Master de Statistique

Université de Strasbourg

UFR Mathématique-Informatique

Du 3 juin au 3 août 2014

Comparaison des complications infectieuses des cathéters centraux posés en voie jugulaire et sous-clavière : analyse causale à l'aide du score de propension

Présenté par Mathieu EURY

Le 25 août 2014

Unité de Biostatistique et Recherche Clinique

Centre Hospitalier Universitaire

Avenue de la côte de Nacre

CS 30001

14033 CAEN cedex 9

Responsable de l'unité : M. Rémy MORELLO

Maître de stage : M. Jean-Jacques PARIENTI

Responsable du Master : Mme Armelle GUILLOU

Table des matières

Remerciements **4**

Résumé **5**

1 Introduction **6**

1.1 Présentation des cathéters 6
1.2 Rappels sur les infections nosocomiales 7
1.3 Objectif du stage 7

2 Outils mathématiques utilisés **8**

2.1 Régression logistique 8
2.1.1 Principe et description 8
2.1.2 Evaluation 10

2.2 Score de propension 12

2.3 Utilisation du score de propension 13
2.3.1 Appariement 13
2.3.2 Pondération 14
2.3.3 Stratification 14

2.4 Analyse de survie 15
2.4.1 Fonction de survie et risque instantané 15
2.4.2 La censure à droite 16
2.4.3 Estimateur de Kaplan-Meier 16
2.4.4 Modèle de Cox 17

3 Présentation des données **19**

4 Application sur la base de données 21

4.1 Comparaison entre les deux groupes ……21
4.2 Calcul du score de propension par la régression logistique ……23
4.3 Appariement sur le score de propension et comparaison entre les deux groupes après appariement ……25
4.4 Analyse des critères de jugement ……25
4.5 Commentaire des résultats ……28

Conclusion 29

Bibliographie 30

Annexes A et B 31

A – Codes SAS ……31
B – Tableau et Graphiques ……35

Remerciements

Je tiens tout d'abord à remercier mon maître de stage, M. Jean-Jacques Parienti, pour m'avoir accueilli au sein de son service, et pour m'avoir toujours guidé et tant enseigné au cours de ces deux mois.

Je remercie ensuite Stéphanie Levesque pour m'avoir laissé travailler dans son bureau, pour son aide et son soutien.

Je tiens également à remercier toute l'Unité de Biostatistique et Recherche Clinique pour leur accueil convivial et leur bonne humeur.

Pour finir, je remercie mes collègues de Master pour l'entraide et la bonne ambiance au sein du groupe.

Résumé

En utilisant des données provenant de deux études sur plusieurs centres, nous avons comparé les cathéters posés en voie jugulaire avec ceux posés en voie sous-clavière pour les risques de colonisation de cathéter, d'infection majeure du cathéter, et d'infection du sang liée au cathéter.

Pour effectuer ce travail, nous avons employé des méthodes statistiques très utilisées en recherche clinique, tel que les modèles 'MSM' (Marginal Structural Models) avec pondération sur l'inverse de la probabilité de traitement, ainsi que l'appariement sur le score de propension pour ajuster sur le biais, pour ensuite procéder à une analyse de survie. Notre étude se porte sur 2065 patients, avec un cathéter par patient.

Les résultats obtenus nous permettent d'avancer une forte présomption de différence en faveur des cathéters posés en voie sous-clavière. Nous n'avons pas trouvé de différence significative d'infection majeure du cathéter ni d'infection du sang liée au cathéter. Cependant, la colonisation était plus importante pour les cathéters posés en voie jugulaire.

Chapitre 1

Introduction

1.1 Présentation des cathéters

Un cathéter veineux central (CVC) est un tube mince et flexible (cathéter) qu'on met dans une grosse veine au-dessus du cœur, habituellement par une veine du cou, du thorax ou du bras. On l'appelle aussi voie veineuse centrale ou voie centrale.

On peut avoir recours au cathéter veineux central pour administrer des agents chimiothérapeutiques. Certains CVC possèdent plus d'une lumière (ouverture) et peuvent être utilisés pour l'administration de plus d'un médicament à la fois. On peut les laisser en place pendant plusieurs semaines, voire des mois. Si on pose un CVC à une personne, elle n'aura pas besoin d'autant d'injections et ses veines seront moins endommagées. Les principales complications du CVC sont l'infection (bactérienne liée au CVC) et la formation de caillots sanguins (thrombose veineuse profonde).

Les cathéters qui ferons l'objet de notre étude sont ceux posés en voie jugulaire (Figure 1.11) et ceux posés en voix sous-clavière (Figure 1.12).

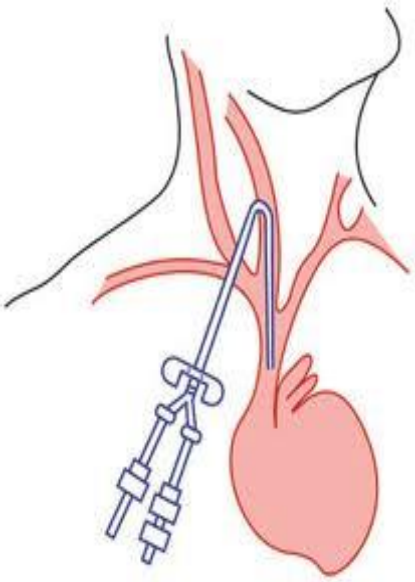


Figure 1.11 : Cathéter en voie jugulaire

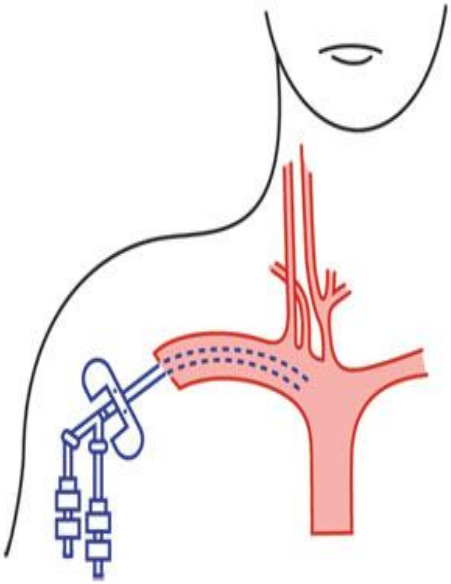


Figure 1.12 : Cathéter en voie sous-clavière

1.2 Rappels sur les infections nosocomiales

Une infection nosocomiale désigne une infection contractée au cours d'une hospitalisation, infection qui n'existait pas auparavant ni, d'ailleurs, durant les 48 premières heures à l'hôpital. Par leur fréquence et leur mortalité, les infections nosocomiales liées au CVC font partie des trois principales infections acquises en réanimation. L'infection se propage de l'extrémité du cathéter qui a été en contact avec la peau du patient jusqu'à l'intérieur de la circulation veineuse.

Les trois phénomènes que nous chercherons à analyser sont :

- La colonisation de cathéter, c'est-à-dire le développement d'un organisme depuis la pointe d'un cathéter retiré du patient.
- L'infection majeure du cathéter.
- L'infection du sang liée au cathéter, c'est-à-dire l'infection que l'on retrouve dans le sang du patient.

1.3 Objectif du stage

L'objectif du stage était double :

- Réaliser un travail original à partir d'une base de données issue de deux essais randomisés français multicentriques en employant des méthodes biostatistiques avancées permettant de limiter les biais. Ce sont les résultats de cette étude que nous détaillerons ici.
- Participer à l'activité du service de Biostatistique et de Recherche Clinique du CHU de Caen.

Chapitre 2

Outils mathématiques utilisés

Dans le cadre de l'analyse de données du travail, nous avons utilisé successivement plusieurs modèles statistiques que nous décrivons brièvement ci-après.

2.1 Régression logistique

Cette partie a pour références [1], [2] et [3] (voir bibliographie).

2.1.1 Principe et description

Le modèle de régression logistique est le modèle utilisé lorsque l'on cherche à expliquer un variable qualitative, notamment binaire, très utilisé dans le domaine médical (par exemple présence/absence de maladie). A l'inverse de la régression linéaire, la régression logistique n'impose pas de condition de normalité ni d'homoscédasticité des résidus, ce qui fait d'elle une méthode de choix pour le statisticien.

Le but de la régression logistique, dans le cas binaire, est de prédire une variable Y = {0,1} dont une des issues est le succès (Y=1) et l'autre un échec (Y=0). Pour ce faire, on cherche à modéliser la probabilité conditionnelle $\mathrm{P}(\mathrm{Y} = 1|\mathrm{X})$, où $X = (X_1, \dots, X_n)$ est le vecteur des covariables explicatives (qui peuvent être continues ou binaires), qui est la probabilité d'obtenir le succès.

Pour simplifier la notation, on écrit :

$$\pi(\mathrm{X}) \ = \ \mathrm{P}(\mathrm{Y} = 1|\mathrm{X})$$

On munit ensuite d'un règle de décision et d'un seuil θ tel que :

$$\mathrm{Y} = \begin{cases} 1 \ si \ P(Y = 1) > \theta \\ 0 \ sinon \end{cases}$$

Avec en général θ=0.5

L'hypothèse fondamentale de la régression logistique est la suivante :

$$\ln\left(\frac{P(X|Y=1)}{P(X|Y=0)}\right) = a_0 + a_1 x_1 + \cdots + a_J x_J$$

Cette hypothèse couvre une très large classe de distribution, par exemple les distributions normales, exponentielles, également celles où les variables explicatives sont discrètes, notamment

booléennes. Donc, contrairement à la régression linéaire, la régression logistique n'exige pas que les variables explicatives aient une distribution normale. Le nuage de points dont la variable π est une variable à valeur dans [0,1] ne représente pas une droite mais une fonction en forme de S, il s'agit de la *fonction logistique* (ou sigmoïde) (Figure 2.11). La fonction logistique s'écrit :

$$\pi(\mathrm{X}) = \frac{e^{\alpha+\beta\prime\mathrm{X}}}{1+e^{\alpha+\beta\prime\mathrm{X}}}$$

Où α est un coefficient, et β un vecteur des coefficients de la régression

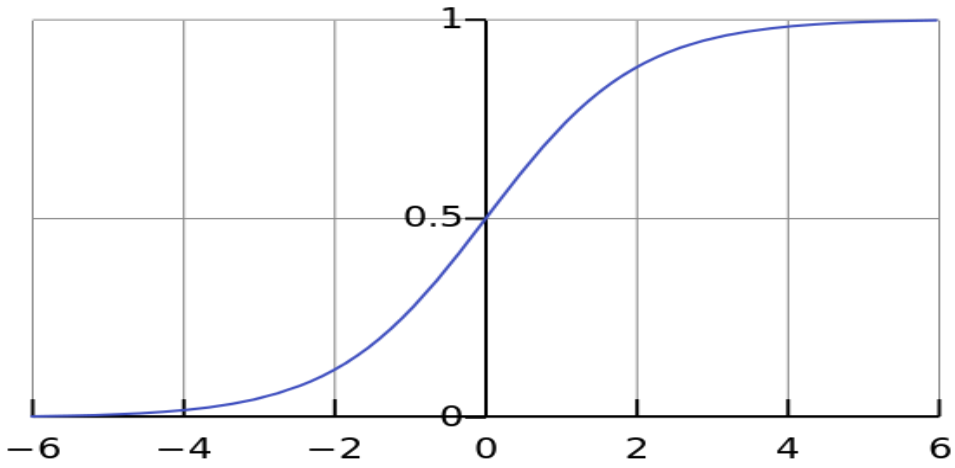


Figure 2.11 : Exemple de fonction logistique

On voit apparaître la version linéaire $\alpha + \beta'\mathrm{X}$. Pour projeter le problème dans un espace linéaire, on introduit donc la fonction *logit* telle que :

$$logit(x) = \ln(\frac{x}{1-x})$$

On a donc par la transformation *logit* le modèle logistique simple :

$$logit\big(\pi(\mathrm{X})\big) = \ln\left(\frac{\pi(\mathrm{X})}{1-\pi(X)}\right) = \ln\left(\frac{\frac{e^{\alpha+\beta'\mathrm{X}}}{1+e^{\alpha+\beta'\mathrm{X}}}}{1-\frac{e^{\alpha+\beta'\mathrm{X}}}{1+e^{\alpha+\beta'\mathrm{X}}}}\right) = \ln\left(\frac{\frac{e^{\alpha+\beta'\mathrm{X}}}{1+e^{\alpha+\beta'\mathrm{X}}}}{\frac{1+e^{\alpha+\beta'\mathrm{X}}}{1+e^{\alpha+\beta'\mathrm{X}}}-\frac{e^{\alpha+\beta'\mathrm{X}}}{1+e^{\alpha+\beta'\mathrm{X}}}}\right)$$

$$= \ln\left(\frac{\frac{e^{\alpha+\beta'\mathrm{X}}}{1+e^{\alpha+\beta'\mathrm{X}}}}{\frac{1}{1+e^{\alpha+\beta'\mathrm{X}}}}\right) = \ln\big(e^{\alpha+\beta'\mathrm{X}}\big) = \alpha+\beta'\mathrm{X}$$

L'importance de cette transformation est que $logit(\pi(\mathrm{X}))$ a beaucoup de propriétés souhaitables d'un modèle de régression linéaire. Le logit, $logit(\pi(\mathrm{X}))$, est linéaire en ses paramètres, peut être continu, et peut décrire]- ∞ ; +∞ [, selon l'étendue de X. Le quotient $\frac{\pi(\mathrm{X})}{1-\pi(X)}$ qui permet d'étendre l'intervalle de définition [0 ,1] et d'envisager la transformation est appelé 'odd'.

2.1.2 Evaluation

Matrice de confusion

On souhaite produite un modèle pour prédire avec le plus de précision possible les valeurs prises par la variable réponse Y. Une approche privilégiée est de comparer les valeurs effectivement prédites par le modèle avec les valeurs observées dans les données et les reporter dans une *matrice de confusion :*

	Prédit 1	Prédit 0
Observé 1	**Correct (VP)**	Faux négatif
Observé 0	Faux positif	**Correct (VN)**

On se définit ensuite le score d'exactitude (accuracy) comme : $ACC = \frac{VP+VN}{N}$

N étant le nombre total d'observation.

On peut aussi construire cette matrice sur un échantillon à part, qui n'aura pas été utilisé pour la construction du modèle, et ainsi obtenir une évaluation qui ne soit pas biaisée. En effet, lorsque l'on construit la matrice de confusion sur les données qui ont servi à construire le modèle, le taux d'erreur est souvent trop optimiste, et par conséquent ne représente pas les performances réelles du modèle dans la population.

Courbe ROC

On peut de même s'intéresser à la capacité du classifieur à discriminer les vrais positifs des faux positifs. On se définit le taux de vrais positifs ou *sensibilité* (Se) et le taux de vrais négatifs ou *spécificité* (Sp) :

$$Se = \frac{VP}{VP + FN} \qquad Sp = \frac{VN}{VN + FP}$$

On représente ensuite la sensibilité en fonction de 1-spécificité en faisant varier le seuil θ, ce qui change la matrice de confusion. On obtient ainsi une courbe ROC (Figure 2.1.21).

Le critère d'évaluation du modèle est alors l'AUC (Area Under Curve), qui est l'aire sous la courbe. Plus l'AUC (compris entre 0 et 1) est grand, plus le classifieur est capable de discriminer les vrais positifs des faux positifs. Un modèle dont l'AUC est de 1 est un modèle qui ne contient aucun faux négatif et aucun faux positif. Un AUC est considéré comme satisfaisant lorsqu'il est supérieur à 0.7 .

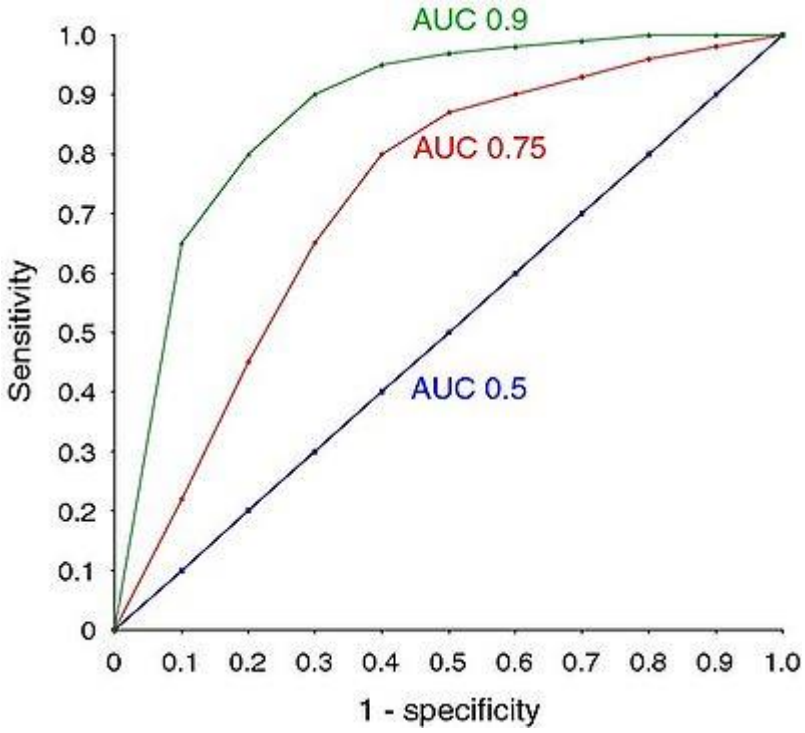


Figure 2.1.21 : Exemple de Courbes ROC avec AUC

Test de Hosmer-Lemeshow

Pour tester l'adéquation du modèle nous présentons également le test de Hosmer-Lemeshow. Ce test évalue si les taux de succès observés valent les taux des succès prédit dans des sous-groupes de la population. On subdivise les données en G sous-groupes en se basant sur les quantiles (par exemple les quantiles d'ordre 4 correspondent aux quartiles, les quantiles d'ordre 10 aux déciles, etc…). Les modèles pour lesquels les taux de succès observés sont similaires aux taux de succès prédits sont dits bien calibrés.

La statistique de test est :

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)}$$

Où dans chaque groupe g d'effectif N_g :

- O_g = nombre de positifs observés
- E_g = nombre de positifs prédits
- π_g = moyenne des scores dans le groupe g

La statistique de test H suit une loi du χ^2 à G-2 degrés de liberté. Lorsque la p-value est supérieure au risque α choisi, le modèle issu de la régression logistique est accepté.

2.2 Score de propension

Cette partie a pour références [4] et [5] (voir bibliographie).

Le score de propension (propensity score) d'un individu se définit comme la probabilité pour cet individu de recevoir un traitement (plutôt qu'un autre) conditionnellement à l'ensemble des caractéristiques de cet individu. Cette méthode peut être utilisée pour rendre deux groupes recevant chacun un traitement différent, plus comparables. Le score de propension est utilisé pour réduire le biais dans des études observationnelles notamment en recherche clinique en rendant la comparaison de deux groupes plus valide.

Dans une étude randomisée, le tirage au sort des traitements garantie la comparabilité des groupes. Cependant, dans une étude observationnelle (non-randomisée) nous n'avons aucun contrôle sur l'affectation du traitement, ainsi des comparaisons brutes de l'efficacité des traitements peuvent être biaisées. Cette erreur systématique peut être en partie évitée si une information sur les covariables observées est incorporée dans l'étude (par exemple par l'échantillonnage apparié 'matched sampling') ou dans l'estimation de l'effet du traitement (par exemple par la stratification). Les méthodes d'ajustement traditionnelles (appariement, stratification, analyse multivariée) sont souvent limitées car elles utilisent un nombre limité de covariables. Cependant, le score de propension, qui fournit un résumé scalaire de l'information de la covariable, n'a pas cette limite.

Formellement, le score de propension d'un individu est la probabilité d'être traité conditionnellement à la valeur de ses covariables. Intuitivement, le score de propension est une mesure de la probabilité qu'une personne a d'être traitée connaissant les valeurs de ses covariables. Rosenbaum et Rubin [5] ont montré que le score de propension était un équilibrage et pouvait être utilisé dans les études observationnelles pour réduire le biais par la méthode d'ajustement mentionnée précédemment.

Pour des données complètes, Rosenbaum et Rubin ont introduit le score de propension de l'individu i (i = 1,…,N) comme la probabilité conditionnelle d'être assigné à un traitement particulier $(Z_i = 1)$ contre la référence $(Z_i = 0)$ connaissant un vecteur des covariables observées, x_i:

$$e(x_i) = \mathrm{P}(Z_i = 1)\,|X_i = x_i)$$

où on suppose, les Xi et Zi étant indépendants, que :

$$\mathrm{P}(Z_1 = z_1, \dots, Z_n = z_n | X_1 = x_1, \dots, X_n) = \prod_{i=1}^{N} e(x_i)^{z_i}\{1 - e(x_i)\}^{1-z_i}$$

Le score de propension est la 'fonction la plus grossière' des covariables qui soit un équilibrage, où un équilibrage b(X) est défini comme 'une fonction des covariables observées X telle que la distribution conditionnelle de X sachant b(X) est la même pour les individus traités (Z=1) et les individus non-traités (Z=0)'. Pour une valeur spécifique du score de propension, la différence entre la moyenne 'traitement' et la moyenne 'non-traitement' pour tous les individus avec ce score de propension est une estimation non-biaisée de l'effet traitement moyen à ce score de propension, si l'affectation du traitement est fortement négligeable, connaissant les covariables. Par conséquent, l'ajustement par appariement ou stratification sur le score de propension a tendance à fournir des estimations non-biaisées des effets du traitement lorsque l'affectation du traitement est fortement négligeable. L'affectation du traitement est considérée comme fortement négligeable si l'affectation

du traitement, Z, et la réponse, Y, sont conditionnellement indépendants connaissant les covariables, X (c'est-à-dire, quand $Y \perp Z|X$).

Quand les covariables ne contiennent pas de valeur manquante, on peut estimer le score de propension en utilisant l'analyse discriminante ou la régression logistique. Ces deux méthodes conduisent à des estimations de probabilités d'affectation au traitement conditionnellement aux covariables observées. Formellement, on suppose que les covariables observées ont une distribution normale multivariée (conditionnellement à Z) quand on utilise l'analyse discriminante, alors que dans le cas de la régression logistique cette hypothèse n'est pas nécessaire.

Le principe du score de propension est d'utiliser la probabilité qu'a un individu d'être traité pour ajuster l'estimation de l'effet du traitement, on créé une expérience 'quasi-randomisée'. Ainsi, si on a deux sujets, un dans le groupe traité et un dans le groupe non-traité, avec le même score de propension, alors on pourrait imaginer que ces deux sujets ont été 'aléatoirement' affecté à chaque groupe dans le sens d'être également probable d'être traité. Dans une expérience contrôlée, la randomisation, qui affecte des paires d'individus aux groupes 'traités' et 'non-traités', est meilleure parce qu'elle ne dépend pas du conditionnement de l'investigateur sur une famille de covariables particulière ; elle s'applique plutôt à n'importe qu'elle famille de covariables observées ou non-observées. Bien que les résultats de l'utilisation du score de propension sont conditionnels aux covariables observées seules, si on a la possibilité de mesurer un grand nombre de covariables que l'on pense liées à l'affectation du traitement, alors on peut assurément obtenir des estimations approximativement non-biaisées de l'effet traitement.

2.3 Utilisation du score de propension

Cette partie a pour références [4], [5], [6] et [7] (voir bibliographie).

2.3.1 Appariement

Souvent, les investigateurs sont confrontés à des études où le nombre de patients dans le groupe traités est limité et le nombre de patients dans le groupe référence est beaucoup plus grand. L'appariement (matching) est une méthode communément utilisée pour sélectionner les sujets référence qui sont 'appariés (matched)' aux sujets traités sur des covariables jugées pertinentes. Bien que l'idée de former des paires semble simple, il est souvent difficile de trouver des sujets similaires (autrement dit, qui peuvent être appariés) en particulier lorsque le nombre de covariables augmente.

L'appariement sur le score de propension résout ce problème en permettant à l'investigateur de contrôler plusieurs covariables simultanément en appariant sur une seule variable scalaire. On range aléatoirement les sujets traités et non-traités, puis on sélectionne le premier sujet traité et on cherche le sujet non-traité qui a le score de propension le plus proche. Les deux sujets sont ensuite retirés de la liste et on sélectionne le sujet traité suivant. On continue ainsi jusqu'à ce qu'il n'y ait plus de sujets traités et non-traités ayant des score de propension assez semblables pour être appariés, l'échantillon des paires de sujet traités et non-traités qui ont été retirés est alors notre échantillon apparié. Il s'agit d'un échantillon dans lequel les différences pour les covariables selectionnées dans le score de propension sont réduites. On peut donc ensuite procéder aux

comparaisons efficacité sur le critère de jugement choisi dans ce nouvel échantillon a priori non-biaisé.

2.3.2 Pondération

La pondération (weighting) créé une pseudo-population dans laquelle l'affectation du traitement est indépendante des variables parasites mesurées. Cette pseudo-population est le résultat de l'affectation de chaque sujet à un poids qui est, informellement, proportionnel à la probabilité qu'a ce sujet de recevoir son traitement (ou son non-traitement). Dans une telle pseudo-population, on peut faire une régression des critères de jugement sur l'affectation du traitement en utilisant un modèle de régression conventionnel qui n'inclue pas les variables parasites comme covariables. Rentrer un modèle dans la pseudo-population est équivalent à rentrer un modèle pondéré dans la population de l'étude. On parle alors de *MSM (Marginal Structural Model)*, qui peut être utilisé pour estimer l'effet traitement moyen dans la population de l'étude.

La méthode de pondération utilise l'inverse du score de propension comme poids pour les sujet traités (ps=1/Prob , Prob étant le score de propension), et l'inverse de un moins le score de propension comme poids pour les sujet non-traités (ps=1/(1-Prob)). Chaque individu est donc pondéré par l'inverse de la probabilité qu'il a d'être dans son groupe. On donne ainsi plus de poids aux individus peu représentés dans l'échantillon.

2.3.3 Stratification

La stratification consiste à grouper les sujets en strates déterminées par les caractéristiques générales observées. Une fois que les strates sont définies, les sujets traités et non-traités qui sont dans la même strate sont comparés directement. Beaucoup de problèmes surviennent lors de la stratification ainsi que lors de l'appariement quand le nombre de covariables augmente. Cochran a fait remarquer que lorsque le nombre de covariables augmente, le nombre de strates augmente de façon exponentielle. En effet, si toutes les covariables sont binaires, alors il y a 2^k sous-classes pour k covariables. Si k est grand, alors certaines strates peuvent contenir des sujets du groupe traité uniquement, ce qui rendrait impossible d'estimer l'effet traitement dans cette strate. Le score de propension est alors très utile, car il est un résumé scalaire de toutes les covariables générales observées. La stratification sur lui seul peut alors équilibrer la distribution des covariables dans les groupes traités et non-traités sans l'augmentation exponentielle du nombre de strates.

Rosenbaum et Rubin affirment que la stratification sur le score de propension équilibre toutes les k covariables utilisées pour estimer le score de propension, et que souvent cinq strates basées sur le score de propension suppriment plus de 90 pourcent du biais dans chacune de ces covariables.

2.4 Analyse de survie

Cette partie a pour références [8] à [14] (voir bibliographie).

2.4.1 Fonction de survie et risque instantané

Soit une variable aléatoire T qui désigne le *temps de survie* d'un être vivant ou d'un système inanimé. On l'appelle également *âge au décès* ou *âge à l'échec*. La distribution du temps de survie est représentée par la fonction :

$$S(t) = 1 - F(t) = P(T > t) \quad avec\ t \geq 0$$

F est appelée *fonction d'incidence cumulée*, elle représente la probabilité de décès (ou réalisation de l'évènement) avant l'instant t. On a $F(t) = P(T \leq t)$.

S est appelée *fonction de survie*.

Si F est différentiable, alors sa dérivée est la fonction de densité de la distribution de survie, que l'on note f :

$$f(t) = F'(t) = \frac{d}{dt}F(t) = \lim_{h\to 0}(\frac{F(t+h) - F(t)}{h})$$

Une notion très important dans l'analyse de survie est la fonction de risque instantané (hazard function) h définie par :

$$h(t) = \lim_{h\to 0}(\frac{P(t \leq T < t + h | T \geq t)}{h}) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

La fonction de risque instantané modélise la probabilité de décès à l'instant t pour un patient sachant que ce patient a survécu jusqu'à l'instant t.

Les fonctions f, F, S et h décrivent de manière mathématiquement équivalente la distribution de T. En effet, ces fonctions sont très liées, puisque $f(t) = -S'(t)$, nous avons que :

$$h(x) = -\frac{d}{dx}\ln(S(x))$$

ainsi,

$$\ln\big(S(x)\big)\,|_0^t = -\int_0^t h(x)dx$$

et comme S(0)=1,

$$S(t) = \exp(-\int_0^t h(x)dx)$$

Il peut aussi être utile de définir la fonction de risque cumulé :

$$H(t) = \int_0^t h(x)dx$$

Qui est donc également liée à la fonction de survie, en effet, $S(t) = \exp(-H(t))$. Si S(∞)=0, alors H(∞)=∞. Finalement, on a immédiatement que :

$$f(t) = h(t)\exp(-\int_0^t h(x)dx)$$

2.4.2 La censure à droite

La censure à droite, selon laquelle seule la borne inférieure du temps de survie est disponible pour un individu, peut se produire pour différentes raisons. Elle peut être prévue, comme lorsque la décision est prise d'achever une analyse de survie avant que tous les sujets aient subit le décès, ou imprévue, comme quand un sujet est 'perdu de vue' lors d'une étude (par exemple, parce qu'il a quitté la région où a lieu l'étude).

Pour traduire au mieux ce type d'observation, on représente les données de survie comme des paires de variables (T,δ), où T est le temps écoulé depuis l'entrée dans l'étude et δ est une indicatrice de décès, en supposant que δ=1 si l'évènement est observé et δ=0 si la durée est censurée. Formellement, on suppose que U est la durée de vie réelle que l'on souhaite connaitre mais ne peut pas toujours observer, et V est le temps potentiel jusqu'à la censure. La durée de vie observée est alors :

$$T = \min(U, V) \quad \text{et} \quad \delta = \begin{cases} 1 \; si \; U \leq V \\ 0 \; si \; U > V \end{cases}$$

Ainsi, T=U seulement si l'observation n'est pas censurée.

2.4.3 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur de la fonction de survie. Il s'agit de l'estimation non-paramétrique du maximum de vraisemblance de S(t) défini par :

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

Où :

- les t_i sont les instants auxquels un décès a lieu
- n_i est nombre de survivants moins le nombre de perdus de vue juste avant l'instant t_i. Ce sont les survivant qui sont toujours observés, les non-censuré ou *sujets à risque*
- d_i est le nombre de décès à l'instant t_i

On peut, d'après l'estimateur défini ci-dessus, construire la courbe de Kaplan-Meier qui donne une représentation de l'estimation de la fonction de survie (Figure 2.4.31).

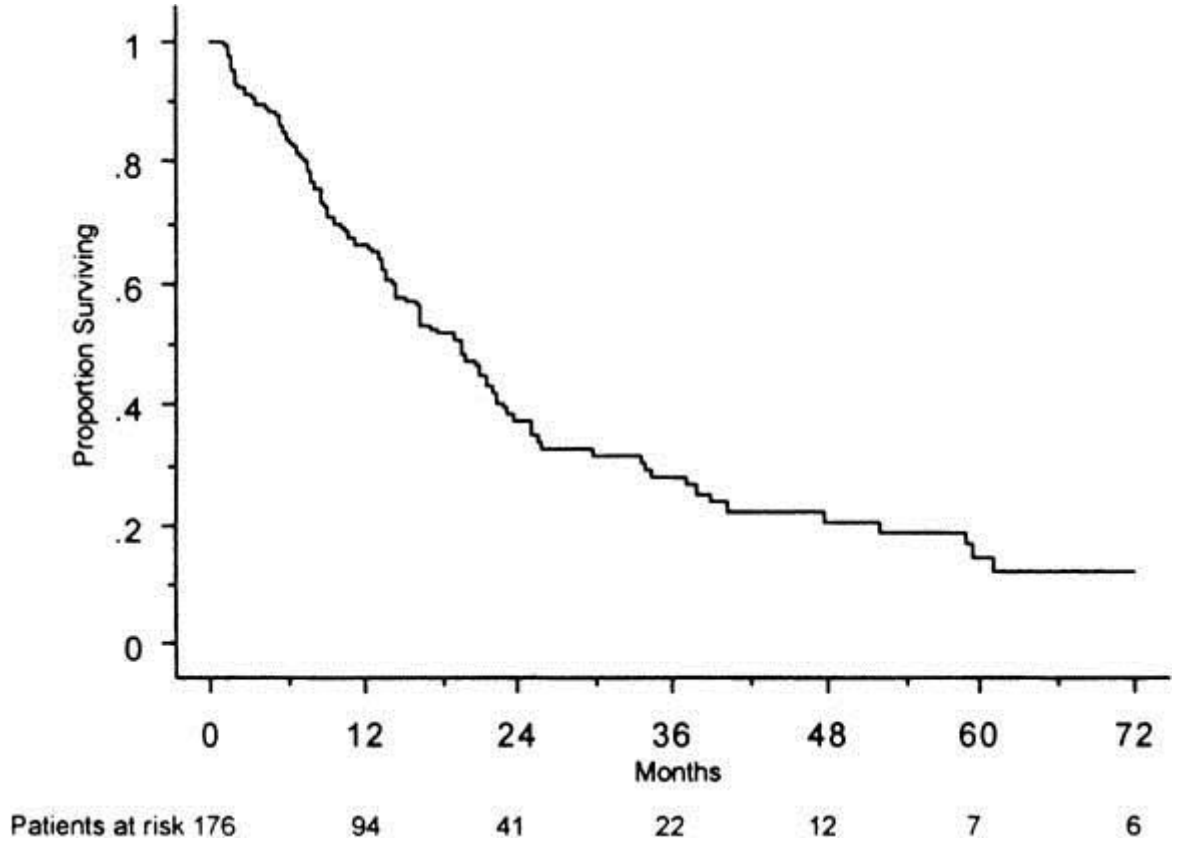


Figure 2.4.31 : Exemple de courbe de Kaplan-Meier pour la survie

2.4.4 Modèle de Cox

Le modèle de Cox (modèle à risques proportionnels) offre une méthode pour estimer l'effet des covariables sur le temps de survie. Il exprime la fonction de risque instantané en fonction du temps t et du vecteur $X = (X_1, \dots, X_n)$ des covariables. On a :

$$h(t|X) = h_0(t)\exp(\beta' X)$$

Où h_0 est le risque instantané de référence et β est le vecteur des coefficients de régression.

On voit que la formule impose au modèle une condition, qui est une hypothèse essentielle du modèle de Cox, celle des *risques proportionnels*. Pour deux individus i et j, le rapport des fonctions de risque instantané (hazard ratio) par :

$$\frac{h(t,j)}{h(t,i)} = \frac{h_0(t)\exp(\beta' X_j)}{h_0(t)\exp(\beta' X_i)} = \exp\left(\beta' X_j - \beta' X_i\right) = \exp(\beta'\left(X_j - X_i\right))$$

Le rapport est donc indépendant du temps et les fonctions de risque instantané des individus i et j sont proportionnelles. A tout instant t, l'individu j a un risque instantané de mourir $\exp(\beta'\left(X_j - X_i\right))$ fois plus élevé que celui de l'individu i.

Cox suggère d'estimer les paramètres de la régression en maximisant la vraisemblance partielle en β :

$$L(\beta) = \prod_{i \in D} \frac{\exp(\beta' X_i)}{\sum_{j:T_j \geq T_i} \exp(\beta' X_j)}$$

Où D est l'ensemble des individus i décédés (tels que δ=1) et T_i la durée de vie observée définie précédemment.

Chapitre 3

Présentation des données

Nous disposons d'une base de données issue de deux études, dans lesquelles les types de pansement du patient ('Groupe'), ainsi que la fréquence de changement du pansement ('Rythme_pansement') ont été tirés au sort. Le choix du site d'insertion du cathéter (jugulaire ou sous-clavière) est par ailleurs laissé au médecin.

La base est composée de 2065 patients, 933 ayant reçu un cathéter en voie jugulaire, et 1132 en voie sous-clavière. Chaque patient dispose d'un seul cathéter. Il y a relativement peu de valeurs manquantes (15 au total), de plus nous les considéreront manquantes aléatoirement (MAR). Nous ferons donc les analyses sans en tenir compte. Après l'importation des données et le recodage de certaines variables (voir A.0 dans l'annexe), nous considérons les variables explicatives quantitatives suivantes :

Age : L'âge du patient.

Saps2 : Simplified Acute Physiologic Score. Il s'agit d'un score pour mesurer la sévérité de maladie du patient. Il prédit la mortalité en réanimation.

Sofa : Sequential Organ Failure Assessment. Score pour mesurer l'état du patient pendant la période de soins intensifs.

Ainsi que les variables explicatives qualitatives :

Sexe : Le sexe du patient : 'H' ou 'F'.

Vm_a_la_pose : Ventilation mécanique à la pose du cathéter : 'oui' ou 'non'.

Inotrope_a_la_pose : Prise d'inotrope (médicament) à la pose du cathéter : 'oui' ou 'non'.

Chlorex : Utilisation de chlorhexidine : 'oui' ou 'non'.

Atb_insertion : Utilisation d'antibiotique a l'insertion du cathéter : 'oui' ou 'non'.

Lipides : Alimentation parentérale : 'oui' ou 'non'.

Heparines : Prise d'héparines (anticoagulant) : 'oui' ou 'non'.

Groupe : Type de pansement : 'Biopatch', 'Tegaderm CHG', 'Tegaderm HP' ou 'Tegaderm standard'.

Ventilationinvasive : Ventilation invasive : 'oui' ou 'non'.

Ventilationinvasiveavecpeep : Ventilation invasive avec peep : 'oui' ou 'non'.

Ventilationnoninvasive : Ventilation non-invasive : ‘oui’ ou ‘non’.

Base_dressing : Base de laquelle est issu le patient : 1 ou 2.

Centre : Centre où était hospitalisé le patient : de C1 à C16.

Rythme_pansement : Fréquence de changement du pansement : 3 ou 7 jours.

Maccabe : Score qui prédit le décès du patient : ‘Non prévu’, ‘Prévu 1 à 5 ans’ ou ‘Prévu dans année’.

Cat_admission : Catégorie d’admission : ‘Médical’, ‘Chirurgical_programme’ ou ‘Chirurgical_urgent’.

Motifadm : Motif d’admission du patient en réanimation, recodage de ‘motif_adm’ avec moins de modalités : ‘Choc septique’, ‘Choc cardiogénique’, ‘Détresse respiratoire aigüe’, ‘Coma’, ‘Traumatisme’ ou ‘Autre’.

Experience_operateur : Expérience du médecin chargé de poser le cathéter : ‘Junior’, ‘Sénior’ ou ‘Sénior après junior’.

Tunnel : Tunnelisation du cathéter sous la veine : ‘oui’ ou ‘non’.

Cote : Coté de pose du cathéter : ‘D’ pour ‘Droite’ et ‘G’ pour ‘Gauche’.

Nb_lum : Nombre de lumières du cathéter : de 1 à 4.

Annee_entree_rea : Année d’entrée en réanimation du patient, obtenue à partir de ‘entrée_reanimation’.

La variable qui permet de comparer les groupes :

Sclav : Type du cathéter (recodage de ‘sous_clav’) : 0 si le cathéter est posé en voie jugulaire, et 1 en voie sous-clavière.

La variable temps utilisée dans le modèle de Cox :

Duree_catheter : Durée du cathéter, de la pose à l’ablation.

Les critères de jugement, variables (qualitatives) que l’on cherche à étudier :

Ktcol : Colonisation du cathéter (recodage de ‘kt_colonise’) : 0 pour ‘non’, 1 pour ‘oui’.

Minf : Infection majeure du cathéter (recodage de ‘major_infection’) : 0 pour ‘non’, 1 pour ‘oui’.

Clab : Infection du sang liée au cathéter (Central-Line Associated Bloodstream infection) (recodage de ‘crbsi’) : 0 pour ‘non’, 1 pour ‘oui’.

Chapitre 4

Application sur base de données

Le choix du site d'insertion du cathéter étant laissé au médecin, nous utiliserons le score de propension pour réduire le biais de notre échantillon. Les risques d'événements nosocomiaux étant liés au temps, nous utiliserons des modèles de Cox pour analyser les critères de jugement.

4.1 Comparaison entre les deux groupes

L'étude n'étant pas randomisée sur le site d'insertion, il est possible qu'il y ait une différence significative (biais) des covariables entre les individus auxquels on a posé un cathéter en voie jugulaire et ceux auxquels on a posé un cathéter en voie sous-clavière. Ainsi, au lieu de passer à l'analyse brute des données, nous allons tout d'abord effectuer des tests pour comparer les deux groupes. Pour les covariables quantitatives, nous effectuerons des tests t de Welch [16] de comparaison de la moyenne (car il ne nécessite pas l'hypothèse d'égalité des variances, contrairement au test classique de Student). Pour les covariables qualitatives, nous effectuerons des tests du Khi-deux d'homogénéité.

Pour le test de Welch la statistique de test est :

$$t = \frac{\overline{X}_\iota - \overline{X}_J}{\sqrt{\frac{{s_i}^2}{N_i} + \frac{{s_j}^2}{N_j}}}$$

Où, respectivement pour les échantillons i et j :

- $\overline{X}_\iota$ et $\overline{X}_J$ sont les moyennes
- ${s_i}^2$ et ${s_j}^2$ sont les variances
- N_i et N_j sont les tailles d'échantillon

Cette statistique suit une loi de Student à v degrés de liberté, avec :

$$v = \frac{(\frac{s_i^2}{N_i} + \frac{s_j^2}{N_j})^2}{\frac{{s_i}^4}{{N_i}^2(N_i - 1)} + \frac{{s_j}^4}{{N_j}^2(N_j - 1)}}$$

L'hypothèse nulle est H_0: $Les\ moyennes\ sont\ égales$

Ce test a pour hypothèse la distribution normale de la variable, mais n'est pas très sensible aux déviations de celle-ci (on peut tout de même faire un test non-paramétrique pour plus de sureté). On l'effectue dans SAS avec la procédure 'proc ttest' (voir annexe A.1).

Pour le test du Khi-deux la statistique de test est :

$$k = \sum_{i,j} \frac{(n_{ij} - n_i p_j)^2}{n_i p_j}$$

Où :

- n_{ij} = effectif observé des individus de l'échantillon i, possédant la modalité j de la variable
- n_i = effectif total observé dans l'échantillon i
- p_j = probabilité d'obtenir une observation possédant la modalité j de la variable lorsqu'on est en présence d'une seule population

Cette statistique suit une loi du Khi-deux à (K-1)(R-1) degrés de liberté, K étant le nombre d'échantillons et R le nombre de modalités de la variable.

L'hypothèse nulle est H_0: $Les\ échantillons\ sont\ issus\ de\ la\ même\ population$

On effectue ce test dans SAS avec la procédure 'proc freq' et l'option 'chisq' (voir annexe A.2).

On utilise la procédure 'proc univariate' pour calculer les médianes pour les variables quantitatives (voir annexes A.3).

On calcule également la différence standardisée SD (standardized difference) [15] à titre de comparaison. Pour deux échantillons i et j, elle s'exprime pour les variables continues par :

$$SD = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{s_i^2 + {s_j}^2}{2}}}$$

$\bar{X}_i$ étant la moyenne et s_i^2 la variance de l'échantillon i, et $\bar{X}_j$ la moyenne et s_j^2 la variance de l'échantillon j.

Et pour les variables discrètes par :

$$SD = \frac{\widehat{p_i} - \widehat{p_j}}{\sqrt{\frac{\widehat{p_i}(1 - \widehat{p_i}) + \widehat{p_j}\left(1 - \widehat{p_j}\right)}{2}}}$$

$\widehat{p_i}$ étant la proportion d'individu observant le caractère dans l'échantillon i, et $\widehat{p_j}$ la proportion d'individu observant le caractère dans l'échantillon j.

On peut alors dresser le tableau de comparaison entre les deux groupes joint en annexe (partie 'Toute la cohorte' de B.1). On remarque alors que la p-value est inférieure à 0.05, donc significative au seuil α=5%, pour les covariables : SAPS2, Inotrope_a_la_pose, Chlorex, Lipides,

Heparines, Base_dressing, Centre, Rythme_pansement, Cat_admission, Motifadm, Operator_experience, Cote et Annee_entree_rea.

On utilise à nouveau la procédure 'proc univariate' avec l'option 'histogram' pour obtenir une représentation des différences de distribution des covariables Age, Saps2 et Sofa entre les groupes jugulaire et sous-clavière (voir A.3 et B.2).

Ces deux groupes n'étant pas homogènes, nous décidons de procéder à un appariement sur le score de propension pour réduire le biais.

4.2 Calcul du score de propension par la régression logistique

Nous allons calculer le score de propension, qui sera la probabilité de recevoir un cathéter en voie jugulaire (c'est-à-dire : P(sclav=0), il s'agit de ce que SAS calcule par défaut), grâce à la régression logistique présentée précédemment. Nous commençons par effectuer une régression logistique sur toutes les variables sans interaction. Dans SAS, on utilise la procédure 'proc logistic', avec l'option 'LACKFIT' pour le test de Hosmer-Lemeshow (voir A.4).

Ce modèle fournit un AUC de 0.781 (Figure 4.21), mais la p-value du test de Hosmer-Lemeshow est de 0.0458 donc significative (Figure 4.22). Le modèle ne convient apparemment donc pas. Par ailleurs, on remarque en regardant le tableau de sélection des variables que les variables les plus influentes dans le modèle sont : Centre, Cote et Annee_entree_rea (Figure 4.23).

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.0	Somers' D	0.562
Percent Discordant	21.8	Gamma	0.563
Percent Tied	0.2	Tau-a	0.278
Pairs	1045656	c	0.781

Figure 4.21 : AUC ('c') donné par SAS

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
15.7681	8	0.0458

Figure 4.22 : Résultat du test de Hosmer-Lemeshow sous SAS

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	centre		15	1	231.5064		<.0001	centre
2	cote		1	2	137.8889		<.0001	cote
3	annee_entree_rea		4	3	68.7767		<.0001	
4	motifadm		5	4	43.3212		<.0001	
5	heparines		1	5	9.1498		0.0025	heparines
6	inotrope_a_la_pose		1	6	7.7089		0.0055	inotrope_a_la_pose
7	cat_admission		2	7	5.6641		0.0589	cat_admission
8	ventilationnoninvasi		1	8	2.2920		0.1300	ventilationnoninvasive
9	VM_a_la_pose		1	9	2.7361		0.0981	VM_a_la_pose
10	SAPS2		1	10	1.9588		0.1616	SAPS2
11	age		1	11	2.2938		0.1299	age
12	atb_insertion		1	12	1.2468		0.2642	atb_insertion
13	chlorex		1	13	0.6185		0.4316	chlorex

Figure 4.23 : Sélection des variables selon la procédure STEPWISE

Au vu de cela, nous décidons de refaire la régression sur toutes les variables, plus les interactions entre Centre et Cote, et entre Centre et Annee_entree_rea (voir A.4).

Cette fois, le modèle fournit un AUC de 0.806 (Figure 4.24) et une p-value de 0.8805 (donc non significative) pour le test de Hosmer-Lemeshow (Figure 4.25). On accepte donc le modèle.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.5	Somers' D	0.613
Percent Discordant	19.3	Gamma	0.614
Percent Tied	0.2	Tau-a	0.303
Pairs	1045656	c	0.806

Figure 4.24 : AUC (‘c’) pour la deuxième régression

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
3.7320	8	0.8805

Figure 4.25 : Résultat du test de Hosmer-Lemeshow pour la deuxième régression

4.3 Appariement sur le score de propension et comparaison entre les deux groupes après appariement

La régression faite ci-dessus fournit la variable Prob, qui est le score de propension, sur lequel nous allons effectuer l'appariement. On utilise pour cela une macro dans SAS prévue à cet effet (macro 'OneToManyMTCH'). Elle nous fournit un sous-échantillon de 'paires' d'effectif 1090 (545 sujets en jugulaire et 545 sujets en sous-clavière). Nous procédons ensuite à l'analyse faite en 4.1, cette fois sur cet échantillon apparié (sous SAS on remplace 'data=cathe.jsc2' par 'data=cathe.matches_1' qui est la table de l'échantillon apparié), ce qui nous permet de dresser un tableau de comparaison entre les deux groupes (partie 'Après appariement sur le score de propension' de B.1).

On remarque que maintenant, toutes les p-values sont non significatives (supérieures à 0.05) au seuil 5%. On a donc bien réduit le biais dû à l'étude non-randomisée sur le site d'insertion. On utilise la procédure 'proc univariate' avec l'option 'histogram' pour obtenir une représentation des différences de distribution de la variable Prob, représentant le score de propension, entre les groupes jugulaire et sous-clavière avant et après appariement (voir A.5 et B.3). Nous allons à présent entreprendre l'analyse des critères de jugement.

4.4 Analyse des critères de jugement

(Pour cette partie, regarder B.4 dans l'annexe)

Nous allons procéder à l'analyse des variables Ktcol, Minf et Clab, à la fois sur cette nouvelle base de donnée que l'on considère non-biaisée, et sur la base de données brute, pour pouvoir faire la comparaison.

On commence par tracer les courbes de Kaplan-Meier de chaque groupe, 'Jugulaire' et 'Sous-clavière', pour les données brutes et pour l'échantillon apparié à l'aide de la procédure 'proc lifetest'

dans SAS (voir A.6). L'option 'lls' dans 'plots' nous donne les courbes log(-log(survie)) sur log(temps de survie) pour les groupes 'Jugulaire' et 'Sous-clavière' (voir B.5). Ces courbes devraient être 'parallèles', avoir même forme si les risques sont proportionnels. Il s'agit de la façon graphique pour vérifier cette hypothèse, nous ferons un test pour plus de sureté ci-dessous.

On modélise ensuite les risques grâce aux modèles de Cox, à l'aide de la procédure 'proc phreg' (voir A.7). Notre variable de temps sera 'Duree_catheter' et notre variable explicative sera 'Sclav'. On commence par faire la régression sur les variables 'Sclav' et 'Sclavt', cette dernière étant l'interaction entre 'Sclav' et le log de 'Duree_catheter', en incluant un test pour vérifier l'hypothèse des risques proportionnels avec 'proportionality_test' (H_0: $Les\ risques\ sont\ proportionnels$). Les tests étant tous non-significatifs (Figures 4.41 à 4.43), l'hypothèse essentielle des risques proportionnels n'est pas rejetée. On fait alors la régression avec 'Sclav' pour seule variable explicative, on obtient alors nos modèles.

Pour l'échantillon apparié, on utilise un modèle de Cox classique. Il s'agit du modèle robuste, qui nous donne les rapports des fonctions de risque instantané (hazard ratio), 'Jugulaire' sur 'Sous-clavière', que l'on a nommé rHR. Pour les données brutes, on utilise d'abord un modèle de Cox brut, lequel fournit les hazard ratios que l'on nomme HR. Enfin, pour ces mêmes données, on modélise en pondérant inversement sur le score de propension. Pour ce faire, nous créons un nouveau jeu de données avec la variable 'Ps', qui, pour chaque individu, est l'inverse de la probabilité qu'il a de recevoir son cathéter. On appelle les hazard ratios obtenus wHR.

Linear Hypotheses Testing Results
with Sandwich Variance Estimate

	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.4576	1	0.4987

Linear Hypotheses Testing Results
with Sandwich Variance Estimate

	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	1.1912	1	0.2751

Linear Hypotheses Testing Results
with Sandwich Variance Estimate

	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.0046	1	0.9461

Figure 4.41 : Résultats des tests de proportionnalité des risques dans le modèle robuste, respectivement pour : 'Ktcol', 'Minf' et 'Clab'

Linear Hypotheses Testing Results				
	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.0559	1	0.8130

Linear Hypotheses Testing Results				
	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.1909	1	0.6622

Linear Hypotheses Testing Results				
	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.0488	1	0.8252

Figure 4.42 : Résultats des tests de proportionnalité des risques dans le modèle brut, respectivement pour : 'Ktcol', 'Minf' et 'Clab'

Linear Hypotheses Testing Results				
	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.5022	1	0.4785

Linear Hypotheses Testing Results				
	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.8736	1	0.3500

Linear Hypotheses Testing Results				
	Label	Wald Chi-Square	DF	Pr > ChiSq
	proportionality_test	0.0433	1	0.8352

Figure 4.43 : Résultats des tests de proportionnalité des risques dans le modèle pondéré, respectivement pour : 'Ktcol', 'Minf' et 'Clab'

Ces modèles nous fournissent également des tests de significativité globale qui testent la nullité des coefficients de la régression (H_0: $\beta_1 = \cdots = \beta_n = 0$). Il s'agit de définir si 'les deux courbes de survie sont différentes'.

On joint aux graphiques des courbes les hazard ratios, leurs intervalles de confiance au seuil 95%, et les p-values P des tests de nullité des coefficients β de la régression.

4.5 Commentaire des résultats

Dans l'analyse faite en 4.4, le risque de colonisation s'est montré plus important dans le groupe des patients ayant reçu un cathéter en voie jugulaire, avec des résultats semblables pour tous les modèles. Les p-values étant significatives, elles nous permettent de fortement soupçonner un effet du type de cathéter sur la survie.

Les différences entre les groupes 'Jugulaire' et 'Sous-clavière' pour l'infection majeure du cathéter (Minf) et l'infection du sang liée au cathéter (Clab) ne se sont quant à elles pas montrées significatives. On remarque que la p-value dans le modèle pondéré est significative pour 'Minf', mais ce résultat ne converge pas avec les résultats obtenus pour les autres modèles. Cela ne nous permet donc pas de conclure à une différence entre les groupes, la courbe de Kaplan-Meier de 'Jugulaire' étant malgré tout en dessous de celle de 'Sous-clavière'.

Nous avons donc réussit à rendre ces deux groupes 'Jugulaire' et 'Sous-clavière' comparables grâce à ces méthodes statistiques, avec en conclusion une forte présomption de différence en faveur des cathéters 'Sous-clavière'.

On conclut également que le score de propension peut donner des résultats discordant (par exemple pour 'Minf' dans notre étude), d'où l'intérêt de faire plusieurs analyses.

Conclusion

Pendant ces deux mois de stage, j'ai pu participer activement à l'activité du service de Biostatistique et de Recherche Clinique, d'abord dans le cadre de mon sujet de stage, et par la suite en travaillant avec l'équipe sur d'autres projets.

En m'investissant dans ce travail, j'ai réalisé que mes connaissances théoriques acquises au cours de cette année de Master 1 devaient s'accompagner de beaucoup de documentation afin de pouvoir accomplir les tâches demandées, le manque d'expérience étant également un frein. J'ai donc eu l'opportunité d'approfondir mes acquis et de m'initier à de nouvelles notions théoriques et à des méthodes statistiques. Travailler sur SAS quotidiennement m'a également permis de me familiariser avec ce logiciel.

L'insertion en milieu hospitalier, notamment dans la recherche clinique, a été très intéressante et m'a apporté une vision plus concrète du métier de biostatisticien et de l'utilité de celui-ci. J'ai pu me rendre compte des différences entre la théorie et la pratique, et prendre conscience des difficultés que l'on peut rencontrer.

Pour conclure, ce stage a été une expérience très enrichissante, tant dans la théorie que dans la pratique. Il m'a permis d'apprendre des méthodes statistiques et de comprendre l'importance de la recherche permanente de documentation.

Bibliographie

[1] Crabbé B. (2007). Régression Logistique - http://www.linguist.univ-paris-diderot.fr/~bcrabbe/LingExp/cours7.pdf

[2] Rakotomala R. Régression logistique, une approche pour rendre calculable P(Y|X) - http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique.pdf

[3] Hosmer DW, Lemeshow S. (2013). Applied logistic regression

[4] D'Agostino RB. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group - Stat Med

[5] Rosenbaum PR, Rubin DB. (1983). The central role of the propensity score in observational studies for causal effects - Biometrika, 70, 41–55

[6] Cole SR, Hernán MA, Robins JM. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models – American Journal of Epidemiology

[7] Rubin DB. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation

[8] Elandt-Johnson RC, Johnson NL. (1999). Survival models and data analysis – New York, Wiley

[9] Lawless JF. (2003). Statistical Models and Methods for Lifetime Data (2nd ed.). Hoboken: John Wiley and Sons

[10] Marubini E, Valsecchi MG. (2004). Analysing survival data from clinical trials and observational studies – Wiley

[11] Kaplan EL, Meier P. (1958). Nonparametric estimation from incomplete observations - Journal of the American statistical association

[12] Costella JP. (2010). A simple alternative to Kaplan–Meier for survival curves

[13] Cox DR. (1995). Partial likelihood – Biometrika

[14] Tsiatis AA. (1981). A large sample study of Cox's regression model – The Annals of Statistics, JSTOR

[15] Yang D, Dalton JE. A unified approach to measuring the effect size between two groups using SAS®

[16] Welch BL. (1947). The generalization of "Student's" problem when several different population variances are involved - Biometrika

Annexes

A – Codes SAS

A.0 Importation des données et recodage

Création de la librairie :

```
libname cathe 'P:\CatheJugSousclav'; /* creation lib*/
```

Importation des données :

```
proc import out=cathe.jsc
datafile = "P:\CatheJugSousclav\firstKT with probs.xls"
dbms = xls replace;
run;
```

Recodage des variables :

```
data cathe.jsc2;
set cathe.jsc;
annee_entree_rea=year(entree_reanimation);
if kt_colonise="oui" then ktcol=1;
if kt_colonise="non" then ktcol=0;
if major_infection="oui" then minf=1;
if major_infection="non" then minf=0;
if crbsi="non" then clab=0;
if crbsi="oui" then clab=1;
if sous_clav="Jug" then sclav=0;
else sclav=1;

if motif_adm="choc_septique" then motifadm="Choc septique";
else motifadm="Autre";
if motif_adm="choc_cardiogenique" then motifadm="Choc cardiogenique";
if motif_adm="detresse_respiratoire_aigue" then motifadm="Detresse
respiratoire aigue";
if motif_adm="coma" then motifadm="Coma";
if motif_adm="traumatisme" then motifadm="Traumatisme";
run;
```

A.1 Tests de Welch

```
proc ttest data=cathe.jsc2;
class sous_clav;
var age saps2 sofa;
run;
```

A.2 Tests du Khi-deux

```
proc freq data=cathe.jsc2;
table sexe*sous_clav vm_a_la_pose*sous_clav inotrope_a_la_pose*sous_clav
chlorex*sous_clav atb_insertion*sous_clav lipides*sous_clav
heparines*sous_clav groupe*sous_clav ventilationinvasive*sous_clav
ventilationinvasiveavecpeep*sous_clav ventilationnoninvasive*sous_clav
base_dressing*sous_clav centre*sous_clav rythme_pansement*sous_clav/chisq;
run;

proc freq data=cathe.jsc2;
table maccabe*sous_clav cat_admission*sous_clav motifadm*sous_clav
experience_operateur*sous_clav tunnel*sous_clav cote*sous_clav
nb_lum*sous_clav annee_entree_rea*sous_clav/chisq;
run;
```

A.3 Univariate

Calcule de la médiane du total :

```
proc univariate data=cathe.jsc2;
var age saps2 sofa;
run;
```

Et des médianes dans les groupes :

```
proc univariate data=cathe.jsc2;
class sous_clav;
var age saps2 sofa;
run;
```

Les histogrammes de distribution :

```
proc univariate data=cathe.jsc2 normal;
class sous_clav;
var age saps2 sofa;
histogram age saps2 sofa;
inset q1 median mean q3 normal min max p5 p95/position=NE noframe;
run;
```

A.4 Régressions logistiques

Régression logistique :

```
proc logistic data=cathe.jsc2 nosimple;
class sexe vm_a_la_pose inotrope_a_la_pose chlorex atb_insertion lipides
heparines groupe ventilationinvasive ventilationinvasiveavecpeep
ventilationnoninvasive base_dressing centre rythme_pansement maccabe
cat_admission motifadm experience_operateur tunnel cote nb_lum
annee_entree_rea;
model sclav = age saps2 sofa sexe vm_a_la_pose inotrope_a_la_pose chlorex
atb_insertion lipides heparines groupe ventilationinvasive
ventilationinvasiveavecpeep ventilationnoninvasive base_dressing centre
```

```
rythme_pansement maccabe cat_admission motifadm experience_operateur tunnel
cote nb_lum annee_entree_rea
/ selection=stepwise include=0 LACKFIT RSQUARE PARMLABEL slentry=0.5
sls=0.5;
output out=cathe.preds pred=prob;
run;
```

Régression logistique avec les interactions :

on ajoute 'centre*cote' et 'centre*annee_entree_rea' dans le modèle

A.5 Histogrammes de la distribution de la variable Prob

Avant appariement :

```
proc univariate data=cathe.jsc2;
class sclav;
var prob;
histogram prob/endpoints=0 to 1 by 0.05 normal;
inset n q1 median mean q3 normal min max p5 p95/position=NE noframe;
run;
```

Après appariement :

on remplace 'data=cathe.jsc2' par 'data=cathe.matches_1'

A.6 Courbes de Kaplan-Meier (pour Ktcol)

Dans l'échantillon d'origine :

```
proc lifetest data=cathe.jsc2 method=KM
 plots=(s(atrisk=0 to 50 by 5),lls) cs=none maxtime=51;
    Time duree_catheter*ktcol(0);
                strata sous_clav;
                symbol1 v=none h=1 color=red line=1;
                symbol2 v=none h=1 color=blue line=2;
                       label duree_catheter="Durée catheter (jours)";
run;
```

Dans l'échantillon apparié :

on remplace 'data=cathe.jsc2' par 'data=cathe.matches_1'

A.7 Modèles de Cox (pour Ktcol)

Modèle robuste :

```
proc phreg covs(aggregate) covm data=cathe.matches_1; /*modèle pour test de
proportionnalité*/
class sclav;
model duree_catheter*ktcol(0)=sclav sclavt / rl ties=efron;
sclavt = sclav*log(duree_catheter);
proportionality_test: test sclavt;
id match_1;
run;

proc phreg covs(aggregate) covm data=cathe.matches_1; /*modèle final*/
class sclav;
model duree_catheter*ktcol(0)=sclav /selection=s
include=1 slentry=0.2 sls=0.05 rl ties=efron;
id match_1;
run;
```

Modèle brut :

```
proc phreg data=cathe.jsc2; /*modèle pour test de proportionnalité*/
class sclav;
model duree_catheter*ktcol(0)=sclav sclavt/ rl ties=efron;
sclavt = sclav*log(duree_catheter);
proportionality_test: test sclavt;
run;quit;

proc phreg data=cathe.jsc2; /*modèle final*/
class sclav;
model duree_catheter*ktcol(0)=sclav/ rl ties=efron;
run;quit;
```

Modèle pondéré :

```
data cathe.MSM;
set cathe.preds;
if sclav=0 then ps=1/prob;
if sclav=1 then ps=1/(1-prob);
run;quit;

proc phreg data=cathe.msm; /*modèle pour test de proportionnalité*/
class sclav;
model duree_catheter*ktcol(0)=sclav sclavt/ rl ties=efron;weight ps;
sclavt = sclav*log(duree_catheter);
proportionality_test: test sclavt;
run;quit;

proc phreg data=cathe.msm; /*modèle final*/
class sclav;
model duree_catheter*ktcol(0)=sclav/ rl ties=efron;weight ps;
run;quit;
```

B – Tableau et graphiques

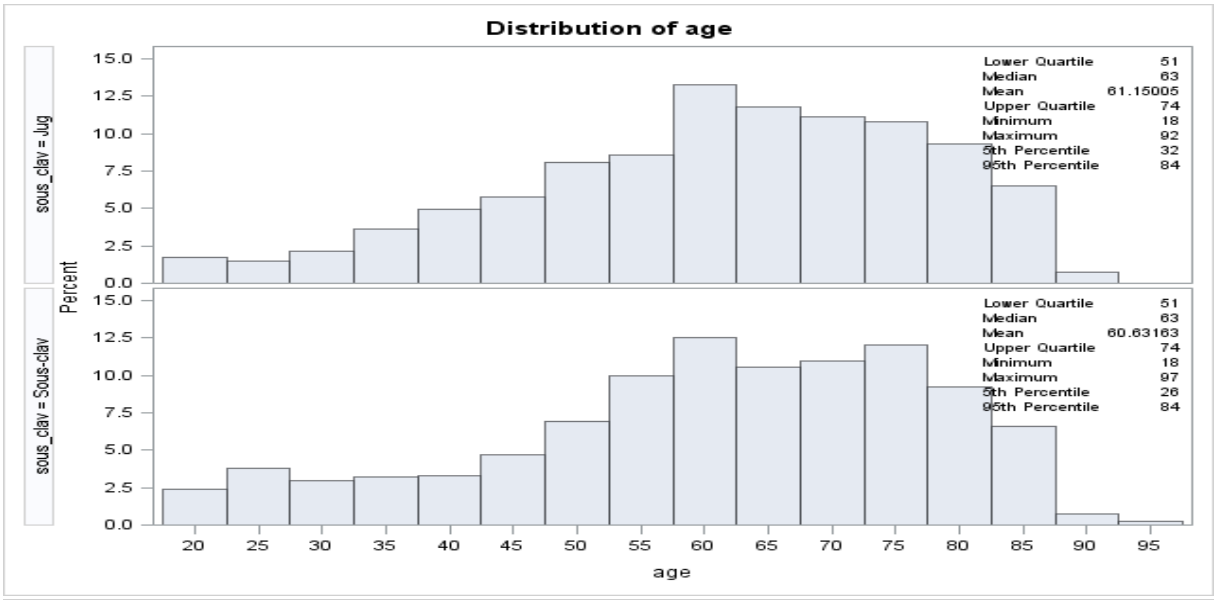
B.1 Tableau de comparaison des covariables

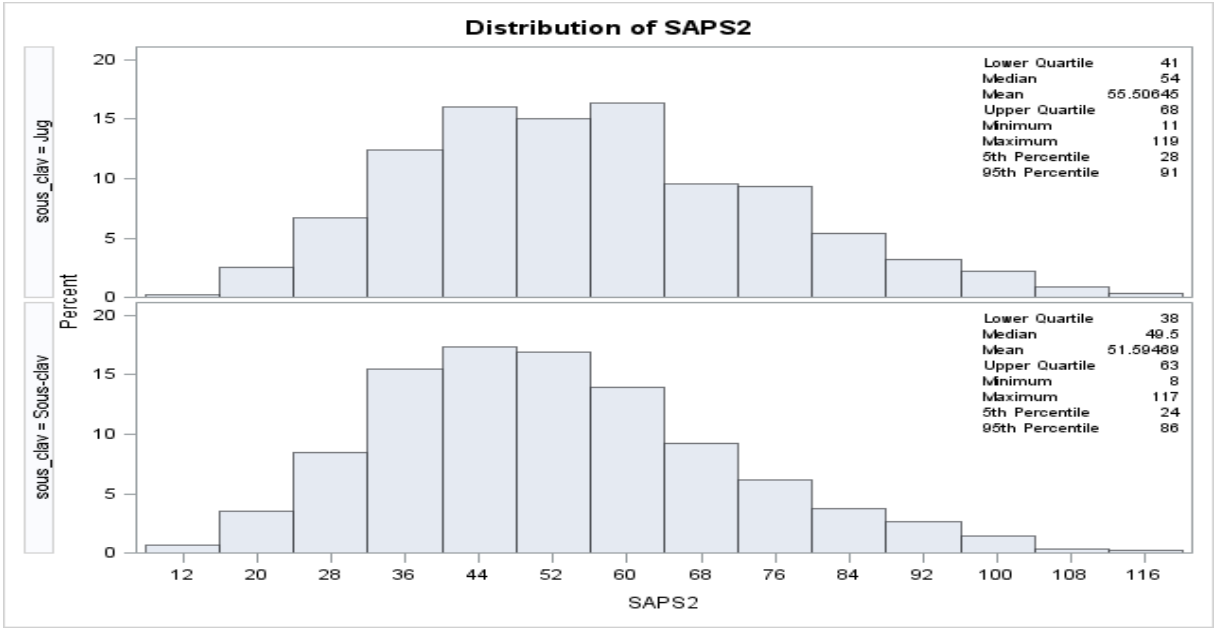
	Toute la cohorte					Après appariement sur le score de propension				
	Tous (n=2065)	Jugulaire (n=933)	Sous-clavière (n=1132)	SD*	p	Tous (n=1090)	Jugulaire (n=545)	Sous-clavière (n=545)	SD	p
Age, moyenne (ET*)	60.9 (16.5)	61.2 (15.8)	60.6 (17.1)	0.04	0.48	61.0 (15.9)	60.6 (16.0)	61.5 (15.8)	0.06	0.36
medianne (EI*)	63 (51-74)	63 (51-74)	63 (51-74)			62.5 (51-73)	62 (51-73)	63 (53-73)		
Hommes, n(%)	1372 (66.4%)	619 (66.4%)	753 (66.5%)	0.002	0.94	706 (64.8%)	354 (65.0%)	352 (64.6%)	0.008	0.90
SAPS 2, moyenne (ET)	53.4 (19.0) MV=5	55.5 (19.4) MV=3	51.6 (18.5) MV=2	0.21	<0.0001	53.9 (18.9)	53.7 (19.3)	54.1 (18.4)	0.02	0.75
medianne (EI)	51 (39-65) MV=5	54 (41-68) MV=3	49.5 (38-63) MV=2			52 (40-66)	51 (39-65)	52 (42-66)		
SOFA, moyenne (ET)	10.3 (4.4) MV=4	10.4 (4.5) MV=2	10.2 (4.3) MV=2	0.05	0.34	10.3 (4.4)	10.3 (4.5)	10.2 (4.3)	0.02	0.78
medianne (EI)	10 (7-13) MV=4	10 (7-13) MV=2	10 (7-13) MV=2			10 (7-13)	10 (7-13)	10 (7-13)		
VM pose, n(%)	1660 (80.4%)	761 (81.6%)	899 (79.4%)	0.06	0.23	865 (79.4%)	432 (79.3%)	433 (79.5%)	0.01	0.95
Inotrope pose, n(%)	1027 (49.7%)	500 (53.6%)	527 (46.6%)	0.14	0.002	548 (50.3%)	273 (50.1%)	275 (50.5%)	0.01	0.91
Chlorex, n(%)	636 (30.8%)	309 (33.1%)	327 (28.9%)	0.09	0.04	371 (34.0%)	182 (33.4%)	189 (34.7%)	0.03	0.66
Atb insertion, n(%)	1628 (78.8%)	738 (79.1%)	890 (78.6%)	0.01	0.79	874 (80.2%)	436 (80.0%)	438 (80.4%)	0.01	0.88
Lipides, n(%)	949 (46.0%) MV=1	394 (42.3%) MV=1	555 (49.0%)	0.13	0.003	506 (46.4%)	255 (46.8%)	251 (46.1%)	0.01	0.81
Heparines, n(%)	686 (33.2%) MV=1	354 (38.0%) MV=1	332 (29.3%)	0.18	<0.0001	381 (35.0%)	192 (35.2%)	189 (34.7%)	0.01	0.85
Groupe, n(%)					0.39					0.91
Biopatch	551 (26.7%)	243 (26.1%)	308 (27.2%)	0.02		283 (26.0%)	139 (25.5%)	144 (26.4%)	0.02	
Tegaderm CHG	482 (23.3%)	231 (24.8%)	251 (22.2%)	0.06		258 (23.7%)	127 (23.3%)	131 (24.0%)	0.02	
Tegaderm HP	249 (12.1%)	118 (12.7%)	131 (11.6%)	0.03		135 (12.4%)	66 (12.1%)	69 (12.7%)	0.02	
Tegaderm standard	783 (37.9%)	341 (36.6%)	442 (39.1%)	0.05		414 (38.0%)	213 (39.1%)	201 (36.9%)	0.05	
Ventilation invasive,	1673 (81.0%)	743 (79.6%)	930 (82.2%)	0.07	0.15	884 (81.1%)	441 (80.9%)	443 (81.3%)	0.01	0.88
Ventilation invasive avec peep, n(%)	752 (36.4%)	348 (37.3%)	404 (35.7%)	0.03	0.45	395 (36.2%)	203 (37.3%)	192 (35.2%)	0.04	0.49
Ventilation non invasive, n(%)	296 (14.3%)	138 (14.8%)	158 (14.0%)	0.02	0.60	171 (15.7%)	82 (15.1%)	89 (16.3%)	0.03	0.56
Dressing, n(%)					0.02					0.91
1	1074 (52.0%)	457 (49.0%)	617 (54.5%)	0.11		554 (50.8%)	278 (51.0%)	276 (50.6%)	0.01	
2	991 (48.0%)	476 (51.0%)	515 (45.5%)	0.11		536 (49.2%)	267 (49.0%)	269 (49.4%)	0.01	
Centre, n(%)					<0.0001					0.99
C1	66 (3.2%)	41 (4.4%)	25 (2.2%)	0.12		24 (2.2%)	13 (2.4%)	11 (2.0%)	0.03	
C2	123 (6.0%)	36 (3.9%)	87 (7.7%)	0.16		55 (5.1%)	33 (6.1%)	22 (4.0%)	0.10	
C3	185 (9.0%)	98 (10.5%)	87 (7.7%)	0.10		118 (10.8%)	58 (10.6%)	60 (11.0%)	0.01	
C4	329 (15.9%)	228 (24.4%)	101 (8.9%)	0.43		183 (16.8%)	90 (16.5%)	93 (17.1%)	0.02	
C5	63 (3.1%)	34 (3.6%)	29 (2.6%)	0.06		41 (3.8%)	22 (4.0%)	19 (3.5%)	0.03	
C6	62 (3.0%)	21 (2.3%)	41 (3.6%)	0.08		38 (3.5%)	21 (3.9%)	17 (3.1%)	0.04	
C7	128 (6.2%)	33 (3.5%)	95 (8.4%)	0.21		55 (5.1%)	28 (5.1%)	27 (5.0%)	0.005	
C8	73 (3.5%)	23 (2.5%)	50 (4.4%)	0.10		49 (4.5%)	23 (4.2%)	26 (4.8%)	0.03	
C9	317 (15.4%)	137 (14.7%)	180 (15.9%)	0.03		182 (16.7%)	90 (16.5%)	92 (16.9%)	0.01	
C10	132 (6.4%)	42 (4.5%)	90 (8.0%)	0.14		67 (6.2%)	34 (6.2%)	33 (6.1%)	0.004	
C11	40 (1.9%)	15 (1.6%)	25 (2.2%)	0.04		32 (2.9%)	15 (2.8%)	17 (3.1%)	0.02	
C12	82 (4.0%)	64 (6.9%)	18 (1.6%)	0.27		30 (2.8%)	12 (2.2%)	18 (3.3%)	0.07	
C13	68 (3.3%)	46 (4.9%)	22 (1.9%)	0.17		40 (3.7%)	20 (3.7%)	20 (3.7%)	0	
C14	246 (11.9%)	71 (7.6%)	175 (15.5%)	0.25		103 (9.5%)	48 (8.8%)	55 (10.1%)	0.04	
C15	10 (0.5%)	5 (0.5%)	5 (0.4%)	0.01		3 (0.3%)	1 (0.2%)	2 (0.4%)	0.04	
C16	141 (6.8%)	39 (4.2%)	102 (9.0%)	0.19		70 (6.4%)	37 (6.8%)	33 (6.1%)	0.03	

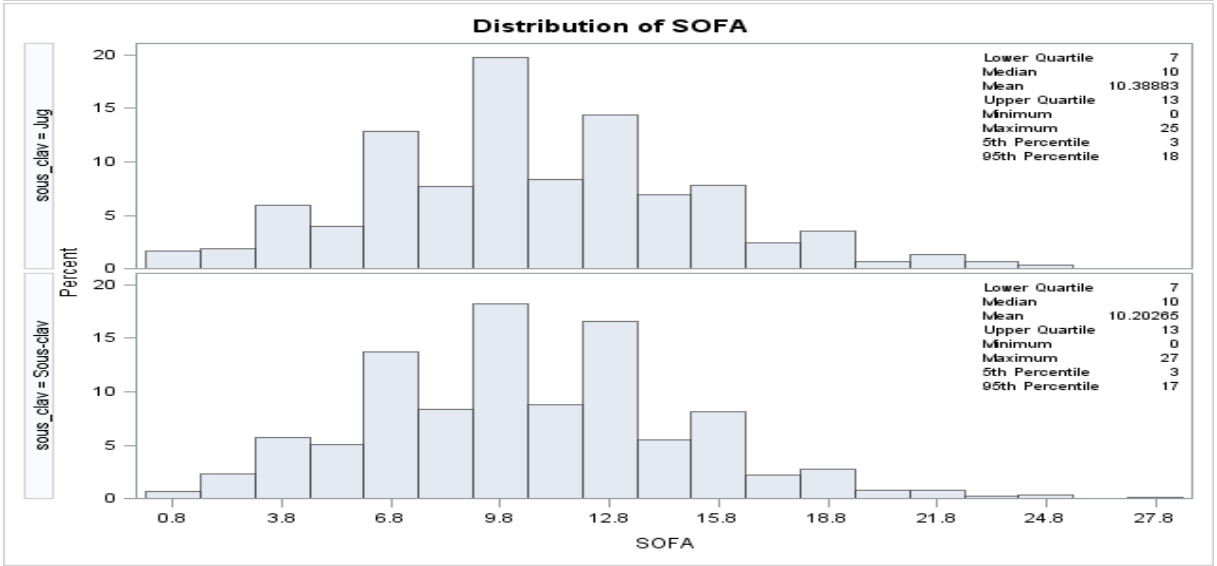
	Toute la cohorte					Après appariement sur le score de propension				
	Tous (n=2065)	Jugulaire (n=933)	Sous-clavière (n=1132)	SD	p	Tous (n=1090)	Jugulaire (n=545)	Sous-clavière (n=545)	SD	p
Rythme pansement,					<0.0001					0.95
Trois	733 (35.5%)	281 (30.1%)	452 (39.9%)	0.21		375 (34.4%)	188 (34.5%)	187 (34.3%)	0.004	
Sept	1332 (64.5%)	652 (69.9%)	680 (60.1%)	0.21		715 (65.6%)	357 (65.5%)	358 (65.7%)	0.004	
Maccabe, n(%)					0.53					0.69
Non prévu	1260 (61.0%)	565 (60.6%)	695 (61.4%)	0.02		641 (58.8%)	324 (59.5%)	317 (58.2%)	0.03	
Prévu 1 a 5 ans	622 (30.1%)	278 (29.8%)	344 (30.4%)	0.01		356 (32.7%)	172 (31.6%)	184 (33.8%)	0.05	
Prévu dans année	183 (8.9%)	90 (9.7%)	93 (8.2%)	0.05		93 (8.5%)	49 (9.00%)	44 (8.1%)	0.03	
Cat admission, n(%)					0.005					0.88
Medical	1410 (68.3%)	669 (71.7%)	741 (65.5%)	0.13		759 (69.6%)	378 (69.4%)	381 (69.9%)	0.01	
Chirurgical programme	155 (7.5%)	56 (6.0%)	99 (8.8%)	0.11		85 (7.8%)	41 (7.5%)	44 (8.1%)	0.02	
Chirurgical urgent	500 (24.2%)	208 (22.3%)	292 (25.8%)	0.08		246 (22.6%)	126 (23.1%)	120 (22.0%)	0.03	
Motif admission, n(%)					<0.0001					0.54
Choc septique	460 (22.3%)	229 (24.5%)	231 (20.4%)	0.10		229 (21.0%)	114 (20.9%)	115 (21.1%)	0.005	
Choc cardiogénique	132 (6.4%)	70 (7.5%)	62 (5.5%)	0.08		73 (6.7%)	35 (6.4%)	38 (7.0%)	0.02	
Détresse respiratoire aigüe	486 (23.5%)	233 (25.0%)	253 (22.6%)	0.06		258 (23.7%)	129 (23.7%)	129 (23.7%)	0	
Coma	215 (10.4%)	76 (8.2%)	139 (12.3%)	0.14		126 (11.6%)	57 (10.5%)	69 (12.7%)	0.07	
Traumatisme	179 (8.7%)	37 (4.0%)	142 (12.5%)	0.31		61 (5.6%)	37 (6.8%)	24 (4.4%)	0.10	
Autre	593 (28.7%)	288 (30.9%)	305 (26.9%)	0.09		343 (31.5%)	173 (31.7%)	170 (31.2%)	0.01	
Expérience opérateur, n(%)	MV=1		MV=1		<0.0001					0.55
Junior	1042 (50.5%)	530 (56.8%)	512 (45.3%)	0.23		580 (53.2%)	296 (54.3%)	284 (52.1%)	0.04	
Senior	1001 (48.5%)	394 (42.2%)	607 (53.7%)	0.23		499 (45.8%)	245 (45.0%)	254 (46.6%)	0.03	
Senior apres junior	21 (1.0%)	9 (1.0%)	12 (1.1%)	0.01		11 (1.0%)	4 (0.7%)	7 (1.3%)	0.06	
Tunnel, n(%)	3 (0.2%)	2 (0.2%)	1 (0.1%)	0.03	0.46	2 (0.2%)	1 (0.2%)	1 (0.2%)	0	1
Coté, n(%)					<0.0001					0.45
D	1270 (61.5%)	702 (75.2%)	568 (50.2%)	0.54		722 (66.4%)	367 (67.3%)	355 (65.1%)	0.05	
G	795 (38.5%)	231 (24.8%)	564 (49.8%)	0.54		368 (33.8%)	178 (32.7%)	190 (34.9%)	0.05	
Nb lum, n(%)					0.41					0.86
1	29 (1.4%)	10 (1.1%)	19 (1.7%)	0.05		11 (1.0%)	5 (0.9%)	6 (1.1%)	0.02	
2	192 (9.3%)	92 (9.9%)	100 (8.8%)	0.04		112 (10.3%)	55 (10.1%)	57 (10.5%)	0.01	
3	1731 (83.8%)	775 (83.1%)	956 (84.5%)	0.04		901 (82.7%)	455 (83.5%)	446 (81.8%)	0.04	
4	113 (5.5%)	56 (6.0%)	57 (5.0%)	0.04		66 (6.1%)	30 (5.5%)	36 (6.6%)	0.05	
Année entrée réa, n(%)	MV=4	MV=3	MV=1		<0.0001					0.999
2006	4 (0.2%)	1 (0.1%)	3 (0.3%)	0.04		2 (0.2%)	1 (0.2%)	1 (0.2%)	0	
2007	760 (36.9%)	329 (35.4%)	431 (38.1%)	0.06		393 (36.1%)	196 (36.0%)	197 (36.2%)	0.004	
2008	306 (14.9%)	124 (13.3%)	182 (16.1%)	0.08		159 (14.6%)	81 (14.9%)	78 (14.3%)	0.02	
2010	466 (22.6%)	192 (20.7%)	274 (24.2%)	0.08		235 (21.6%)	116 (21.3%)	119 (21.8%)	0.01	
2011	525 (25.5%)	284 (30.5%)	241 (21.3%)	0.21		301 (27.6%)	151 (27.7%)	150 (27.5%)	0.004	

*ET : Ecart-type, EI : Ecart interquartile, SD : Différence standardisée (Standardized difference)

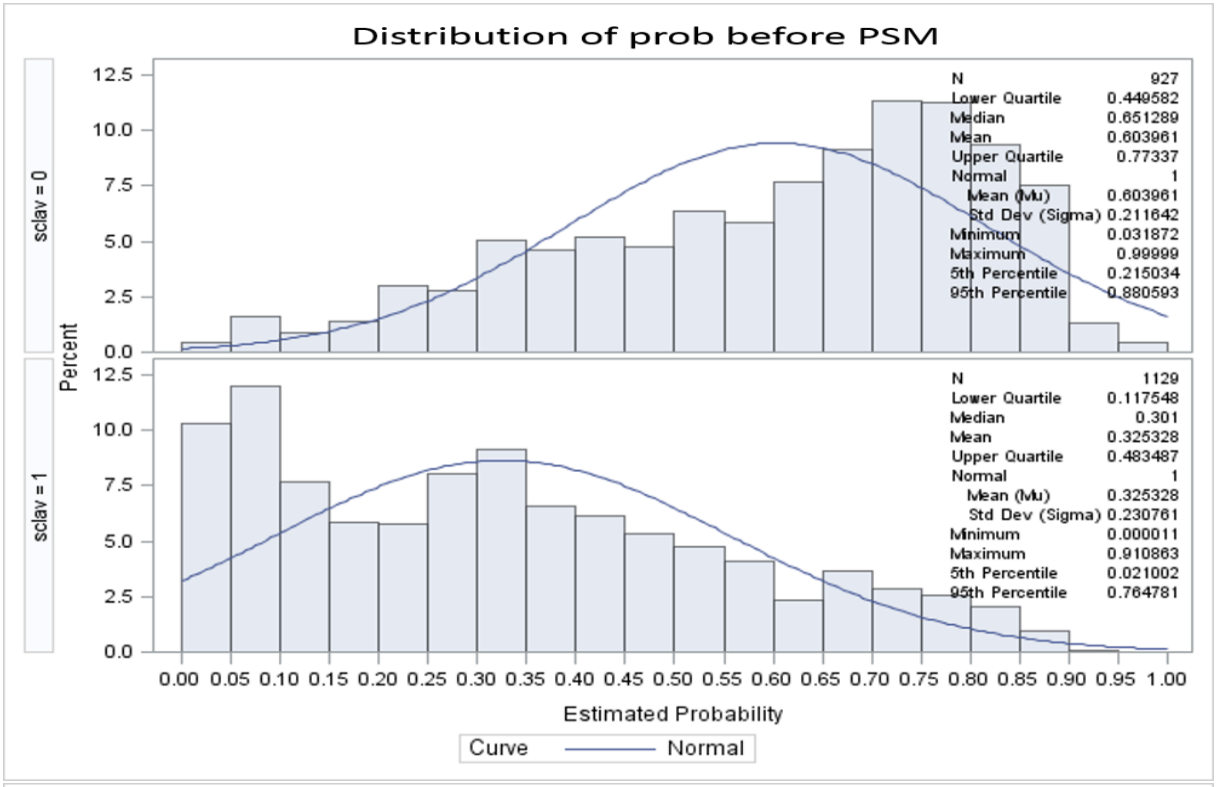
B.2 Distribution des variables Age, SAPS2 et SOFA

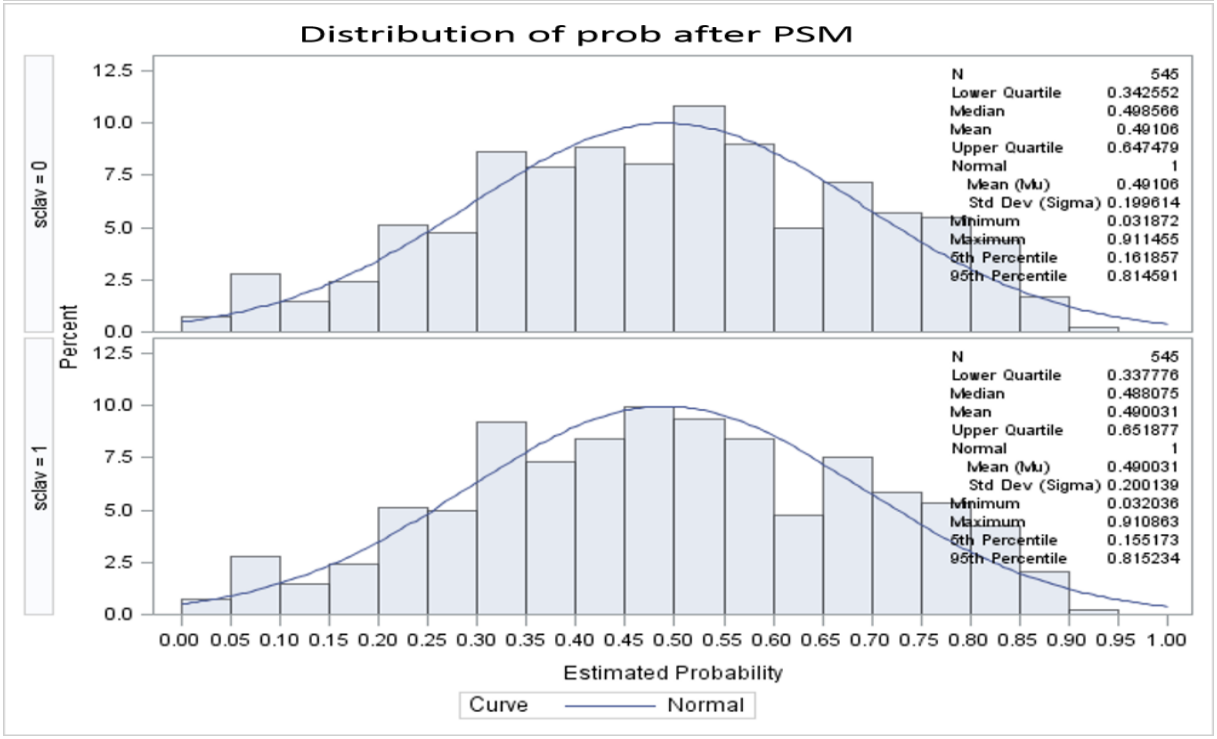

Distribution of age
sous_clav = Jug
Lower Quartile 51
Median 63
Mean 61.15005
Upper Quartile 74
Minimum 18
Maximum 92
5th Percentile 32
95th Percentile 84
sous_clav = Sous-clav
Lower Quartile 51
Median 63
Mean 60.63163
Upper Quartile 74
Minimum 18
Maximum 97
5th Percentile 26
95th Percentile 84
Percent
0.0 2.5 5.0 7.5 10.0 12.5 15.0
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
age


Distribution of SAPS2
sous_clav = Jug
Lower Quartile 41
Median 54
Mean 55.50645
Upper Quartile 68
Minimum 11
Maximum 119
5th Percentile 28
95th Percentile 91
sous_clav = Sous-clav
Lower Quartile 38
Median 49.5
Mean 51.59469
Upper Quartile 63
Minimum 8
Maximum 117
5th Percentile 24
95th Percentile 86
Percent
0 5 10 15 20
12 20 28 36 44 52 60 68 76 84 92 100 108 116
SAPS2


Distribution of SOFA
sous_clav = Jug
Lower Quartile 7
Median 10
Mean 10.38883
Upper Quartile 13
Minimum 0
Maximum 25
5th Percentile 3
95th Percentile 18
sous_clav = Sous-clav
Lower Quartile 7
Median 10
Mean 10.20265
Upper Quartile 13
Minimum 0
Maximum 27
5th Percentile 3
95th Percentile 17
Percent
0 5 10 15 20
0.8 3.8 6.8 9.8 12.8 15.8 18.8 21.8 24.8 27.8
SOFA

B.3 Distribution de la variable Prob avant et après appariement

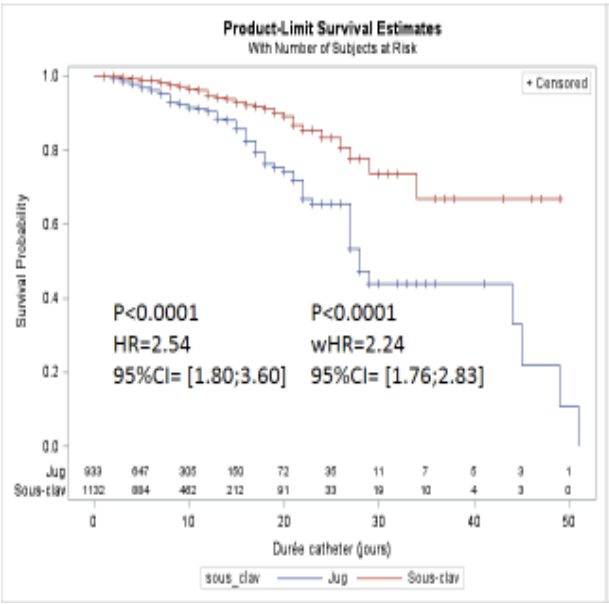




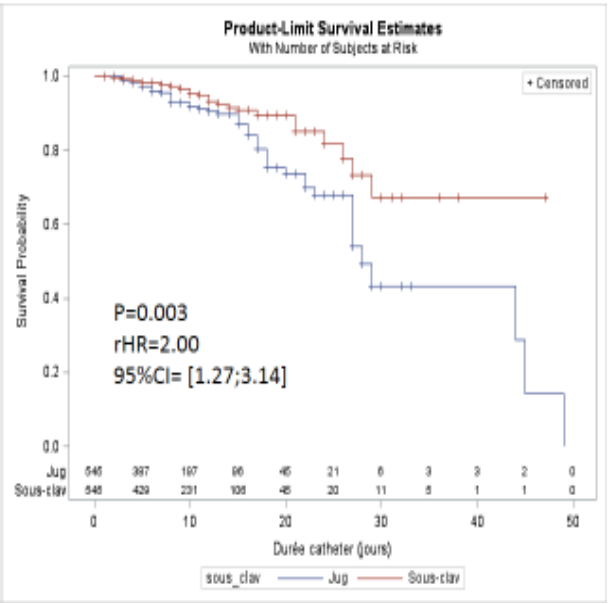
B.4 Courbes de Kaplan-Meier avec hazard ratios

Ktcol

Overall cohort

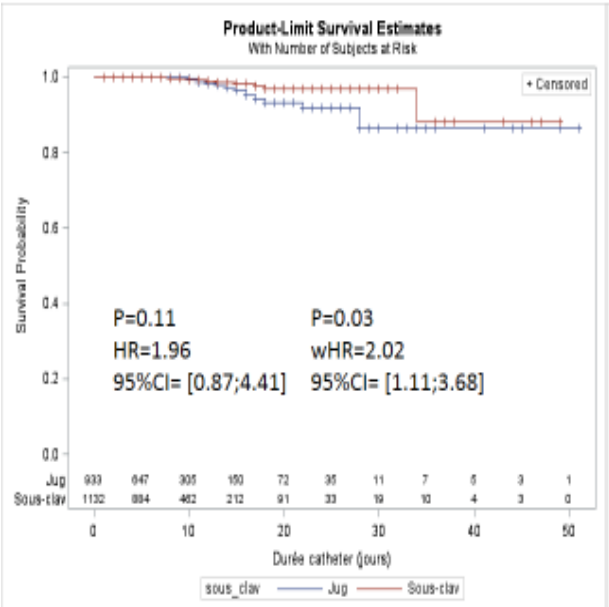


After propensity matching

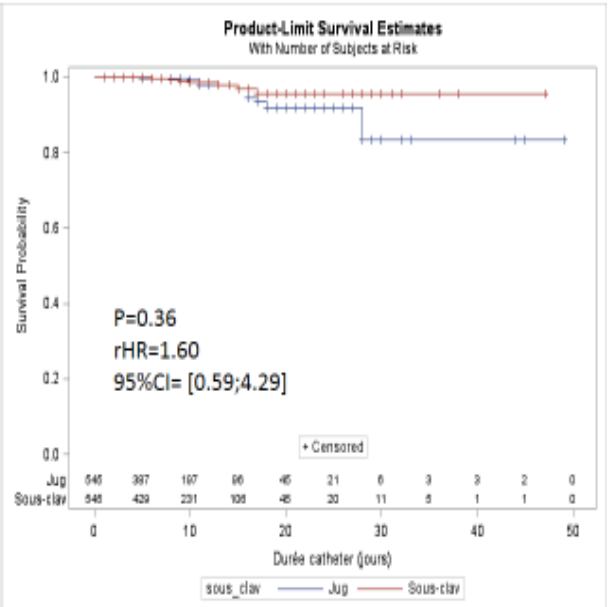


Minf

Overall cohort

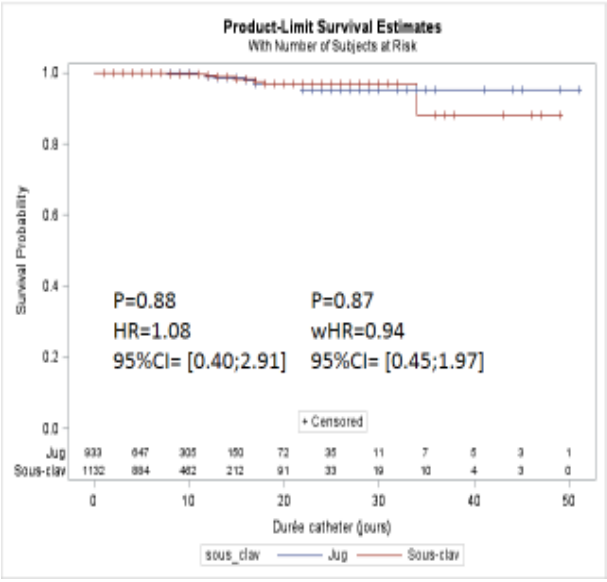


After propensity matching

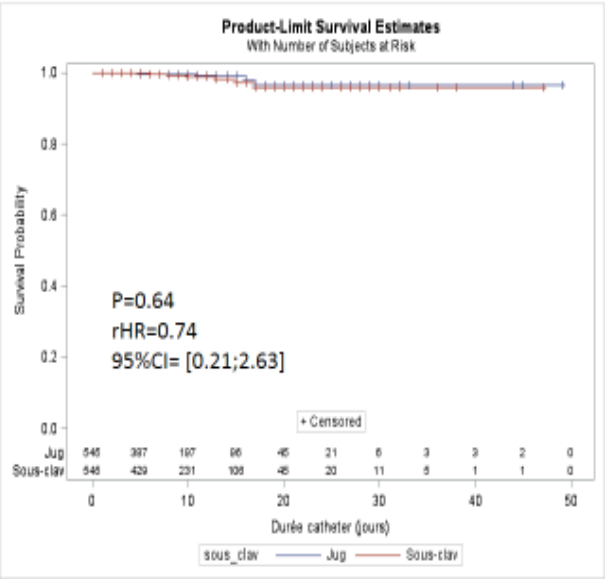


Clab

Overall cohort



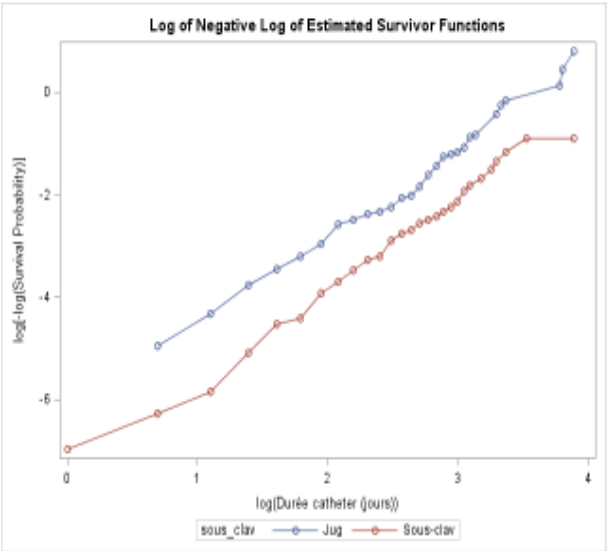
After propensity matching



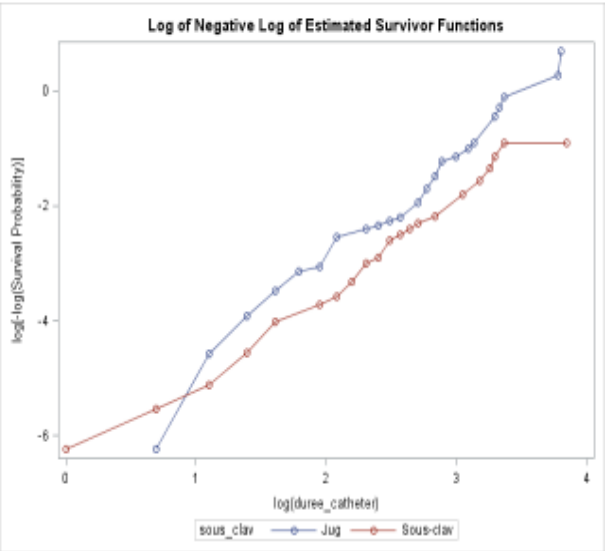
B.5 Courbes de log(-log(survie)) sur log(temps de survie)

Ktcol

Overall cohort

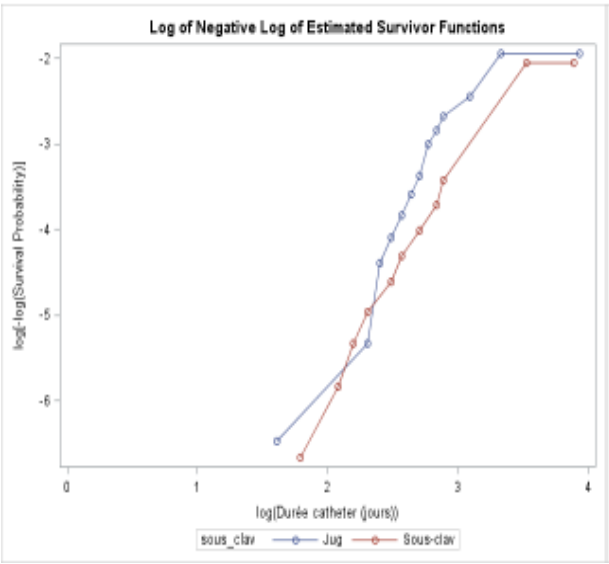


After propensity matching

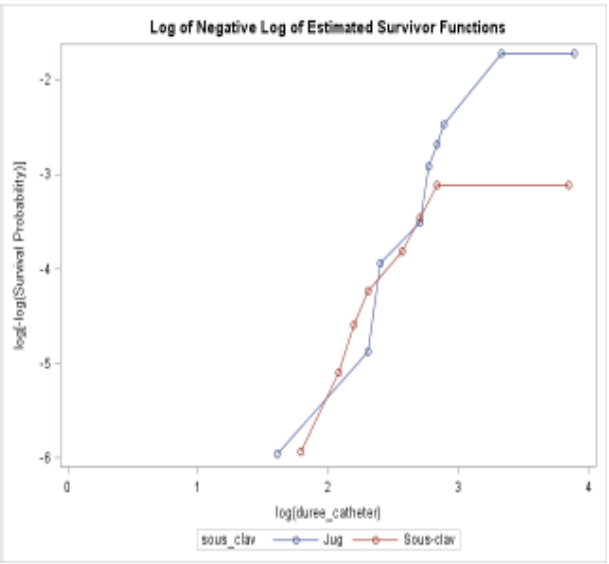


Minf

Overall cohort

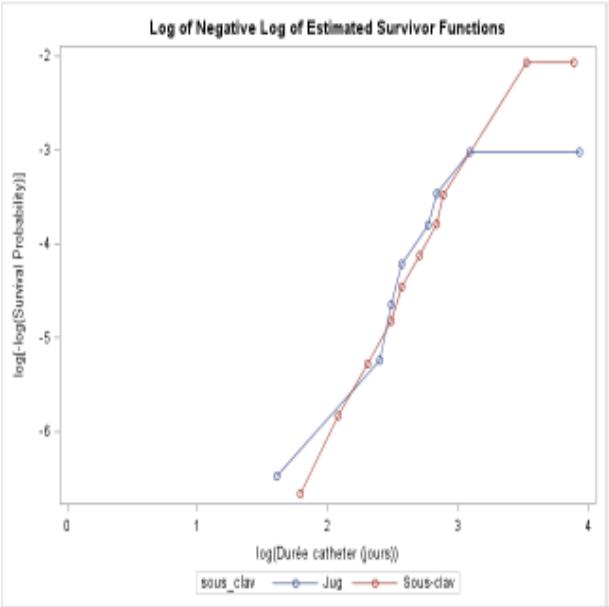


After propensity matching



Clab

Overall cohort



After propensity matching

