

Epidémiologie Interventionnelle



Efficacité du RHC en réanimation

Réalisé par : Mourad do-REGO / Groupe 45

Table des matières

1. Analyse Univariée	3
<i>Data Management</i>	3
<i>Statistique descriptive</i>	9
<i>Gestion des données manquantes</i>	13
2. Analyse bivariée	14
<i>Réflexion</i>	14
<i>Explication de la variable « Pose de sonde de Swan-Ganz »</i>	14
<i>Explication de la variable « Statut vital à J30 »</i>	17
3. Analyse multivariée	19
<i>Score de Propension</i>	19
<i>Ajustement sur le score de Propension</i>	20
<i>Stratification du score de propension</i>	22
<i>Vérification des qualités de l'appariement</i>	23
<i>Pondération sur le score de propension</i>	26
4. Analyse de survie à J30	27
<i>Analyse de survie sur les données brutes</i>	27
<i>Analyse de survie sur les données appariées</i>	28
5. Récapitulation	28
6. Synthèse des résultats	30
7. Annexe	31

1. Analyse Univariée

Data Management

La base comporte 63 avec 5735 observations (lignes). Ci-dessous la description des variables :

Variable	Renommée/supprimée /Imputée	Définition	Remarques
ROWNAMES	-	Numéro de ligne	Gardée comme étant numéro de patient
CAT1	-	Pathologie primaire	-
CAT2	Supprimée	Pathologie secondaire	
CA	-	Cancer à l'admission	Recodée en 0 (absence), 1 (localisé) et 2 (métastatique)
SADMDTE	Renommée "DATE_ENT"	Date d'entrée dans l'étude	Conversion au format date JJ-MM-AAAA. Période d'inclusion: entre le 11/06/1989 et le 23/01/1994
DSCHDTE	Renommée DATE_SORHOP	Date de sortie de l'hospitalisation	Conversion au format date JJ-MM-AAAA
DTHDTE	Renommée DATE_DC	Date de décès	Conversion au format date JJ-MM-AAAA
LSTCTDTE	Renommée DATE_DDN	Date de dernières nouvelles	Conversion au format date JJ-MM-AAAA
DEATH	-	Statut vivant ou décédé jusqu'à J180	Variable binaire codée 0(vivant), 1(décédé)
....HX	-	Présence de comorbidité (cardiaque : CARDIOHX, Insuff cardiaque: CHFHX, Neurologique: DEMENTHX, Psychiatrique: PSYCHHX, respiratoire: CHRPUHX, rénale : RENALHX, hépatique : LIVERHX, hémorragie digestive : GIBLEDHX, néoplasie : MALIGHX, immunologique : IMMUNHX, Infarctus :AMIHX)	Variables binaires codées en 0,1
AGE	-	Age à l'admission	
SEXE	-	Sexe du patient	
EDU	-	Nombre d'années d'étude	Variable quantitative
SURV2MD1	-	Probabilité de survie à 2 mois estimée par le médecin (selon « SUPPORT » model)	
DAS2D3PC	Renommée DASI	Duke Activity Status Index	
T3D30	Renommée JVIE_J30	Nombre de jours de suivi jusqu'à J30	
DTH30	-	Statut vivant ou décédé à J30	Codée en 0(vivant), 1(décédé)
APS1	-	Score APACHE III	
TRANSHX	-	Transfert>24hrs depuis un autre hôpital	Variable binaire, codée en 0,1
SCOMA1	-	Score de Coma, de 0 à 100	C'est un dérivé du score de GLASGOW habituel. Probablement non-linéaire, et très asymétrique : médiane=0. moyenne=21. variable non modifiée.
MEANBP1	Renommée PAM1	Pression artérielle moyenne	Valeur nulle (PAM=0) plausible mais suspecte si FR/FC normales
HRT1	-	Fréquence cardiaque	Valeurs extrêmes plausibles (0 et 250) en réa
WBLC1	-	Taux de leucocytes (g/L)	
RESP1	-	Fréquence respiratoire	Valeur nulle (RESP1=0) plausible mais suspecte si HRT1 normale
CREA1	-	Créatininémie	Changement d'unité de mesure: multipliée par 88.42, pour transformer les g/dL en µmol/L
TEMP1	-	Température (°C)	
SOD1	-	Natrémie en mmol/L	
PAF1	-	Rapport PAO2/FiO2	
ALB1	-	Albuminémie	Changement d'unité de mesure : g/L → g/dL pour retrouver des chiffres « habituels »
HEMA1	-	Hématocrite en %	Valeurs extrêmes plausibles (HEMA1<5%)
SOD1	-	Natrémie en mmol/L	-
BIL11	-	Bilirubine	Changement d'unité de mesure: multipliée par 17.1
POT1	-	Kaliémie en mmol/L	
PACO21	-	PaCO2 en mmHg	
PH1	-	PH	
SWANG1	-	Pose ou non d'une Swan Ganz	variable binaire, tantôt codé en «0, 1», tantôt codé en « No RHC, RHC ».
WTKILO1	Imputée	Poids en Kg	515 données manquantes. Imputation multiple.
DNR1	-	Consignes de non réanimation	Variable binaire, codée en 0,1
NINSCLAS	Renommée ASSURANCE	Type d'assurance	Variable catégorielle à 6 classes
RESP/CARD/NEURO/GASTRO...	-	Diagnostic à l'admission	Variable binaire codée en 0, 1
ALD3P	Supprimée	Echelle ADL	4296 données manquantes. Suppression
URIN1	Supprimée	Diurèse	3028 données manquantes : suppression

Epidémiologie

Race	-	Origine ethnique	Variable catégorielle à 3 classes
Income		Revenues	Variable ordonnée à 4 classes
PTID	Supprimée	Identifiant Patient	Pour assurer l'anonymat (doublon avec ROWNAMES)

Tableau 1: Variables présentes dans le jeu de données et les actions principales du Data Management

ROWNAMES : variable quantitative donnant le numéro de la ligne. Pas de donnée aberrante ni manquante. Nous gardons donc cette variable pour s'assurer des doublons.

CAT1 : variable catégorielle à 9 modalités décrivant la pathologie principale : « ARF » (acute respiratory failure) ; « CHF » (congestive heart failure) ; « cirrhosis » ; « colon cancer » ; « coma » ; « COPD » (chronic obstructive pulmonary disease) ; « lung cancer » ; « MOSF (multi organ system failure) with malignancy » ; « MOSF with sepsis ». Pas de donnée manquante ni de donnée aberrante.

CAT2 : variable catégorielle à 7 modalités décrivant une éventuelle pathologie secondaire associée à la pathologie principale : « cirrhosis » ; « colon cancer » ; « coma » ; « lung cancer » ; « MOSF with malignancy » ; « MOSF with sepsis ». On observe 4535 sujets pour lesquels la donnée est manquante, soit 79% des sujets. On choisit donc de supprimer cette variable de l'analyse.

CA : variable catégorielle à 3 modalités donnant le statut vis-à-vis du cancer codé en 0 « non » ; 1 « oui » ; 2 « métastatique ». Il n'y a pas de donnée manquante ni de donnée aberrante.

SADMDTE : variable catégorielle à 1401 modalités, indiquant la date d'inclusion dans l'étude. On transforme cette variable en une variable de type « DATE_ENT » pour permettre des manipulations ultérieures. Il n'y a pas de donnée manquante ni de donnée aberrante, tous les patients ayant bien été inclus entre juin 1989 et janvier 1994. On supprime ensuite la variable « SADMDTE ».

DSCHDTE : variable catégorielle à 1487 modalités indiquant la date de sortie d'hôpital. On transforme cette variable en une variable de type « date » pour permettre des manipulations ultérieures. La donnée est manquante pour une observation. Il n'y a pas de donnée aberrante, toutes les sorties ayant eu lieu entre juin 1989 et mai 1994. On crée une nouvelle variable « DATE_SORHOP » correspondant au délai entre l'inclusion et la sortie d'hôpital, qui comprend donc elle aussi une donnée manquante, mais aucune donnée aberrante. On supprime ensuite la variable « DSCHDTE ».

DTHDTE : variable catégorielle à 1593 modalités indiquant la date de décès. On transforme cette variable en une variable de type « date » pour permettre des manipulations ultérieures. Il y a 2013 données manquantes, qui correspondent aux patients non décédés. Il n'y a pas de donnée aberrante, tous les décès ayant eu lieu entre juin 1989 et décembre 1994. On crée une nouvelle variable « DATE_DC » correspondant au délai entre l'inclusion et le décès, qui comprend elle aussi 2013 données manquantes. On vérifie l'absence de donnée aberrante en vérifiant que le délai jusqu'au décès est toujours supérieur ou égal au délai jusqu'à la sortie, ce qui est bien le cas ici. On supprime ensuite la variable « DTHDTE ».

LSTCTDTE : variable catégorielle à 932 modalités indiquant la date de dernier contact avec le patient. On transforme cette variable en une variable de type « date » pour permettre des manipulations ultérieures. Il n'y a pas de donnée manquante. Il semble y avoir des données aberrantes, 3 sujets ayant une date de dernier contact postérieure à leur date de décès. Pour ces 3 observations ([2538], [2900] et [4886]), on choisit de remplacer la date de dernier contact par la date de décès. On crée ensuite une nouvelle variable « DATE_DDND » correspondant au délai entre l'inclusion et la date de dernier contact, qui ne comprend pas de donnée manquante ni de donnée aberrante. On supprime ensuite la variable « LSTCTDTE ».

DEATH : variable catégorielle à 2 modalités indiquant le décès dans les 180 jours : « oui »

ou « non ». Il n'y a pas de donnée manquante. Pour vérifier l'absence de données aberrantes, on croise cette variable avec le délai entre l'inclusion et le décès (« DATE_DC », créée précédemment). On constate alors que tous les décès survenus pendant la période de suivi sont codés « oui », même si le délai jusqu'au décès est supérieur à 180 jours. Les patients codés « non » sont bien ceux pour lesquels le décès n'a pas été observé au cours du suivi. On choisit de conserver cette variable telle quelle, en la considérant non pas comme le « décès dans les 180 jours » comme c'est indiqué dans le dictionnaire, mais comme le « décès au cours du suivi ».

CARDIOHX : variable quantitative indiquant une comorbidité cardiovasculaire (de type infarctus du myocarde, maladie vasculaire périphérique, symptômes cardiologiques sévères ou très sévères) absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

CHFHX : variable quantitative indiquant une comorbidité de type insuffisance cardiaque congestive absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

DEMENTHX : variable quantitative indiquant une comorbidité de type neurologique (démence, AVC, maladie de Parkinson) absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

PSYCHHX : variable quantitative indiquant une comorbidité de type psychiatrique (antécédent psychiatrique, psychose ou dépression sévère) absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

CHRPULHX : variable quantitative indiquant une comorbidité de type pulmonaire absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

RENALHX : variable quantitative indiquant une comorbidité de type néphrologique (insuffisance rénale chronique, dialyse) absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

LIVERHX : variable quantitative indiquant une comorbidité de type hépatique (cirrhose, insuffisance hépatique) absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

MALIGHX : variable quantitative indiquant une comorbidité de type tumeur (tumeur solide, cancer métastatique, leucémie aiguë ou chronique, lymphome) absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

IMMUNHX : variable quantitative indiquant une comorbidité de type immunodépression (transplanté d'organe, VIH, diabète) absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

TRANSHX : variable quantitative indiquant un transfert depuis un autre hôpital dans un délai de plus de 24h absent ou présent, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

AMIHX : variable quantitative indiquant une comorbidité de type infarctus du myocarde confirmé, absente ou présente, codée en 0 ou 1. Transformation en variable catégorielle à 2 modalités. Pas de donnée manquante, pas de donnée aberrante.

AGE : variable quantitative donnant l'âge du patient en années. Pas de données aberrantes ni manquante.

SEXE : variable catégorielle à deux modalités, « homme » ou « femme ». Pas de donnée manquante, pas de donnée aberrante.

EDU : variable quantitative donnant le nombre d'années d'éducation. Il n'y a pas de donnée manquante. Il y a des durées d'éducation semblant très longues, jusqu'à 30 ans, mais cela ne semble pas impossibles si l'on considère la totalité de la scolarisation et des études. On conserve donc ces données.

SURV2MD1 : variable quantitative donnant une estimation de la probabilité de survie à 2 mois estimée par un modèle statistique. Les données varient entre 0 et 0.962. Pas de donnée manquante, pas de donnée aberrante.

DAS2D3PC : variable quantitative donnant le DASI (Duke Activity Status Index). Il s'agit d'un score qui évalue la capacité fonctionnelle des patients atteints de maladie cardiovasculaire, à l'aide de 12 items évaluant les capacités des patients dans la vie quotidienne. Il varie de 0 à 58.2, plus il est élevé meilleure est la capacité fonctionnelle du patient. Il s'agit aussi d'un score pronostique et d'un bon prédicteur de mortalité. Ici, il n'y a pas de donnée manquante et pas de donnée aberrante, toutes les valeurs étant comprises entre 11 et 33.

T3D30 : variable quantitative variant entre 2 et 30 et renommée en « JVIE_J30 ». Cette variable n'est pas fournie dans le dictionnaire, mais en la croisant avec les variables DTH30 et DATE_DC, on constate qu'il s'agit du nombre de jours de survie sur les 30 premiers jours. Il n'y a pas de donnée manquante et pas de donnée aberrante.

DTH30 : variable catégorielle à 2 modalités, indiquant le décès dans les 30 premiers jours, codée en « oui » ou « non ». Il n'y a pas de donnée manquante. On vérifie que cette variable est cohérente avec la variable DATE_DC donnant la durée de suivi entre l'inclusion et le décès, ce qui est bien le cas ici.

APS1 : variable quantitative donnant le score APACHE du patient. Il s'agit d'un score de gravité à l'admission utilisé en réanimation. Il inclut entre autres les items suivants : âge, température, pression artérielle moyenne, pH, fréquence respiratoire, fréquence cardiaque, natrémie, kaliémie, créatininémie, hématokrite, globules blancs, score de Glasgow, PaO₂. Il existe plusieurs versions du score : le score Apache II qui varie entre 0 et 71, et le score Apache III entre 0 et 299, la sévérité et le risque de mortalité étant d'autant plus grands que le score est élevé. Dans cette étude, le score varie entre 3 et 147, on peut donc supposer qu'il s'agit du score Apache III. Pour conforter notre hypothèse, on vérifie rapidement graphiquement, sans test statistique, que le score Apache semble en moyenne plus élevé dans le groupe des patients décédés à J30. Il n'y a pas de valeur manquante pour cette variable et pas de donnée aberrante.

SCOMA1 : variable quantitative donnant le score de Glasgow, un indicateur de l'état de conscience qui est fortement corrélé à la sévérité du patient. En pratique clinique, le score de Glasgow varie de 3 (coma profond) à 15 (personne parfaitement consciente). Ici, les valeurs varient entre 0 et 100, avec 11 valeurs possibles (0, 9, 26, 37, 41, 44, 55, 61, 89, 94, 100) et 3105 patients, soit 54%, ont un score égal à 0. On constate que le score de Glasgow a été recodé de façon différente dans cette étude. Pour connaître comment, on observe graphiquement la distribution du score de Glasgow en fonction de la mortalité à J30 [Figure 2], et on remarque que le score de Glasgow semble avoir une moyenne plus élevée dans le groupe des patients décédés à J30. De même, le coefficient de corrélation entre score de Glasgow et durée de survie est de -0.17, ce qui semble indiquer (bien que la force de l'association soit faible) qu'une augmentation du score de Glasgow est associée à une diminution de la durée de survie. L'hypothèse la plus valable ici est qu'un score de Glasgow

proche de 0 correspond à un état de conscience normal, et un score de Glasgow élevé correspond à un trouble de la conscience de plus en plus important. Compte tenu de cette inversion d'échelle, nous choisissons de conserver la valeur donnée. Il n'y a pas de donnée manquante ni de donnée aberrante.

MEANBP1 : variable quantitative indiquant la pression artérielle moyenne en mmHg. On constate qu'il existe des valeurs extrêmes, dont 80 fois la valeur 0, qu'on choisit de recoder en donnée manquante. On choisit par ailleurs de conserver intactes les autres données extrêmes du fait du contexte réanimatoire qui expose à des chiffres de pression artérielle très bas ou très élevés en fonction du contexte.

WBLC1 : variable quantitative indiquant le nombre de leucocytes en G/L. On constate qu'il existe des données extrêmes, qu'on choisit de conserver puisqu'il est possible en cas d'hémopathie d'avoir des valeurs extrêmement hautes ou basses, voire nulles.

HRT1 : variable quantitative indiquant la fréquence cardiaque en battements par minute. On constate qu'il existe des valeurs extrêmes, dont 159 fois la valeur 0, qu'on choisit de recoder en donnée manquante. On choisit de conserver intactes les autres données extrêmes toujours du fait du contexte réanimatoire qui rend possible des bradycardies ou des tachycardies extrêmes.

RESP1 : variable quantitative indiquant la fréquence respiratoire par minute. On constate qu'il existe des valeurs extrêmes, dont 136 fois la valeur 0, qu'on choisit de recoder en donnée manquante. On considère également comme donnée aberrante toute valeur supérieure à 70/minute, qu'on recode en données manquantes. On choisit de conserver intactes les autres données, même extrêmes, toujours du fait du contexte réanimatoire qui rend possible des bradypnées ou des polypnées majeures.

TEMP1 : variable quantitative indiquant la température en degrés Celsius. Il existe là aussi des valeurs extrêmes (valeurs variant de 27°C à 43°C) qu'on choisit de conserver car compatibles avec un contexte réanimatoire. Il n'y a pas de donnée manquante.

PAFI1 : variable quantitative indiquant le rapport PaO₂/FiO₂ en mmHg. Il existe des valeurs extrêmement basses qu'on choisit de conserver car compatibles avec un contexte réanimatoire. Il n'y a pas de donnée manquante.

ALB1 : variable quantitative indiquant l'albuminémie, en g/dL. Il existe deux valeurs extrêmement hautes, supérieures à 10 g/dL, qui semblent aberrantes même en contexte réanimatoire.

HEMA1 : variable quantitative indiquant le taux d'hématocrite en pourcentage. Il existe là aussi des valeurs extrêmes hautes et basses (valeurs variant de 2 à 66.19) qu'on choisit de conserver car compatibles avec un contexte réanimatoire. Il n'y a pas de donnée manquante.

BILI1 : variable quantitative indiquant la bilirubinémie, probablement en mg/dL est convertie en en µmol/L. Il existe là aussi des valeurs élevées qu'on choisit de conserver car tout à fait compatibles avec le contexte réanimatoire. Il n'y a pas de donnée manquante.

CREA1 : variable quantitative indiquant la créatininémie, probablement en mg/dL est convertie en en µmol/L. Il existe là aussi des valeurs élevées qu'on choisit de conserver car tout à fait compatibles avec un contexte réanimatoire. Il n'y a pas de donnée manquante.

SOD1 : variable quantitative indiquant la natrémie en mmol/L. Il existe là aussi des valeurs anormalement élevées ou basses qu'on choisit de conserver car tout à fait compatibles avec un contexte réanimatoire. Il n'y a pas de donnée manquante.

POT1 : variable quantitative indiquant la kaliémie en mmol/L. Il existe là aussi des valeurs anormalement élevées ou basses qu'on choisit de conserver car tout à fait compatibles avec un contexte réanimatoire. Il n'y a pas de donnée manquante.

PACO21 : variable quantitative indiquant la PaCO₂ en mmHg. Il existe là aussi des valeurs anormalement élevées ou basses qu'on choisit de conserver car tout à fait compatibles avec un contexte réanimatoire. Il n'y a pas de donnée manquante.

PH1 : variable quantitative indiquant le pH. Il existe là aussi des valeurs anormalement élevées ou basses qu'on choisit de conserver car tout à fait compatibles un contexte réanimatoire. Il n'y a pas de donnée manquante.

SWANG1 : variable catégorielle à 2 modalités indiquant si la pose d'une sonde de Swan-Ganz ou Right Heart Catheterization (RHC) a été réalisée (codage « RHC ») ou non (codage « No RHC »). Il n'y a pas de donnée manquante ni de donnée aberrante.

WTKILO1 : variable quantitative indiquant le poids en kg. On constate qu'il existe des valeurs extrêmes, dont 515 fois la valeur 0, qu'on choisit de recoder en donnée manquante car ces valeurs correspondent probablement à des sujets pour lesquels le poids n'a pas été renseigné. Les autres valeurs, bien que certaines soient basses ou élevées, ne semblent pas aberrantes.

DNR1 : variable catégorielle à 2 modalités indiquant le statut non réanimatoire à J1 : patient réanimatoire (« No ») ou non réanimatoire (« Yes »). Il n'y a pas de donnée manquante ni de donnée aberrante.

NINSCLAS : variable catégorielle à 6 modalités indiquant le type d'assurance médicale : Medicaid, Medicare, Medicaid et Medicare, Assurance privée, Assurance privée et Medicare, Absence d'assurance. Il n'y a pas de donnée manquante ni de donnée aberrante.

RESP : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie respiratoire à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

CARD : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie cardiologique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

NEURO : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie neurologique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

GASTR : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie gastro-entérologique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

RENAL : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie néphrologique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

META : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie métabolique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

HEMA : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie hématologique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

SEPS : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de sepsis à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

TRAUMA : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou l'absence (« No ») de pathologie traumatologique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

ORTHO : variable catégorielle à deux modalités indiquant la présence (« Yes ») ou

l'absence (« No ») de pathologie orthopédique à l'admission. Il n'y a pas de donnée manquante ni de donnée aberrante.

ADLD3P : variable quantitative indiquant le score ADL qui est une échelle d'autonomie. On constate que cette donnée est manquante pour 4296 sujets, soit 75% des observations, on choisit donc de supprimer cette variable.

URIN1 : variable quantitative indiquant la diurèse en mL. Cette donnée est manquante pour 3028 sujets, soit 53% des observations. On choisit donc de supprimer cette variable.

RACE : variable catégorielle à 3 modalités indiquant l'ethnie : « blanc », « noir », ou « autre ». Il n'y a pas de donnée manquante ni de donnée aberrante.

INCOME : variable catégorielle à 4 modalités indiquant le montant des revenus : « 50k\$ ». Il n'y a pas de donnée manquante ni de donnée aberrante.

PTID : variable quantitative indiquant l'identité du patient par un numéro. On vérifie l'absence de doublon. Cette variable présente peu d'intérêt pour l'analyse, on supprime donc cette variable.

Statistique descriptive

Les statistiques de bases des variables sont synthétisées dans les tableaux qui suivent. Les variables quantitatives ont été décrite à l'aide de la moyenne, la médiane et de l'écart type. Les variables catégorielles ont été décrite quant à elle par le nombre et le pourcentage.

	Variable	Sous-catégories	% de patients (n=5735)
Socio-éco	Sexe	Male	55.6% (3192)
		Female	44.4% (2543)
	Assurance Médicale	Medicaid	11.1% (647)
		Medicare	25.4% (1458)
		Medicare et Medicaid	6.5% (374)
		Pas d'assurance	5.6% (322)
		Privée	29.6% (1698)
		Privée et Medicare	21.5% (1236)
		Noir	16% (920)
	Race	Blanc	77.7% (4460)
		Autres	6% (355)
	Revenu	> \$50k	7.8% (451)
		\$11-\$25k	20% (1165)
		\$25-\$50k	15.6% (893)
		sous \$11k	56% (3226)
Données diagnostiques	Cancer	0 (absence)	76% (4379)
		1 (localisé)	17% (972)
		2 (métastatique)	6.7% (384)
Catégorie de maladie principale	CAT1	Insuffisance respiratoire aigue	43.4% (2490)
		Défaillance cardiaque congestive	7.95% (456)
		Cirrhose	3.9% (224)
		Cancer du Colon	0.1% (7)
		Coma	7.6% (436)
		COPD	8% (457)
		Cancer du poumon	0.6% (39)
		MOSF et/ Malignité	6.9% (399)
		MOSF et/ Septicité	21.4% (1227)

Tableau 2: Résumé des statistiques descriptives pour les variables catégorielles

Epidémiologie

Ce tableau montre que la population d'étude présente une légère prédominance masculine avec 55,6 % d'hommes contre 44,4 % de femmes. Sur le plan ethnique, la cohorte est majoritairement composée de patients de race blanche (77,7 %). On observe une précarité financière marquée, puisque plus de la moitié des sujets (56 %) disposent de revenus annuels limités, inférieurs à 11 000 \$. Concernant la couverture santé, bien que les situations soient variées, l'assurance privée (incluant les combinaisons avec Medicare) reste la catégorie la plus représentée. La grande majorité des patients (76 %) ne présentait aucun antécédent de cancer au moment de l'inclusion. Les affections respiratoires prédominent largement, l'insuffisance respiratoire aiguë étant la cause d'admission la plus fréquente (43,4 %), suivie des défaillances multiviscérales (MOSF) avec septicité (21,4 %).

Variable		% de patients (n=5735)	
		Non	Oui
Comorbidités (Antécédents)	Maladie cardiologique	82,3% (4722)	17,7% (1013)
	Défaillance cardiaque congestive	82,2% (4714)	17,8% (1021)
	Démence, Parkinson	90,2% (5171)	9,8% (564)
	Antécédents psychiatriques	93,3% (5349)	6,7% (386)
	Maladie pulmonaire chronique	81,0% (4646)	19,0% (1089)
	Maladie rénale chronique	95,6% (5480)	4,4% (255)
	Cirrhose	93,0% (5334)	7,0% (401)
	Saignement gastrointestinal	96,8% (5550)	3,2% (185)
	Maladie métastatique	3,2% (4419)	22,9% (1316)
	Immunodépression	73,1% (4192)	26,9% (1543)
	Transfert d'un autre hôpital	26,9% (5073)	26,9% (662)
	Infarctus du myocarde	96,5% (5535)	3,5% (200)
Diagnostic à l'admission	Respiratoire	63,2% (3622)	36,8% (2113)
	Cardiovasculaire	66,3% (3822)	33,7% (1913)
	Neurologique	87,9% (5042)	12,1% (693)
	Gastrointestinal	83,6% (4793)	16,4% (942)
	Rénal	94,9% (5440)	5,1% (295)
	Métabolique	95,4% (5470)	4,6% (265)
	Hématologique	93,8% (5381)	6,2% (354)
	Septicité	82,0% (4704)	18,0% (1031)
	Traumatique	99,1% (5683)	0,9% (52)
	Orthopédique	99,9% (5728)	0,1% (7)
Statut DNR	Ne pas réanimer	88,6% (5081)	11,4% (654)

Tableau 3 Résumé des statistiques descriptives pour les variables qualitatives (Oui/Non)

Les antécédents les plus fréquents sont l'immunodépression (26,9 %) et les maladies métastatiques (22,9 %). Les pathologies respiratoires chroniques concernent 19,0 % des sujets, tandis que la sphère cardiovasculaire est également très représentée avec 17,7 % de maladies cardiologiques et 17,8 % d'insuffisances cardiaques congestives.

Au moment de l'entrée en unité de soins intensifs (ICU), les défaillances respiratoires (36,8 %) et cardiovasculaires (33,7 %) sont les motifs de prise en charge prépondérants. La présence de sepsis concerne 18,0 % de la population.

Une décision de limitation de soins (« Ne pas réanimer ») avait déjà été établie dès le premier jour pour 11,4 % des patients (N = 654\$). Ce taux reflète la sévérité du pronostic vital engagé ou le caractère terminal de certaines comorbidités préexistantes.

Donc environ un quart de la population souffre d'un cancer métastatique ou d'immunodépression, et près d'une cinquième présente des antécédents cardiovasculaires majeurs. Ce profil de fragilité extrême explique une probabilité moyenne de survie à deux mois relativement faible 0,59 (tableau suivant), confirmant le pronostic sévère de cette cohorte.

	Variable	Moyenne	Ecart-type	Médiane
Socio-éco	Age	61.4	16.7	64.05
	Années d'études	11.7	3.15	12
Survie	Probabilité de survie	0.59	0.20	0.63
	Temps de survie (30 J maxi)	23.6	10.1	30
Mesures Biologiques	Score DASI	20.5	5.3	19.75
	Score APACHE	54.67	20.0	54
	Score de coma de Glasgow	21.0	30.2	0
	Poids	74.5	20.64	72.2
	Pression Sanguine	78.52	38.04	63
	Taux de globules blancs	15.7	11.9	14.1
	Rythme cardiaque	115	41.24	124
	Rythme respiratoire	28.09	14.08	30
	Température	37.62	1.77	38.09
	PaO2/FIO2	222.3	115.0	202
	Albumine	30.9	7.8	35
	Hématocrites	31.9	8.4	30
	Bilirubine	38.7	17.26	82.1
	Créatinine	2.1	2.0	1.5
	Sodium	136.8	7.7	136
	Potassium	4.1	1.0	3.8
	PaCo2	38.8	13.2	37
	PH	7.39	0.1	7.4

Tableau ❶. Résumé des statistiques descriptives pour les variables quantitatives

L'analyse des variables quantitatives confirme la gravité extrême de cette cohorte. Le score APACHE moyen de 54,67 et le score de Glasgow inversé de 21 témoignent d'un risque de mortalité de base élevé et de troubles profonds de la conscience. L'instabilité physiologique est marquée par une tachycardie (115 bpm) et une polypnée (28 cycles/min), tandis que le rapport PaO₂/FiO₂ abaissé à 222,3 révèle une oxygénation sous-optimale. Les atteintes organiques sont visibles via l'élévation des globules blancs (15,7 G/L) et de la créatinine (2,1 mg/dL). Ce profil critique corrèle avec un pronostic réservé, affichant une probabilité de survie de seulement 0,59. Enfin, la moyenne de survie de 23,6 jours sur un mois souligne l'importance des décès précoces dans cette population.

Les valeurs d'intérêt de l'intervention et des paramètres de survie sont représentées dans les deux tableaux suivants :

Variable	% de patients (n=5735)	
	Non	Oui
Mort au cours du suivi	35,1% (2013)	64,9% (3722)
Mort au cours de 30 premiers jours	66,6% (3817)	33,4% (1918)
Statut RHC	61,9% (3551)	38,1% (2184)

Tableau 5: Intervention et données de survie de la population, variables qualitatives

Variable	Moyenne	Ecart-type	Médiane
Délai entre inclusion et sortie de l'hôpital (jours)	21.49	25.4	14
Délai entre inclusion et décès (jours)	162.2	295.63	28
Délai entre inclusion et dernière nouvelle (jours)	142.5	146.08	163
Jours de survie parmi les 30 premiers jours	23.61	10.05	30

Tableau 6: Intervention et données de survie de la population, variables quantitatives

Le cathétérisme cardiaque droit (RHC) a été pratiqué chez 38,1 % des patients dès les 24 premières heures. La mortalité à 30 jours atteint 33,4 %, avec une survie moyenne de 23,61 jours durant ce premier mois. Sur l'ensemble du suivi, le taux de décès s'élève à 64,9 %, confirmant la sévérité extrême des pathologies. Bien que le délai moyen de sortie soit de 21,49 jours, la moitié des décès survient très précocement, comme l'indique une médiane de survie de seulement 28 jours pour les patients décédés. Cette forte concentration de décès initiaux et une mortalité globale dépassant les 60 % démontrent un pronostic vital très engagé. Une telle

urgence vitale justifie l'évaluation rigoureuse du bénéfice réel de l'intervention RHC face à ses risques potentiels.

Gestion des données manquantes

Les variables « diurèse » (URIN1) et « score d'autonomie » (ADLD3P) ont été retirées de l'analyse en raison d'un taux de données manquantes trop élevé, s'élevant respectivement à 53 % et 75 % des observations. Leur maintien aurait induit un risque de biais de sélection majeur. Quant à la variable « poids », elle présentait un pourcentage de valeurs initialement codées à 0 (n=515), traitées comme des données manquantes. Nous avons procédé à une imputation multiple générant 5 jeux de données complets.

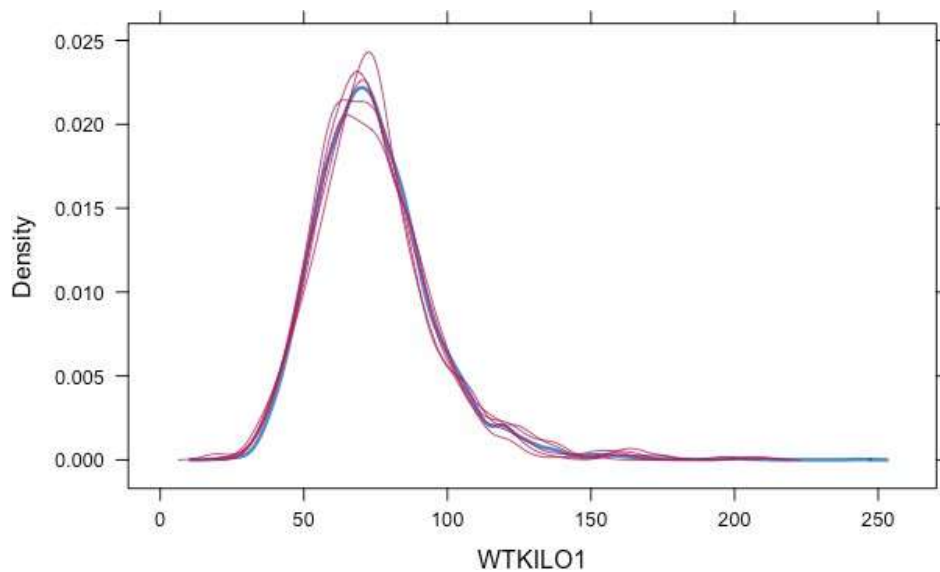
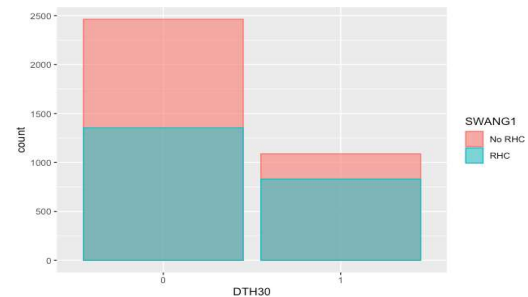
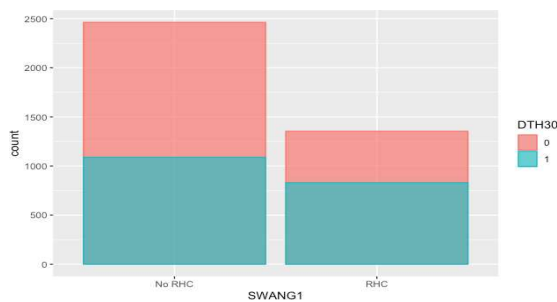


Figure 1: Courbes de densité de la variable POIDS, (lors des 5 imputations multiples)

Comme l'illustre la Figure 1, les courbes de densité des 5 itérations d'imputation (en violet) se superposent étroitement à la distribution des données observées avant imputation (en bleu). Cette stabilité de la distribution, conservant une moyenne de 74,5 kg et une médiane de 72,2 kg, confirme que le processus d'imputation n'a pas déformé la structure originale des données.

2. Analyse bivariée

Réflexion



Ce qui nous intéresse ici est de savoir si le taux de mortalité diffère entre le groupe Swan Ganz et le groupe sans. Une façon « exploratoire » sera de mener un tableau de contingence 2x2 et de réaliser un test de Chi2 à 1ddl.

	Sans Swan Ganz	Avec Swan Ganz	Total
Vivant à J30	2463	1354	3817
Mort à J30	1088	830	1918
Total	3551	2184	5735

Le tableau de contingence montre que le taux de mortalité est de 38,0 % (830/2184) dans le groupe "Avec Swan Ganz", contre 30,6 % (1088/3551) dans le groupe "Sans Swan Ganz".

Un test du χ^2 à 1 degré de liberté révèle une valeur p extrêmement faible ($9,5 \cdot 10^{-9}$), confirmant que la mortalité est statistiquement plus élevée dans le groupe ayant reçu l'intervention.

Ce résultat, bien que significatif, ne permet pas de conclure à une causalité. Il existe potentiellement un biais d'indication majeur : les patients les plus graves sont plus susceptibles de recevoir une sonde de Swan-Ganz et présentent intrinsèquement un risque de décès plus élevé. Pour corriger ce biais, nous allons utiliser l'analyse bivariée comme filtre exploratoire pour construire un score de propension. Elle nous aide également à différencier les patients qui ont eu la sonde de Swan-Ganz des patients décédés. Ceci nous donne une représentation du contexte. Plutôt que d'inclure toutes les variables disponibles, nous allons identifier les facteurs de confusion patents et potentiels. Un facteur de confusion doit être lié à la fois à l'exposition (pose de la sonde) et à l'issue (décès à J30).

Les résultats de l'analyse bivariée selon la pose ou pas d'une sonde de SWAN-GANZ et selon le statut du patient vivant ou décédé à J30 sont représentés dans les 2 tableaux suivants. Conformément aux bonnes pratiques, nous n'avons réalisés aucun test statistique lors de cette phase bivariée exploratoire afin de ne pas "consommer" inutilement le risque alpha global de l'étude par des comparaisons multiples.

Explication de la variable « Pose de sonde de Swan-Ganz »

Pour évaluer la comparabilité des patients ayant reçu une sonde de Swan-Ganz ($n = 2184$) par rapport à ceux n'en ayant pas reçu ($n = 3551$), nous avons utilisé la différence

Epidémiologie

standardisée moyenne (SMD). Cette mesure permet de quantifier le déséquilibre des caractéristiques initiales indépendamment de la taille de l'échantillon.

Conformément aux recommandations méthodologiques, un seuil de **SMD > 0,1** a été retenu pour identifier une imbalance significative entre les deux groupes.

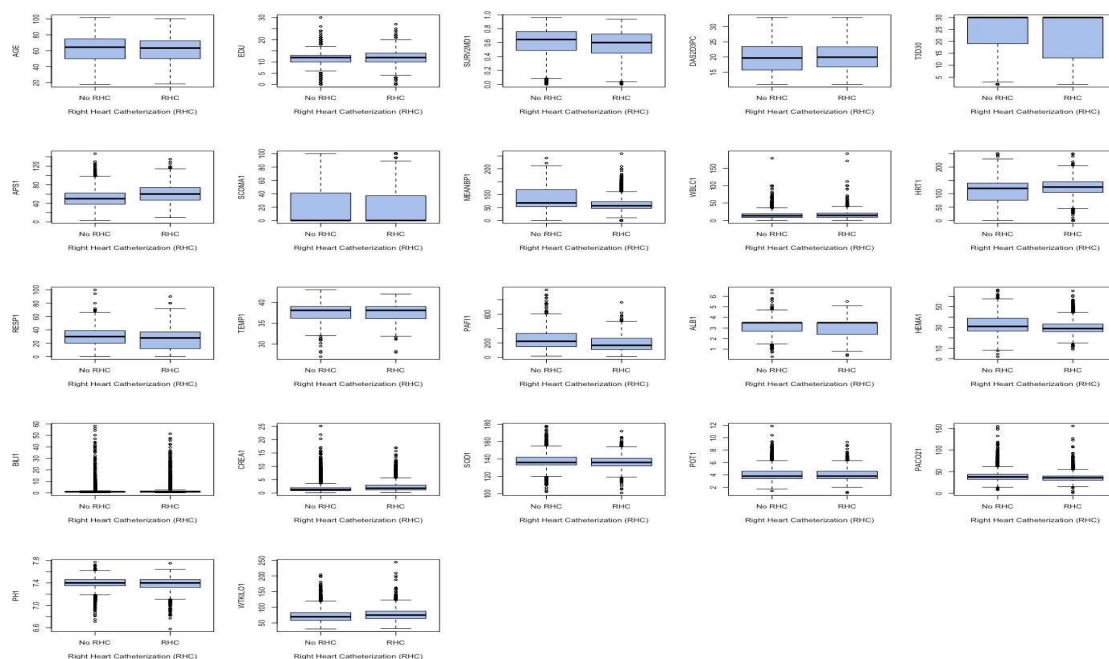
	Stratification selon Swan-Ganz		
	Sans Swan-Ganz (n = 3551)	Avec Swan-Ganz (n = 2184)	SMD
AGE (mean (SD))	61.76 (17.29)	60.75 (15.63)	0.061
SEX = Male (%)	1914 (53.9)	1278 (58.5)	0.093
RACE (%)			0.036
black	585 (16.5)	335 (15.3)	
other	213 (6.0)	142 (6.5)	
white	2753 (77.5)	1707 (78.2)	
EDU (mean (SD))	11.57 (3.13)	11.86 (3.16)	0.091
INCOME (mean (SD))	1.70 (0.96)	1.83 (1.02)	0.135
ASSURANCE (%)			0.194
Medicaid	454 (12.8)	193 (8.8)	
Medicare	947 (26.7)	511 (23.4)	
Medicare & Medicaid	251 (7.1)	123 (5.6)	
No insurance	186 (5.2)	136 (6.2)	
Private	967 (27.2)	731 (33.5)	
Private & Medicare	746 (21.0)	490 (22.4)	
CAT1 (%)			0.583
ARF	1581 (44.5)	909 (41.6)	
CHF	247 (7.0)	209 (9.6)	
Cirrhosis	175 (4.9)	49 (2.2)	
Colon Cancer	6 (0.2)	1 (0.0)	
Coma	341 (9.6)	95 (4.3)	
COPD	399 (11.2)	58 (2.7)	
Lung Cancer	34 (1.0)	5 (0.2)	
MOSF w/Malignancy	241 (6.8)	158 (7.2)	
MOSF w/Sepsis	527 (14.8)	700 (32.1)	
DASI (mean (SD))	20.37 (5.48)	20.70 (5.03)	0.063
DNR1 = Yes (%)	499 (14.1)	155 (7.1)	0.228
CA (mean (SD))	0.33 (0.61)	0.27 (0.55)	0.105
SURV2MD1 (mean (SD))	0.61 (0.19)	0.57 (0.20)	0.198
APS1 (mean (SD))	50.93 (18.81)	60.74 (20.27)	0.501
SCOMA1 (mean (SD))	22.25 (31.37)	18.97 (28.26)	0.110
WTKILO1 (mean (SD))	72.57 (21.16)	77.97 (20.72)	0.258
TEMP1 (mean (SD))	37.63 (1.74)	37.59 (1.83)	0.021
PAM (mean (SD))	84.87 (38.87)	68.20 (34.24)	0.455
RESP1 (mean (SD))	28.98 (13.95)	26.65 (14.17)	0.165
HRT1 (mean (SD))	112.87 (40.94)	118.93 (41.47)	0.147
PAFI1 (mean (SD))	240.63 (116.66)	192.43 (105.54)	0.433
PACO21 (mean (SD))	39.95 (14.24)	36.79 (10.97)	0.249
PH1 (mean (SD))	7.39 (0.11)	7.38 (0.11)	0.120
WBLC1 (mean (SD))	15.26 (11.41)	16.27 (12.55)	0.084
HEMA1 (mean (SD))	32.70 (8.79)	30.51 (7.42)	0.269
SOD1 (mean (SD))	137.04 (7.68)	136.33 (7.60)	0.092
POT1 (mean (SD))	4.08 (1.04)	4.05 (1.01)	0.027
CREA1 (mean (SD))	170.09 (179.07)	218.70 (181.52)	0.270
BILI1 (mean (SD))	34.15 (75.67)	46.27 (91.13)	0.145

Epidémiologie

ALB1 (mean (SD))	31.64 (6.72)	29.78 (9.25)	0.230
------------------	--------------	--------------	-------

RESP = Yes (%)	1481 (41.7)	632 (28.9)	0.270
CARD = Yes (%)	1007 (28.4)	924 (42.3)	0.295
NEURO = Yes (%)	575 (16.2)	118 (5.4)	0.353
GASTR = Yes (%)	522 (14.7)	420 (19.2)	0.121
RENAL = Yes (%)	147 (4.1)	148 (6.8)	0.116
META = Yes (%)	172 (4.8)	93 (4.3)	0.028
HEMA = Yes (%)	239 (6.7)	115 (5.3)	0.062
SEPS = Yes (%)	515 (14.5)	516 (23.6)	0.234
TRAUMA = Yes (%)	18 (0.5)	34 (1.6)	0.104
ORTHO = Yes (%)	3 (0.1)	4 (0.2)	0.027
CARDIOHX (mean (SD))	0.16 (0.37)	0.20 (0.40)	0.116
CHFHx (mean (SD))	0.17 (0.37)	0.19 (0.40)	0.069
DEMENTHX (mean (SD))	0.12 (0.32)	0.07 (0.25)	0.163
PSYCHHX (mean (SD))	0.08 (0.27)	0.05 (0.21)	0.143
CHRPULHX (mean (SD))	0.22 (0.41)	0.14 (0.35)	0.192
RENALHX (mean (SD))	0.04 (0.20)	0.05 (0.21)	0.032
LIVERHX (mean (SD))	0.07 (0.26)	0.06 (0.24)	0.049
GIBLEDHX (mean (SD))	0.04 (0.19)	0.02 (0.16)	0.070
MALIGHX (mean (SD))	0.25 (0.43)	0.20 (0.40)	0.101
IMMUNHX (mean (SD))	0.26 (0.44)	0.29 (0.45)	0.080
TRANSHX (mean (SD))	0.09 (0.29)	0.15 (0.36)	0.170
AMIHX (mean (SD))	0.03 (0.17)	0.04 (0.20)	0.074

Nous remarquons que 32 variables parmi les 50 ont un SMD>0.1, ce qui montre un déséquilibre des caractéristiques des patients entre les 2 groupes. Cela nous permet de dire que les patients qui ont eu la pose de Swan Ganz n'ont pas les mêmes caractéristiques des patients qui ne l'ont pas eue.



Les graphiques en boîtes à moustaches (boxplots) confirment visuellement l'existence d'un déséquilibre significatif entre les patients ayant reçu une sonde de Swan-Ganz (RHC) et ceux

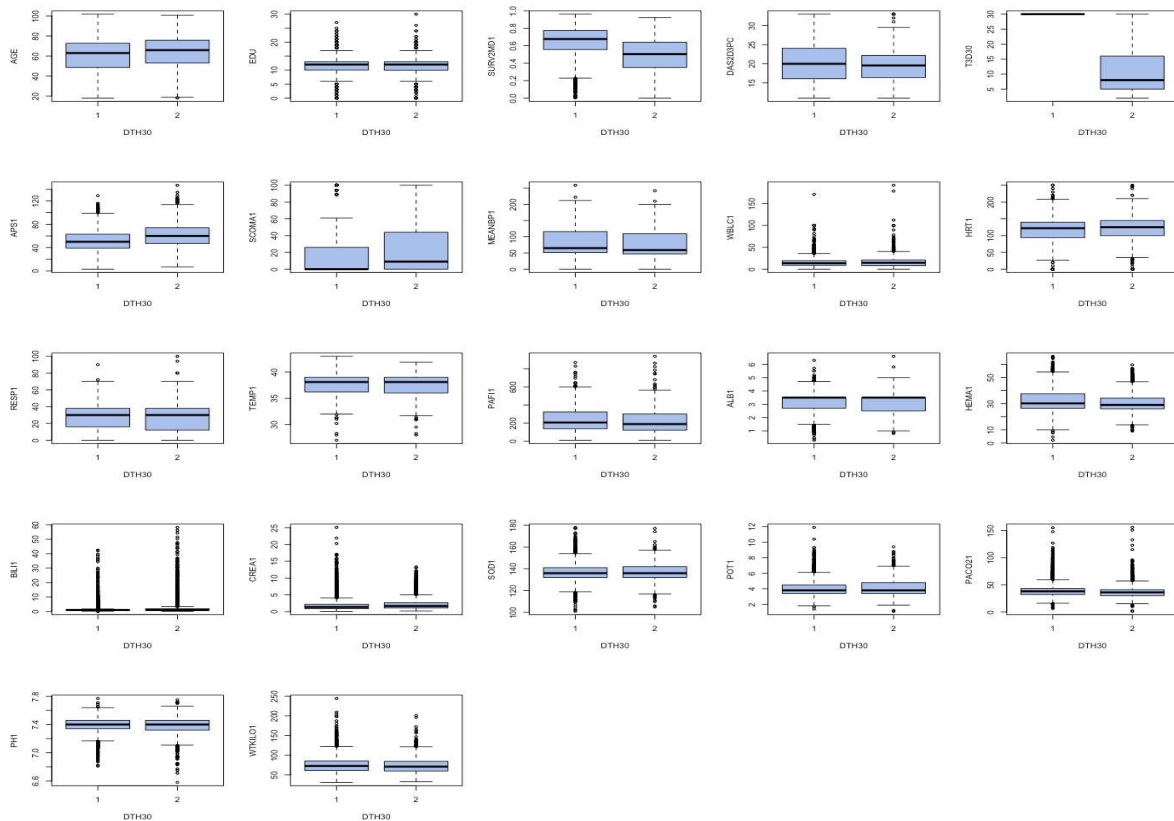
n'en ayant pas reçu. Ce sont des hommes plus jeunes avec des revenus annuels moyen et possédant une assurance médicale privée. Cette hétérogénéité des caractéristiques initiales valide l'observation des différences standardisées moyennes (SMD) supérieures à 0,1 pour la majorité des variables cliniques.

Explication de la variable « Statut vital à J30 »

	Stratification selon statut vital à J30		SMD
	vivant (n = 3817)	décédé (n = 1918)	
AGE (mean (SD))	60.23 (16.77)	63.66 (16.27)	0.208
SEX = Male (%)	2116 (55.4)	1076 (56.1)	0.013
RACE (%)			0.017
black	615 (16.1)	305 (15.9)	
other	231 (6.1)	124 (6.5)	
white	2971 (77.8)	1489 (77.6)	
EDU (mean (SD))	11.71 (3.13)	11.61 (3.17)	0.032
INCOME (mean (SD))	1.79 (1.00)	1.67 (0.96)	0.117
ASSURANCE (%)			0.152
Medicaid	473 (12.4)	174 (9.1)	
Medicare	930 (24.4)	528 (27.5)	
Medicare & Medicaid	251 (6.6)	123 (6.4)	
No insurance	208 (5.4)	114 (5.9)	0.596
Private	1173 (30.7)	525 (27.4)	
Private & Medicare	782 (20.5)	454 (23.7)	
CAT1 (%)			
ARF	1744 (45.7)	746 (38.9)	0.143
CHF	398 (10.4)	58 (3.0)	
Cirrhosis	130 (3.4)	94 (4.9)	
Colon Cancer	6 (0.2)	1 (0.1)	
Coma	150 (3.9)	286 (14.9)	0.415
COPD	377 (9.9)	80 (4.2)	
Lung Cancer	24 (0.6)	15 (0.8)	
MOSF w/Malignancy	183 (4.8)	216 (11.3)	
MOSF w/Sepsis	805 (21.1)	422 (22.0)	0.239
DASI (mean (SD))	20.74 (5.72)	20.01 (4.36)	
DNR1 = Yes (%)	257 (6.7)	397 (20.7)	
CA (mean (SD))	0.26 (0.55)	0.40 (0.65)	
SURV2MD1 (mean (SD))	0.65 (0.17)	0.48 (0.20)	0.874
APS1 (mean (SD))	51.46 (18.52)	61.06 (21.14)	0.483
SCOMA1 (mean (SD))	16.12 (26.05)	30.72 (35.32)	0.470
WTKILO1 (mean (SD))	75.22 (21.51)	73.44 (20.38)	0.085
TEMP1 (mean (SD))	37.66 (1.72)	37.53 (1.87)	0.070
PAM (mean (SD))	81.08 (37.76)	73.42 (38.11)	0.202
RESP1 (mean (SD))	28.31 (13.67)	27.66 (14.86)	0.045
HRT1 (mean (SD))	114.87 (39.67)	115.79 (44.22)	0.022
PAFI1 (mean (SD))	225.94 (112.52)	214.97 (119.34)	0.095
PACO21 (mean (SD))	39.56 (13.43)	37.14 (12.52)	0.186
PH1 (mean (SD))	7.39 (0.10)	7.38 (0.12)	0.067
WBLC1 (mean (SD))	15.38 (10.97)	16.16 (13.47)	0.064
HEMA1 (mean (SD))	32.35 (8.59)	30.91 (7.81)	0.175
SOD1 (mean (SD))	136.69 (7.52)	136.93 (7.91)	0.032
POT1 (mean (SD))	4.06 (1.01)	4.09 (1.07)	0.028
CREA1 (mean (SD))	181.61 (186.31)	202.52 (170.82)	0.117
BILI1 (mean (SD))	29.09 (53.94)	58.03 (117.54)	0.316
ALB1 (mean (SD))	31.34 (8.10)	30.11 (7.21)	0.160
RESP = Yes (%)	1459 (38.2)	654 (34.1)	0.086
CARD = Yes (%)	1345 (35.2)	586 (30.6)	0.100
NEURO = Yes (%)	398 (10.4)	295 (15.4)	0.148
GASTR = Yes (%)	559 (14.6)	383 (20.0)	0.141

Epidémiologie

RENAL = Yes (%)	184 (4.8)	111 (5.8)	0.043
META = Yes (%)	184 (4.8)	81 (4.2)	0.029
HEMA = Yes (%)	172 (4.5)	182 (9.5)	0.196
SEPS = Yes (%)	638 (16.7)	393 (20.5)	0.097
TRAUMA = Yes (%)	40 (1.0)	12 (0.6)	0.046
ORTHO = Yes (%)	6 (0.2)	1 (0.1)	0.032
CARDIOHX (mean (SD))	0.19 (0.39)	0.14 (0.35)	0.134
CHFHX (mean (SD))	0.20 (0.40)	0.14 (0.35)	0.148
DEMENTHX (mean (SD))	0.09 (0.28)	0.12 (0.32)	0.094
PSYCHHX (mean (SD))	0.08 (0.26)	0.05 (0.22)	0.097
CHRPULHX (mean (SD))	0.20 (0.40)	0.16 (0.37)	0.112
RENALHX (mean (SD))	0.04 (0.20)	0.05 (0.21)	0.010
LIVERHX (mean (SD))	0.06 (0.23)	0.10 (0.29)	0.145
GIBLEDHX (mean (SD))	0.03 (0.16)	0.04 (0.20)	0.074
MALIGHX (mean (SD))	0.20 (0.40)	0.30 (0.46)	0.232
IMMUNHX (mean (SD))	0.27 (0.44)	0.27 (0.45)	0.011
TRANSHX (mean (SD))	0.12 (0.32)	0.11 (0.31)	0.021
AMIHX (mean (SD))	0.04 (0.19)	0.03 (0.17)	0.043



D'après les résultats des tests obtenus et des boxplots, nous pouvons dire que les patients qui décèdent à 30 jours sont en général plus âgés, avec des revenus <\$11k et des pathologies supplémentaires.

Ce contraste marqué entre les deux populations souligne que l'intervention est préférentiellement prescrite aux patients les plus critiques, rendant l'analyse bivariée brute insuffisante pour conclure sur l'efficacité réelle de la sonde. L'ajustement par score de propension est donc impératif pour neutraliser ces biais de sélection.

3. Analyse multivariée

D'après les analyses bivariées nous nous sommes rendu compte que les patients ayant une sonde Swan-Ganz ou pas, n'ont pas les mêmes caractéristiques. Il existe sans doute des biais d'indication, c'est pour cette raison que nous allons utiliser par la suite un score de propension.

Score de Propension

Le score de propension est défini comme la probabilité conditionnelle pour un patient de recevoir le traitement (la pose d'une sonde de Swan-Ganz) compte tenu de ses caractéristiques initiales observées. Il agit comme une « métavariable » résumant l'ensemble des facteurs de confusion, permettant ainsi de traiter les données d'une étude observationnelle comme celles d'un essai randomisé en équilibrant les groupes.

Construction du score

La construction repose généralement sur une régression logistique multivariée où la variable dépendante est l'exposition au traitement (SWANG1). Une fois estimé, ce score est utilisé via différentes méthodes d'ajustement (Appariement, Stratification, Pondération...).

Vérification des qualités du score

- Condition du support commun

Le support commun c'est la zone de superposition des deux groupes sur l'ensemble des valeurs du score de propension. Le support commun du score de propension permet de s'assurer qu'il est possible, pour chaque individu du groupe traité, de trouver au moins un participant du groupe témoin ayant le même score de propension.

L'utilisation du score de propension n'est adéquate que pour les individus situés dans cette zone. On va vérifier la condition du support commun en se basant sur la comparaison de la distribution du score de propension dans les deux groupes (surface commune sous les courbes de densité, et des boxplots).

- Propriété équilibrante du score

L'objectif premier du score est de neutraliser le biais d'indication en équilibrant la distribution des covariables. La qualité de cet équilibrage est validée par la réduction de la différence standardisée des moyennes (SMD), idéalement pour atteindre un seuil inférieur à 0,1 pour l'ensemble des variables. Et aussi par l'observation graphique (boxplots) montrant une homogénéisation des caractéristiques entre les groupes après ajustement.

- Estimation de l'effet du traitement

L'étape finale consiste à estimer l'effet propre de la sonde de Swan-Ganz sur la mortalité à J30. On utilise pour cela un modèle de régression (logistique ou de Cox) incluant l'exposition au traitement et le score de propension (sous forme de variable continue, de quantiles ou de poids) afin d'obtenir un Odds Ratio (OR) ou une Relative Hazard (RH) ajusté.

Ajustement sur le score de Propension

Construction du score de propension

Nous avons estimé le score de propension à l'aide d'un modèle de régression logistique en incluant les variables liées au critère de jugement (avoir une SWAN-GANZ) et donc celles potentiellement pronostiques de la survenue de décès à 30 jours.

Construction du score de propension à l'aide des 50 variables suivantes :

Groupe1 : Variables socio-démographiques (6 covariables) Age, Sex, Education, Revenus, Catégorie d'assurance, Race

Groupe2 : Présentation à l'admission (7 covariables) CAT1, DNR1, WTKILO, HRT1, PAM1, RESP1, TEMP1

Groupe3 : Catégories de diagnostic à l'admission (10 covariables) CARDIO, GASTRO, HEMATO, META, NEURO, ORTHO, RENAL, RESP, SEPS, TRAUMA

Groupe4 : Comorbidités et statut de transfert (13 covariables) AMIHX, CA, CARDIOHX, CHFHX, CHRPULHX, DEMENTHX, GIBLEDHX, IMMUNHX, LIVERHX, MALIGHX, PSYCHHX, RENALHX, TRANSHX

Groupe5 : Mesures résumant la sévérité à la présentation (4 covariables) APS1, DASI, SCOMA1, SURV

Groupe6 : Résultat Labo J1 (10 covariables)

ALB1, BILI1, CREA1, HEMA1, PACO21, PAFI1, POT1, SOD1, WBLC1

Avec le modèle logistique, le score de propension qu'on a obtenu a la distribution suivante :

Min	1 ^{re} quartile	Médiane	Moyenne	3 ^e quartile	Max
0.004	0.164	0.352	0.381	0.573	0.982

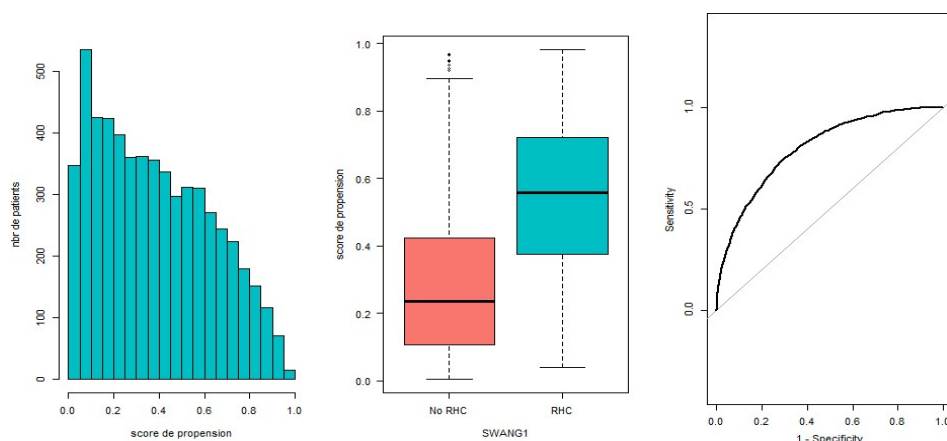


Figure 4: Score de propension. À gauche: La distribution du score de propension de tous les individus. Au centre: La distribution du score de propension (en boîte à moustaches) selon si groupe Swan ou non. À droite : Courbe ROC pour vérifier « l'adéquation » du score.

L'aire sous la courbe (AUC) est de **0,799**, ce qui indique une très bonne capacité du modèle à différencier les patients qui recevront la sonde de ceux qui ne la recevront pas. La boîte à moustaches et le *dot plot* (Figure 7) révèlent que le score de propension est significativement plus élevé dans le groupe RHC. Le graphique du groupe "No RHC" montre une base large (forte

densité de scores faibles) s'affinant au sommet, confirmant que ces patients avaient une probabilité initialement plus faible d'être appareillés.

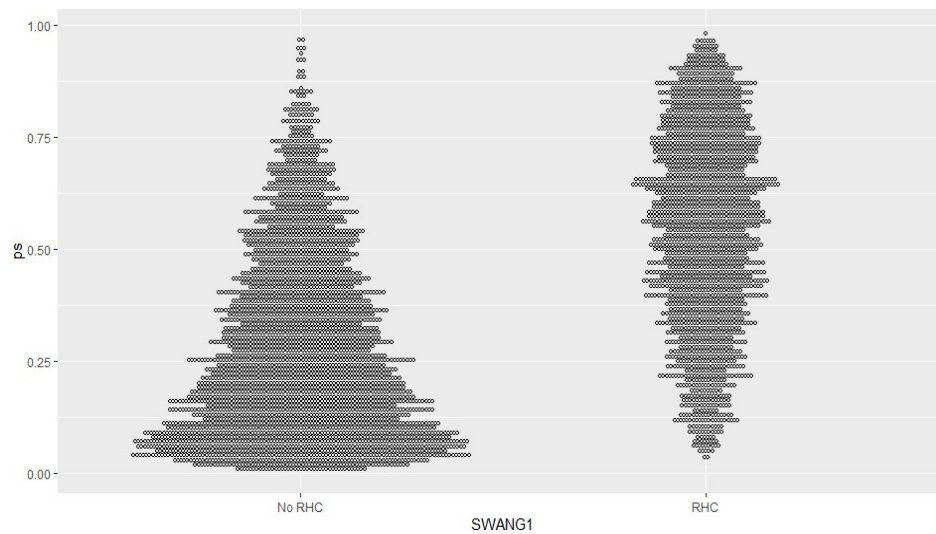


Figure 5: Score de propension. Dotplot : un point=1 patient, selon son score de propension et selon la pose ou pas de SWAN-GANZ. Le graphique de « No RHC » est large dans sa base, et plutôt fin au sommet.

Linéarisation du score

À savoir qu'on peut utiliser une forme linéaire de notre score de propension (linps : linear ps=logit ps). Cette forme linéaire peut avoir des propriétés intéressantes (distributions symétriques,) surtout lorsque les modélisations utilisées vont se baser sur des modèles linéaires.

La distribution d'un éventuel score de propension linéaire devient :

Min	1 ^{re} quartile	Médiane	Moyenne	3 ^e quartile	Max
-5.384	-1.626	-0.606	-0.686	0.290	3.987

Ce score de propension linéaire semble moins intuitif car on ne peut pas l'interpréter en termes de probabilités, c'est pour cela qu'on va se basé sur le score de propension (tel qu'il est) dans nos analyses.

Vérification des qualités su score de propension

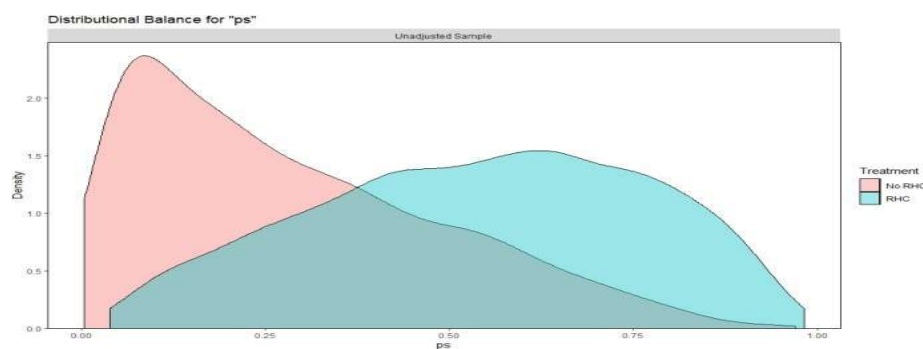


Figure 6: La distribution du score de propension dans les groupes RHC (Swan-Ganz) et le groupe No RHC (Pas de Swan-Ganz)

Les courbes de densité du score de propension (ps) montrent une zone de recouvrement significative entre le groupe "RHC" (bleu) et "No RHC" (rose). Cette superposition confirme l'existence d'un support commun : pour une large gamme de probabilités d'intervention, il existe des individus présentant des caractéristiques similaires dans les deux bras de l'étude. La présence de cette surface commune assure qu'il est statistiquement possible de comparer des patients "échangeables", c'est-à-dire ayant les mêmes probabilités cliniques de recevoir une sonde de Swan-Ganz, indépendamment du traitement réellement administré.

- On modélise la variable du décès à J30 par une régression logistique avec comme variables d'ajustement la variable SwanGanz (bien-sûr) et le score de propension.

$$\text{logit Décès à J30} = a + b (\text{Swan Ganz}) + c (\text{Score de Propension})$$

- Le ratio événements/variables est largement respecté (bien au-delà du seuil de 5 à 10 événements par variable), garantissant la stabilité des estimations. Bien que ce modèle confirme la tendance observée dans l'analyse bivariée, cet ajustement "brut" suppose une relation linéaire entre le score de propension et le logit de la mortalité, une hypothèse forte qui nécessite d'être complétée par des méthodes plus robustes comme l'appariement ou la pondération.

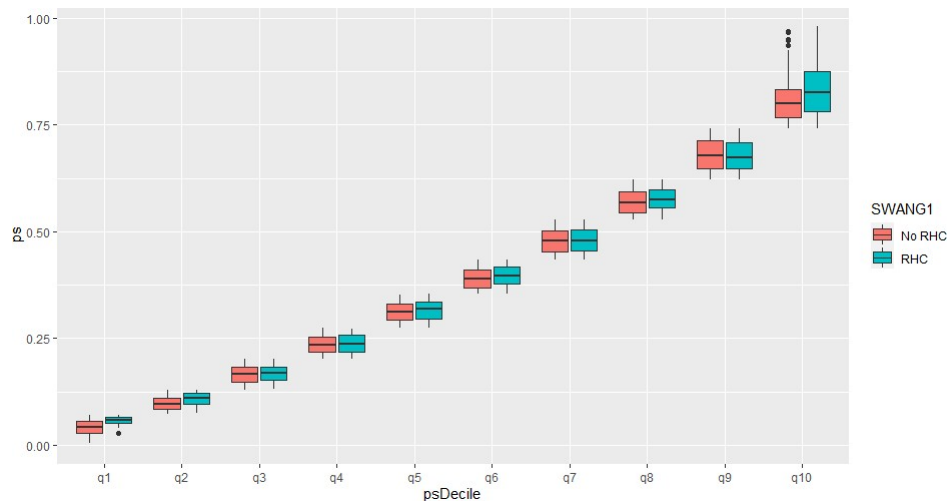
	Coefficient de régression		OR et son IC 95%
Swan Ganz (1=Oui, 0=Non)	0.261 (0.06638)	8.36e-05	1.298 [1,139;1,478]

- La pose d'une sonde de Swan-Ganz reste associée à une augmentation du risque de décès à 30 jours, avec un Odds Ratio (OR) de 1,29.

Stratification du score de propension

Construction du score de propension

L'ajustement direct sur le score de propension suppose une relation linéaire entre ce dernier et le risque de décès, une hypothèse souvent trop restrictive. Pour pallier cette limite, nous avons divisé la population en déciles de score de propension. Cette approche garantit que, dans chaque strate, les patients présentent des probabilités d'intervention comparables, neutralisant ainsi localement les biais d'indication.



Comme l'illustre la Figure 8 ci-dessus, la distribution des scores de propension entre le groupe "No RHC" et "RHC" est parfaitement équilibrée au sein de chaque décile (q1 à q10).

Estimation de l'effet de la sonde Swan-Ganz

	Coefficient de régression (+/- standard error)	p-value	OR et son IC 95%
Swan Ganz (1=Oui, 0=Non)	0.268 (0.066)	5.5e-05	1.307 [1.148 -1.489]

La variable « score de propension en déciles » étant ici une variable catégorielle à de nombreuses modalités (=10), on va modéliser le décès à 30j par une régression logistique conditionnelle aux strates de déciles. Cette méthode permet une analyse plus fine que la régression logistique simple et notamment permet de s'affranchir du nombre élevé de modalités du score de propension, pris ici comme variable à 10 catégories.

- L'effet global du cathétérisme droit est donc une association statistiquement significative à une surmortalité, avec un OR de 1.307.

Vérification des qualités de l'appariement

La propriété équilibrante

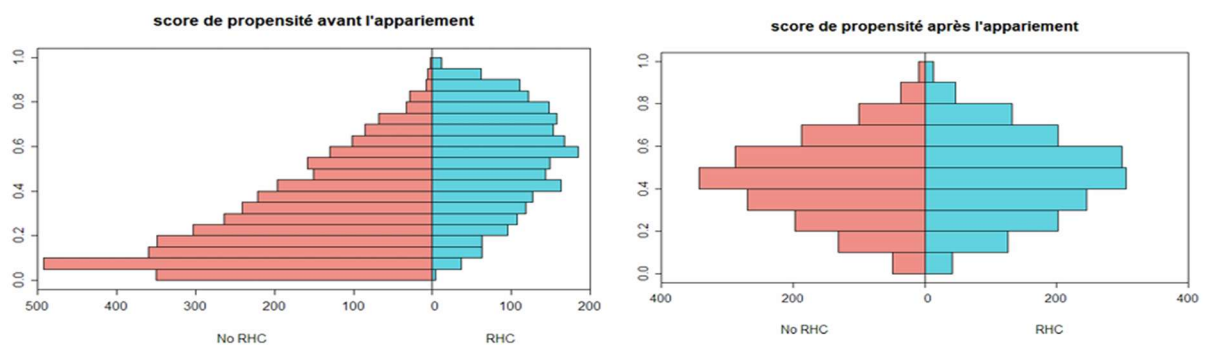


Figure 8: La distribution du score de propension dans les 2 groupes, avant et après l'appariement

Après appariement, le score de propension est équilibré dans les 2 groupes de traitement. Cet équilibre est illustré avec des histogrammes mais aussi avec des boxplots (figure 8). Ce sont surtout les « loveplot » qui nous permettent de souligner la minimisation du SMD (figure 9).

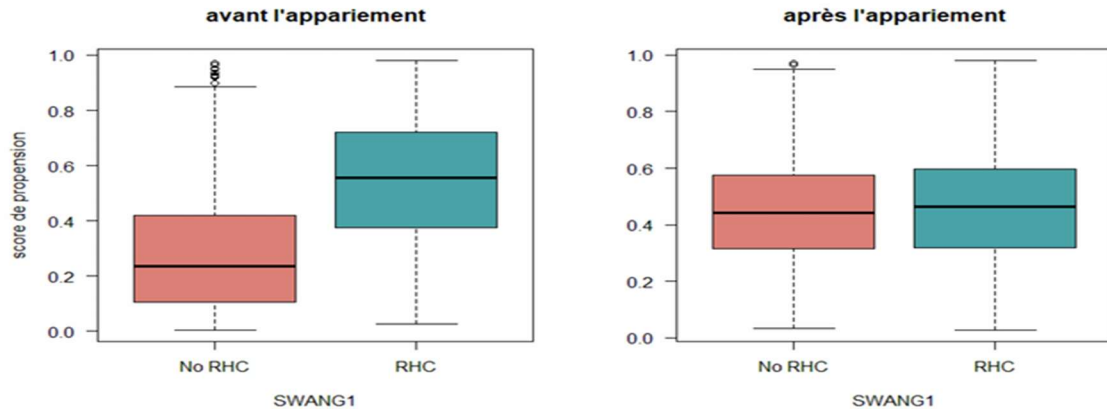
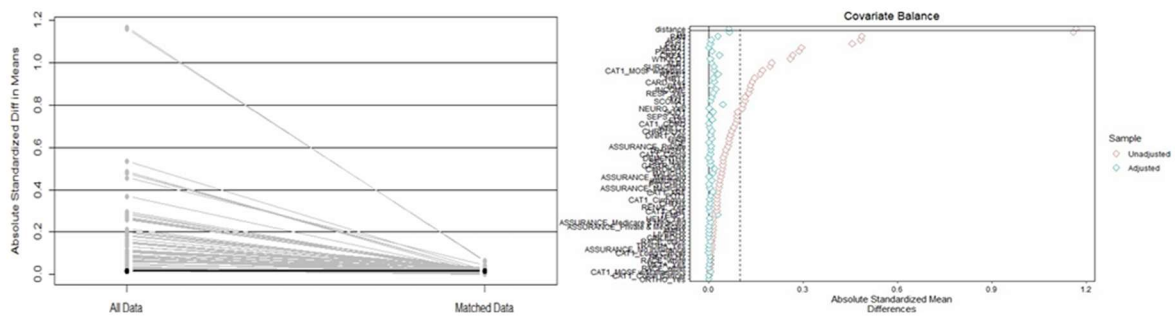


Figure 9: La distribution du score de propension dans les 2 groupes, avant et après l'appariement

On trouve que les SMD (absolute standardized differences in means) après l'appariement sont toutes inférieures à 0.1, donc l'équilibre est très satisfaisant.



La condition du support commun

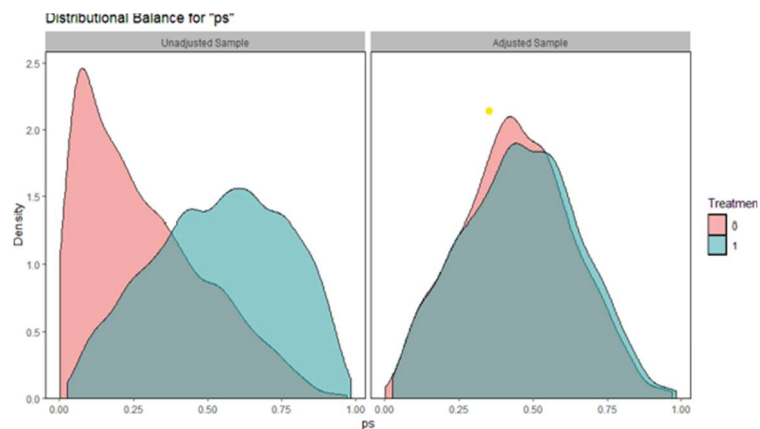


Figure 11: Courbes de densité, avant et après l'appariement

Les deux courbes de densité du score de propension, après l'appariement, ont une grande surface commune (figure 11). La distribution du score de propension dans les groupes appariés (traitement et contrôle appariés) sont « superposables », selon le graphique des points et selon les histogrammes (figures 12 et 13).

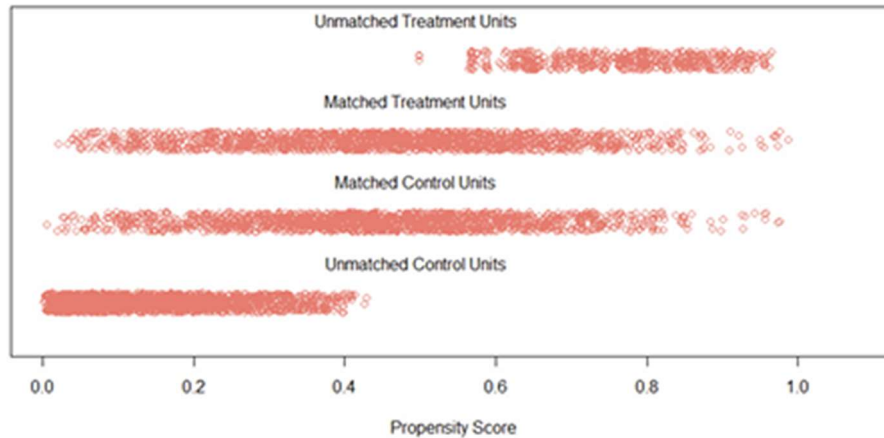


Figure 12: Distribution des scores de propension dans les groupes « traitement » (RHC) et « Contrôle » (No RHC)

On remarque que les 2 sous-groupes appariés (matchés) ont une distribution superposable. Les sous- groupes non-matchés ont des scores de propension qui ne se chevauchent pas.

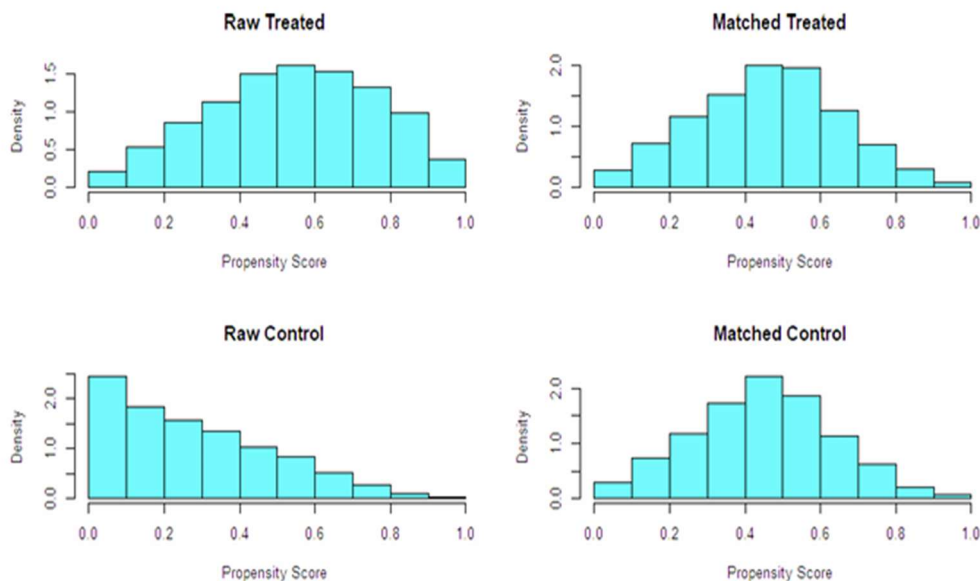


Figure 13: Histogrammes des scores de propension dans les groupes traitement et contrôle, avant appariement (Raw) et après appariement (matched)

Estimation de l'effet de la sonde Swan-Ganz

À travers un modèle logistique spécial (Weighted GLM) où l'on rentre le poids d'appariement, on arrive à estimer l'effet de Swan Ganz.

	Coefficient de régression	p-value	OR et son IC95%
Swan Ganz (1=Oui, 0=Non)	0.270 (0.074)	0.000261	1.316 [1.133 ; 1.515]

La pose de Swan Ganz est significativement associée à un OR plus élevé de mortalité à J30.

Pondération sur le score de propension

Construction des pondérations

Une pondération intéressante est celle de l'IPTW (inverse probability of treatment weighting) avec laquelle chaque individu reçoit une pondération inverse de la probabilité d'avoir reçu le traitement, probabilité ici estimée par le score de propension. Plus spécifiquement, pour les individus traités, la pondération est l'inverse du score de propension, c'est-à-dire $(PS)^{-1}$, tandis que, pour les individus du groupe témoin, la pondération est $(1-PS)^{-1}$.

En plus, on s'intéresse à l'effet marginal de la Swan-Ganz dans la population effectivement traitée (donc ATT: Average Treatment effect on the Treated).

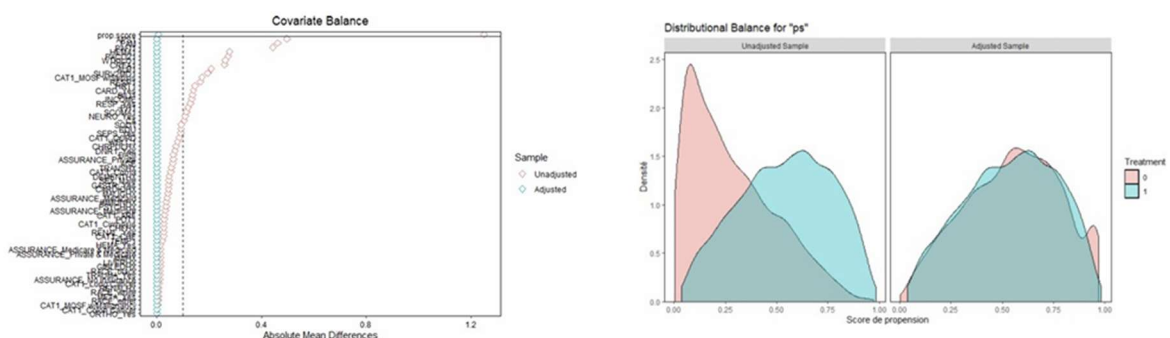
La distribution du poids de pondération qu'on retient est la suivante :

Min	Q1	Médiane	Moyenne	Q3	Max	Coefficient de Var
1.004	1.19	1.47	1.98	2.08	37.4	1.01

	Coefficient de variance du poids de pondération	Taille de l'échantillon « pondéré »
Groupe « Swan-Ganz »	1.017	1073
Groupe témoin (Pas de Swan)	0.86	2035

Vérification des qualités de la pondération

On remarque que les qualités équilibrantes et la surface de support commun avec notre pondération sont satisfaisantes grâce aux représentations ci-dessous :



Estimation de l'effet de la sonde Swan-Ganz

Avec un modèle logistique prenant en compte le poids de pondération, on estime l'effet de Swan Ganz, avec un OR de 1.296 pour une mortalité significativement plus élevée.

	Coefficient de régression	OR et son IC95%
Swan Ganz (1=Oui, 0=Non)	0.06	1,296 [1,199 - 1,401]

4. Analyse de survie à J30

Analyse de survie sur les données brutes

L'analyse par la méthode de Kaplan-Meier (Figure 16) révèle que la survie à 30 jours est significativement plus faible dans le groupe ayant reçu une sonde de Swan-Ganz (\$p < 0,0001\$).

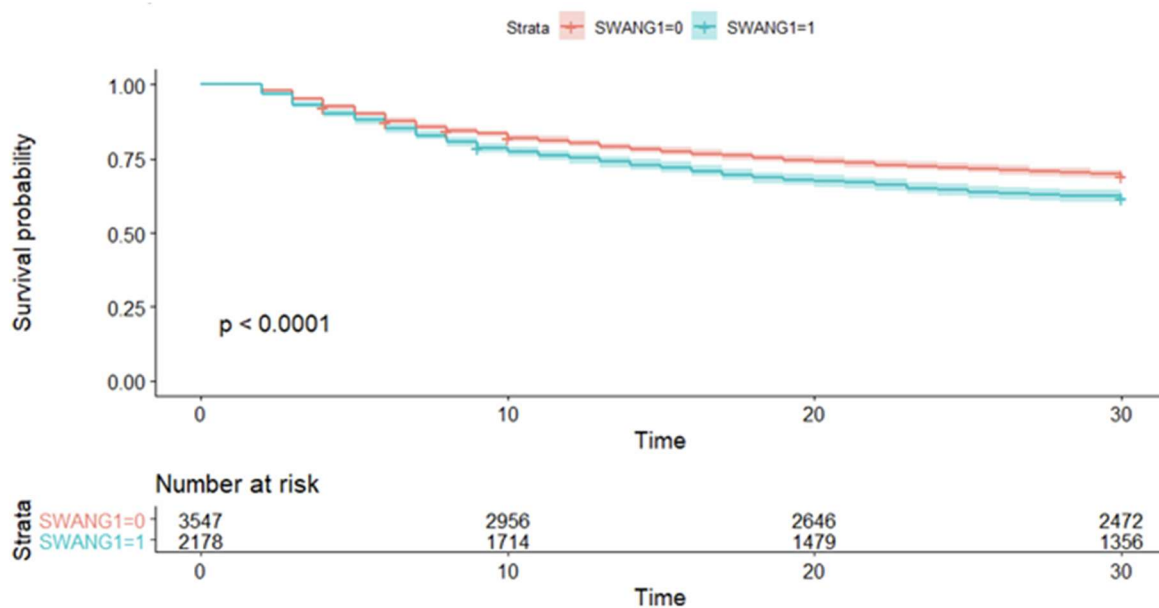


Figure 16 : Courbe survie (méthode Kaplan Meier) entre les groupes avec et sans sonde sur données brutes

Le modèle de Cox non ajusté confirme cette tendance avec un Hazard Ratio (HR) de 1,309 [1.196 ; 1.433]. Ce résultat est cohérent avec les Odds Ratios obtenus précédemment lors des analyses par score de propension. Les conditions d'application du modèle de Cox, notamment l'hypothèse des risques proportionnels, ont été vérifiées et validées pour cette analyse.

Afin d'affiner l'estimation et de corriger les biais d'indication identifiés lors de la phase bivariable, nous avons introduire le score de propension comme variable d'ajustement dans le modèle de Cox. Ce qui permet de neutraliser l'effet des facteurs de confusion résumés dans cette métavariable.

Le Hazard Ratio ajusté est de 1,235 [1.112 ; 1.373].

Bien que l'ajustement par le score de propension réduise légèrement l'ampleur de l'effet (passant d'un HR de 1,309 à 1,235), l'association entre la pose de la sonde de Swan-Ganz et l'augmentation du risque de décès à 30 jours demeure statistiquement significative. Ces résultats confirment que, même à caractéristiques cliniques et pronostiques égales, l'intervention est associée à une survie moindre dans cette cohorte.

Analyse de survie sur les données appariées

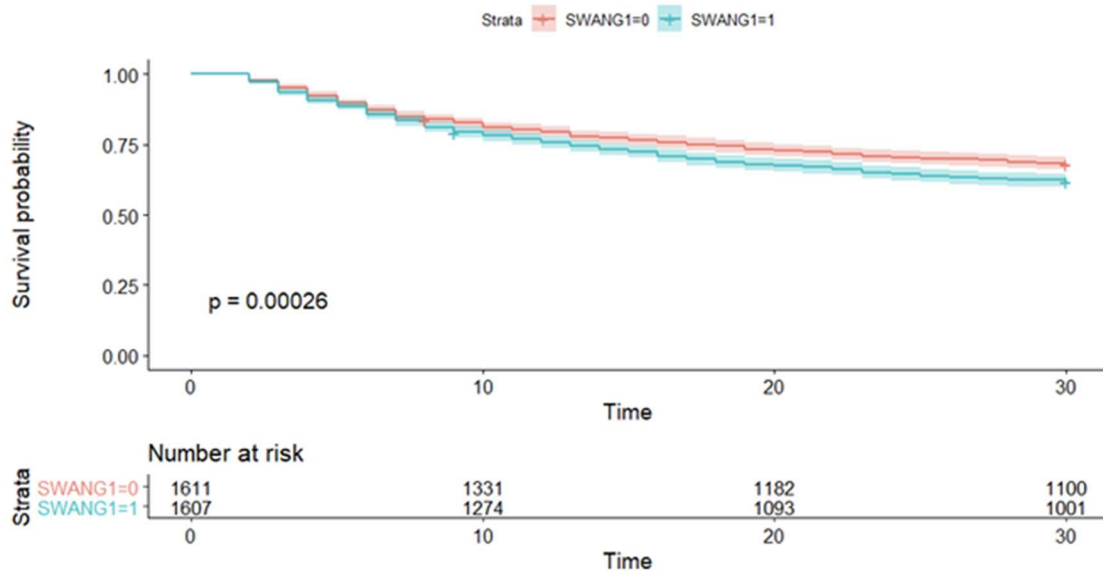


Figure 17 : Courbe survie (méthode Kaplan Meier) entre les groupes avec et sans sonde sur données appariées

Les courbes de Kaplan-Meier obtenues après appariement (Figure 17) confirment visuellement la persistance d'un écart de survie entre les deux bras. La survie reste systématiquement plus faible chez les patients ayant bénéficié de la pose d'une sonde de Swan-Ganz par rapport à leurs "jumeaux statistiques" non exposés.

L'application du modèle de Cox sur cette population sélectionnée permet d'isoler l'effet de l'intervention en minimisant les biais de confusion.

Le risque de décès est estimé avec un Hazard Ratio (HR) de 1,281 [1,139 ; 1,442]. On voit que cet intervalle ne contient pas la valeur 1, ce qui confirme que cette augmentation du risque est statistiquement significative.

5. Récapitulation

On récapitule tous les résultats obtenus permettant d'estimer l'effet de la Swan Ganz sur la mortalité à J30 (critère de jugement principal) et à J180 (critère de jugement secondaire), avec les 4 méthodes basées de score de propension et les modèles de survie.

- Mortalité à J30

	OR	IC95%
Méthode1 : Ajustement sur le score de propension	1.300	[1.140 ; 1.480]
Méthode2 : Stratification sur le score de propension (modèle logistique conditionnel)	1.307	[1.148 ; 1.489]
Méthode3 : Appariement sur le score de propension (modèle weighted GLM)	1.316	[1.133 ; 1.515]
Méthode4 : Pondération sur le score de propension (modèle weighted GLM)	1.296	[1.199 ; 1.401]

- **Mortalité à J180**

	OR	IC95%
Méthode1 : Ajustement sur le score de propension	1.314	[1.154 ; 1.499]
Méthode2 : Stratification sur le score de propension (modèle logistique conditionnel)	1.328	[1.165 ; 1.513]
Méthode3 : Appariement sur le score de propension (modèle weighted GLM)	1.317	[1.138 ; 1.526]
Méthode4 : Pondération sur le score de propension (modèle weighted GLM)	1.314	[1.216 ; 1.421]

- **Modèle de survie**

	HR	IC95%
Méthode1 : Modèle sur données brutes	1.314	[1.154 ; 1.499]
Méthode2 : Modèle sur données brutes avec ajustement sur score de propension	1.235	[1.112 ; 1.373]
Méthode2 : Modèle sur données appariées	1.281	[1.139 ; 1.442]

Toutes les méthodes sont en accord avec une augmentation de l'OR de mortalité chez les patients ayant eu la pose de SWAN – GANZ.

6. Synthèse des résultats

Le score de propension nous a permis dans cette étude de nous affranchir des biais d'indication potentiels. Nous avons trouvé que la pose de la sonde Swan-Ganz dans les 24h après une admission en réanimation était associée au décès des patients dans les 30 jours.

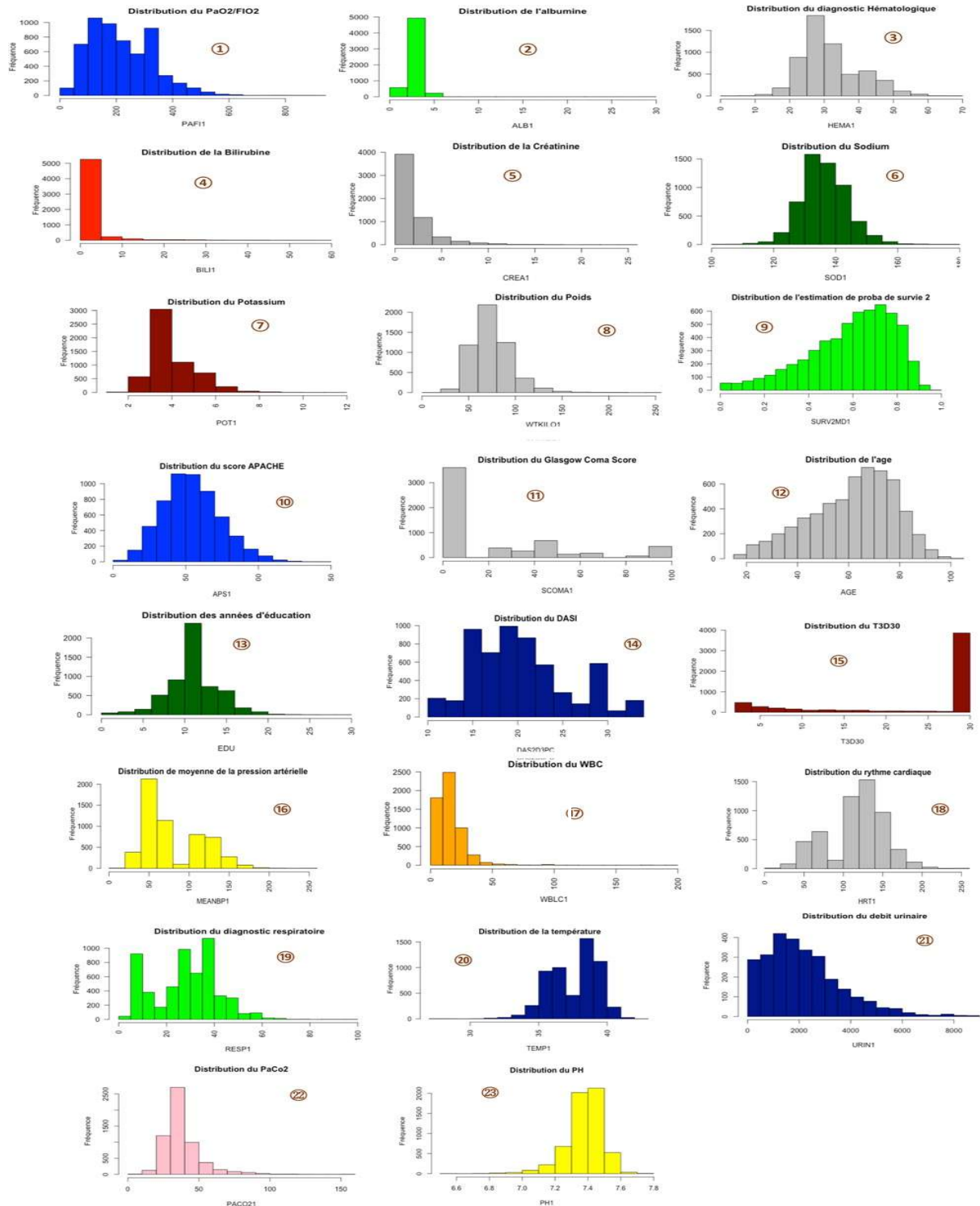
On peut donc dire qu'il est très probable l'existence d'une association entre la pose de la sonde Swan-Ganz dans les 24h premières heures d'hospitalisation en réanimation et le décès à J30 qui reste significatif après appariement sur un score de propension et après ajustement.

Il n'est cependant pas possible de conclure que les patients recevant la sonde Swan-Ganz décèdent plus en réanimation indépendamment de la sévérité de leur pathologie à l'admission car cette étude reste une étude observationnelle et non randomisée.

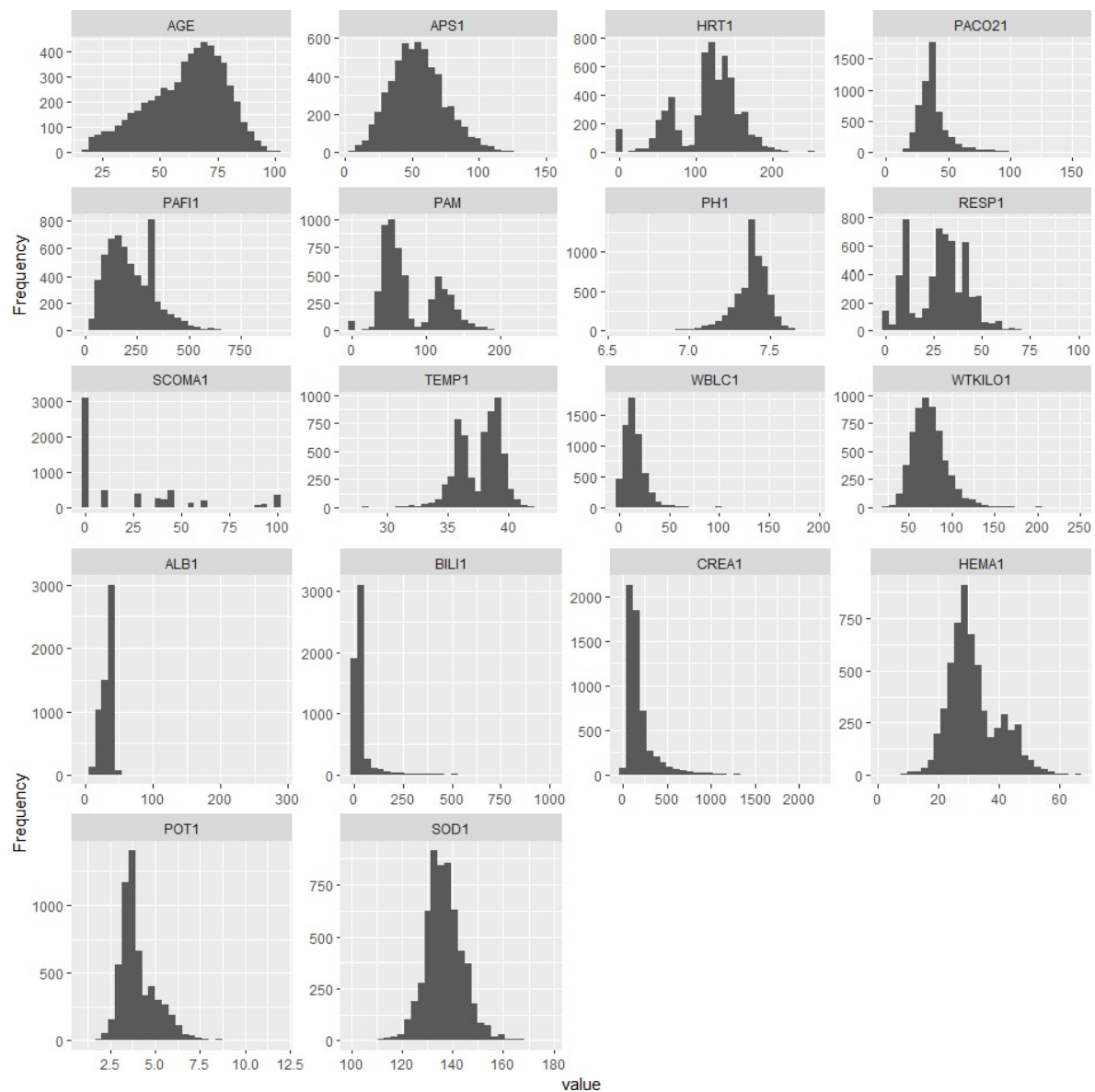
La limite des méthodes basées sur le score de propension c'est que, contrairement à la randomisation, elles ne permettent d'ajuster (équilibrer) que sur les variables mesurées. Même si le score de propension ne remplace pas la randomisation et donc les essais cliniques, son utilisation est intéressante lorsque la randomisation ne peut avoir lieu.

7. Annexe

Représentation graphique des variables quantitatives.



Représentation graphique des variables qualitatives



Epidémiologie

Data Management

#Chargement des librairies

```
library(scales) #pour trouver de "couleurs" vives
show_col(hue_pal()(5)) #pour choisir la couleur qu'on désire
library(prettyR)
library(ggplot2)
library(gplots)
library(doParallel)
registerDoParallel(cores=8)
```

#Importation du jeu de données

```
#rhc: nom du jeu de données, dans lequel on va faire le data management
rhc <- read.csv2("~/Downloads/rhc devoir épidémio-RC csv - rhc devoir épidémio-RC csv .csv") View(rhc)
str(rhc) ## 63 variables, 5735 observations
summary(rhc)
#Recherche de doublons
table(duplicated(rhc)) # il n'y a pas de doublon de sujet -> 5735
#Valeurs manquantes
table(is.na(rhc))
prop.table(table(is.na(rhc)))*100 #3.84% NA
sapply(rhc[,1:63], function(i){table(is.na(i))})
#Variables ayant les valeurs manquantes: CAT2, DSCHDTE, DTHDTE, ADLD3P, URIN1
```

#Suppression de PTID pour assurer l'anonymat

```
rhc$PTID <- NULL
```

#Mise en facteurs

```
rhc$CARDIOHX <- as.factor(rhc$CARDIOHX)
rhc$CHFHX <- as.factor(rhc$CHFHX)
rhc$DEMENTHX <- as.factor(rhc$DEMENTHX)
rhc$PSYCHHX <- as.factor(rhc$PSYCHHX)
rhc$CHRPULHX <- as.factor(rhc$CHRPULHX)
rhc$RENALHX <- as.factor(rhc$RENALHX)
rhc$LIVERHX <- as.factor(rhc$LIVERHX)
rhc$GIBLEDHX <- as.factor(rhc$GIBLEDHX)
rhc$MALIGHX <- as.factor(rhc$MALIGHX)
rhc$IMMUNHX <- as.factor(rhc$IMMUNHX)
rhc$TRANSHX <- as.factor(rhc$TRANSHX)
rhc$AMIHX <- as.factor(rhc$AMIHX)
```

#pour la variable AGE

```
summary(rhc$AGE)
describe.numeric(rhc$AGE, num.desc = c("min","max","mean","median","sd","valid.n")) rhc$ROWNAMES[which(rhc$AGE
>= 100)]
#donc 3 patients agés de >100 ans: plausible
hist(rhc$AGE,col="gray", las=1, main = "Distribution de l'Age", xlab="Age",ylab="Nbr de patients",ylim=c(0,800))
#Création d'une variable AGE_CAT pour répartir les patients en catégories : <40 / 40-60 /60-80 / > 80
rhc$AGE_CAT <- ifelse(rhc$AGE < 40, "< 40 ans", ifelse(rhc$AGE >= 40 & rhc$AGE <60, "40-60 ans", ifelse(rhc$AGE >=60 &
rhc$AGE < 80, "60-80 ans", ">=80 ans")))
describe.factor(rhc$AGE_CAT)
pie(table(rhc$AGE_CAT), col=c("#F8766D","#00BFC4"), main="Catégorie d'Age")
```

#pour la variable SEX

```
summary(rhc$SEX,useNA="always")
table(rhc$SEX)
tab <- table(rhc$SEX)
prop.table(tab)
```

Epidémiologie

```
barplot(sort(table(rhc$SEX), decreasing=TRUE), col= "red", xlab="Sexe", ylab="Fréquence")
```

#pour la variable RACE

```
barplot(table(rhc$RACE))
describe.factor(rhc$RACE)
barplot(sort(table(rhc$RACE),decreasing=TRUE), col="dark gray",xlab=" RACE ", ylab="Fréquence")
pie(table(rhc$RACE), col=hue_pal()(4),main="origine ethnique")
```

#pour la variable EDU

```
describe.numeric(rhc$EDU, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
summary(rhc$EDU)
hist(rhc$EDU,col="dark green",main="Années d'éducation",xlab="nbr d'années d'études", ylab="Nbr de patients")
table(rhc$EDU<1) #31 patients qui n'ont jamais fait d'années d'éducation
rhc$ROWNAMES[which(rhc$EDU<1)]
```

#graphiques démographiques annexes

#entre income, age, et sexe

```
ggplot(rhc, aes(INCOME,AGE, fill=SEX)) + geom_boxplot()
```

#entre income, age et race

```
ggplot(rhc, aes(INCOME,AGE, fill=RACE)) + geom_boxplot()
```

#pour la variable INCOME

```
summary(rhc$INCOME)
table(rhc$INCOME,useNA="always")
rhc$INCOME <- ifelse(rhc$INCOME == "Under $11k",1, ifelse(rhc$INCOME == "$11-$25k",2, ifelse(rhc$INCOME == "$25-$50k",3,4)))
table(rhc$income)
prop.table(table(rhc$INCOME))
barplot(sort(table(rhc$INCOME),decreasing=TRUE), col="blue",xlab=" INCOME ", ylab="Fréquence")
pie(table(rhc$INCOME),col=hue_pal()(4),main="revenus")
```

#pour la variable NINSCLAS (type assurance)

```
rhc$ASSURANCE <- rhc$NINSCLAS
table(rhc$ASSURANCE, useNA = "always")
rhc$NINSCLAS <- NULL
describe.factor(rhc$ASSURANCE)
table(rhc$INCOME,rhc$ASSURANCE)
```

#pour la variable CAT1

```
table(rhc$CAT1,useNA="always") #9 facteurs dans CAT 1
length(rhc$CAT1)
barplot(sort(table(rhc$CAT1), decreasing=TRUE), col="green", xlab=" CAT1", ylab="Fréquence")
```

#pour la variable CAT2

```
table(rhc$CAT2, useNA="always") #6 facteurs dans CAT 2
length(rhc$CAT2) 5735-4535 #seulement 1200 patients ont eu deux diagnostics lors du codage.
rhc$CAT2 <- NULL #on décide de virer cette variable, bcp de données manquantes.
```

palettacol<-c("#F8766D", "#00BFC4") #nuances de couleurs

#pour les variables binaires indicatrices d'antécédants (comorbidités)

```
barplot(prop.table(table(rhc$CARDIOHX, useNA="ifany"))*100, main="CARDIOHX", col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$CHFHX, useNA="ifany"))*100, main="CHFHX", col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$DEMENTHX, useNA="ifany"))*100, main="DEMENTHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$PSYCHHX, useNA="ifany"))*100, main="PSYCHHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$CHRPULHX, useNA="ifany"))*100, main="CHRPULHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$RENALHX, useNA="ifany"))*100,main="RENALHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$LIVERHX, useNA="ifany"))*100,main="LIVERHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$GIBLEDHX, useNA="ifany"))*100, main="GIBLEDHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$MALIGHX, useNA="ifany"))*100, main="MALIGHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$IMMUNHX, useNA="ifany"))*100,main="IMMUNHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$TRANSHX, useNA="ifany"))*100,main="TRANSHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
```

```
barplot(prop.table(table(rhc$AMIHX, useNA="ifany"))*100,main="AMIHX",col= palettacol,ylim=c(0,100), ylab="Fréquence (%)")
```

Epidémiologie

#pour les variables binaires indicatrices d'admission

```
par(mfrow=c(2,5))
barplot(prop.table(table(rhc$RESP, useNA="ifany"))*100, main="RESP", col=c("#F8766D", "#00BFC4"),ylim=c(0,100),
  ylab="Fréquence (%)")
barplot(prop.table(table(rhc$CARD, useNA="ifany"))*100, main="CARD", col=c("#F8766D", "#00BFC4"),ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$NEURO, useNA="ifany"))*100, main="NEURO", col=c("#F8766D", "#00BFC4"),ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$GASTR, useNA="ifany"))*100, main="GASTR", col=c("#F8766D", "#00BFC4"),ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$RENAL, useNA="ifany"))*100, main="RENAL", col=c("#F8766D", "#00BFC4"),ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$META, useNA="ifany"))*100, main="META", col=c("#F8766D", "#00BFC4"),ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$HEMA, useNA="ifany"))*100, main="HEMA", col=c("#F8766D", "#00BFC4"),ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$SEPS, useNA="ifany"))*100, main="SEPS", col=c("#F8766D", "#00BFC4"),ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$TRAUMA, useNA="ifany"))*100, main="TRAUMA", col=c("#F8766D", "#00BFC4"), ylim=c(0,100), ylab="Fréquence (%)")
barplot(prop.table(table(rhc$ORTHO, useNA="ifany"))*100, main="ORTHO", col=c("#F8766D", "#00BFC4"),ylim=c(0, 100), ylab="Fréquence (%)")
```

#pour la variable ADL: activity of Daily Living

```
describe.numeric(rhc$ADLD3P, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
```

summary(rhc\$ADLD3P) #il y a 4296 données manquantes (74,9%)

#Suppression de la variable ADLD3P

```
rhc$ADLD3P <- NULL
```

#pour la variable DASI – Duke Activity Status Index

summary(rhc\$DAS2D3PC) #Les réponses positives sont additionnées pour obtenir un score total, qui varie de 0 à 58,2.

#Pas de données manquants, pas d'aberrants (valeurs théoriques du score comprises entre 0 et 58)

```
rhc$DASI <- rhc$DAS2D3PC
```

```
rhc$DAS2D3PC <- NULL
```

```
summary(rhc$DASI)
```

#pour la variable DNR Status

```
summary(rhc$DNR1)
```

table(rhc\$DNR1) #Pas de données manquantes

#pour la variable cancer

```
summary(rhc$CA)
```

```
table(rhc$CA)
```

table(rhc\$CA, rhc\$MALIGHX) #donc peut-être 40 patients

```
rhc$CA <- ifelse(rhc$CA=="No",0,ifelse(rhc$CA=="Yes",1,2))
```

```
table(rhc$CA)
```

```
rhc$CA <- as.factor(rhc$CA)
```

```
par(mfrow=c(1,1))
```

#pour la variable SURV2MD1: SUPPORT model estimate of the probability of surviving 2 months

```
describe.numeric(rhc$SURV2MD1, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
```

```
summary(rhc$SURV2MD1)
```

hist(rhc\$SURV2MD1,col=palettecol) #On remarque qu'il n'y a pas de valeurs aberrantes

#pour la Variable Swang

```
table(rhc$SWANG1)
```

```
tab1 <- table(rhc$SWANG1)
```

```
prop.table(tab1)
```

```
barplot(sort(table(rhc$SWANG1),decreasing=TRUE), col=" dark green", xlab="Swang1", ylab="Fréquence")
```

#pour la variable APACHE

summary(rhc\$APS1) #pas de valeurs aberrantes

```
describe.numeric(rhc$APS1, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
```

boxplot(APS1~DTH30,rhc,ylab="Score Apache",xlab="Décès à 30 jours",main="Score Apache en fonction de la mortalité à 30 jours",col="grey")

#pour la variable SCOMA

```
table(rhc$SCOMA1, useNA="always")
```

Epidémiologie

```
describe.numeric(rhc$SCOMA1, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
hist(rhc$SCOMA1)
#Valeurs inhabituelles: normalement le score de Glasgow est compris entre 3 et 15, mais ici il est indiqué entre 0 et 100
selon l'article de Annals of Internal Medicine (Knaus et al, The SUPPORT prognostic model), le score de Glasgow a une
relation non linéaire avec le décès et donc a été relinéarisé.
boxplot(SCOMA1~DTH30,rhc ,ylab="Score de Glasgow",xlab="Décès à 30 jours",main="Score de Glasgow en fonction de la
mortalité à 30 jours", col="grey")
```

```
#pour la variable DEATH
summary(rhc$DEATH)
ifelse(rhc$DEATH=="Yes",1,0)
table(rhc$DEATH)
round(prop.table(table(rhc$DEATH,rhc$SCOMA1)),2)
tab2<-table(rhc$DEATH)
prop.table(tab2)
barplot(sort(table(rhc$DEATH), decreasing=TRUE), col=" blue", xlab=" DEATH ", ylab="Fréquence")
```

```
#pour la variable weight
summary(rhc$WTKILO1)
table(rhc$WTKILO1==0)
hist(rhc $WTKILO1, col="gray",xlab="WTKILO1",ylab="Fréquence", main=" Distribution du Poids ")
rhc$WTKILO1 <- ifelse(rhc$WTKILO1==0, NA, rhc$WTKILO1)
hist(rhc $WTKILO1, col="gray",xlab="WTKILO1",ylab="Fréquence", main=" Distribution du Poids ")
summary(rhc$WTKILO1) #515 données manquantes
#515 données en 0 transformées en manquantes, on aura par la suite recours à une imputation multiple
```

```
#pour la variable température
summary(rhc$TEMP1)
sd(rhc$TEMP1)
hist(rhc$TEMP1, col="blue",xlab=" Temp1 ",ylab="Fréquence", main=" Distribution de la température ")
#pas de données manquantes, ni aberrante.
```

```
#pour la variable MEANBP1 (Moyenne de pression cardiaque)
rhc$PAM <- rhc$MEANBP1
rhc$MEANBP1 <- NULL
describe.numeric(rhc$PAM, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
hist(rhc$PAM, col="yellow",xlab="PAM",ylab="Fréquence", main=" Distribution de moyenne de la pression artérielle ")
rhc$ROWNAMES[which(rhc$PAM > 220)]
# 1 patient a une PAM > 250, on conserve ce patient isolé
table(rhc$PAM == 0) # 80 patients ont une tension 0
describe.numeric(rhc$HRT1, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
table(rhc$HRT1 == 0) # Il y a des valeurs extrêmes 0 et 250, plausibles en réanimation
table(rhc$PAM==0,rhc$HRT1==0,deparse.level = 2)
table(subset(rhc,PAM==0,c(HRT1))) #20 patients ont une PAM 0 mais une fréquence cardiaque mesurable...
```

```
#pour la variable RESP1
describe.numeric(rhc$RESP1, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
table (rhc$RESP1 == 0)
```

```
#pour la variable PAFI1 (Ratio PAO2/FIO2)
describe.numeric(rhc$PAFI1, num.desc = c("min", "max", "mean", "median", "sd", "valid.n"))
table(rhc$PAFI1>500) #une personne normale peut avoir une valeur de 500
```

```
#pour la variable PACO2
describe.numeric(rhc$PACO21,num.desc=c(num.desc = c("min", "max", "mean", "median", "sd", "valid.n")))
table(rhc$PACO21<5) #2 patients ont des valeurs extrêmes (<5), mais ce n'est pas impraticable.
```

```
#pour la variable PH
describe.numeric(rhc$PH,num.desc=c(num.desc = c("min", "max", "mean", "median", "sd", "valid.n")))
hist(rhc $PH1, col="yellow",xlab="PH1",ylab="Fréquence", main=" Distribution du PH ")
table(subset(rhc, PACO21>100,PH1)) #les patients à PACO2>100 avaient un pH faible
```

Epidémiologie

#pour la variable WBLC1 (White Blood cells)

```
summary(rhc$WBLC1)
describe.numeric(rhc$WBLC1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
hist(rhc$WBLC1, col="gray",xlab="WBLC1",ylab="Fréquence", main=" Distribution du WBC")
#les valeurs extrêmes peuvent correspondre à des leucémies, ou hémopathies, etc..
#table(rhc$WBLC1>100) #5 patients ont des valeurs très élevés de WBLC
rhc$ROWNAMES[which(rhc$WBLC1>100)]
table(rhc$WBLC1==0) #4 patients ont 0 globules blancs...
```

#pour la variable hématocrite

```
describe.numeric(rhc$HEMA1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
hist(rhc $HEMA1, col="gray",xlab="HEMA1",ylab="Fréquence", main=" Distribution de l'hématocrite ")
table(rhc$HEMA1<5) #2 patients ont des valeurs d'hématocrite trop basse et c'est plausible
rhc$ROWNAMES[which(rhc$HEMA1<5)]
```

#pour la variable sodium

```
describe.numeric(rhc$SOD1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
hist(rhc $SOD1, col="dark green",xlab="SOD1",ylab="Fréquence", main=" Distribution du Sodium ")
table(rhc$SOD1<110) #pas de données manquantes ou aberrantes, distribution normale.
```

#pour la variable T3D30: temps de vie en jours jusqu'à 30 jours

```
describe.numeric(rhc$T3D30) #on renomme la variable en JVIE_J30 pour plus de compréhension
rhc$JVIE_J30 <- rhc$T3D30
rhc$T3D30 <- NULL
describe.numeric(rhc$JVIE_J30, num.desc = c("min","max","mean","median","sd", "valid.n")) #pas de données aberrantes
a priori, personnes > 30 jours, pas de NA, moyenne à 23.61, médiane 30, minimum 2, maximum 30 jours.
hist(rhc$T3D30, col="dark red",xlab="T3D30",ylab="Fréquence", main=" Distribution du T3D30 ")
```

#pour la variable potassium

```
summary(rhc$POT1)
describe.numeric(rhc$POT1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
hist(rhc $POT1, col="dark red",xlab="POT1",ylab="Fréquence", main=" Distribution du Potassium ")
table(rhc$POT1<1.2) #2 patients ont des valeurs de potassium anormalement trop basses
table(rhc$POT1>9) #5 patients ont des valeurs de potassium anormalement trop élevées
```

#pour la variable créatinine

```
summary(rhc$CREA1)
describe.numeric(rhc$CREA1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
#conversion en µmol/L
rhc$CREA1 <- rhc$CREA1*88.42
describe.numeric(rhc$CREA1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
table(subset(rhc, POT1>9,CREA1)) #parmi les 5 patients qui avaient une hyperkaliémie très importante,
4 avaient une insuffisance rénale.
```

#pour la variable BILIRUBINE

```
summary(rhc$BILI1)
rhc$BILI1 <- rhc$BILI1*17.1 #conversion en µmol/L
describe.numeric(rhc$BILI1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
hist(rhc $BILI1, col="red",xlab="BILI1",ylab="Fréquence", main=" Distribution de la Bilirubine ")
```

#pour la variable ALBUMINE

```
summary(rhc$ALB1)
rhc$ALB1 <- rhc$ALB1*10
describe.numeric(rhc$ALB1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
table(rhc$ALB1>100)
rhc$ROWNAMES[which(rhc$ALB1>100)] #2 patients ont des valeurs d'albumine anormalement trop élevées.
```

#pour la variable URINE

```
summary(rhc$URIN1)
describe.numeric(rhc$URIN1,num.desc=c(num.desc = c("min","max","mean","median","sd","valid.n")))
#les valeurs extrêmes sont plausibles en réa (hypertrigycéridémie, diabète...) #mais 3028 données manquantes..
```

Epidémiologie

#suppression de la variable Urine

```
rhc$URIN1 <- NULL
```

#Imputation multiple pour les données NA du weight

```
rhc2<-rhc [,!colnames(rhc)%in%c("SADMDTE","DTHDTE","ADLD3P","LSTCTDTE","ROWNAME S","DSCHDTE")]
```

```
summary(rhc2)
```

```
library(mice)
```

```
rhc3 <- mice (rhc2,m=5)
```

```
library(micemd)
```

```
rhc3 <- mice.par(rhc2, m = 5)
```

densityplot(rhc3) # La densité des données imputées pour chaque ensemble de données imputées est indiquée en rose

#tandis que la densité des données observées est indiquée en bleu.

```
rhc4 <- complete(rhc3)
```

```
summary(rhc4)
```

```
hist(rhc4$WTKILO1)
```

```
rhc.copie <- rhc
```

```
rhc$WTKILO1 <- rhc4$WTKILO1
```

```
describe.numeric(rhc4$WTKILO1, num.desc=c("min","max","mean","median","sd","valid.n"))
```

```
summary(rhc4$WTKILO1)
```

#pour les variables DATES

#les variables dates sont mises comme facteur; on va les convertir en format date

##Date d'entrée dans l'étude (SADMDTE)

```
str(rhc$SADMDTE)
```

```
length(which(is.na(rhc$SADMDTE))) #Pas de valeur manquante
```

```
rhc$SADMDTE <- as.Date(rhc$SADMDTE,"%d/%m/%Y")
```

```
summary(rhc$SADMDTE)
```

#Période d'inclusion: entre le 11/06/1989 et le23/01/1994

```
describe.factor(rhc$DEATH)
```

```
length(which(is.na(rhc$DTHDTE))) #2013 valeurs manquantes qui correspondent soit aux non décédés ou perdus de vue =>
```

ces 2013 dates manquantes correspondent exclusivement aux survivants.

##Date de décès (DTHDTE)

```
str(rhc$DTHDTE)
```

```
rhc$DTHDTE <- as.Date(rhc$DTHDTE,"%d/%m/%Y")
```

```
summary(rhc$DTHDTE) #Patients décédés entre le 14/06/1989 et le 31/12/1994
```

```
length(which(is.na(rhc$DTHDTE))) #2013 valeurs manquantes qui correspondent soit aux non décédés ou perdus de vue =>
```

ces 2013 dates manquantes correspondent exclusivement aux survivants.

#Date de dernières Nouvelles (LSTCTDTE)

```
str(rhc$LSTCTDTE)
```

```
rhc$LSTCTDTE <- as.Date(rhc$LSTCTDTE,"%d/%m/%Y")
```

```
length(which(is.na(rhc$LSTCTDTE))) #Pas de valeur manquante
```

```
summary(rhc$LSTCTDTE)
```

```
length(rhc$LSTCTDTE)
```

#dernière nouvelle: 14/08/1994

#dernier décès: 31/12/1994

#quelle est la date de last contact?

```
rhc$LSTCTDTE[which(rhc$DTHDTE=="1994-12-31")]
```

```
rhc$DTHDTE[which(rhc$LSTCTDTE=="1994-08-14")]
```

##Date de sortie de l'hôpital (DSCHDTE)

```
str(rhc$DSCHDTE)
```

```
rhc$DSCHDTE <- as.Date(rhc$DSCHDTE,"%d/%m/%Y")
```

```
length(which(is.na(rhc$DSCHDTE))) #1 valeur manquante
```

```
rhc$ROWNAME[is.na(rhc$DSCHDTE)] #Valeur manquante : observation num. 4683
```

```
summary(rhc$DSCHDTE) #Patients sortant de l'hôpital entre le 14/06/1989 et le 22/05/1994
```

```
length(which(rhc$DSCHDTE>rhc$LSTCTDTE))
```

#1055 sujets ont la date de sortie d'hospitalisation postérieure à la date de dernières nouvelles.

```
length(which(rhc$DTHDTE>rhc$LSTCTDTE))
```

Epidémiologie

#2151 patients dont la date de dernières nouvelles est antérieure à la date de décès. Ce sont des patients qu'on n'a pas pu bien les suivre, mais on a eu leur état mort plus tard. #c'est la date de dernières nouvelles qui pose problème.

#On remplace par la date de décès si patient décédé ou par la date de sortie d'hospitalisation si date de dernières nouvelles antérieure à la date de sortie d'hospitalisation, sinon par des NA

```
length(which(rhc$DTHDTE < rhc$LSTCTDTE | rhc$SADMDTE == rhc$LSTCTDTE | rhc$DTHDTE > rhc$LSTCTDTE |  
rhc$DSCHDTE > rhc$LSTCTDTE)) #2164 dates de dernières nouvelles incohérentes  
rhc$LSTCTDTE[rhc$DTHDTE < rhc$LSTCTDTE | rhc$SADMDTE == rhc$LSTCTDTE | rhc$DTHDTE > rhc$LSTCTDTE |  
rhc$DSCHDTE > rhc$LSTCTDTE] <- NA
```

#Vérification

```
length(which(is.na(rhc$LSTCTDTE))) #On obtient aussi 2164 NA
```

#Les variables des dates : on les renomme pour qu'elles soient plus simple

```
rhc$DATE_ENT <- rhc$SADMDTE  
rhc$SADMDTE <- NULL  
rhc$DATE_SORHOSP <- rhc$DSCHDTE  
rhc$DSCHDTE <- NULL  
rhc$DATE_DC <- rhc$DTHDTE  
rhc$DTHDTE <- NULL  
rhc$DATE_DDN <- rhc$LSTCTDTE  
rhc$LSTCTDTE <- NULL
```

#Création des variables "temps" et "delai"

```
rhc$temps.suivi<-as.numeric(rhc$DATE_DDN-rhc$DATE_ENT)  
rhc$temps.survie<-as.numeric(rhc$DATE_DC-rhc$DATE_ENT)  
rhc$temps.hosp<-as.numeric(rhc$DATE_SORHOSP-rhc$DATE_ENT)  
rhc$delai.sortie.deces<-as.numeric(rhc$DATE_DC-rhc$DATE_SORHOSP)  
rhc$delai.ddn.deces<-as.numeric(rhc$DATE_DC-rhc$DATE_DDN)  
cor(rhc$SCOMA1,rhc$temps.survie,use="complete.obs") #-0.166 une corrélation négative entre Coma score  
et temps de survie  
rhc$DTH30 <- ifelse (rhc$DTH30 == "Yes", 1, 0)  
rhc$DTH30 <- as.factor(rhc$DTH30)  
summary(rhc$DTH30) #1918 patients morts à J30. 3817 vivants à J30  
sum(is.na(rhc))
```

Analyse Bivariée

#Réflexion

```
table(rhc$DTH30,rhc$SWANG1,useNA = "always")  
chisq.test(rhc$DTH30,rhc$SWANG1)
```

```
quantitatives <- rhc[,c("AGE","EDU","SURV2MD1","DAS2D3PC","T3D30","APS1",  
"SCOMA1","PAM","WBLC1","HRT1","RESP1","TEMP1","PAFI1","ALB1","HEMA1",  
"BILI1","CREA1","SOD1","POT1","PACO21","PH1","WTKILO1")]  
nomquanti=colnames(quantitatives)
```

#Chargement des librairies

```
library(dplyr)  
library(tableone)  
vars <- c("AGE","AGE_CAT", "SEX","RACE","EDU","INCOME","ASSURANCE","CAT1","DASI","DNR1",  
"CA","SURV2MD1","APS1","SCOMA1","WTKILO1","TEMP1","PAM","RESP1",  
"HRT1","PAFI1","PACO21","PH1","WBLC1","HEMA1","SOD1","POT1","CREA1",  
"BILI1","ALB1","RESP","CARD","NEURO","GASTR","RENAL","META","HEMA",  
"SEPS","TRAUMA","ORTHO","CARDIOHX","CHFHX","DEMENTHX","PSYCHHX",  
"CHRPULHX","RENALHX","LIVERHX","GIBLEDHX","MALIGHX","IMMUNHX","TRANSHX","AMIHX","DTH30","DEATH",  
"JVIE_J30", "temps.suivi", "temps.survie", "temps.hosp")
```

Epidémiologie

#Distribution des variables entre deux groupes (Swan et Non-Swan)

```
tableaucontingence <- CreateTableOne(vars = vars, strata = 'SWANG1', data = rhc, test=FALSE)
print(tableaucontingence,noSpaces = TRUE)
tableau<- print(tableaucontingence, showAllLevels = TRUE, smd =TRUE)
#Exporter le tableau en csv
write.csv2(tableau, file = "myTable.csv")

#graph
par(mfrow = c(5,5)) for (i in 1:25)
{
boxplot(quantitatives[,i]~rhc$SWANG1,col="#AAC0EC", ylab=nomquantiti[i], xlab="Right Heart Catheterization (RHC)")
}
```

```
vars1<- c("AGE","AGE_CAT", "SEX","RACE","EDU","INCOME","ASSURANCE","CAT1","DASI","DNR1",
"CA","SURV2MD1","APS1","SCOMA1","WTKILO1","TEMP1","PAM","RESP1",
"HRT1","PAFI1","PACO21","PH1","WBLC1","HEMA1","SOD1","POT1","CREA1",
"BILI1","ALB1","RESP","CARD","NEURO","GASTR","RENAL","META","HEMA",
"SEPS","TRAUMA","ORTHO","CARDIOHX","CHFHX","DEMENTHX","PSYCHHX",
"CHRPULHX","RENALHX","LIVERHX","GIBLEDHX","MALIGHX","IMMUNHX", "TRANSHX","AMIHX", "SWANG1")
```

#Distribution des variables selon décès à J30

```
vars1<- c("AGE","AGE_CAT", "SEX","RACE","EDU","INCOME","ASSURANCE","CAT1","DASI","DNR1",
"CA","SURV2MD1","APS1","SCOMA1","WTKILO1","TEMP1","PAM","RESP1",
"HRT1","PAFI1","PACO21","PH1","WBLC1","HEMA1","SOD1","POT1","CREA1",
"BILI1","ALB1","RESP","CARD","NEURO","GASTR","RENAL","META","HEMA",
"SEPS","TRAUMA","ORTHO","CARDIOHX","CHFHX","DEMENTHX","PSYCHHX",
"CHRPULHX","RENALHX","LIVERHX","GIBLEDHX","MALIGHX","IMMUNHX", "TRANSHX","AMIHX", "SWANG1")
tableaucontingence1 <- CreateTableOne(vars = vars1, strata = 'DTH30', data = rhc, test=FALSE)
print(tableaucontingence1,noSpaces = TRUE,)
tableau1<- print(tableaucontingence1, showAllLevels = TRUE, smd =TRUE)
#Exporter le tableau 1 en csv
write.csv2(tableau1, file = "myTable1.csv")
```

#graph boxplot (pour comparaison quanti) par(mfrow = c(5,5))

```
for (i in 1:22)
{
boxplot(quantitatives[,i]~rhc$DTH30,col="#AAC0EC", ylab=nomquantiti[i], xlab="DTH30")
}
```

```
table(rhc$DTH30,rhc$SWANG1,useNA = "always")
library(ggplot2)
ggplot(rhc, aes(x=SWANG1, fill=DTH30, col=DTH30)) +geom_bar(position = "identity" , alpha=0.6)
```

```
library(ggplot2)
ggplot(rhc, aes(x=SWANG1, fill=DTH30, col=DTH30)) +geom_bar(position = "identity" , alpha=0.3)
ggplot(rhc, aes(x=DTH30, fill=SWANG1, color=SWANG1)) +geom_bar(position = "identity" , alpha=0.3)
plot(x=rhc$CAT1,y=rhc$SWANG1,las=1,col=myPalette,xlab="les diagnostics à l'admission (CAT1)")
dev.off()
```

#Création du score de propension

```
ps <- glm(formula = SWANG1 ~ AGE+SEX+EDU+INCOME+ASSURANCE+CAT1+DASI+DNR1+CA+SURV2MD1+APS1+SCOMA1+
WTKILO1+TEMP1+PAM+RESP1+HRT1+PAFI1+PACO21+PH1+WBLC1+HEMA1+SOD1+POT1+CREA1+BILI1+ALB1+RESP+CARD
+NEURO+GASTR+RENAL+META+HEMA+SEPS+TRAUMA+ORTHO+CARDIOHX+CHFHX+DEMENTHX+PSYCHHX+CHRPULHX+RE
NALHX+LIVERHX+ GIBLEDHX+MALIGHX+IMMUNHX+TRANSHX+AMIHX,family = binomial(link="logit"), data = rhc)
summary(ps)
rhc$ps <- predict(ps, type="response")
summary(rhc$ps)
```

Epidémiologie

#Exploration du score

```
par(mfrow=c(1,2))
hist(rhc$ps,col="#00BFC4",xlab="score de propension",ylab="nbr de patients",main="")
boxplot(ps~SWANG1,data=rhc,col=c("#F8766D","#00BFC4"),ylab="score de propension", las=1) #Distribution du score de propension selon les groupes SWANG = 1 et SWANG = 0
```

#Chargement de la librairie

```
library(cobalt)
ggplot(rhc, aes(x=ps, fill=SWANG1)) + geom_density(alpha=.3)
```

#Chargement de la librairie

```
library(pROC)
plot(rocPsModel,legacy.axes=TRUE,ylim=c(0,1))
rocPsModel
```

#Dans R: ps in 3551 controls (SWANG1 No RHC) < 2184 cases (SWANG1 RHC) #Area under the curve: 0.7995

```
par(mfrow=c(1,1))
```

```
summary(rhc$SWANG1)
```

#Première méthode: ajustement sur le score de propension (en "quantitatif", linéaire)

```
mod1 <- glm((DTH30==1)~(SWANG1=="RHC")+ps,data=rhc,family="binomial")
summary(mod1)
```

```
exp(coefficients(mod1))
```

exp(confint(mod1)) #donc après avoir ajuster sur le score de propension, Swan est significativement associée au décès ; OR= 1.287 avec IC [1.14-1.48]

#2ème méthode: stratification= ajustement sur le score de propension en "deciles"

```
decilePoints <- quantile(x=rhc$ps,probs=seq(from=0.1,to=0.9, by=0.1))
```

```
decilePoints <- c(0,decilePoints,1)
```

```
table(decilePoints)
```

```
rhc$psDecile <- cut(x=rhc$ps, breaks=decilePoints,labels=paste0("q",1:10))
```

```
bal.plot(SWANG1~ps,data=rhc,var.name="ps",which="both",subclass="psDecile",which.sub = c("q1","q2","q3"))
```

```
bal.plot(SWANG1~ps,data=rhc,var.name="ps",which="both",subclass="psDecile",which.sub = c("q4","q5","q6","q7"))
```

```
bal.plot(SWANG1~ps,data=rhc,var.name="ps",which="both",subclass="psDecile",which.sub = c("q8","q9","q10"))
```

```
ggplot(rhc, aes(x=psDecile, y=ps, fill=SWANG1)) + geom_boxplot()
```

```
mod2 <- glm((DTH30==1)~(SWANG1=="RHC"))
```

```
as.factor(psDecile),data=rhc, family="binomial")
```

```
summary(mod2)
```

```
exp(coefficients(mod2))
```

```
exp(confint(mod2)) #OR= 1.3077532 [1.1479752 1.4897141]
```

```
?clogit
```

```
library(survival)
```

```
mod3 <- clogit(as.numeric(DTH30)~(SWANG1)+strata(psDecile),data=rhc)
```

```
summary(mod3)
```

```
exp(coefficients(mod3))
```

```
exp(confint(mod3)) #OR= 1.30714 [1.147609 1.488838]
```

#méthode 3: Appariement sur le score de propension

```
library(Matchit)
```

#faut coder ici SWANG1 en numérique 1,0

```
summary(rhc$SWANG1)
```

```
rhc$SWANG1 <- ifelse(rhc$SWANG1=="RHC",1,0)
```

#il faut enlever les colonnes DATES (car il y a des données manquantes destabilisant Matchit)

```
names(rhc)
```

```
DE_COTE <- rhc[,colnames(rhc)%in%c("ROWNAMES","DATE_ADM","DATE_SORHOP","DATE_DC",  
"DATE_DDN","temps.suivi","temps.survie","temps.hosp","delai.sortie.deces","delai.ddn.deces")]
```

```
rhc<-rhc[,!colnames(rhc)%in%c("DATE_ADM","DATE_SORHOSP","DATE_DC","DATE_DDN")]
```

Epidémiologie

```
, "temps.suivi", "temps.survie", "temps.hosp", "delai.sortie.deces", "delai.ddn.deces"]])
m.out<- matchit(SWANG1~AGE+SEX+RACE+EDU+INCOME+ASSURANCE+CAT1+DASI+DNR1+CA+SU
RV2MD1+APS1+SCOMA1+WTKILO1+TEMP1+PAM+RESP1+HRT1+PAFI1+PACO21+PH1+WBLC
1+HEMA1+SOD1+POT1+CREA1+BILI1+ALB1+RESP+CARD+NEURO+GASTR+RENAL+META+H
EMA+SEPS+TRAUMA+ORTHO+CARDIOHX+CHFHX+DEMENTHX+PSYCHHX+CHRPULHX+RENALH
X+LIVERHX+GIBLEDHX+MALIGHX+IMMUNHX+TRANSHX+AMIHX+ps,
data=rhc, method="nearest", caliper = 0.1, ratio=1)
#exploration de l'appariement
s.out <- summary(m.out, standardize=TRUE)
plot(s.out)
summary(m.out, standardize=TRUE) #1612 paires
#Vérification de la propriété équilibrante
balMatch<- bal.tab(m=m.out,
formula=SWANG1~AGE+SEX+RACE+EDU+INCOME+ASSURANCE+CAT1+DASI+DNR1+CA+SURV2MD1+APS1+SCOMA1+WTKI
LO1+TEMP1+PAM+RESP1+HRT1+PAFI1+PACO21+PH1+WBLC1+HEMA1+SOD1+POT1+CREA1+BILI1+ALB1+RESP+CARD+NEU
RO+GASTR+RENAL+META+HEMA+SEPS+TRAUMA+ORTHO+CARDIOHX+CHFHX+DEMENTHX+PSYCHHX+CHRPULHX+RENALH
X+LIVERHX+GIBLEDHX+MALIGHX+IMMUNHX+TRANSHX+AMIHX+ps,un=TRUE,disp.v.ratio=TRUE,s.d.denom="treated",
data=rhc)
love.plot(balMatch,stat="mean.diffs",abs=TRUE, var.order = "unadjusted", threshold = 0.1, shape=23)
bal.plot(m.out,formula=SWANG1~AGE+SEX+RACE+EDU+INCOME+ASSURANCE+CAT1+DASI+DNR1+CA+SURV2MD1+APS1+S
COMA1+WTKILO1+TEMP1+PAM+RESP1+HRT1+PAFI1+PACO21+PH1+WBLC1+HEMA1+SOD1+POT1+CREA1+BILI1+ALB1+RE
SP+CARD+NEURO+GASTR+RENAL+META+HEMA+SEPS+TRAUMA+ORTHO+CARDIOHX+CHFHX+DEMENTHX+PSYCHHX
+CHRPULHX+RENALHX+LIVERHX+GIBLEDHX+MALIGHX+IMMUNHX+TRANSHX+AMIHX,data=rhc,var.name =
"ps",which="both")
#Support commun avant et après appariement
plot(m.out, type="jitter", interactive=FALSE, col="#F8766D")
plot(m.out, type="hist",col="#00BFC4")
rhc_match <- match.data(m.out) #dataset des données appariées
rhc$SWANG1<-ifelse(rhc$SWANG1=="1", "RHC", "No RHC")
rhc_match$SWANG1<-ifelse(rhc_match$SWANG1=="1", "RHC", "No RHC") par(mfrow=c(1,2))
#Distribution du score de propension selon les groupes SWANG = 1 et SWANG = 0
boxplot(ps~SWANG1,data=rhc,col=c("#F8766D", "#00BFC4"),ylab="score de propension", las=1, main="avant
l'appariement") boxplot(ps~SWANG1,data=rhc_match,col=c("#F8766D", "#00BFC4"),ylab="", las=1, main="après
l'appariement")
library(Hmisc,quietly=TRUE)
?histbackback
out<-histbackback(split(rhc$ps,rhc$SWANG1),main="score de propensité avant l'appariement",xlab=c("No RHC", "RHC"))
barplot(out$left, col="#F8766D", horiz=TRUE, space=0, add=TRUE, axes=FALSE)
barplot(out$right, col="#00BFC4", horiz=TRUE, space=0, add=TRUE, axes=FALSE)
out2<-histbackback(split(rhc_match$ps,rhc_match$SWANG1),main="score de propensité après l'appariement",xlab=c("No
RHC", "RHC"))
barplot(-out2$left, col="#F8766D", horiz=TRUE, space=0, add=TRUE, axes=FALSE)
barplot(out2$right, col="#00BFC4", horiz=TRUE, space=0, add=TRUE, axes=FALSE)
TABLEAU3 <- CreateTableOne(vars = vars, strata = "SWANG1", data=rhc_match, test=FALSE)
a3 <- print(TABLEAU3,noSpaces = TRUE,smd=TRUE) a3
addmargins(table(ExtractSmd(TABLEAU3) > 0.1))

#mesure de l'effet: modèle weighted GLM
rhc_match$SWANG1 <- as.factor(rhc_match$SWANG1)
mod4 <- glm(rhc_match$DTH30~rhc_match$SWANG1,family=binomial,weights=rhc_match$weights)
summary(mod4)
exp(coefficients(mod4)) #1.316
exp(confint(mod4)) #1.133 1.515
#lors de l'appariement 1:1, on a perdu des sujets, donc on a perdu de puissance

#Appariement 1 cas 2 témoins
m.out2<-
matchit(SWANG1~AGE+SEX+RACE+EDU+INCOME+ASSURANCE+CAT1+DASI+DNR1+CA+SU
RV2MD1+APS1+SCOMA1+WTKILO1+TEMP1+PAM+RESP1+HRT1+PAFI1+PACO21+PH1+WBLC
```

Epidémiologie

```
1+HEMA1+SOD1+POT1+CREA1+BILI1+ALB1+RESP+CARD+NEURO+GASTR+RENAL+META+H
EMA+SEPS+TRAUMA+ORTHO+CARDIOHX+CHFHX+DEMENTHX+PSYCHHX+CHRPULHX+RENALH
X+LIVERHX+GIBLEDHX+MALIGHX+IMMUNHX+TRANSHX+AMIHX+ps,
data=rhc, method="nearest", caliper = 0.1, ratio=2)
s.out2 <- summary(m.out2, standardize=TRUE)
plot(s.out2)
summary(m.out2, standardize=TRUE)
rhc_match2 <- match.data(m.out2)
summary(rhc_match2)
TABLEAU4 <- CreateTableOne(vars = vars, strata = 'SWANG1', data
=rhc_match2, test=FALSE)
a4 <- print(TABLEAU4, noSpaces = TRUE, smd=TRUE)
addmargins(table(ExtractSmd(TABLEAU4) > 0.1))
summary(rhc_match2$DTH30)
rhc_match2$SWANG1 <- as.factor(rhc_match2$SWANG1)

mod4c <-
glm(DTH30~SWANG1+CAT1+APS1+PAM+CARD, data=rhc_match2, family=binomial, weights=rhc_match2$weights)
summary(mod4c)
exp(coefficients(mod4c)) #1.312
exp(confint(mod4c)) #1.139 1.510

#Méthode 4: pondération #poids IPTW
rhc <- within(rhc, {iptw <- NA ## 1/PS for treated
iptw[SWANG1 == "RHC"] <- 1/ps[SWANG1 == "RHC"]
#1/(1-PS) for untreated (1/probability of not treated)
iptw[SWANG1 != "RHC"] <- (1/(1 - ps))[SWANG1 != "RHC"]})
summary(rhc$iptw)
#Poids w2
rhc$w2 <- with(rhc, SWANG1/rhc$ps + (1 - SWANG1)/(1 - rhc$ps))
summary(rhc$w2)
#Poids pondération
rhc <- mutate(rhc, PONDERATION = case_when(SWANG1 == "1" ~ (1 / ps), SWANG1 == "0" ~ (1 / (1 - ps))))
summary(rhc$PONDERATION)
#w2 et pondération sont équivalents à IPTW avec une légère troncation # Vérification de la propriété équilibrante après
pondération library("WeightIt")

bal.w <- weightit(formula = SWANG1~AGE+SEX+RACE+EDU+INCOME+ASSURANCE+CAT1+DASI+DNR1+CA+SURV2MD1+A
PS1+SCOMA1+WTKILO1+TEMP1+PAM+RESP1+HRT1+PAFI1+PACO21+PH1+WBL1+HEMA1+
SOD1+POT1+CREA1+BILI1+ALB1+RESP+CARD+NEURO+GASTR+RENAL+META+HEMA+SEPS
+TRAUMA+ORTHO+CARDIOHX+CHFHX+DEMENTHX+PSYCHHX+CHRPULHX+RENALHX+LIVERH
X+GIBLEDHX+MALIGHX+IMMUNHX+TRANSHX+AMIHX+ps, data= rhc, method =
"ps", weights="PONDERATION", threshold=0.1, s.denom="treated")

bal.plot.w <- bal.plot(obj = SWANG1~AGE+SEX+RACE+EDU+INCOME+ASSURANCE
+CAT1+DASI+DNR1+CA+SURV2MD1+APS1+SCOMA1+WTKILO1+TEMP1+PAM+RESP1+HRT1+
PAFI1+PACO21+PH1+WBL1+HEMA1+SOD1+POT1+CREA1+BILI1+ALB1+RESP+CARD+NEU
RO+GASTR+RENAL+META+HEMA+SEPS+TRAUMA+ORTHO+CARDIOHX+CHFHX+DEMENTHX+PS
YCHHX+CHRPULHX+RENALHX+LIVERHX+GIBLEDHX+MALIGHX+IMMUNHX+TRANSHX+AMIH
X+ps, data= rhc, var.name="ps", weights = "PONDERATION", method = "weighting", which = "both")

bal.plot.w + xlab ("Score de propension") + ylab ("Densité")

#"SuperLearner" is not an allowable estimand for propensity score weighting with binary treatments.
#Only "ATT", "ATC", "ATE", "ATO", and "ATM" are allowed
bal.plot(bal.w, which = "both", data= rhc, weights = "PONDERATION", method = "weighting")
print(bal.w) summary(bal.w)
bal.tab(bal.w, un=TRUE, threshold=0.1, disp.v.ratio=TRUE, weights="PONDERATION", s.d.denom = "treated")
```

Epidémiologie

#Représentations graphiques

```
love.plot(x= bal.w, abs = TRUE, var.order = "unadjusted", threshold = 0.1, shape = 23)
```

#Modèle avec pondération

```
rhc$DTH30<-as.factor(rhc$DTH30)
rhc$SWANG1 <- as.factor(rhc$SWANG1)
summary(rhc$SWANG1)
WPS <- glm(DTH30~SWANG1, family = binomial, weights=rhc$PONDERATION, data = rhc)
summary(WPS)
exp(coefficients(WPS)) #OR 1,296
exp(confint(WPS)) #IC95% 1,199 - 1,401
```

#Analyses sur la variable binaire 'DEATH' (mort à J180)

##Ajuster sur ps

```
modA<-glm(DEATH~SWANG1+ps, data=rhc, family=binomial)
summary(modA)
exp(coefficients(modA)) #1.315
exp(confint(modA)) #1.154 - 1.499
```

##Stratifier sur ps

```
modB <- glm(DEATH~SWANG1+as.factor(psDecile), data =rhc, family = binomial)
exp(coef(modB)) #1.329
exp(confint(modB)) # 1.1663 - 1.515
```

```
modC <- clogit(as.numeric(DEATH)~SWANG1+strata(psDecile), data = rhc)
exp(coef(modC)) #1.328
exp(confint(modC)) #1.165 - 1.514
```

##Apparier sur ps

```
modD <- glm(rhc_match$DEATH~rhc_match$SWANG1,family=binomial,weights=rhc_matc h$weights)
exp(coefficients(modD)) #1.317361
exp(confint(modD))
```

##Pondérer sur ps

```
WPS2 <- glm(DEATH~SWANG1, weights=rhc$PONDERATION, family="binomial",data = rhc)
summary(WPS2)
exp(coefficients(WPS2)) #1.314
exp(confint(WPS2)) #1.216 1.421
```

##Survie

#Rajouter les dates dans notre jeu de données

```
#DE_COTE <- rhc[,colnames(rhc)%in%c("SADMDTE","DTHDTE","ADLD3P","LSTCTDTE","ROWNAMES ","DSCHDTE")]
rhc <- cbind(rhc,DE_COTE)
names(rhc)
rhc[62]<-NULL #ROWNAMES apparait deux fois
```

##Kaplan Meier décès 30 jours en fonction de RHC, sans ajustement

```
#Création de la variable temps.survie2 en remplaçant les NA par temps.suivi
summary(rhc$temps.survie) #valid.n =3722
rhc$temps.survie2<-ifelse(is.na(rhc$temps.survie),rhc$temps.suivi, rhc$temps.survie)
summary(rhc$temps.survie2)
#Prendre 30 jours comme date de point
rhc$temps.survie3<-ifelse(rhc$temps.survie2>30, 30, rhc$temps.survie2)
summary(rhc$temps.survie3)
```