre ces variables.

2.3 Analyse de corrélation

L'analyse de corrélation est une méthode statistique utilisée pour examiner la relation entre deux variables. Elle permet de déterminer si, et dans quelle mesure, les variations d'une variable sont associées aux variations d'une autre.[3]

2.3.1 Matrice de corrélation

Une matrice de corrélation est un outil statistique qui résume les relations entre plusieurs variables dans un ensemble de données. Elle est représentée sous forme de tableau carré où chaque ligne et colonne correspond à une variable, et chaque cellule contient une valeur appelée coefficient de corrélation. Ce coefficient indique la force et la direction de la relation entre deux variables.Il est généralement compris entre -1 et +1.

- +1 : Corrélation positive parfaite (quand une variable augmente, l'autre augmente proportionnellement).
- 0 : Aucune corrélation (les variables ne sont pas liées).
- -1 : Corrélation négative parfaite (quand une variable augmente, l'autre diminue proportionnellement).
- time et death (-0.63) : Une corrélation négative modérément forte. Cela suggère que plus le temps passe, moins la probabilité de décès est élevée, ou inversement. Cela pourrait être un indice sur l'effet des traitements ou des soins reçus.

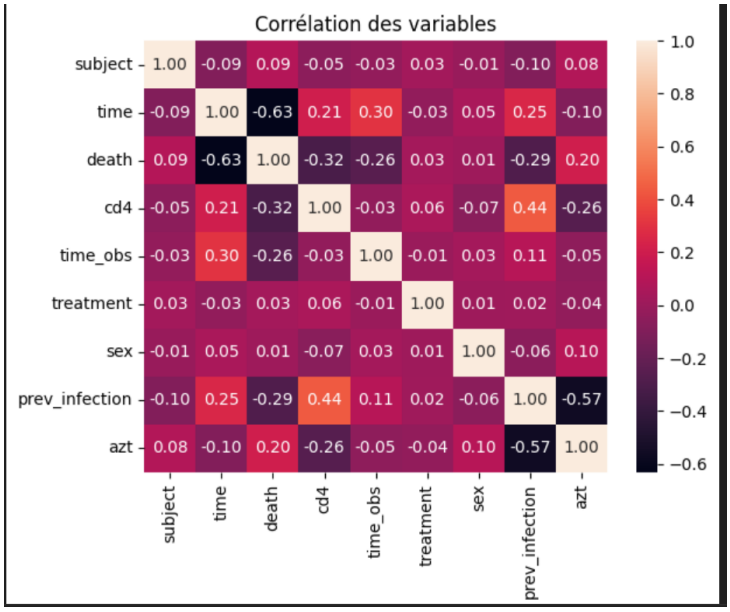


FIGURE 5 – Matrice de corrélation

— prev-infection et cd4 (0.44) : Une corrélation positive modérée. Les patients ayant eu des infections antérieures ont des niveaux de CD4 (marqueur immunitaire) légèrement plus élevés.
— prev-infection et azt (-0.57) : Une corrélation négative modérée. Les patients avec des infections précédentes semblent moins susceptibles de recevoir un traitement à base d'AZT (antirétroviral).
— cd4 et death (-0.32) : Une légère corrélation négative. Les patients avec un taux plus faible de CD4 (indiquant une immunodéficience) semblent avoir un risque accru de décès.
— azt et death (0.20) : Corrélation faible et positive. Cela pourrait indiquer que l'AZT est administré plus souvent aux patients gravement malades, mais cette hypothèse nécessite une exploration plus approfondie.
— La plupart des corrélations sont proches de 0, ce qui signifie qu'il n'y a pas de relation linéaire forte entre ces variables. Par exemple :traitement et les autres variables ; Aucune corrélation significative n'est observée. Et sex et les autres variables ; Le sexe ne semble pas être un facteur influent dans les relations entre ces variables.

2.3.2 Interprétation Globale

— Facteur clé time et death : La forte corrélation négative entre ces deux variables pourrait indiquer que les interventions (comme les traitements) augmentent les chances de survie au fil du temps.
— Effets de l'AZT : L'usage de ce traitement semble lié aux antécédents d'infection et potentiellement à la survie.
— Rôle du système immunitaire (cd4) : Comme prévu dans les maladies immunitaires comme le VIH, le taux de CD4 joue un rôle important dans la survie.

2.4 Visualisation de la fonction de survie

Dans cette section on va analyser les fonctions de survie en fonction selons les genres des patients et le traitement.

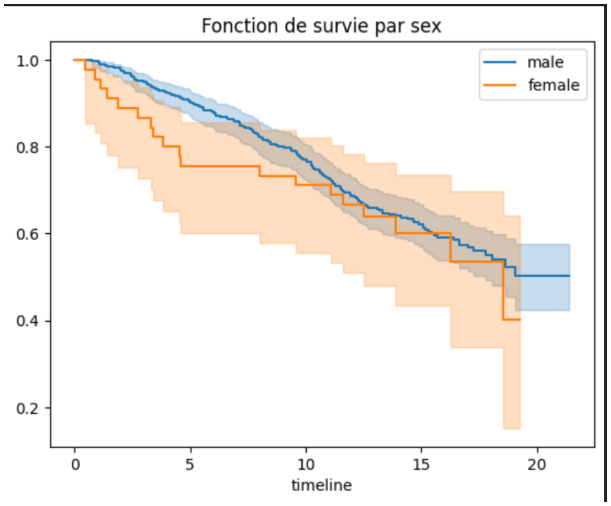


FIGURE 6 – survie par genres

2.4.1 Fonction de survie par sexe

- Ce graphique représente une courbe de survie par sexe, avec deux groupes comparés : hommes (male) en bleu et femmes (female) en orange.
- La fonction de survie montre la probabilité de survie (en ordonnée) à un moment donné (en abscisse).
- Au temps 0, la probabilité de survie est de 1 , car tout le monde est en vie au départ.
- Les courbes décroissent avec le temps, indiquant qu'à mesure que le temps avance, un certain nombre d'individus décèdent ou quittent l'étude.
- La courbe orange (femmes) diminue plus rapidement, suggérant une survie plus faible chez les femmes dans cette étude.
- Les zones ombrées autour des courbes représentent des intervalles de confiance. Ces intervalles montrent l'incertitude des estimations de survie. Plus les intervalles sont larges, plus l'incertitude est grande.
- La courbe bleue (hommes) reste au-dessus de la courbe orange (femmes) sur presque toute la durée, ce qui indique que les hommes dans cette étude ont une probabilité de survie plus élevée.
- Les deux courbes chutent avec le temps, ce qui est attendu puisque la survie diminue naturellement avec le temps.

2.4.2 interprétation globale

Les hommes ont une survie globale plus élevée par rapport aux femmes dans cette étude. Les variations peuvent être influencées par des facteurs sous-jacents comme les caractéristiques biologiques, les conditions initiales, ou d'autres biais propres à la population étudiée.

2.4.3 Fonction de survie par traitement

Ce graphique représente une courbe de survie par traitement, avec deux groupes comparés : ddc (1) en bleu et ddI (2) en orange. Les courbes décroissent avec le temps, indiquant qu'à mesure que le temps avance, un certain nombre d'individus décèdent ou quittent l'étude, ce qui est attendu puisque la survie diminue naturellement avec le temps.

- Le groupe ddI (orange) semble avoir une survie inférieure à celle du groupe ddc (bleu) à la plupart des moments, car sa courbe est en dessous de celle du groupe ddc

— Vers la fin de l'échelle temporelle (à droite), l'incertitude augmente car les bandes ombrées deviennent plus larges. Cela peut être dû au fait qu'il reste moins d'individus à cette période, ce qui augmente la variabilité des estimations.
— Les zones d'incertitude (intervalles de confiance) montrent un chevauchement notable entre les deux groupes, ce qui peut indiquer que la différence de survie pourrait ne pas être statistiquement significative.
— Vers la fin de l'étude (aux alentours de 20 unités de temps), les échantillons deviennent plus petits (zones élargies), ce qui réduit la précision des estimations.

2.4.4 interprétation globale

Les courbes montrent un comportement relativement similaire, indiquant que les deux traitements ont une efficacité globalement proche, même si ddC semble avoir un léger avantage. Cette différence semble modeste et pourrait nécessiter une analyse statistique pour confirmer sa significativité.

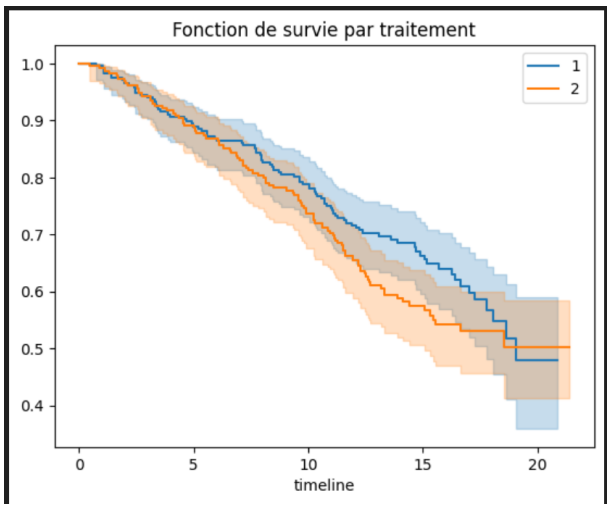


FIGURE 7 – Survie par traitement

2.5 Modélisation et analyse statistiques

Dans ce paragraphe, nous allons effectuer des tests statistiques qui nous permettront de confirmer les hypothèses posées dans les paragraphes précédents.

2.5.1 Log-Rank Test

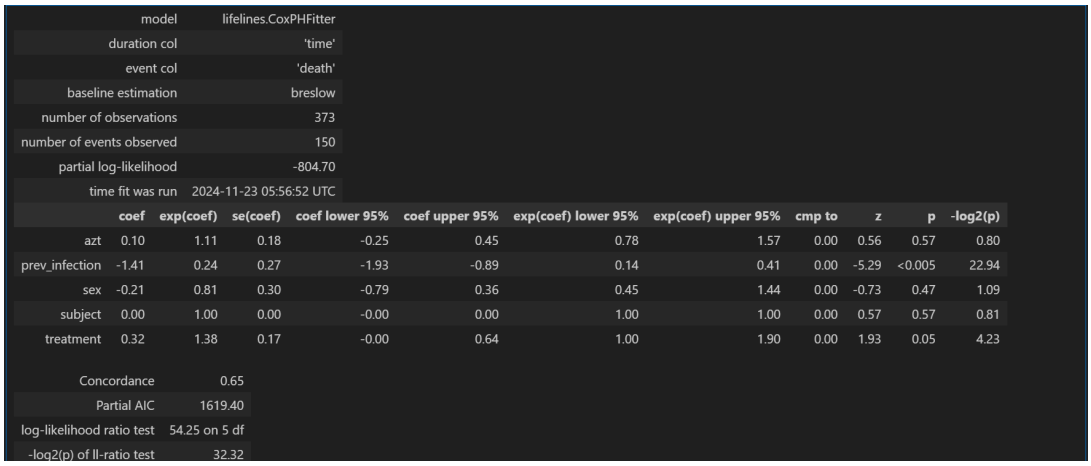
Le Log-Rank Test est une méthode statistique non paramétrique utilisée pour comparer les courbes de survie entre deux ou plusieurs groupes. Il est souvent employé dans l'analyse de survie pour déterminer si la différence observée entre les courbes est statistiquement significative.[2]

— Pour la courbe de survie par traitement, le test de log-rank nous donne une p-value de 0.15, ainsi une p-value de 0,15 indique qu'il n'y a pas de preuve statistiquement significative (au seuil de 0,05) pour rejeter l'hypothèse nulle.Bien que les courbe de survie semble soutenir l'hypothèse selon laquelle le traitement par DDC est plus efficace que celui par DDI, cette différence n'est pas statistiquement significative.
— Pour la courbe de survie par sex, le test de log-rank nous donne une p-value de 0.13. cette résultat indique qu'il n'y a pas de preuve statistiquement significative (au seuil de 0,05) pour rejeter l'hypothèse nulle.

On constate que le test de log-rank ne permet pas de confirmer les hypothèses posées dans les paragraphes précédents. Ainsi, nous allons utiliser le modèle de Cox, qui pourrait fournir des informations supplémentaires.

2.5.2 Modèle de cox

Le modèle de Cox, est une méthode statistique largement utilisée pour analyser les données de survie. Ce modèle est particulièrement utile pour évaluer l'effet de plusieurs variables explicatives sur le temps jusqu'à la survenue d'un événement d'intérêt, comme la mort, la rechute, ou une autre condition mesurable.



model	lifelines.CoxPHFitter
duration col	'time'
event col	'death'
baseline estimation	breslow
number of observations	373
number of events observed	150
partial log-likelihood	-804.70
time fit was run	2024-11-23 05:56:52 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
azt	0.10	1.11	0.18	-0.25	0.45	0.78	1.57	0.00	0.56	0.57	0.80
prev_infection	-1.41	0.24	0.27	-1.93	-0.89	0.14	0.41	0.00	-5.29	<0.005	22.94
sex	-0.21	0.81	0.30	-0.79	0.36	0.45	1.44	0.00	-0.73	0.47	1.09
subject	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	0.57	0.57	0.81
treatment	0.32	1.38	0.17	-0.00	0.64	1.00	1.90	0.00	1.93	0.05	4.23

Concordance	0.65
Partial AIC	1619.40
log-likelihood ratio test	54.25 on 5 df
-log2(p) of ll-ratio test	32.32

FIGURE 8 – Resultat du modèle de cox

2.5.3 interprétation

Le modèle a une concordance de 0.65, indiquant une capacité modérée à prédire les rangs de survie. Bien que cette performance soit acceptable, elle laisse une marge pour améliorer le modèle.

- (prev-infection) : Les individus ayant eu une infection précédente présentent un risque de décès significativement réduit (76 % de réduction du risque). Cela peut indiquer que ces individus ont développé une certaine immunité ou résilience qui leur confère une meilleure survie. Ce facteur est hautement significatif ($p < 0.005$), ce qui confirme que cet effet est robuste et fiable.
- treatment : Le traitement étudié semble augmenter le risque de décès de 38%, ce qui peut surprendre. Cependant, la signification statistique est marginale ($p = 0.05$). Cela pourrait indiquer un effet potentiellement négatif du traitement, mais il faut interpréter ce résultat avec prudence.
- AZT : Le traitement AZT n'a pas d'effet statistiquement significatif sur le risque de décès ($p = 0.57$). Cela signifie qu'il n'y a pas suffisamment de preuves pour conclure qu'AZT influence la survie dans cette population.
- Sexe : Aucune différence significative entre les sexes en termes de survie n'est observée ($p = 0.47$). Cela suggère que le sexe n'est pas un facteur clé dans la variation du risque de décès.

2.5.4 Modèle Random Survival Forest et Weibull

L'utilisation du modèle Random Survival Forest nous donne des résultats presque identiques, avec un C-index de 0.60, et nous permet d'arriver aux mêmes interprétations concernant les variables explicatives et leur influence sur la survie des patients. De la même manière, le modèle

Weibull fournit des résultats qui renforcent les précédents, avec une concordance de 0.65, similaire à celle du modèle de Cox. Les coefficients des variables restent également presque identiques. Cela nous permet de conclure de la manière suivante.

3 Conclusion

Ce projet d'analyse de survie offre des insights précieux sur les facteurs influençant la survie des patients atteints par le VIH. Les résultats soulignent l'importance des infections antérieures comme facteur protecteur, tout en soulevant des interrogations sur certains traitements. Ces observations constituent une base solide pour affiner les pratiques cliniques et guider les recherches futures, avec pour objectif ultime d'améliorer la qualité et la durée de vie des patients vivant avec le VIH.

Bibliographie

[1] Jean Bouyer. Statistique en médecine et en biologie : exercices corrigés et commentés. *Paris : Flammarion Médecine-Sciences , DL 1994*, 1994.

[2] http ://www.youtube.com/@epimedopencourse. Introduction à l'analyse de survie, courbe de kaplan-meier. 2021.

[3] Jean-François Viel Jean-François Mercier, Mariette Morin. Biostatistique et probabilités. *Paris : Ellipses , DL 2011*, 2011.