Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

Méthodologie probabiliste des études épidémiologiques

Analyse critique de l'Etude [Huss]

Société de Calcul Mathématique SA

rédaction : Bernard Beauzamy et Manon Baradat

juillet 2009

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel : 01 42 89 10 89. Fax : 01 42 89 10 69. www.scmsa.com
Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 7219Z

Nous réalisons ici une analyse critique de l'étude :

Residence Near Power Lines and Mortality From Neurodegenerative Diseases: Longitudinal Study of the Swiss Population

Anke Huss, Adrian Spoerri, Matthias Egger, and Martin Röösli

American Journal of Epidemiology Advance Access published November 5, 2008

I. Présentation générale

Cette étude concerne la maladie d'Alzheimer et la démence sénile : sont-elles plus fréquentes au sein de la population suisse vivant à proximité des lignes HT ?

Du point de vue de la présentation des données, l'étude est très bien faite (à la différence de [Draper]) : les auteurs prennent soin en effet d'évaluer la population "à risque", en fonction de la durée d'exposition. Cependant, si l'on prend en compte les affirmations de l'INSERM (citées plus haut) selon lesquelles la moitié seulement des cas d'Alzheimer est recensée, disons clairement que cette étude n'aurait jamais dû être lancée. Mais c'est là une réserve qui concerne avant tout l'épidémiologiste. Dans notre critique mathématique, nous ferons comme si les données étaient les bonnes.

Les auteurs concluent à un excès de risque au voisinage des lignes : "Overall, the adjusted hazard ratio for Alzheimer's disease in persons living within 50 m of a 220–380 kV power line was 1.24 (95% confidence interval (CI): 0.80, 1.92) compared with persons who lived at a distance of 600 m or more."

II. Un défaut de bon sens

Malheureusement, les chiffres bruts qu'ils utilisent contredisent leur conclusion ; celle-ci, une fois encore, repose sur une utilisation inappropriée de modèles statistiques (en l'occurrence le modèle de Cox). On ne peut que citer à nouveau Henri Poincaré : il suffit de regarder les chiffres pour avoir du bon sens.

Reproduisons en effet le tableau 2 issu de [Huss] :

Cause of Death	Distance to 220–380 kV Power Line, m	No. Of cases	No. Of Person-Years	Proba
Entire study population				
Alzheimer's disease	0–<50	20	58 423	0,000342
	50–<200	130	363 460	0,000358
	200–<600	572	1 688 323	0,000339
	>= 600	8 506	20 711 618	0,000411
Senile dementia	0–<50	60	58 423	0,001027
	50–<200	371	363 460	0,001021
	200–<600	1 702	1 688 323	0,001008
	>=600	26 155	20 711 618	0,001263
Individuals living at least 15 years at the identical place of residence				
Alzheimer's disease	0–<50	15	22 320	0,000672
	50–<200	63	145 148	0,000434
	200–<600	259	641 017	0,000404
	>= 600	3 861	7 698 419	0,000502
Senile dementia	0–<50	33	22 320	0,001478
	50–<200	169	145 148	0,001164
	200–<600	819	641 017	0,001278
	>= 600	11 930	7 698 419	0,001550

Tableau 1 : Données [Huss]

Nous avons simplement ajouté à droite une colonne "proba" qui est calculée de la manière suivante : c'est le nombre de la colonne 3 (nombre de cas) divisé par le nombre de la colonne 4 (nombre de personnes × années).

C'est en effet la probabilité de mourir de la maladie considérée, sachant que l'on est dans la catégorie correspondante (en termes de distance à la ligne). Pour la première ligne, par exemple (Alzheimer, distance inférieure à 50 m), nous avons 58 423 personnes × années d'exposition, et 20 cas recensés : cela nous fait $\frac{20}{58\,423} = 0{,}000342331$ cas par personne × année d'exposition.

Eh bien, on constate que, pour la maladie d'Alzheimer, toutes ces probabilités, pour toutes les distances à la ligne, sont inférieures à la probabilité de référence (personnes à plus de 600 m), qui est $0{,}000410687$! Il en est de même de la démence sénile.

Si maintenant on considère les populations ayant vécu 15 années à la même place, on constate que, pour la démence sénile, toutes les probabilités concernant les populations proches des lignes sont inférieures à la probabilité de référence.

Il en est de même pour la maladie d'Alzheimer, pour ces mêmes populations sédentaires, sauf pour celles à proximité immédiate des lignes : ici nous avons une probabilité de $0{,}000672$ pour les personnes vivant au voisinage immédiat des lignes, et de $0{,}00050$ pour la population de référence.

Prenons donc une probabilité de référence :

$$p_R = 0{,}0005$$

et voyons à quoi on peut s'attendre pour une population à risque constituée de $N = 22\,320$ personnes. L'intervalle de confiance à 95 %, vu plus haut, nous donne :

$$I = \left[Np_R - 2\sqrt{Np_R(1-p_R)}, Np_R - 2\sqrt{Np_R(1-p_R)} \right]$$

soit avec les valeurs présentes :

$$I = \left[4.5, 17.88\right]$$

La valeur observée, ici 15, est compatible avec cet intervalle et peut s'expliquer par le seul fait du hasard. Pour toutes les autres distances, la probabilité observée est inférieure à la probabilité de référence.

III. Calcul de la probabilité qu'une zone test soit plus dangereuse que la référence

Comme nous l'avons expliqué, les méthodes probabilistes développées dans [BB1] permettent de répondre à la question suivante : étant donnés une population et un nombre d'accidents dans une zone test et une zone de référence, quelle est la probabilité que la zone test soit plus dangereuse que la zone de référence ? Voici les résultats (dernière colonne du tableau) :

Cause of Death	Distance to 220–380 kV Power Line, m	No. Of cases	No. Of Person-Years	Proba	Proba + dgrx que référence
Entire study population					
Alzheimer's disease	0–<50	20	58 423	0,000342	0,24314
	50–<200	130	363 460	0,000358	0,06122
	200–<600	572	1 688 323	0,000339	0,00000
	>= 600	8 506	20 711 618	0,000411	
Senile dementia	0–<50	60	58 423	0,001027	0,05760
	50–<200	371	363 460	0,001021	0,00001
	200–<600	1 702	1 688 323	0,001008	0,00000
	>=600	26 155	20 711 618	0,001263	
Individuals living at least 15 years at the identical place of residence					
Alzheimer's disease	0–<50	15	22 320	0,000672	0,89622
	50–<200	63	145 148	0,000434	0,13872
	200–<600	259	641 017	0,000404	0,00029
	>= 600	3 861	7 698 419	0,000502	
Senile dementia	0–<50	33	22 320	0,001478	0,43720
	50–<200	169	145 148	0,001164	0,00006
	200–<600	819	641 017	0,001278	0,00000
	>= 600	11 930	7 698 419	0,001550	

Tableau 2 : probabilité que la zone test soit plus dangereuse

Dans chacun des quatre cas, la zone de référence est la zone située à plus de 600 m des lignes.

IV. Données globales de l'étude [Huss]

Travaillons maintenant sur données globales, pour Alzheimer, et non plus par tranche de distance. Nous faisons la somme des trois premières lignes du tableau 1. Nous obtenons :

Nombre de cas : 20 + 130 + 572 = 722

Nombre de personnes × années : 58 423 + 363 460 + 1 688 323 = 2 110 206

Si on prend pour référence la zone à plus de 600 m, elle comporte 8 506 cas pour 20 711 618 personnnes × années, soit une probabilité de $4{,}1\times10^{-4}$. Pour 2 110 206 personnes × années, cela devrait nous faire 866 cas, alors que nous n'en comptons que 722 !

V. Erreurs méthodologiques de l'étude [Huss]

L'erreur commise tient à l'utilisation du modèle de Cox, qui réclame la proportionnalité des données, d'une situation à l'autre. Les auteurs affirment avoir testé cette proportionnalité : "We tested our models successfully for the proportionality assumption using Nelson-Aalen survivor functions and statistical tests based on Schoenfeld residuals". Mais, manifestement, le test de proportionnalité était défectueux !

Nous donnons en Annexe 3 un exemple très simple qui montre que le modèle de Cox, utilisé hors de ses hypothèses, conduit à des conclusions absurdes. Mais dans le cas présenté en Annexe, les données "passent" le test d'utilisation du modèle : le test statistique affirme que Cox peut être utilisé, alors que ce n'est pas le cas !

VI. Que retenir de cette étude ?

Cette étude est faite avec beaucoup de sérieux et les données (nombre de décès et population concernée) sont manifestement recueillies avec grand soin, dans chacun des cas qui sont traités.

Ces données montrent un déficit de mortalité dans les zones à risque, comme nous l'avons vu, et contrairement à ce qu'affirment les auteurs de l'étude. Ce déficit de mortalité, dans la mesure où les informations ont été recueillies avec grand soin, peut être considéré comme acquis.

Peut-on en déduire que les lignes HT protègent contre certaines maladies ? Évidemment non, et nous avons expliqué pourquoi au cours du premier chapitre : qu'il y ait moins de morts, soit, mais ce peut être dû à bien d'autres causes, et en premier lieu, tout simplement, parce que la population est plus jeune !

L'étude [Huss] ne fournit aucune indication sur la pyramide des âges des populations concernées, mais seulement sur l'âge au moment du décès. Nous voyons ici une illustration particulière de ce que nous avons présenté plus haut : il s'agit d'une faute de logique ; les auteurs ont tout simplement oublié que la population, dans son ensemble, n'était pas immortelle.

Annexe 1

La physique du problème : lignes HT et champ magnétique

Les auteurs des études statistiques que nous analysons ici semblent faire reposer leurs raisonnements uniquement sur les statistiques (qu'ils maîtrisent mal), et non sur la physique, qu'ils ne maîtrisent pas du tout. La question de savoir si les lignes HT sont dangereuses est tout à fait légitime, encore faut-il qu'elle soit abordée avec les connaissances appropriées.

Une ligne électrique crée un champ électrique et un champ magnétique. Les auteurs semblent vouloir se limiter aux effets du champ magnétique. Pour celui-ci, les faits physiques sont les suivants :

L'intensité du champ magnétique (mesurée en Tesla) est proportionnelle à l'intensité du courant dans la ligne (et non à la tension !!) ; pour une ligne rectiligne de grandes dimensions (cas usuel), l'intensité du champ magnétique s'exprime par la formule :

$$B = c\frac{I}{d} \qquad (1)$$

où c est une constante, I l'intensité du courant dans la ligne (en Ampères) et d la distance à la ligne (en mètres).

La formule est démontrée plus bas.

Par conséquent, si vous êtes à 100 m d'une ligne transportant 500 Ampères, vous recevez le même champ magnétique que si vous êtes à 1 m du câble alimentant une cuisinière (5 Ampères). La fréquence des deux champs est la même (50 Hz).

Dans l'étude [Draper], diverses hypothèses sont faites en ce qui concerne la dépendance du champ par rapport à la distance : on y rencontre $1/d$, $1/d^2$, $1/d^3$; cette ignorance du phénomène physique en cause est tout de même étonnante, car enfin si l'effet était en $1/d^3$, à 100 m il n'en reste plus que le millionième !

Démontrons la formule (1). Elle est claire intuitivement, car dans le cas d'un conducteur rectiligne infini (ou simplement de grande longueur par rapport à la distance à l'observateur) le champ a nécessairement une symétrie cylindrique. La quantité reçue dans une portion de l'espace à distance r et d'épaisseur dr est proportionnelle au périmètre du cercle, soit $2\pi r dr$.

Donnons aussi une démonstration complète, issue de la formule de Biot et Savart :

$$B(M) = \frac{\mu_0}{4\pi} \int_C I \frac{\overrightarrow{dl} \wedge \overrightarrow{SM}}{\|SM\|^3}$$

où $B(M)$ est le champ magnétique en M, μ_0 la perméabilité magnétique du vide, C le conducteur (ici l'axe des x), $\overrightarrow{dl}$ l'élément de longueur sur l'axe Ox, et S le point courant.

Mettant le point M sur l'axe Oy avec l'ordonnée d, nous obtenons :

$$B(M) = \frac{\mu_0 I}{4\pi} \int_{-\infty}^{+\infty} \frac{(x^2 + d^2)^{1/2} \sin(\vartheta) dx}{(x^2 + d^2)^{3/2}}$$

où ϑ désigne l'angle (Ox, SM), et donc $\sin(\vartheta) = \frac{|x|}{\sqrt{x^2 + d^2}}$. Reportant dans l'expression précédente, nous obtenons :

$$B(M) = \frac{\mu_0 I}{2\pi} \int_0^{+\infty} \frac{x dx}{(x^2 + d^2)^{3/2}}$$

Le changement de variable $x = d \cdot y$ donne :

$$B(M) = \frac{\mu_0 I}{2\pi d} \int_0^{+\infty} \frac{y dy}{(1^2 + y^2)^{3/2}} = \frac{\mu_0 I}{2\pi d},$$

comme annoncé.

Annexe 2

Utilisation inappropriée du modèle de Cox : un exemple simple

I. Description de la population

Prenons deux populations, chacune de 60 personnes, atteintes d'une même maladie et soumises à des traitements différents A et B. Chaque mois, des individus décèdent dans chacune des populations. Après 110 mois, il n'y a plus d'individus en vie.

Durée de traitement (en mois)	Nombre de décès parmi la population qui suit le traitement A	Nombre de décès parmi la population qui suit le traitement B
entre 0 et 10 mois	1	10
entre 10 et 20 mois	3	8
entre 20 et 30 mois	5	6
entre 30 et 40 mois	7	4
entre 40 et 50 mois	9	2
entre 50 et 60 mois	10	0
entre 60 et 70 mois	9	2
entre 70 et 80 mois	7	4
entre 80 et 90 mois	5	6
entre 90 et 100 mois	3	8
entre 100 et 110 mois	1	10

Tableau 1 : Nombre de décès, pour les personnes ayant suivi les traitements A ou B par intervalle de temps

La figure ci-dessous représente le nombre de décès par population en fonction du nombre de mois :

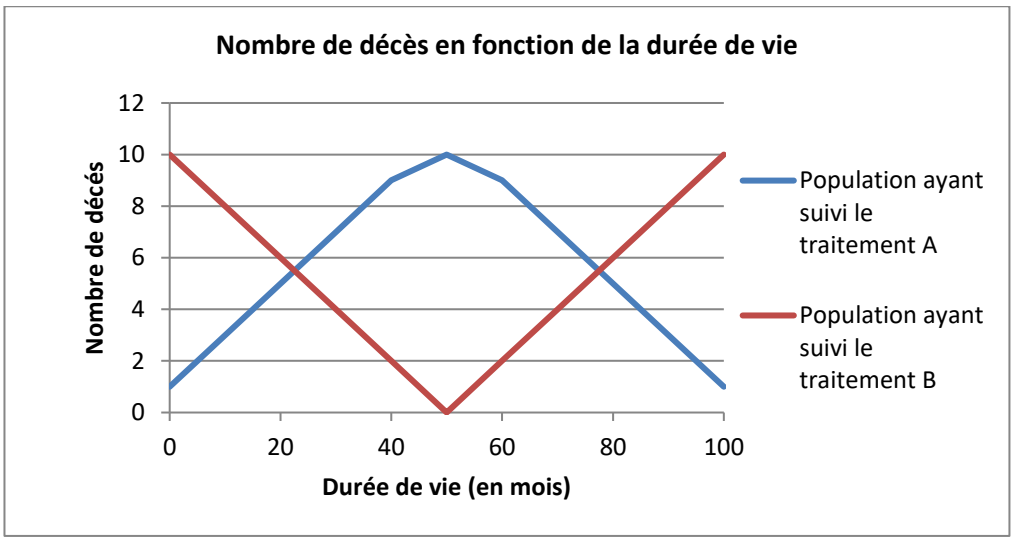

Figure 2 : Loi de probabilité des décès pour la population ayant suivi le traitement A

Les deux figures ci-dessous illustrent les lois de probabilité correspondantes.

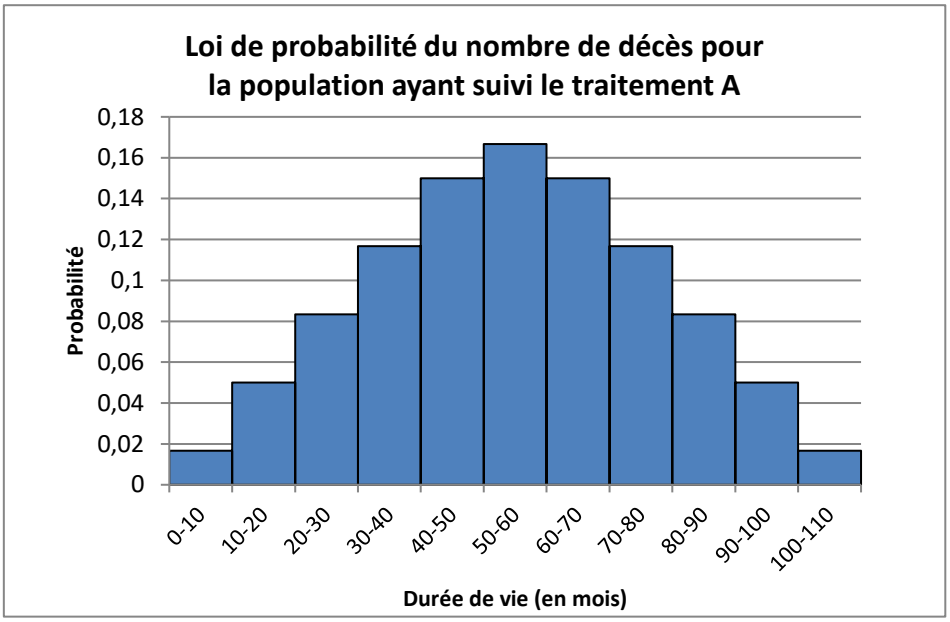

Figure 3 : Loi de probabilité des décès pour la population ayant suivi le traitement A

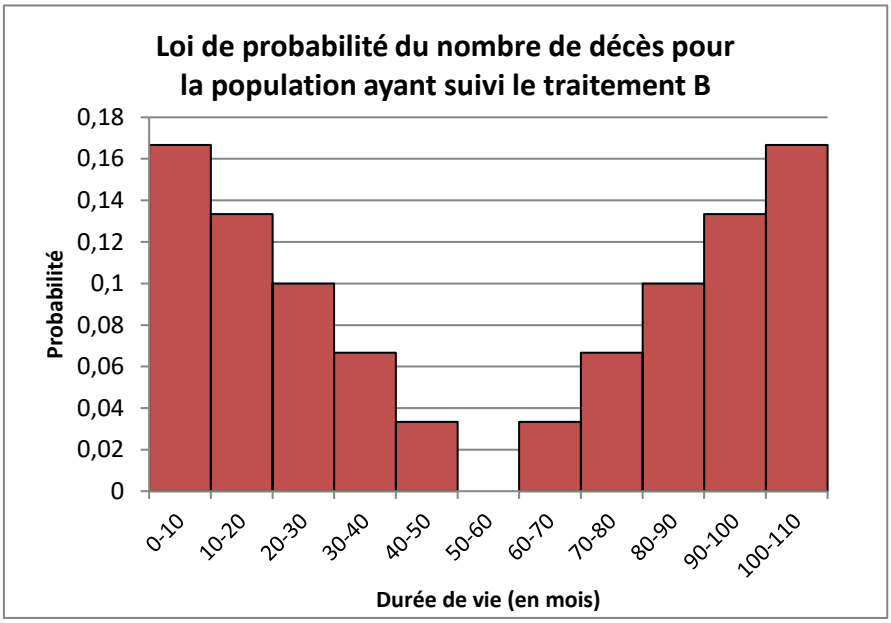

Figure 4 : Loi de probabilité des décès pour la population ayant suivi le traitement B

II. Utilisation du modèle de Cox

On regarde s'il existe une influence du type de traitement sur la durée de vie des individus. Pour cela, on regarde l'influence du type de traitement *B* par rapport au type de traitement *A*. Ainsi, les résultats fournis par le modèle correspondent au traitement *B*.

On utilise le logiciel *R*, qui est est un langage de programmation et un environnement mathématique utilisé pour le traitement de données et l'analyse statistique. Il permet notamment de simuler le modèle de Cox.

On fournit en entrée du logiciel trois vecteurs de taille 120 (une valeur par individu) :

- La durée de vie : 0, 0 …10, 10, 10…20, 20, 20… 100, 100, 100 ;
- Le type de traitement suivi par l'individu : on affecte 1 si l'individu a suivi le traitement B et 0 si l'individu a suivi le traitement A ;
- L'indicateur de décès : on affecte 1 à un individu mort et 0 à un individu perdu de vue. Dans notre cas, tous les individus sont morts.

En appliquant le modèle de Cox aux données, le logiciel *R* donnes les résultats suivants :

e^{β}	Intervalle de confiance	Degré de signification du test
0.66	[0,44 ; 0,96]	0.033 (< 0.05)

Tableau 5 : résultats

Le coefficient e^{β} et son intervalle de confiance sont strictement inférieurs à 1. De plus, le degré de signification du test est bien inférieur à 5%. Ainsi, le modèle de Cox conclut que le traitement B a une influence néfaste certaine sur la durée de vie et que, inversement, le traitement A a une influence bénéfique sur la durée de vie : mais ceci n'est pas correct.

III. Utilisation des probabilités conditionnelles

Pour évaluer l'impact d'un traitement sur la durée de survie d'un individu, nous utilisons une méthode probabiliste. Cette méthode consiste à tracer la fonction de répartition de la durée de survie dans deux cas de figure : pour la population ayant suivi le traitement A, et pour la population ayant suivi le traitement B.

La durée de survie est découpée en intervalles. Nous déterminons l'effectif cumulé de chaque intervalle, c'est-à-dire le nombre d'individus dont la durée de survie est supérieure à chaque borne.

En divisant l'effectif cumulé d'un intervalle par l'effectif du cas de figure (traitement A ou traitement B), nous obtenons le pourcentage d'individus dont la durée de vie dépasse le seuil fixé.

Les courbes ci-dessous sont obtenues en traçant les pourcentages correspondant à chaque intervalle en fonction du type de traitement.

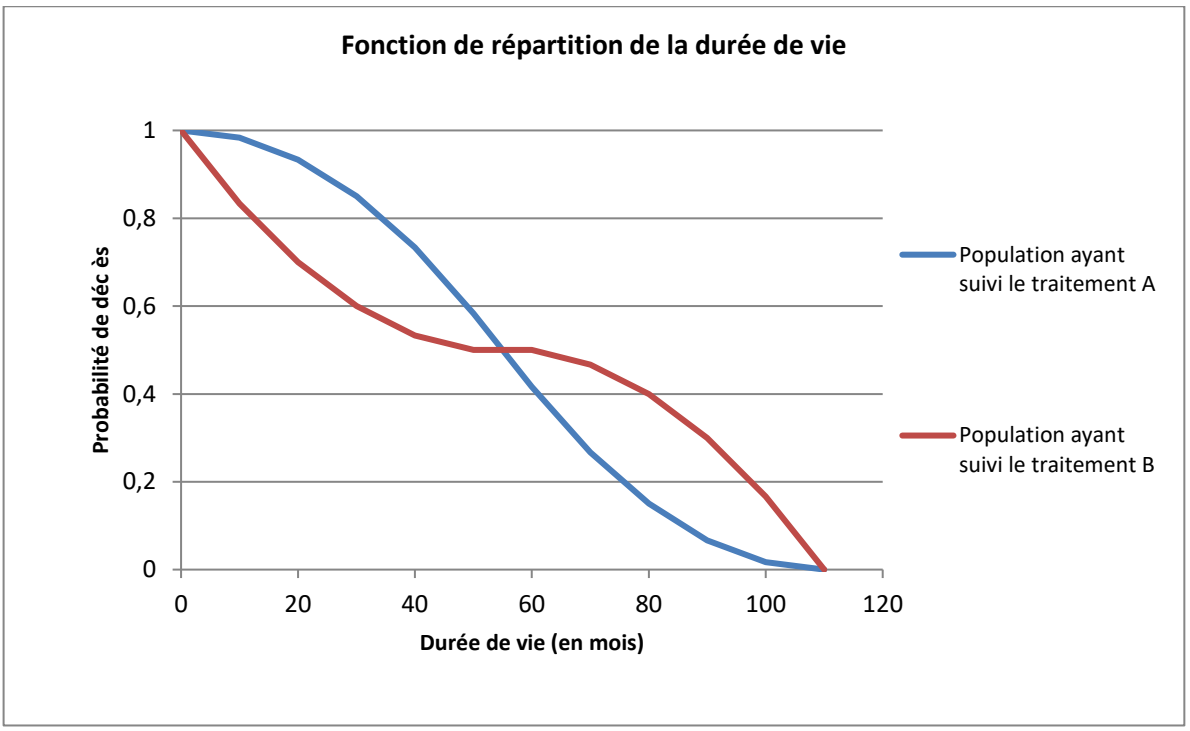

Figure 6 : Fonctions de répartition du de la durée de survie des individus sachant qu'ils ont une fonction rénale normale ou sachant qu'ils ont une fonction rénale anormale

La courbe rouge correspond à la fonction de répartition de la durée de vie pour les individus qui ont suivi le traitement *B* ; la courbe bleue correspond à la fonction de répartition de la durée de vie pour les individus qui suivi le traitement *A*.

Considérons une durée de vie de 20 mois, le pourcentage associé à la courbe rouge est 70 %, le pourcentage associé à la courbe bleue est 96 %. Ceci signifie que 96 % des individus ayant suivi le traitement *A* vivent plus de 20 mois, alors que seulement 70 % des individus ayant suivi le traitement *B* dépassent cette durée de vie.

Considérons maintenant une durée de vie de 80 mois, le pourcentage associé à la courbe rouge est 40 %, le pourcentage associé à la courbe bleue est 12 %. Ceci signifie que 12 % des individus ayant suivi le traitement *A* vivent plus de 80 mois, alors que 40 % des individus ayant suivi le traitement *B* dépassent cette durée de vie.

Ainsi, nous observons deux tendances :

- Le traitement *A* a un effet protecteur dans un premier temps, puis il a un effet néfaste dans un second temps ;
- Le traitement *B* a un effet néfaste dans un premier temps, puis il a un effet protecteur dans un second temps ;

IV. Comparaison des résultats des deux méthodes

Pour notre exemple, les résultats des deux méthodes sont différents : la méthode des probabilités conditionnelles montre qu'il y a deux situations distinctes, que le modèle de Cox n'identifie pas.

V. Autre remarque sur le modèle de Cox

Les résultats du modèle de Cox dépendent de l'échelle de temps choisie. Dans l'exemple développé ci-dessus, les données sont exprimées en mois : 0, 10, 20…, 110, 120 et le modèle de Cox conclut à une influence néfaste du traitement B. Si on prend les mêmes données en changeant l'échelle de temps, par exemple en prenant 0, 1, 2…,11, 12, le modèle de Cox ne conclut pas.

Les résultats sont donc différents si on exprime les données en jour, en mois, en années.

En revanche, dans notre modèle utilisant les lois de probabilités conditionnelles, les résultats sont indépendants de l'échelle de temps.