IUT de Paris - Rives de Seine
Université de Paris

# Rapport informatique intermédiaire

CAI Matteo
DE RIDDER Mathias
LEFEBVRE William
TRAORE Makam

Le 01/03/2024
BUTSD1 - Groupe 14
Groupe de travail n°3

## Introduction

La cathétérisassions cardiaque droite, également connue sous le sigle RHC, est une procédure médicale cruciale utilisée pour évaluer et surveiller la fonction cardiaque chez les patients présentant une variété de pathologies cardiovasculaires. En insérant un cathéter dans les vaisseaux sanguins, notamment dans les cavités cardiaques droites, cette technique permet une évaluation précise des pressions, des débits sanguins et d'autres paramètres hémodynamiques vitaux. Depuis son introduction dans les années 1970, le RHC est devenu un outil essentiel en médecine, offrant des informations précieuses pour guider le diagnostic et la gestion des maladies cardiovasculaires complexes. Cependant, son utilisation nécessite une expertise technique et comporte des risques, ce qui soulève des questions concernant son application appropriée et son efficacité clinique.

Quels sont les principaux facteurs de risque de décès ? En quoi la cathétérisation cardiaque droite/ Right Heart Catheterization (RHC) influence-t-elle ces risques ?

Pour cette analyse, nous possédons une base de données provenant d'une étude observationnelle multicentrique réalisée aux Etats-Unis. Cette étude porte sur 5 500 patients adultes admis en unité de soins intensifs entre juin 1989 et janvier 1994, et suivis pendant 6 mois.

La base de données est composée de 67 variables, regroupées en 7 catégories distinctes. Ces dernières sont "Diagnostic à l'admission" reflétant les diverses affections médicales des patients, " Evénements" qui correspond au traitement et à l'état vital du patient, " Scores Cliniques", "Caractéristiques physiologiques", "Caractéristiques sociodémographiques", "Maladies associées" et "Dates".

# Table des matières

# 1. Description des données brutes

## Diagnostic à l'admission

Nous détaillons ici la population cible pour le diagnostic d'admission.

Parmi la principale catégorie de maladie (CAT1) 9 maladies sont présentes, ARF étant la plus rencontrée, avec 2388 individus touchés. Il en va de même pour les maladies présente dans Catégorie secondaire de maladie (CAT2), cela dit, le pourcentage de personnes touchées est moindre en comparaison à CAT1 (20,82% de malades contre 100% en CAT1).

Pour les autres catégories cancer (CA), maladies respiratoires (RESP), problèmes cardiovasculaires (CARD), problèmes neurologiques (NEURO), problèmes gastro-intestinaux (GASTR), problèmes hématologiques (HEMA)) elles représentent entre 1000 et 1500 malades. Cela dit pour les problèmes de type rénaux (RENAL), métaboliques (META), infectieux (SEPS) ou traumatiques (TRAUMA) il y a entre 50 et 400 touchés.

Nous décrirons ici en détail les différents types de patients dans les catégories de maladies, chacune des réponses ayant reçu 5 500 réponses selon l'enquête, Il n'y a pas de valeurs manquantes et également il n'y a pas de valeurs aberrantes. Le problème respiratoire (RESP) est la maladie la plus fréquente, représentant 36% des réponses fournies. Ensuite, les maladies cardiovasculaires (CARD) sont légèrement moins courantes, avec 1854 personnes, soit 32%. Les maladies qui se situent autour de 10% de la distribution sont les maladies infectieuses (SEPS), les cancers (CA) et les troubles gastro-intestinaux (GASTR). Les autres catégories, telles qu'ORTHO et TRAUMA, représentent moins de 5% chacune, ce qui indique qu'elles ne sont pas très représentatives dans cet ensemble de données et peuvent être négligées.

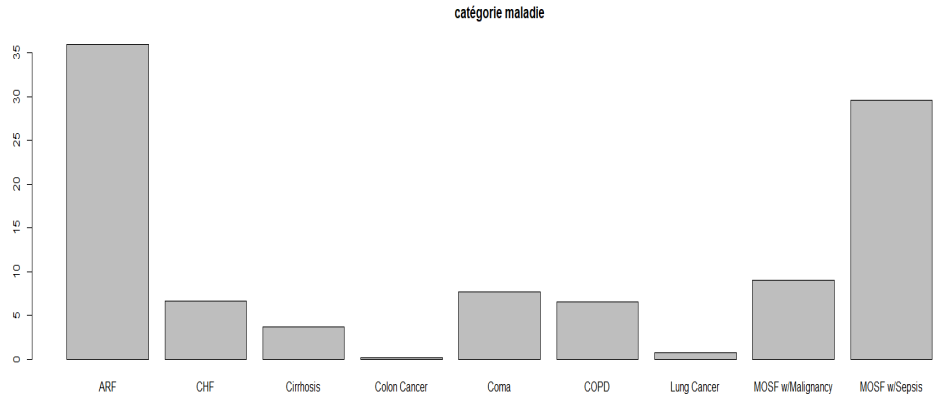


*Figure 1 - Répartition des maladies par catégorie*

En examinant maintenant les catégories de maladies et chaque patient peut avoir jusqu'à deux catégories, nous observons un total de 9 catégories. La catégorie la plus prédominante est ARF, représentant 36% des réponses. Ensuite, la catégorie la plus fréquente est celle du MOSF avec sepsis, englobant environ 30% des réponses. En revanche, certaines catégories, telles que le cancer du côlon (0,14%) et le cancer du poumon (0,76%), sont moins représentées. Ces deux catégories présentent les proportions les plus faibles parmi toutes les catégories examinées. Nous pouvons également négliger les autres catégories inférieures à 10%.

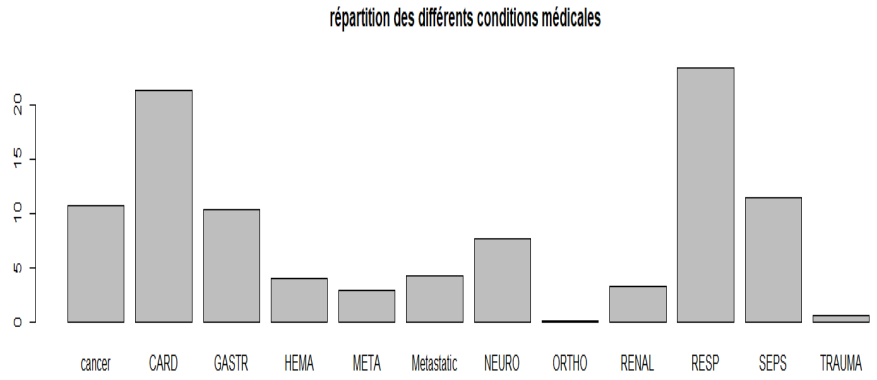


*Figure 2 - Répartition des différentes conditions médicales*

## Maladies associées

Les maladies avec les plus grands taux de contamination (IMMUNHX, MALIGHX, CHRPULHX, CHFHX, CARDIOHX) sont entre 17% et 27% avec l'immunosuppression, le VIH, le diabète (IMMUNHX) qui a le plus haut taux de contamination : 26,96%

Pour les moins contagieuse (DEMENTHX, PSYCHHX, RENALHX, LIVERHX, GIBLEDHX, TRANSHX, AMIHX) elle gravite entre 3 et 11% avec les Hémorragie gastro-intestinale (GIBLEDHX) et les Infarctus du myocarde (AMIHX) qui ont respectivement 3,18% et 3,47% de contamination.

## Scores Cliniques

La moyenne du Score DASI (20.83) semble être impactée par des valeurs élevées, comme indiqué par le fait que le troisième quartile (23.91) est plus proche de la médiane (19.91) que de la moyenne. La plage étendue (11.00 à 49.00) suggère une variabilité significative dans les scores.

Pour le Score Apache 3, la médiane (54.00) est plus proche du premier quartile (41.00) que du troisième quartile (68.00), suggérant une possible asymétrie dans la distribution des données. La présence d'une valeur maximale très élevée (499.00) peut indiquer la présence de valeurs aberrantes ou d'extrêmes dans le jeu de données.

Pour le Score Glasgow, La moyenne (23.34) est nettement inférieure à la médiane (0.00), suggérant une distribution fortement asymétrique avec une présence significative de valeurs zéro.

La dispersion des données est étendue, avec un écart important entre le premier quartile (0.00) et le troisième quartile (41.00).

Nous nous intéressons maintenant aux Scores Cliniques. Seule la variable SURV2MD1 ne nécessite pas de traitement spécifique. Nous ne prenons pas en compte la variable ADL3P (le score ADL), car elle présente un nombre important de données manquantes, environ 75%. Il est à noter que les données manquantes représentent une proportion significative. Pour les autres variables, il existe des données aberrantes.

Nous constatons un écart important dans les probabilités de survie entre différentes populations, allant de 0% à 0.96%. Il est intéressant d'analyser les raisons de cette disparité. La médiane, qui est proche de 0.6, est similaire à la moyenne qui est également proche de 0.5, suggérant ainsi une distribution relativement uniforme des probabilités. De plus, il est à noter que la probabilité peut descendre jusqu'à 0, ce qui signifie qu'il y a des cas de décès dans les deux mois suivant l'hospitalisation.

En ce qui concerne le score DASI, les valeurs semblent bien réparties, avec peu de différence entre la médiane et la moyenne (19.78 - 20.51).

Pour le score APACHE, le score théorique maximal peut atteindre 299. La valeur maximale observée est de 147 et le minimum est de 3. Les données semblent être uniformément réparties car la moyenne et la médiane sont toutes deux proches de 54.

Quant aux données sur le score de Glasgow, les scores sont compris entre 0 et 100. Il s'agit d'une distribution asymétrique avec une médiane de 0 et une moyenne de 21.17. La moitié de la population a obtenu un score de 0 pour le score de Glasgow.

## Caractéristiques physiologiques

Certaines variables physiologiques portent sur des indicateurs médicaux très technique qui semble demander une expertise toute particulière que nous ne possédons pas.

Le poids minimum observé est de 19.50kg et 244kg au maximum. 25% des individus ont un poids inférieur ou égal à 60.40 kg. La moitié de la population étudiée a un poids inférieur ou égal à 72.10 kg. 75% des individus ont un poids inférieur ou égal à 85.20 kg. La moyenne est de 74.49kg.

La température minimale observée est de 27.00 °C et 43°C au max. 25% des individus ont une température inférieure ou égale à 36. 09°C.La moitié de la population étudiée a une température inférieure ou égale à 38.09 °C. 75% des individus ont une température inférieure ou égale à 39.00°C. La moyenne de la température est de 37.62°C.

La tension artérielle minimale observée est de 0 mm Hg et tension maximale vaut 259mm Hg. 25% des individus ont une tension artérielle inférieure ou égale à 50.00 mm Hg. La moitié de la population étudiée a une tension artérielle inférieure ou égale à 63.00 mm Hg. 75% des individus ont une tension artérielle inférieure ou égale à 115.00 mm Hg.

La fréquence respiratoire minimale observée est de 0 et 100 au max. respirations par minute ce qui pourrait s'agir de personnes déjà décédées. 25% des individus ont une fréquence respiratoire inférieure ou égale à 14.00 respirations par minute. La moitié de la population étudiée a une fréquence respiratoire inférieure ou égale à 30.00 respirations par minute. La moyenne de la fréquence respiratoire est de 28.15 respirations par minute. 75% des individus ont une fréquence respiratoire inférieure ou égale à 38.00 respirations par minute.

La fréquence cardiaque minimale et maximale observée est de 0-250. Bpm. 25% des individus ont une fréquence cardiaque inférieure ou égale à 97bpm. La moitié de la population étudiée a une fréquence cardiaque inférieure

ou égale à 124. 75% des individus ont une fréquence cardiaque inférieure ou égale à 142bpm. La moyenne de la fréquence cardiaque est de 115.3.

Après quelques recherches, les taux de la pression partielle du dioxyde de carbone (PACO21), le PH (PH1), d'Hématocrite (HEMA1), de sodium dans le sang (SOD1) sont dans les normes recommandées.

Pour ce qui est des taux de créatinine (CREA1), d'albumine (ALB1), de bilirubine (BILI1) et de de potassium (POD1) ils semblent globalement être légèrement en dessous des normes.

Bien qu'aucune donnée ne semble être aberrante et que le pourcentage de valeurs manquante n'est pas excessif (entre 30 et 50 en moyenne) nous n'utiliserons probablement pas ces indicateurs dans notre étude.

La variable URIN1 quant à elle possède 53% de valeurs manquante, de plus, les valeurs existantes ne semblent pas sortir des normes, sauf exceptions, nous n'utiliserons pas cet indicateur dans notre étude.

## Evénements

La variable SWANG1 indique si un patient a bénéficié ou non de la technique de réanimation expérimentale (RHC). 38.2 % d'entre eux en ont bénéficié.

La variable DEATH indique si un patient est décédé durant la période de l'étude. 65.1% d'entre eux le sont.

## Dates

La composante « Dates » de la base de données est composée de 12 variables. Ces variables fonctionnent par trois (jour, mois et année) et communiquent de l'information sur :

- La date d'inclusion dans l'étude du patient (d_SADMDTE, m_SADMDTE et y_SADMDTE)
- La date de sa sortie (vivant ou non) de l'hôpital (d_DSCHDTE, m_DSCHDTE et y_DSCHDTE)
- La date de décès éventuel du patient (d_DTHDTE, m_DTHDTE, y_DTHDTE)
- La date de dernière nouvelle du patient (d_LSTCTDTE, m_LSTCTDTE, y_LSTCTDTE)

Toutes les dates sont comprises entre le 11 juin 1989 et le 31 décembre 1994. Une donnée est manquante pour la date de sortie de l'hôpital et 1920 entrées n'ont pas de date de décès, en effet elles ne sont pas mortes.

Ce qui est intéressant à étudier pour les dates, sont les durées. Nous avons créé plusieurs agrégats qui quantifient la durée du séjour à l'hôpital (duree_sejour), la durée de survie après l'inclusion dans l'enquête (duree_survie_inclusion), la durée de survie après la sortie de l'hôpital (duree_survie_sortie) et la durée du contact avec le patient après que celui-ci est sorti de l'hôpital (duree_contact_apres_survie).

- Durée du séjour (sortie – inclusion) : les patients séjournent 24 jours en moyenne ou 14 jours en médiane, il y a des séjours qui sont particulièrement longs. Le séjour le plus court est de 2 jours, le plus long est de 342 jours.
- Durée de survie (décès – inclusion) : les patients meurent 161 jours en moyenne ou 28 jours en médiane, il y a à nouveau des durées de survie particulièrement longues. La survie la plus courte et la survie la plus longue est de 1943 jours. (« pour les patients qui meurent » ?)
- Durée de survie après sortie (décès – inclusion) : les patients survivent en moyenne 141 jours à leur sortie de l'hôpital ou 0 jours en médiane, ainsi, au moins 50% (54,8%) des patients décèdent à l'hôpital.

La survie après sortie la plus longue est de 1341 jours. Les individus qui ne décèdent pas à l'hôpital, survivent 169 jours en médiane et 312 jours en moyenne.

- Durée de contact après la sortie (dernier contact – sortie) : au moins 25% des patients ont une date de dernier contact précédant leur sortie de l'hôpital, les variables indiquant le dernier contact avec le patient sont donc à remettre en cause.

## Caractéristiques sociodémographiques

L'étude démographique révèle une répartition de 55,6% d'hommes et 44,4% de femmes parmi les participants. Sur ce total, 77,45% sont de race blanche, 16,25% sont d'origine noire, et 6,29% appartiennent à d'autres groupes ethniques. Les patients inclus dans cette étude ont un âge varié, s'étalant de 18 à 102 ans, avec une prédominance significative de personnes âgées de plus de 50 ans, représentant 76% de l'échantillon, dont 24,3% se situent dans la tranche d'âge de 60 à 69 ans.

Sur le plan éducatif, la majorité des individus (61%) ont suivi entre 10 et 15 années d'études (de BAC à BAC+5). En ce qui concerne le revenu annuel, 56% des patients gagnent moins de 11 000€. Les données sur l'assurance médicale indiquent une répartition de 29,67% pour l'assurance privée, 25,38% pour Medicare, et 21,62% pour une combinaison d'assurances privée et Medicare.

# 2. Création d'une base de données propre à partir des données brutes

Dans l'objectif de créer une base de données propre pour répondre à notre problématique, après avoir chargé les données dans un DataFrame et en prenant en compte les valeurs manquantes, nous avons effectué différents traitements concernant la gestion de données manquantes ou de données aberrantes, du regroupement de modalités ou encore de jointures.

```
file_path <-'Tables/Epi_Clin2024.txt'
donnees <- read.table(file = file_path, sep = '\t', header = TRUE, na.strings = c("NA", "", "No"))
```

*Figure 3 - Lecture de la base de données et traitement des valeurs manquantes*

## Description des traitements

Sur chacune des variables, nous avons effectué des analyses à l'aide de fonctions Table et Summary ou encore de graphiques, pour nous permettre de les décrire. A partir de ces résultats, nous avons pu vérifier la qualité des données ou bien proposer des traitements à appliquer.

```
cancer_1<-table(donnees$CA, useNA = 'always',dnn = "cancer")
barplot(cancer_1)
```

*Figure 4 - Exemple d'une fonction table. Appliquée à chacune des variables qualitatives, cette fonction nous permet d'obtenir le nombre d'occurrence de chacune des modalités. Lorsque cela nous semble nécessaire, nous utilisons Barplot ou Boxplot, pour créer des graphiques.*

## Diagnostic à l'admission

Pas de valeurs aberrantes pour CAT1 et CAT2, pour les catégories de maladies. Aucun traitement nécessaire.

```
cat_maladie <- table(c(donnees$CAT1,donnees$CAT2),useNA = 'no',dnn = "categorie maladie ")
```

*Figure 5 - Pour CAT1 et CAT2, les variables ont été rassemblées pour compter les maladies, indistinctement du fait qu'elles soient principales ou secondaires.*

Le contenu des autres variables du diagnostic à l'admission (CA, RESP, CARD, NEURO, GASTR, RENAL, META, HEMA, SEPS, TRAUMA et ORTHO) a été renommées, les « Yes » ont été remplacé par le nom des maladies pour faciliter de futures analyses. Pas de valeurs aberrantes, aucun traitement supplémentaire.

## Maladies associées

Pas de valeurs aberrantes concernant les variables maladies associées. Aucun traitement nécessaire.

## Scores Cliniques

Les scores cliniques sont censés avoir des valeurs situées dans des bornes prédéfinies, comme indiquées dans le document Descriptif_Data_20232024.docx. Le score Apache 3 par exemple est censé être 0 et 33, la variable

SURV2MD1 est une probabilité est doit être comprise entre 0 et 1. Après analyse, il y a bien des valeurs aberrantes sur certaines de ces variables.

DAS2D3PC, APS1 et SCOMA1 possèdent des valeurs qui dépassent les bornes. Ces valeurs ne concernent pas plus d'une centaine de patients par variables, et le reste des variables de ces patients semblent être de qualité. Ainsi, nous prenons la même décision que nous avons prise pour la Tension artérielle et la Fréquence respiratoire, c'est-à-dire, ne pas faire de traitement immédiat pour supprimer les patients mais les retirer lorsque nous ferons des analyses sur ces 3 variables.

La variable ADLD3P contient 4113 valeurs manquantes (près de 75%). Nous retirons cette variable du reste de notre étude.

## Caractéristiques physiologiques

Les variables Poids (WTKILO1), Température corporelle (TEMP1), Tension Artérielle (MEANBP1) et Fréquence respiratoire (RESP1) ont été sujette à réflexion concernant des valeurs qui détonnent. Les deux premières variables possèdent des valeurs particulièrement faibles ou particulièrement élevées. Nous pouvons par exemple citer les 3 personnes qui font moins de 25kg ou bien les personnes qui ont une tension inférieure à 30. Après délibération collective, il a été décidé que ces valeurs étaient extrêmes mais pas nécessairement aberrantes. Nous pouvons notamment mentionner la personne la plus légère du monde pesait 12kg et a vécu jusqu'à 75ans. De plus, les entrées des patients correspondants ne semblaient pas particulièrement anormales. Ainsi, aucun traitement n'a été effectué sur les variables WKILO1 et TEMP1. Concernant la tension (MEANBP1) et la fréquence respiratoire (RESP1), certains patients ont des entrées égales à 0, 76 patients pour la tension et 130 pour la fréquence respiratoire. Ces valeurs sont de toute évidence anormales et résultent probablement d'erreurs de saisies dans la base de données. Cependant, le reste des variables des patients concernés semblent être de qualité, nous avons préféré ne pas supprimer ces patients de la base pour ne pas perdre des informations. Nous retirerons les patients qui ont des entrées égales à 0 uniquement lors des analyses sur la tension ou la fréquence, ainsi nous pouvons préserver la fiabilité de ces deux variables sans perdre d'informations sur les autres. Aucun traitement, immédiat n'a donc été réalisé sur MEANBP1 et RESP1.

Les autres variables décrivant les caractéristiques physiologiques sont très techniques et bien qu'elle semble de qualité correcte, elles dépassent nos compétences d'analyses. C'est pour cela que nous nous dirigeons vers une option qui consisterait à retirer les colonnes trop complexes de notre étude. Pour les citer, PACO21, PH1, WBLC1, HEMA1, SOD1, POT1, CREA1, BILI1 et ALB1 ne seraient pas prises en compte dans cette enquête.

URIN1, elle, sera retiré pour cause de valeurs manquantes (53%).

## Evènements

Les variables d'évènements n'ont pas nécessité de traitement. SWANG1 et DEATH ne nous ont pas mené à faire des modifications sur la base.

## Dates

Les variables dates ont requis des traitements au niveau de la mise en forme, de l'analyse puis a des traitements correctifs mineurs.

Après des vérifications préalables (mois compris entre 1 et 31 ; pas de concentration des jours de la semaine sur un jour particuliers…), les variables ont été regroupées entre elles pour former des variables au format « Date » sur R. (voir le code ci-dessous)

```
date_sortie_hopital = as.Date(paste(donnees$d_DSCHDTE, donnees$m_DSCHDTE, donnees$y_DSCHDTE,  sep="/"), format = "%d/%m/%Y")
```

*Figure 6 - Exemple de remise en forme des dates. Une fois les chaines de caractères concaténées (fonction Paste) de manière convenable, la nouvelle chaîne de caractère est alors convertie au format Date puis ajoutée dans la base de données finale.*

Dans un second temps, grâce au package lubridate, des variables « durées » ont été créées pour nous permettre de faire des analyses. Ainsi, duree_sejour décrira le nombre de jour passé à l'hôpital, duree_survie_inclusion décrira le nombre de jours survécu après avoir été inclus dans l'étude, duree_survie_sortie décrira le nombre de jours survécu après la sortie de l'hôpital et duree_contact_apres_sortie décrira combien de temps l'hôpital est resté en contact avec le patient après sa sortie de l'hôpital (voir code). Ces agrégats ont été ajoutés à la base que nous utiliserons dans notre enquête.

```
duree_survie_sortie = as.period(dates$date_deces - dates$date_sortie_hopital)
```

*Figure 7 - Création de variables durées. A partir de deux dates au format Date et du package lubridate, il est possible de calculer une durée, qui nous permet de faire des analyses. Chacune des durées créées est stockée dans un data frame puis est ajouté dans un data frame puis est ajouté dans la base de données finale.*

Enfin, nous avons pu, notamment grâce à ces durées vérifier la qualité des variables. Concernant la variable date_sortie_hopital, le patient 4487 n'est jamais sorti de l'hôpital. En revanche, il a une date de décès en sachant que tous les patients décédés sortent de l'hôpital le même jour (à vérifier), nous pouvons corriger cette valeur manquante en y indiquant sa date de décès. Concernant la variable date_deces possède 1920 valeurs manquantes, cela correspond au nombre de personnes survivantes, aucun traitement n'est requis.

Avec la variable duree_contact_apres_sortie, nous avons pu nous rendre compte que certaines dates de sortie de l'hôpital étaient postérieures à leur date de dernière nouvelle correspondante. Cette remarque remet en cause la fiabilité de la variable date_derniere_nouvelle, en effet, cela sous entendrait que des patients sont hospitalisés dans l'établissement mais que le personnel ne parvient pas à les joindre. De plus, les informations apportées par cette variable ne nous semblent pas suffisamment cruciales à notre étude pour justifier des traitements lourds. Nous avons choisi de retirer cette variable de notre étude.

# Conclusion :

Après avoir exploré, analysé et décrit chacune des données de la base de données avec soin grâce à R et aux packages lubridate. Nous avons pu avoir une meilleure idée de la composition des données, déterminer une problématique pour notre projet et avons pu effectuer des traitements (remise en forme variables, corrections, suppressions…) variables pour obtenir une base de données propre.

Ce processus n'a pas été sans ses défis, notamment lors de la recherche d'informations techniques et médicales. Nous avons trouvé cette expérience à la fois enrichissante et amusante, en découvrant de nouveaux aspects des données et en repoussant nos limites.

# Annexes :

| Variable | Type | Traitement | Libellé | Variable | Type | Traitement | Libellé |
|---|---|---|---|---|---|---|---|
| PTID | Texte | . | Identifiant Patient | ***Caractéristiques physiologiques*** *(dans les 24h)* | | | |
| ***Diagnostic à l'admission*** | | | | WTKILO1 | Num. | . | Poids (kg) |
| CAT1 | Texte | . | Catégorie principale de maladie | TEMP1 | Num. | . | Température corporelle (°C) |
| CAT2 | Texte | . | Catégorie secondaire de maladie | MEANBP1 | Num. | Traitement différé | Tension Artérielle (mm Hg) |
| CA | Texte | Remise en forme | Cancer | RESP1 | Num. | Traitement différé | Fréquence respiratoire (resp/mn) |
| RESP | Texte | Remise en forme | Respiratoire | HRT1 | Num. | . | Fréquence cardiaque (Bpm) |
| CARD | Texte | Remise en forme | Cardiovasculaire | PAFI1 | Num. | . | Ratio PaO2/FIO2 (mm Hg) |
| NEURO | Texte | Remise en forme | Neurologique | ***Caractéristiques sociodémographiques*** | | | |
| GASTR | Texte | Remise en forme | Gastro-intestinal | AGE | Num. | . | Age (année) |
| RENAL | Texte | Remise en forme | Rénal | SEX | Texte | . | Sexe |
| META | Texte | Remise en forme | Métabolique | RACE | Texte | . | Ethnie |
| HEMA | Texte | Remise en forme | Hématologique | EDU | Num. | . | Année d'étude (année) |
| SEPS | Texte | Remise en forme | Infectieux | INCOME | Texte | . | Revenu |
| TRAUMA | Texte | Remise en forme | Traumatique | NINSCLAS | Texte | . | Type d'assurance médicale |
| ORTHO | Texte | Remise en forme | Orthopédique | | | | |
| ***Maladies associées*** | | | | | | | |
| CARDIOHX | Num. | . | Atteinte vasculaire, cardiovasc. | | | | |
| CHFHX | Num. | . | Crise cardiaque | ***Evénements*** | | | |
| DEMENTHX | Num. | . | Démence, infarctus céréb., Parkinson | SWANG1 | Texte | . | Right Heart Catheterization (RHC) |
| PSYCHHX | Num. | . | Psychose, dépression | DEATH | Texte | . | Décès durant le suivi |
| CHRPULHX | Num. | . | Atteinte pulmonaire | ***Dates*** | | | |
| RENALHX | Num. | . | Atteinte rénale | date_inclusion_etude | Num. | Fusion | Date inclusion étude (jour / mois / année) |
| LIVERHX | Num. | . | Cirrhose, atteinte hépatique | date_sortie_hopital | Num. | Fusion + correction | Date de sortie de l'hôp (jour / mois / année) |
| GIBLEDHX | Num. | . | Hémorragie gastro-intestinale Haute | date_deces | Num. | Fusion | Date du décès (jour / mois / année) |
| MALIGHX | Num. | . | Tumeur solide, hémopathie maligne | date_dernière_nouvelle | Num. | Fusion | Date de dernière nouvelle (jour / mois / année) |
| IMMUNHX | Num. | . | Immunosuppression, VIH, diabète… | ***Durées*** | | | |
| TRANSHX | Num. | . | Transfert d'un autre hôpital (>24h) | Duree_sejour | Num. | | Durée du séjour dans l'hôp |
| AMIHX | Num. | . | Infarctus du myocarde | Duree_survie_inclusion | Num. | | Date de survie 1 |
| ***Scores Cliniques*** | | | | duree_survie_sortie | Num. | | Date de survie 2 |
| SURV2MD1 | Num. | . | Probabilité Estimée de survie à 2 mois | | | | |
| DAS2D3PC | Num. | Traitement différé | Score DASI (range | | | | |
| APS1 | Num. | Traitement différé | Score Apache 3 | | | | |
| SCOMA1 | Num. | Traitement différé | Score Glasgow | | | | |

*Tableau 1 - Résumé des variables du jeu de données propre.*