DEVOIR EPIDEMIOLOGIE : UTILISATION DU SCORE DE PROPENSION

Sarah F FELDMAN

27 février 2017

# INTRODUCTION

Le cathétérisme cardiaque droit (CCD) est un examen invasif utilisé en soins intensif pour mesurer directement la fonction cardiaque, ce qui permettrait selon certains médecins une meilleure prise en charge du patient et donc une augmentation de la survie. Cependant pour montrer une telle causalité il faudrait un essai randomisé. Or nous avons ici les données d'une cohorte prospective, non interventionnelle, de patients hostpitalisés en réanimation, certains ayant été cathétérisés, d'autres non. La cathétérisation n'a pas été randomisée, il y a donc lieu de penser que l'examen a été réalisé préférentiellement chez certains types de patients, potentiellement les patients les plus graves qui ont donc plus de risque de décéder. Cela constitue un biais d'indication qui rend l'analyse de l'efficacité du CCD ininterprêtable. Pour pouvoir analyser l'efficacité du CCD par sonde de swan Ganz, nous allons prendre en compte les biais d'indication propables à l'aide d'un score de propension, nous permettant d'améliorer le niveau de causalité de la relation CCD/Décès si elle existe.

# DESCRIPTION DE LA BASE DE DONNEES

[1] 5735 63

[1] 5735

Base de données de 5735 lignes et 63 colonnes. Ce sont des données transversales avec une ligne par patient (pas de doublon), chaque variable ayant été mesurées une seule fois.

# doublons

## Data management

FALSE   
 5735

FALSE TRUE   
 3719 3

0 1  
 FALSE 2005 1804  
 TRUE 8 1918

0 1   
3809 1918

0 1   
3817 1918

0% 25% 50% 75% 100%   
 2 7 12 22 338

0% 25% 50% 75% 100%   
 2 8 16 31 342

3 patients pour lesquels la date de décès est antérieure à la date de derniere nouvelle --> 2151 pour lesquels date de dernières nouvelles est antérieure au décès. -->

Avant d'utiliser la variable DTH30 dans une régression logistique, je vérifie que tous les sujets non décédés sont bien suivis au moins 30j. Si non, alors ils sont NA pour le décès et non pas 0 car on ne tient pas compte du temps. 8 sujets sont perdus de vu avant 30 jours et non décédés, ils sont donc NA pour le décès et non pas 0 comme noté dans la variable DTH30 (ils sont cependant censor = 0 pour une analyse de survie)

## Je réalise une analyse descriptive préliminaire afin de repérer les incohérences

### Présentation des variables de l'examen clinique et biologique:

Activities of daily living (ADL) : échelle d'autonomie de 0 à 12 (score>6 signe une dépendance)  
Duke Activity Status Index (DASI) : auto-questionnaire de 12 items mesurant l'activité fonctionnelle. Le score va de 0 à 58.2, plus le score est élevé, meilleur est l'activité fonctionnelle. DNR pour Do not resuscitate : 1 pour une interdiction de réanimation cardiopulmonaire, 0 sinon.  
Probabilité de survie à 2 mois, estimée par model : de 0 à 1  
APACHE III Acute Physiology scores : score de prediction du risque de mortalité de patients hospitalisés en soins intensifs. Plus le score est élevé plus le risque de mortalité est important. LE score va de 0 à 299 mais ici seule la partie physiologie du score est utilisée.  
Le score de Glasgow évalue l'état de conscience du patient: un score de 3 équivaut à un coma profond, un score de 15 est un état de conscience normal.  
Le rapport PAO2/FIO2 permet de diagnostiquer une agression pulmonaire aigue(rapport>300), un rapport <200 definissant le syndrome de detresse respiratoire aigue et l'ECMO est envisageable en cas de rapport <50. Un pH normal varie entre 7,38 et 7.42, le taux d'hématocrite normal est comprise entre 41 et 50% environ, La PaCO2 normale varie de 35 à 45mmHg.  
Le taux d'albumine normal varie entre 25 et 44 g/L.

## Données manquantes

0 1 2 3 4   
 269 2171 2711 583 1

ROWNAMES CAT1 CAT2 CA SADMDTE DSCHDTE DTHDTE LSTCTDTE   
 0 0 0 0 0 1 2013 0   
 DEATH CARDIOHX CHFHX DEMENTHX PSYCHHX CHRPULHX RENALHX LIVERHX   
 0 0 0 0 0 0 0 0   
GIBLEDHX MALIGHX IMMUNHX TRANSHX AMIHX AGE SEX EDU   
 0 0 0 0 0 0 0 0   
SURV2MD1 DAS2D3PC T3D30 DTH30 APS1 SCOMA1 MEANBP1 WBLC1   
 0 0 0 0 0 0 0 0   
 HRT1 RESP1 TEMP1 PAFI1 ALB1 HEMA1 BILI1 CREA1   
 0 0 0 0 0 0 0 0   
 SOD1 POT1 PACO21 PH1 SWANG1 WTKILO1 DNR1 NINSCLAS   
 0 0 0 0 0 0 0 0   
 RESP CARD NEURO GASTR RENAL META HEMA SEPS   
 0 0 0 0 0 0 0 0   
 TRAUMA ORTHO ADLD3P URIN1 RACE INCOME PTID SWAN   
 0 0 4296 3028 0 0 0 0   
 SWANT time DTH30b timeb   
 0 0 8 0

0 1 8 2013 3028 4296   
 63 1 1 1 1 1

FALSE TRUE   
0.2509154 0.7490846

FALSE TRUE   
0.4720139 0.5279861

ADLD3P et URIN1 ont respectivivement 75% et 53% de données manquantes tandis que les autres variables n'ont aucune ou moins de 5% de données manquantes. Je ne prendrai donc pas en compte ces deux variables dans le calcul du score de propension et je laisse les autres variables telles quelles pour le moment. Il n'y a pas de doublon.

## Pas de doublons sujets

[1] ROWNAMES CAT1 CAT2 CA SADMDTE DSCHDTE DTHDTE   
 [8] LSTCTDTE DEATH CARDIOHX CHFHX DEMENTHX PSYCHHX CHRPULHX  
[15] RENALHX LIVERHX GIBLEDHX MALIGHX IMMUNHX TRANSHX AMIHX   
[22] AGE SEX EDU SURV2MD1 DAS2D3PC T3D30 DTH30   
[29] APS1 SCOMA1 MEANBP1 WBLC1 HRT1 RESP1 TEMP1   
[36] PAFI1 ALB1 HEMA1 BILI1 CREA1 SOD1 POT1   
[43] PACO21 PH1 SWANG1 WTKILO1 DNR1 NINSCLAS RESP   
[50] CARD NEURO GASTR RENAL META HEMA SEPS   
[57] TRAUMA ORTHO ADLD3P URIN1 RACE INCOME PTID   
[64] SWAN SWANT time DTH30b timeb   
<0 rows> (or 0-length row.names)

## valeurs aberrantes

Pour détecter les valeurs aberrantes je dispose de plusieurs méthodes. Je regarde le tableau descriptif en calculant pour chaque variable quantitative la moyenne, l'intervalle interquartile et surtout le range (voir annexe). Je croise également les variables fréquence cardiaque, pression artérielle et fréquence respiratoire à la recherche d'incohérence. Enfin je regarde la distribution des variables ce qui me permet de réperérer éventuellement d'autres incohérences. (voir annexe) Par exemple nous voyons un pic à 0 pour la variable WTKILO1 (nous l'avions déjà vu grâce au tableau descriptif, mais c'est un deuxième filet de sécurité).

0|0|0 0|0|TA0 0|FR0|0 0|FR0|TA0 FC0|0|0 FC0|0|TA0   
 5509 15 47 5 60 15   
 FC0|FR0|0 FC0|FR0|TA0   
 39 45

HRT1 RESP1 MEANBP1 DTH30 SCOMA1 TEMP1  
29 0 0 44 1 100 34.69531  
313 0 0 47 1 61 39.50000  
333 0 0 43 1 0 39.59375  
383 0 0 38 0 89 38.19531  
438 0 0 46 1 100 34.89844  
511 0 0 125 0 100 36.00000  
519 0 0 49 1 100 35.50000  
557 0 0 47 0 0 39.09375  
791 0 0 56 1 100 34.89844  
885 0 0 47 1 0 34.89844  
1207 0 0 116 0 41 38.79688  
1213 0 0 61 1 0 39.39844  
1232 0 0 29 1 26 38.79688  
1372 0 0 46 1 100 38.69531  
1430 0 0 41 1 100 38.89844  
1622 0 0 47 1 0 39.19531  
1708 0 0 129 0 0 40.09375  
1761 0 0 38 1 100 38.89844  
1809 0 0 45 1 100 37.00000  
1855 0 0 69 1 0 35.69531  
1871 0 0 63 0 94 38.79688  
1926 0 0 147 1 100 38.89844  
2710 0 0 146 0 0 39.29688  
2983 0 0 27 0 0 39.09375  
2984 0 0 51 1 100 36.00000  
3099 0 0 117 0 89 36.39844  
3318 0 0 47 1 100 41.29688  
3493 0 0 50 0 44 38.69531  
3679 0 0 162 1 55 40.69531  
3879 0 0 33 1 94 38.59375  
4239 0 0 118 1 9 39.69531  
4694 0 0 53 1 100 36.09375  
4782 0 0 210 1 100 39.79688  
5122 0 0 53 0 100 34.79688  
5202 0 0 63 1 100 35.29688  
5277 0 0 40 0 100 42.00000  
5415 0 0 33 0 100 34.89844  
5421 0 0 40 1 41 39.59375  
5431 0 0 55 0 0 35.00000

Les scores qui ont été mesurés, semblent tous avoir des résultats plausibles excepté le score de Glasgow qui a été modifié : d'un score allant de 3 à 15 on passe à un score allant de 0 à 100. Le score de 100 est probablement lié à un coma sévère car les patients avec un score plus élevé ont un risque de décès augmenté (alors que dans l'échelle d'origine, 3 correspond au coma sévère, et 15 à une conscience normale) (voir courbe ci desous).  
Je laisse la variable telle quelle car elle a de toute évidence été modifiée, je ne veux pas la remodifier une seconde fois.  
La température corporelle peut aller de 27° en cas d'hypothermie profonde à 42-43° en cas de forte fièvre, donc je ne modifie pas.  
Un patient ne peut pas avoir un poids de 0, ces 515 patients sont recodés en NA pour le poids.  
Un patient peut être en aplasie mais je ne suis pas sûr qu'une hyperleucocytose de plus de 40 leucocytes/10^9L soit possible, cependant dans le doute concernant le taux maximum en cas de leucémie et au vue de la distribution de la variable qui semble plausible, je laisse telle quelle.  
Hematocrite : une anémie chronique peut entrainer une anémie profonde avec une hématocrite de 10%, en dessous de cette valeur je mets NA. (un taux de 66% est possible). Un taux de PAcO2 de 156 mmHg me semble complètement impossible, au delà de 100 mmHg je met NA. Un pH de 6.6 est extrêmement faible et n'est pas viable mais je ne suis pas sûre qu'on ne puisse pas l'observer en réanimation.  
Une créatininémie de 2200umol/L ou 25 mg/dL est possible en cas d'insuffisance rénale terminale. Un volume urinaire de 0 et de 9000mL est possible également.  
Une valeur normale d'albumine se situe entre 3.5 et 5 g/dL. Les patients avec un taux d'albumine supérieur à 10g/dL auront une valeur NA pour l'albumine.

Les valeurs de natrémie, de kaliémie, de fréquence respiratoire, fréquence cardiaque, tension artéirelle et bilirubine semblent plausible

<http://www.srlf.org/wp-content/uploads/2015/11/0505-Reanimation-Vol14-N3-p177_185.pdf>

En observant ensemble la fréquence cardiaque, la fréquence respiratoire et la tension artérielle, je note certaines incohérences : parfois 2 de ces variables peuvent valoir 0 mais pas la troisième, et parfois une seule de ces variables vaut 0. Ainsi 1 patient à une tension artérielle à 0 mais une fréquence cardiaque et une fréquence respiratoire différente de 0, 47 patients ont une fréquence respiratoire à 0 mais une fréquence cardiaque et une tension artérielle différente de 0 et 3 patients ont une fréquence cardiaque à 0 mais une fréquence respiratoire et une tension artérielle différente de 0.

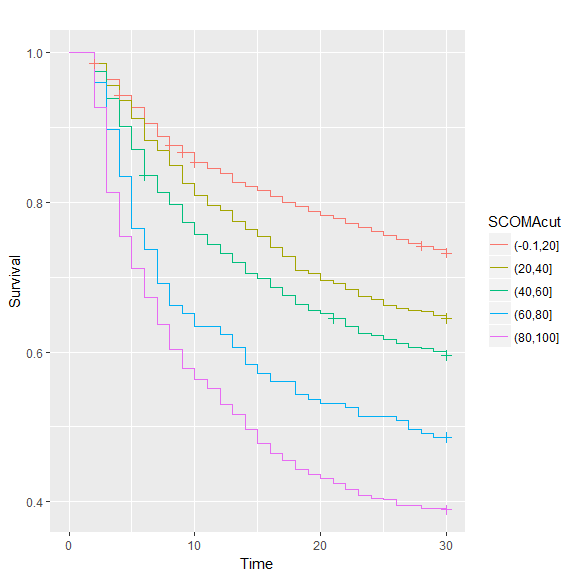
Je modifie donc ces 3 variables selon les règles suivantes : - Si deux de ces variables valait 0, alors la valeur de 0 était systématiquement attribuée à la troisième.  
- Si une seule de ces variables valait 0 alors elle était systématiquement transformée en valeur manquante. Recodage du poids : si le poids est inférieur à 25 je remplace par une valeure manquante.

0 1 2 3 4 5   
 250 1930 2694 796 64 1

ROWNAMES CAT1 CAT2 CA SADMDTE DSCHDTE DTHDTE LSTCTDTE   
 0 0 0 0 0 1 2013 0   
 DEATH CARDIOHX CHFHX DEMENTHX PSYCHHX CHRPULHX RENALHX LIVERHX   
 0 0 0 0 0 0 0 0   
GIBLEDHX MALIGHX IMMUNHX TRANSHX AMIHX AGE SEX EDU   
 0 0 0 0 0 0 0 0   
SURV2MD1 DAS2D3PC T3D30 DTH30 APS1 SCOMA1 MEANBP1 WBLC1   
 0 0 0 0 0 0 15 0   
 HRT1 RESP1 TEMP1 PAFI1 ALB1 HEMA1 BILI1 CREA1   
 60 0 0 0 2 8 0 0   
 SOD1 POT1 PACO21 PH1 SWANG1 WTKILO1 DNR1 NINSCLAS   
 0 0 21 0 0 515 0 0   
 RESP CARD NEURO GASTR RENAL META HEMA SEPS   
 0 0 0 0 0 0 0 0   
 TRAUMA ORTHO ADLD3P URIN1 RACE INCOME PTID SWAN   
 0 0 4296 3028 0 0 0 0   
 SWANT time DTH30b timeb   
 0 0 8 0

0 1 2 8 15 21 60 515 2013 3028 4296   
 57 1 1 2 1 1 1 1 1 1 1

# redecrire nombre de données manquantes!!

 Fig X. Décès en fonction du score de glasgow découpé en quintile.

# Possible erreur des variables T3D30 et DTH30

8 sujets sont notés comme ayant une durée de suivi de 30 jours, or ils sont en fait perdus de vu avant 30 jours. J'ai donc recré une variable durée de suivi, time, qui est la différence entre entre la date des dernières nouvelles (décès ou date de dernier contact) et la date d'admission, et j'ai ensuite crée la variable timeb telle que si cette durée était supérieure à 30, elle prenait alors la valeur 30. en comparant la variable timeb à la variable T3D30, je vois que les variables sont égales 2 à 2 pour chaque sujets exceptés les 8 sujets dont j'ai parlé ci dessus. J'ai donc remplacé la valeur de T3D30 pour ces sujets là par la valeure de timeb. Ces 8 sujets sont notés comme n'ayant pas eu d'évènement à 30 jours. Cela ne pose pas de problème si l'on fait un modèle de Cox qui prend en compte les perdus de vue, mais je souhaitais faire une régression logistique conditionnelle et dans ce cas je ne peux pas dire que les 8 sujets perdus de vue avant 30 jours ne sont pas mort à 30 jours et je choisi donc de mettre une donnée manquante pour ces 8 valeurs. Il s'agit peut être d'erreurs de saisie et je ne peux pas trancher. Dans le doute je prendrai donc les variables T3D30 et DTH30 sans les modifier, en supposant que la date d'admission ou la date des dernières nouvelles est erronée.

# Description des variables

Je décris les variables dans le groupe des patients cathétérisés, dans le groupe des patients non cathétérisés et les deux groupes confondus.

Variables quantitatives: - continues : - discrètes : - binaires variables qualitatives

Pause dans les inclusions

# Imputation des valeurs manquantes

Je décide de ne garder que les variables avec moins de 20% de données manquantes et les sujets avec moins de 50% de données manquantes (excepté la variable date de décès bien sûr). Je ne tiens compte que des variables explicatives (je ne tiens pas compte du décès ni du traitement, qui n'est jamais NA de toutes façons, ni des dates d'entrée dans l'étude ou de sortie de l'hopital). J'imputerai les données manquantes restantes. Pour rappel certaines données sont notées comme manquantes car aberrantes.

Je vais donc dans un premier temps observer les données manquantes par variable et par sujet.

ADLD3P URIN1   
74.90846 52.79861

napat1  
 0 1 2 3 4   
 581 2565 2390 196 3

napat  
 0 1.47058823529412 2.94117647058824 4.41176470588235   
 250 1930 2694 796   
5.88235294117647 7.35294117647059   
 64 1

Le score d'autonomie ADL a environ 75% de données manquantes et la variable volume d'excrétion urinaire a plus de 50% de données manquantes. Je supprime donc ces deux variables de l'analyse. 250 patient patients n'ont aucune donnée manquante, 5485(95.6%) patients ont au moins 1 donnée manquante dont : 1930 patients avec 1 donnée manquante, 2694 patients avec 2 de données manquantes, 796 patients avec 3 données manquantes, 64 patients avec 4 données manquantes et 1 patient avec 5 données manquantes. Tous les patients ont donc moins de 50% de données manquantes. Je les garde tous dans l'analyse.

Je note également qu'il n'y a pas de valeurs manquantes concernant la variable d'intérêt Swan ganz in concernant l'outcome décès à 30 jours.

MEANBP1 HRT1 ALB1 HEMA1 PACO21 WTKILO1   
0.26155187 1.04620750 0.03487358 0.13949433 0.36617262 8.97994769

napat1  
 0 1 2   
5122 605 8

napat  
 0 1.66666666666667 3.33333333333333   
 5122 605 8

En supprimant ces 2 colonnes de l'analyse, j'ai 2397 patients avec au moins une donnée manquante, soit plus de 5 % de patients. Je ne peux donc pas simplement supprimer les patients avec des données manquantes. 6 variables ont au moins une donnée manquante : la tension artérielle (15 NA), la fréquence cardiaque (60 NA), l'albuminémie (2 NA) l'hématocrite (8 NA), la PaCO2 (21 NA) et le poids (515 NA).

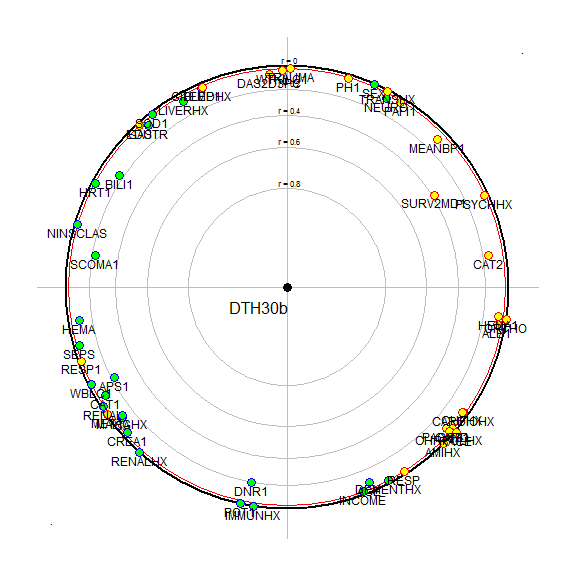
Je réalise ensuite une imputation multiple avec le package mice en faisant l'hypothèse que les données manquantes le sont aléatoirement. Je ne réalise qu'un seul jeu de de données imputé.

iter imp variable  
 1 1 MEANBP1 HRT1 ALB1 HEMA1 PACO21 WTKILO1 DTH30b  
 2 1 MEANBP1 HRT1 ALB1 HEMA1 PACO21 WTKILO1 DTH30b  
 3 1 MEANBP1 HRT1 ALB1 HEMA1 PACO21 WTKILO1 DTH30b  
 4 1 MEANBP1 HRT1 ALB1 HEMA1 PACO21 WTKILO1 DTH30b  
 5 1 MEANBP1 HRT1 ALB1 HEMA1 PACO21 WTKILO1 DTH30b

# Description après imputation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | des\_noSWAN[, 1] | des\_SWAN[, 1] | valeur | missing values |
| SEX\_Female | 1637 (46.1%) | 906 (41.5%) | 2543 (44.3%) |  |
| SEX\_Male | 1914 (53.9%) | 1278 (58.5%) | 3192 (55.7%) | 0 (0%) |
| RACE\_black | 585 (16.5%) | 335 (15.3%) | 920 (16%) |  |
| RACE\_other | 213 (6%) | 142 (6.5%) | 355 (6.2%) | 0 (0%) |
| RACE\_white | 2753 (77.5%) | 1707 (78.2%) | 4460 (77.8%) |  |
| INCOME\_$11-$25k | 713 (20.1%) | 452 (20.7%) | 1165 (20.3%) |  |
| INCOME\_$25-$50k | 500 (14.1%) | 393 (18%) | 893 (15.6%) | 0 (0%) |
| INCOME\_> $50k | 257 (7.2%) | 194 (8.9%) | 451 (7.9%) |  |
| INCOME\_Under $11k | 2081 (58.6%) | 1145 (52.4%) | 3226 (56.3%) |  |
| NINSCLAS\_Medicaid | 454 (12.8%) | 193 (8.8%) | 647 (11.3%) |  |
| NINSCLAS\_Medicare | 947 (26.7%) | 511 (23.4%) | 1458 (25.4%) | 0 (0%) |
| NINSCLAS\_Medicare & Medicaid | 251 (7.1%) | 123 (5.6%) | 374 (6.5%) |  |
| NINSCLAS\_No insurance | 186 (5.2%) | 136 (6.2%) | 322 (5.6%) |  |
| NINSCLAS\_Private | 967 (27.2%) | 731 (33.5%) | 1698 (29.6%) |  |
| NINSCLAS\_Private & Medicare | 746 (21%) | 490 (22.4%) | 1236 (21.6%) |  |
| AGE\* | 64.59 (50.08-74.97) | 63.5 (50.21-72.65) | 64.05 (50.15-73.93) | 0 (0%) |
| EDU\* | 12 (10-13) | 12 (10-14) | 12 (10-13) | 0 (0%) |
| CAT1\_ARF | 1581 (44.5%) | 909 (41.6%) | 2490 (43.4%) |  |
| CAT1\_CHF | 247 (7%) | 209 (9.6%) | 456 (8%) | 0 (0%) |
| CAT1\_Cirrhosis | 175 (4.9%) | 49 (2.2%) | 224 (3.9%) |  |
| CAT1\_Colon Cancer | 6 (0.2%) | 1 (0%) | 7 (0.1%) |  |
| CAT1\_Coma | 341 (9.6%) | 95 (4.3%) | 436 (7.6%) |  |
| CAT1\_COPD | 399 (11.2%) | 58 (2.7%) | 457 (8%) |  |
| CAT1\_Lung Cancer | 34 (1%) | 5 (0.2%) | 39 (0.7%) |  |
| CAT1\_MOSF w/Malignancy | 241 (6.8%) | 158 (7.2%) | 399 (7%) |  |
| CAT1\_MOSF w/Sepsis | 527 (14.8%) | 700 (32.1%) | 1227 (21.4%) |  |
| CAT2\_Cirrhosis | 27 (0.8%) | 11 (0.5%) | 38 (0.7%) |  |
| CAT2\_Colon Cancer | 1 (0%) | 1 (0%) | 2 (0%) | 0 (0%) |
| CAT2\_Coma | 70 (2%) | 20 (0.9%) | 90 (1.6%) |  |
| CAT2\_Lung Cancer | 13 (0.4%) | 2 (0.1%) | 15 (0.3%) |  |
| CAT2\_MOSF w/Malignancy | 171 (4.8%) | 58 (2.7%) | 229 (4%) |  |
| CAT2\_MOSF w/Sepsis | 406 (11.4%) | 420 (19.2%) | 826 (14.4%) |  |
| CAT2\_NoCAT2 | 2863 (80.6%) | 1672 (76.6%) | 4535 (79.1%) |  |
| CA\_No | 2652 (74.7%) | 1727 (79.1%) | 4379 (76.4%) |  |
| CA\_Metastatic | 261 (7.4%) | 123 (5.6%) | 384 (6.7%) | 0 (0%) |
| CA\_Localized | 638 (18%) | 334 (15.3%) | 972 (16.9%) |  |
| CARDIOHX\_1 | 567 (16%) | 446 (20.4%) | 1013 (17.7%) | 0 (0%) |
| CHFHX\_1 | 596 (16.8%) | 425 (19.5%) | 1021 (17.8%) | 0 (0%) |
| DEMENTHX\_1 | 413 (11.6%) | 151 (6.9%) | 564 (9.8%) | 0 (0%) |
| PSYCHHX\_1 | 286 (8.1%) | 100 (4.6%) | 386 (6.7%) | 0 (0%) |
| CHRPULHX\_1 | 774 (21.8%) | 315 (14.4%) | 1089 (19%) | 0 (0%) |
| RENALHX\_1 | 149 (4.2%) | 106 (4.9%) | 255 (4.4%) | 0 (0%) |
| LIVERHX\_1 | 265 (7.5%) | 136 (6.2%) | 401 (7%) | 0 (0%) |
| GIBLEDHX\_1 | 131 (3.7%) | 54 (2.5%) | 185 (3.2%) | 0 (0%) |
| MALIGHX\_1 | 872 (24.6%) | 444 (20.3%) | 1316 (22.9%) | 0 (0%) |
| IMMUNHX\_1 | 907 (25.5%) | 636 (29.1%) | 1543 (26.9%) | 0 (0%) |
| TRANSHX\_1 | 335 (9.4%) | 327 (15%) | 662 (11.5%) | 0 (0%) |
| AMIHX\_1 | 105 (3%) | 95 (4.3%) | 200 (3.5%) | 0 (0%) |
| DAS2D3PC\* | 19.66 (15.73-23.46) | 19.92 (16.71-23.36) | 19.75 (16.06-23.43) | 0 (0%) |
| DNR1\_1 | 499 (14.1%) | 155 (7.1%) | 654 (11.4%) | 0 (0%) |
| SURV2MD1\* | 0.64 (0.49-0.76) | 0.6 (0.45-0.72) | 0.63 (0.47-0.74) | 0 (0%) |
| APS1\* | 50 (38-62) | 60 (47-74) | 54 (41-67) | 0 (0%) |
| SCOMA1\* | 0 (0-41) | 0 (0-37) | 0 (0-41) | 0 (0%) |
| WTKILO1\* | 70 (58.2-83) | 75.5 (64.27-87.92) | 72.4 (60.35-85.2) | 0 (0%) |
| TEMP1\* | 38.09 (36.2-39) | 38.09 (36.09-39) | 38.09 (36.09-39) | 0 (0%) |
| MEANBP1\* | 68 (52-119) | 57 (47-72.25) | 63 (49-114) | 0 (0%) |
| RESP1\* | 30 (20-39) | 28 (12-37) | 30 (14-38) | 0 (0%) |
| HRT1\* | 121 (79-140) | 125 (106-145) | 124 (100-142) | 0 (0%) |
| PAFI1\* | 224 (148.8-333.31) | 168.44 (110-266.62) | 202.5 (133.31-316.62) | 0 (0%) |
| PACO21\* | 38 (32-44) | 36 (30-40) | 37 (31-42) | 0 (0%) |
| PH1\* | 7.4 (7.35-7.46) | 7.4 (7.32-7.46) | 7.4 (7.34-7.46) | 0 (0%) |
| WBLC1\* | 13.6 (8.2-19.4) | 14.7 (8.6-21.2) | 14.1 (8.4-20.05) | 0 (0%) |
| HEMA1\* | 31 (26.6-39) | 29 (26-33.4) | 30 (26.1-36.3) | 0 (0%) |
| SOD1\* | 136 (133-142) | 136 (132-141) | 136 (132-142) | 0 (0%) |
| POT1\* | 3.8 (3.4-4.6) | 3.8 (3.4-4.6) | 3.8 (3.4-4.6) | 0 (0%) |
| CREA1\* | 1.3 (0.9-2) | 1.8 (1.2-3) | 1.5 (1-2.4) | 0 (0%) |
| BILI1\* | 1.01 (0.7-1.2) | 1.01 (1-1.7) | 1.01 (0.8-1.4) | 0 (0%) |
| ALB1\* | 3.5 (2.7-3.5) | 3.5 (2.4-3.5) | 3.5 (2.6-3.5) | 0 (0%) |
| RESP\_1 | 1481 (41.7%) | 632 (28.9%) | 2113 (36.8%) | 0 (0%) |
| CARD\_1 | 1007 (28.4%) | 924 (42.3%) | 1931 (33.7%) | 0 (0%) |
| NEURO\_1 | 575 (16.2%) | 118 (5.4%) | 693 (12.1%) | 0 (0%) |
| GASTR\_1 | 522 (14.7%) | 420 (19.2%) | 942 (16.4%) | 0 (0%) |
| RENAL\_1 | 147 (4.1%) | 148 (6.8%) | 295 (5.1%) | 0 (0%) |
| META\_1 | 172 (4.8%) | 93 (4.3%) | 265 (4.6%) | 0 (0%) |
| HEMA\_1 | 239 (6.7%) | 115 (5.3%) | 354 (6.2%) | 0 (0%) |
| SEPS\_1 | 515 (14.5%) | 516 (23.6%) | 1031 (18%) | 0 (0%) |
| TRAUMA\_1 | 18 (0.5%) | 34 (1.6%) | 52 (0.9%) | 0 (0%) |
| ORTHO\_1 | 3 (0.1%) | 4 (0.2%) | 7 (0.1%) | 0 (0%) |

# Exploratoire



# Tests bivariés

Je dois intégrer dans le score de propension les variables qui sont associées significativement avec le décès à 30 jours ou avec le décès à 30 jours et le traitement par cathétérisme. Pour cela je réalise deux séries de tests bivariés. Une première série testant l'association entre chaque variable et le décès à 30 jours et une deuxième série testant l'association entre chaque variable et le traitement par cathétérisme. Pour rappel, les variables ADLD3P et URIN1 ne sont pas dans l'analyse. J'utilise un modèle de régression logistique, la variable à expliquer étant soit la variable cathétérisme, soit la variable décès à 30 jours (variables toutes deux binaires) et la variable explicative étant chacune des variables à tester.

Les conditions de validité sont toujours respectées car j'ai 9 classes au maximum (variable CAT1), donc la condition des 5 à 10 variables par variable est toujours respectée.

coef pvalue SWAN significatif SWAN coef pvalue DEATH  
RESP 0.000 \* 0.002  
CARD 0.000 \* 0.000  
NEURO 0.000 \* 0.000  
GASTR 0.000 \* 0.000  
HEMA 0.026 \* 0.000  
SEPS 0.000 \* 0.000  
CAT1 0.000 \* 0.000  
CAT2 0.000 \* 0.000  
CA 0.000 \* 0.000  
DAS2D3PC 0.023 \* 0.000  
DNR1 0.000 \* 0.000  
SURV2MD1 0.000 \* 0.000  
APS1 0.000 \* 0.000  
SCOMA1 0.000 \* 0.000  
WTKILO1 0.000 \* 0.002  
TEMP1 0.429 0.011  
MEANBP1 0.000 \* 0.000  
PAFI1 0.000 \* 0.001  
PACO21 0.000 \* 0.000  
PH1 0.000 \* 0.015  
WBLC1 0.002 \* 0.020  
HEMA1 0.000 \* 0.000  
CREA1 0.000 \* 0.000  
BILI1 0.000 \* 0.000  
ALB1 0.000 \* 0.000  
CARDIOHX 0.000 \* 0.000  
CHFHX 0.010 \* 0.000  
DEMENTHX 0.000 \* 0.001  
PSYCHHX 0.000 \* 0.001  
CHRPULHX 0.000 \* 0.000  
LIVERHX 0.075 0.000  
GIBLEDHX 0.012 \* 0.007  
MALIGHX 0.000 \* 0.000  
AGE 0.026 \* 0.000  
INCOME 0.000 \* 0.000  
NINSCLAS 0.000 \* 0.000  
 significatif DEATH select  
RESP \* 1  
CARD \* 1  
NEURO \* 1  
GASTR \* 1  
HEMA \* 1  
SEPS \* 1  
CAT1 \* 1  
CAT2 \* 1  
CA \* 1  
DAS2D3PC \* 1  
DNR1 \* 1  
SURV2MD1 \* 1  
APS1 \* 1  
SCOMA1 \* 1  
WTKILO1 \* 1  
TEMP1 \* 1  
MEANBP1 \* 1  
PAFI1 \* 1  
PACO21 \* 1  
PH1 \* 1  
WBLC1 \* 1  
HEMA1 \* 1  
CREA1 \* 1  
BILI1 \* 1  
ALB1 \* 1  
CARDIOHX \* 1  
CHFHX \* 1  
DEMENTHX \* 1  
PSYCHHX \* 1  
CHRPULHX \* 1  
LIVERHX \* 1  
GIBLEDHX \* 1  
MALIGHX \* 1  
AGE \* 1  
INCOME \* 1  
NINSCLAS \* 1

[1] 36

c("RESP", "CARD", "NEURO", "GASTR", "HEMA", "SEPS", "CAT1", "CAT2",   
"CA", "DAS2D3PC", "DNR1", "SURV2MD1", "APS1", "SCOMA1", "WTKILO1",   
"TEMP1", "MEANBP1", "PAFI1", "PACO21", "PH1", "WBLC1", "HEMA1",   
"CREA1", "BILI1", "ALB1", "CARDIOHX", "CHFHX", "DEMENTHX", "PSYCHHX",   
"CHRPULHX", "LIVERHX", "GIBLEDHX", "MALIGHX", "AGE", "INCOME",   
"NINSCLAS")

|  |
| --- |
| RESP |
| CARD |
| NEURO |
| GASTR |
| HEMA |
| SEPS |
| CAT1 |
| CAT2 |
| CA |
| DAS2D3PC |
| DNR1 |
| SURV2MD1 |
| APS1 |
| SCOMA1 |
| WTKILO1 |
| TEMP1 |
| MEANBP1 |
| PAFI1 |
| PACO21 |
| PH1 |
| WBLC1 |
| HEMA1 |
| CREA1 |
| BILI1 |
| ALB1 |
| CARDIOHX |
| CHFHX |
| DEMENTHX |
| PSYCHHX |
| CHRPULHX |
| LIVERHX |
| GIBLEDHX |
| MALIGHX |
| AGE |
| INCOME |
| NINSCLAS |

[1] "RESP" "CARD" "NEURO" "GASTR" "HEMA" "SEPS"   
 [7] "CAT1" "CAT2" "CA" "DAS2D3PC" "DNR1" "SURV2MD1"  
[13] "APS1" "SCOMA1" "WTKILO1" "MEANBP1" "PAFI1" "PACO21"   
[19] "PH1" "WBLC1" "HEMA1" "CREA1" "BILI1" "ALB1"   
[25] "CARDIOHX" "CHFHX" "DEMENTHX" "PSYCHHX" "CHRPULHX" "GIBLEDHX"  
[31] "MALIGHX" "AGE" "INCOME" "NINSCLAS"

[1] "TEMP1" "LIVERHX"

[1] "RENAL" "TRAUMA" "RESP1" "HRT1" "SOD1" "IMMUNHX" "TRANSHX"  
 [8] "AMIHX" "SEX" "EDU"

2 variables sont liées au décès uniquement : "TEMP1" "LIVERHX"

34 variables sont liées au décès et au traitement par cathétérisme : "RESP" "CARD" "NEURO" "GASTR" "HEMA" "SEPS" "CAT1" "CAT2" "CA"  
"DAS2D3PC" "DNR1" "SURV2MD1" "APS1" "SCOMA1" "WTKILO1" "MEANBP1" "PAFI1" "PACO21"  
"PH1" "WBLC1" "HEMA1" "CREA1" "BILI1" "ALB1" "CARDIOHX" "CHFHX" "DEMENTHX" "PSYCHHX" "CHRPULHX" "GIBLEDHX" "MALIGHX" "AGE" "INCOME" "NINSCLAS"

10 variables sont liées uniquement au cathétérisme : "RENAL" "TRAUMA" "RESP1" "HRT1" "SOD1" "IMMUNHX" "TRANSHX" "AMIHX" "SEX" "EDU"

# Score de Propension

Afin d'équilibrer les groupes, je vais calculer un score de propension qui est la propabilité pour chaque patient d'être traité par la sonde de swan ganz. Pour cela je crée un modèle linéaire avec comme outcome le traitement par swan ganz et comme variables explicatives toutes les variables liées au décès à 30j uniquement(facteurs pronostiques) ou au décès et au traitement(facteurs de confusion). Je n'intègre pas les variables liées uniquement au traitement pour ne pas perdre de puissance à la fin.

Condition de validité : C'est une régression logistique, je dois avoir 5 à 10 évènement par variable. J'ai 36 variables explicatives dans le score de propension dont 5 variables qualitatives : 9 classes pour CAT1, 7 classes pour CAT2, 3 classes pour CA, 4 classes pour INCOME, 6 classes pour NINSCLAS. Ces variables qualitatives seront donc transformées en 9+7+3+4+6-5= 24 variables binaires. Soit l'équivalent de 36-5+24 = 55 variables dans le modèle logistique pour 2184 évènements. J'ai donc plus de 10 évènements par variable explicative, les conditions de validité sont respectées.

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

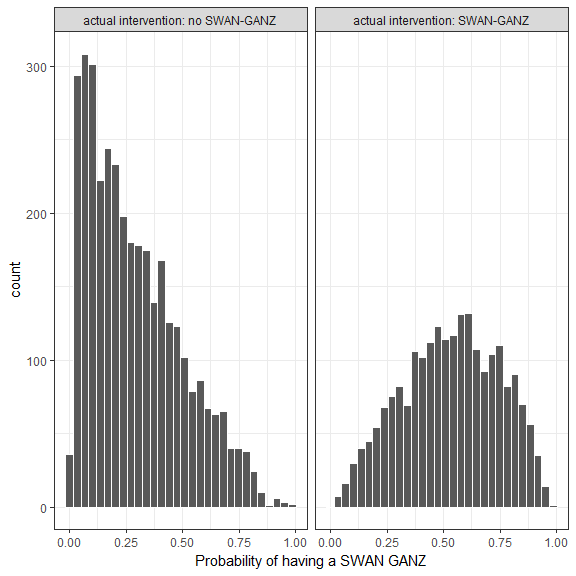


Figure x. Distribution de la probablité d'être cathétérisé selon qu'ils ont effectivement été cathétérisé ou non.

On voit que la distribution du score n'est pas la même parmi les patients ayant été cathétérisé et ceux n'ayant pas été cathétérisé : beaucoup de patient n'ayant pas été cathétérisé avait une faible probabilité d'être cathétérisé (distribution en L), alors que la distribution est plutôt en cloche centré autour de 0.5(distribution en cloche). On comprend donc que nécessairement l'appariemment va supprimer de nombreux patients.

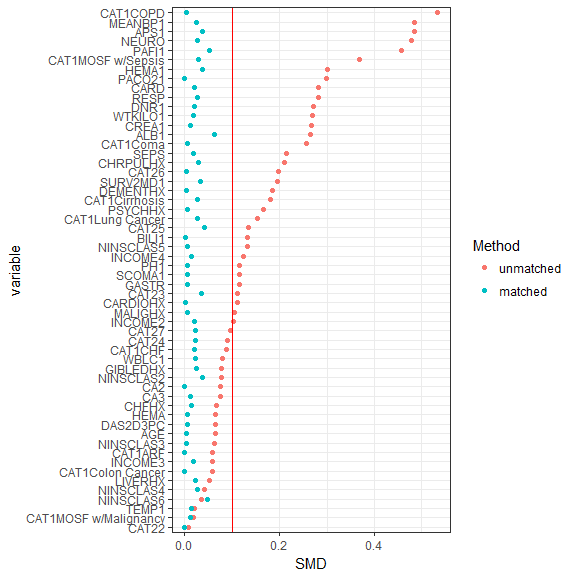
# Appariemment sur le score de propension

A partir du score de propension, je vais apparier un sujet cathétériser avec un sujet non cathétérisé ayant un score de propension proche c'est à dire une probabilité d'être cathétérisé proche. Le sujets non appariés sont écartés de l'analyse. Les sujets devrait donc se ressembler dans les 2 groupes. J'utilise un caliper de 0.2, c'est à dire un seuil d'appariemme,t de 0.2 x sd(logit(score de propension)).

## nombre de patients gardés

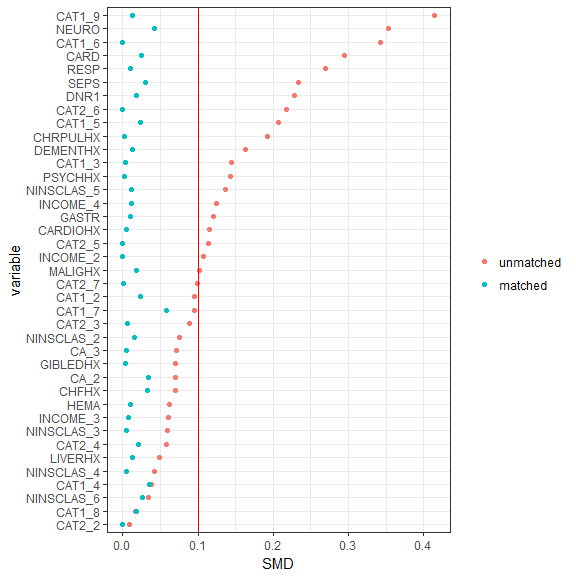
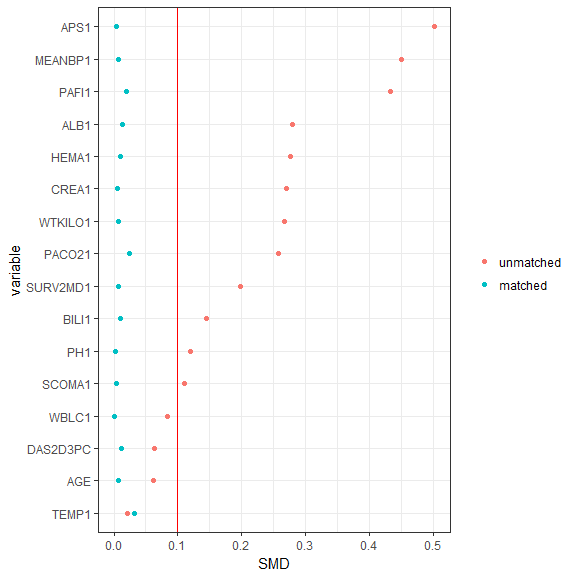
# Vérification de l'équilibre des variables entre les deux groupes après appariemment

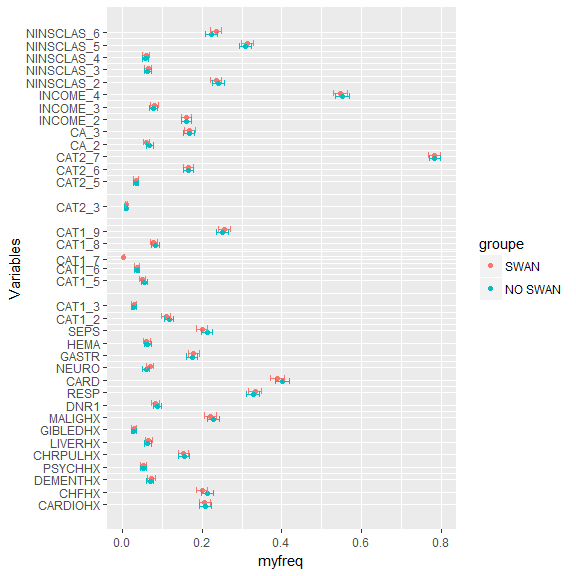
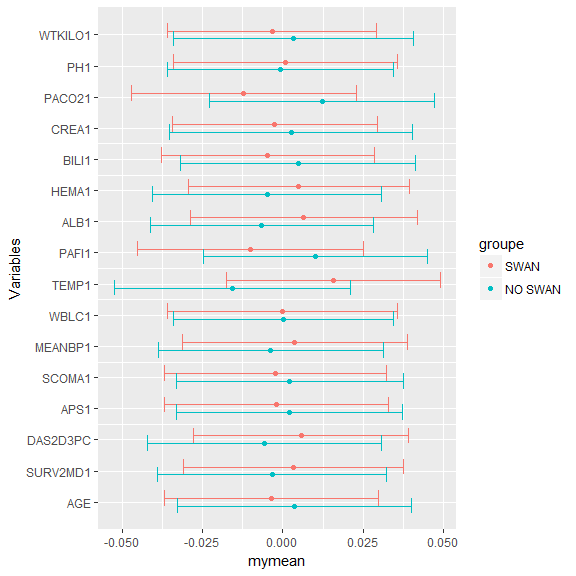
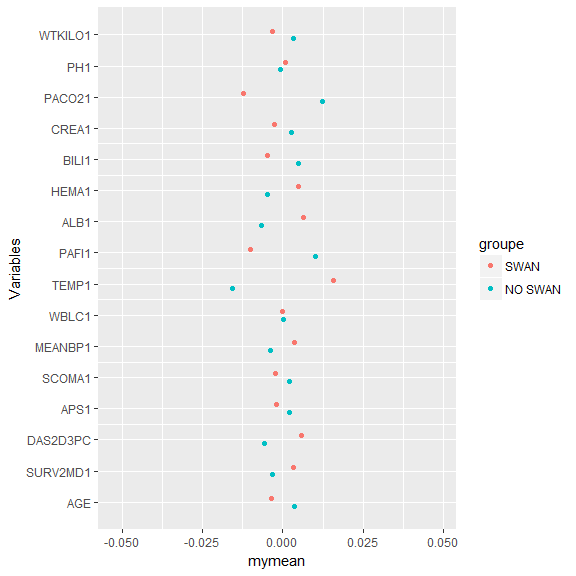
Saving 6 x 6 in image



[1] -0.06135247

[1] -0.06468618





Tout d'abord, je transforme mes variables qualitatives en n-k variables binaires, k étant le nombre de classe de la variable qualitative.  
J'ai ensuite plusieurs méthodes possibles pour regarder si l'appariemment a équilibré la distribution des variables dans le groupe traité et non traité : - je peux regarder la moyenne de chaque variable en fonction du score de popension. Je vois donc pour des patients de ces 2 groupes ayant la même probabilité d'être traité, si la moyenne de la variable est semblable. Bien sûr si les courbes se superposent parfaitement, on peut dire que la distribution est semblable dans les deux groupes.Si elles sont disjointes à certaines valeures d'abscisse, nous pouvont dire que pour les individus ayant tel probabilité d'être traité, la moyenne des variables diffèrent. Ainsi dans les courbes présentées en annexe, on voit ainsi que les individus avec une forte probabilité d'être traité ont une moyenne qui semble différer entre les deux groupes pour INCOME, NINCLAS, RESP et SEPS, et que les individus avec une faible probabilité d'être traité ont une moyenne qui semble différer poyr les variables RESP, HEMA et DNR1. Or il me semble qu'on ne cherche pas à ce que les individus ayant la même probabilité d'être traité soit exactement semblable 2 à 2 (même si bien sûr c'est idéal) mais plutôt que les populations des groupes traités et non traités soit homogène, comme c'est le cas lorsque l'on randomise. Il me semble donc plus pertinent de voir si globablement les distribution sont les mêmes dans les deux groupes, et les deux méthodes présentées ci dessous répondent à cette question.  
- Je regarde la différence standardisée des moyennes(smd), c'est à dire la différence entre la moyenne dans le groupe cathétérisé et la moyenne dans le groupe non cathétérisé divisé par la variance commune. J'ai séparé les schémas en variable binaire et variable quantitative non binaire pour plus de visibilité. En instaurant un seuil de smd à 0.1 comme conseillé dans la littérature, je vois que l'appariemment établi un équilibre entre les deux groupes pour toutes les variables. - Je regarde les intervalles de confiances des différentes variables, j'ai réalisé les intervalles de confiance des moyennes pour les variables quantitatives et des fréquences de 1 pour les variables binaires (qui sont en 0/1). Les variables quantitatives sont standardisées, ce qui permet une lecture plus aisée du graphique des variables quantitatives. Je ne sais pas comment réaliser un équivalent de cette standardisation avec les variables binaires et le graphique est donc moins facillement analysable. A noter que les variables pour lesquelles on n'a pas pu calculer d'intervalle de confiance pour non respect des conditions de validité (np>=5 et n(1-p)>=5) ne sont pas présente dans le graphe. On voit cependant là aussi que les variables sont plutôt bien équilibrée après appariemment.

Dans le futur, je préférerai quand même l'analyse de l'équilibre par la méthode smd, j'ai d'ailleurs pu voir qu'elle était souvent utilisée dans les publications.

# Analyses dans la population appariée

Pour répondre à la question "le cathétérisme cardiaque droit modifie-t-il la survie à 30 jours?", je peux utiliser deux méthodes différentes. Soit j'utilise une régression logistique conditionnelle, soit j'utilise un modèle de Cox. Dans les deux cas, j'utilise mon échantillon apparié et je stratifie sur le numéro de paire.

## Régression logistique conditionnelle

La variable à expliquer est la mort à 30 jours, la variable explicative est le cathétérisme cardiaque et je stratifie sur le paire.

SWAN   
1.188797

2.5 % 97.5 %  
SWAN 1.053171 1.341888

Les conditions de validité sont respectées, j'ai plus de 5 à 10 évènements par variable explicative (la stratification sur la paire n'est pas une variable explicative). Le coefficient du cathétérisme est significatif (pvalue<0.01), le cathétérisme cardiaque droit a donc un effet significatif sur le risaue de décès à 30 jours. le coeffcient vaut 0.17, l'exponentiel du coefficient me permet d'obtenir l'Odds ratio du risque de décès associé au cathétérisme. L'OR est de 1.19 avec un intervalle de confiance à 95% [1.05-1.34]. Le décès n'est pas un évènement rare, je ne peux donc pas interpréter l'OR comme un RR mais je peux dire que le risque de décès est augmenté lorsque le patient est cathétérisé.

## Modèle de Cox

[1] 1569

0 1 2   
0.4380587 0.4501916 0.1117497

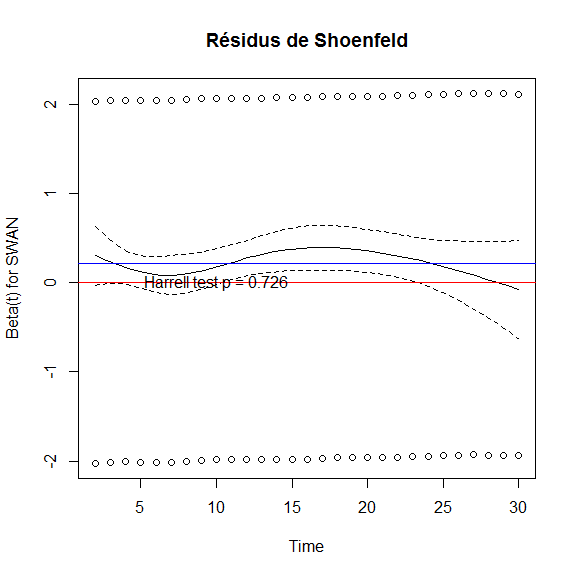
Une deuxième manière de réaliser le calcul est par modèle de cox. Dans ce cas il est problématique de prendre en compte l'appariemment par une stratification sur la paire car FE Harrell nous précise dans son livre Regression Modeling Strategies(2nd Ed) p.482 que si le nombre de strates est très grand par rapport au nombre total d'évènement, on perd en efficience(? "Loss of efficiency" dans le texte). Or on a ici 1569 paires ou strates pour 1055 évènements, ce qui peut être considéré comme un grand nombre de strates relativement au nombre d'évènements bien qu'aucun seuil ne soit précisé. Et de plus FE Harrell nous précise qu'une strate qui ne contient aucun évènement ne contribue pas à l'information et qu'une telle situation doit donc être évitée si possible. Or ici 44% des paires sont sans évènements et ne contribue pas à l'information avec donc j'imagine une perte de puissance. J'utiliserai donc l'option cluster(paire) plutôt que strata(paire) pour prendre en compte l'appariemment dans le modele de cox. Cette option prend en compte le design apparié, et on regarde alors la variance robuste et le test du score robuste. Cette méthode calcule une vraisemblance globale et non pas strate par strate comme lorsque l'on stratifie, évitant donc de perdre de l'information.

A noter que contrairement à la régression logstique conditionnelle, les sujets perdus avant 30 jours sont gérés par le modèle de cox. Cependant je préfère garder le même échantillon de patients que pour la régression logistique conditionnelle en prenant la variable décès à 30 jours modifiée telle que les patients perdus de vu avant 30 jours sont manquant pour le statut(les sujets ayant un statut manquant sont automatiquement retirés de l'analyse, comme pour le modèle de régression logistique).

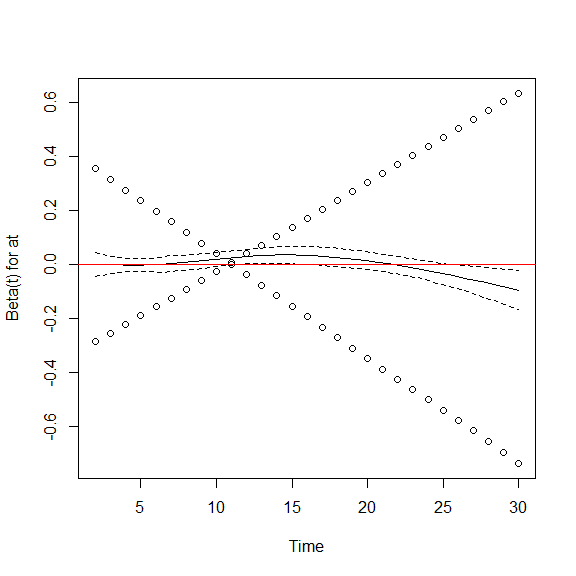
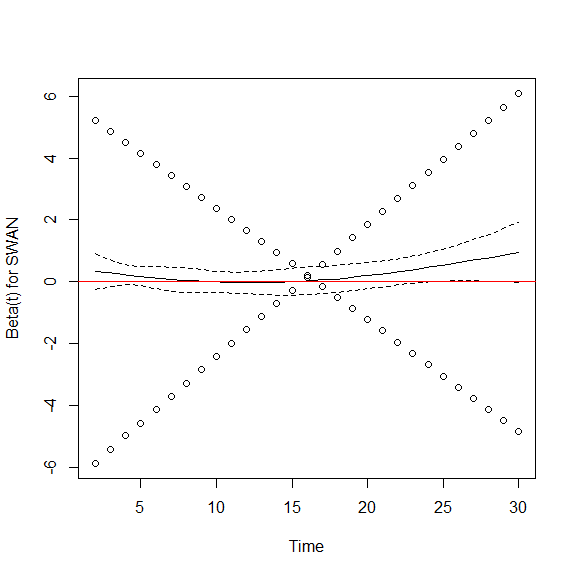
Test de Harrell

rho chisq p  
SWAN 0.0108 0.123 0.726

Test de Harrell p value: 0.726



[1] 0.7977832



pvalue   
0.0006457109

[1] "1.236" "1.233" "1.239"

L'hypothèse des risques proportionnels n'est pas vérifiée. En effet l'intervalle de confiance des résidus de shoenfeld n'incluent pas 0 en tous points. Il faut donc en toute rigueur que j'ajoute une variable SWAN dépendante du temps. Cependant je n'ai pas trouvé de transformation qui convienne, aucun coefficient de paramètre dépendant du temps n'étant significatif lorsqu'on l'ajoute au modèle (les résidus de Shoenfeld des modèles ne doivent donc même pas être regardés). Voici les transformations essayeées : log, racine carrée, \*temps, 1/temps, racine cube, carré, cube et puissance de 0.7. J'analyse donc le modèle de Cox sans ajouter de variable dépendante du temps, mais il faudra avoir en tête que le résultat est erroné.

L'ajout d'un cluster sur la paire me permet d'avoir une variance robuste prenant en compte l'appariemment. J'utilise cette variance robuste pour calculer l'intervalle de confiance à 95% du Hazard ratio: 1.236 [1.233-1.239]. Je regarde le test du score robuste pour conclure : p<0.001. Le risque de décès est donc significativement différent selon que le patient est cathétérisé ou non. Et c'est dans le sens d'un risque plus grand chez les patients cathétérisés avec un risque de décès à 30 jours multiplié par 1.236.

Je retrouve cette information graphiquement en traçant une courbe de survie en fonction du traitement par cathéterisme cardiaque droit.

## Représentation graphique de la survie à 30 jours en fonction du traitement par cathétérisme cardiaque droit

### Méthode

La méthode de Kaplan Meier permet de représenter graphiquement les courbes de survie (une courbe par groupe).

### Conditions de validité

Kaplan Meier :  
- censure indépendante de la probabilité de survenue de l'evenement  
- probabilité de survie indépendante du moment de recrutement dans l'étude  
- censure indépendante du groupe

In group CCD = 0

At 10 days, survival[95%CI] 82% [80% - 84%]  
 At 30 days, survival[95%CI] 69% [67% - 72%]

In group CCD = 1

At 10 days, survival[95%CI] 78% [76% - 81%]  
 At 30 days, survival[95%CI] 63% [61% - 66%]

Scale for 'colour' is already present. Adding another scale for  
'colour', which will replace the existing scale.

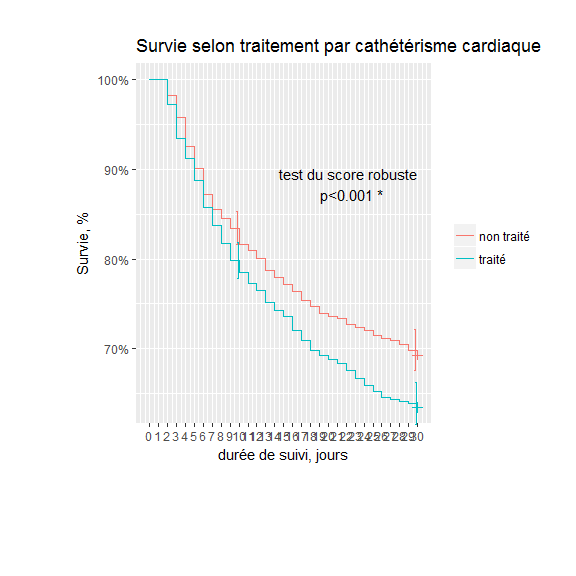


Fig X. Courbes de survie par la méthode de Kaplan Meier de la survie à 30 jours en fonction du traitement par cathétérisme cardiaque droit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | J0 | J10 | J20 | J30 |
| n.risk | 1569 | 1308 | 1159 | 1095 |
| n.event | 0 | 288 | 414 | 482 |
| surv | 100% | 82 % | 74 % | 69 % |

Table X Table de survie des patients non traités par cathétérisme cardiaque droit

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | J0 | J10 | J20 | J30 |
| n.risk | 1569 | 1251 | 1084 | 1000 |
| n.event | 0 | 337 | 489 | 573 |
| surv | 100% | 78 % | 69 % | 63 % |

Table X Table de survie des patients traités par cathétérisme cardiaque droit

On observe 2 courbes de survie, 1 pour le groupe non traité(0) et 1 pour le groupe traité(1). A chaque temps, les "marches d'escalier" représentent les patients qui ont eu l'évènement, faisant diminuer le nombre de personnes à risque d'évènement(c'est à dire les non répondeurs). Les croix représentent à chaque temps la présence de censure (perdus de vue ou exclus vivant). Par exemple, je passe de 75% de répondeurs à J7 à un peu plus de 50% de répondeurs à J14 dans le groupe 0. Dans la table de survie (table 3) je vois que ça correspond à 16 patients ayant eu l'évènement à J14. Pour le groupe 1, je lis sur la courbe que je passe de 90% environ à 75% de répondeurs. Dans la table de survie (table 4) cela correspond à 10 patients ayant eu l'évènement. La courbe de Kaplan Meier nous permet d'observer graphiquement que la survie est meilleure lorsque les patients ne sont pas traités par cathétérisme cardiaque droit.

# Conclusion

Dans cette étude observationnelle, après prise en compte des biais d'indication propables à l'aide d'un score de propension, nous mettons en évidence une association entre le cathétérisme cardiaque droit par sonde de swan ganz et le risque de mortalité à 30 jours : le cathétérisme cardiaque droit est associé à une augmentation du risque de mortalité (bien que l'utilisation du score de propension augmente le niveau de causalité, on ne peut pas conclure avec le même niveau de causalité qu'un essai randomisé).

# Annexes

# Distribution des variables avant modification et imputation