# Représentation et description des parcours patients en chirurgie cardiaque : une approche exploratoire par le clustering

Lucile Trutt

Service Commun de la Documentation
Chemin la censive du Tertre - BP 32211
44322 Nantes cedex 3

www.bu.univ-nantes.fr

# NANTES UNIVERSITÉ

_____

# FACULTÉ DE MÉDECINE

_____

Année : 2022 N°

**THÈSE**

pour le

**DIPLÔME D'ÉTAT DE DOCTEUR EN MÉDECINE**

SANTÉ PUBLIQUE

par

Lucile TRUTT

_____

Présentée et soutenue publiquement le 21 juin 2022

_____

REPRÉSENTATION ET DESCRIPTION DES PARCOURS PATIENTS EN CHIRURGIE CARDIAQUE : UNE APPROCHE EXPLORATOIRE PAR LE CLUSTERING

_____

Président : Madame le Professeur Leïla MORET

Directeur de thèse : Docteur Brice LECLERE

Je remercie chaleureusement toutes les personnes qui ont rendu possible ce travail et qui m'auront accompagnée tout le long de mon internat…

…Brice, à la fois pour le soutien que tu m'as apporté et pour la latitude que tu as accepté de me laisser dans nos différents travaux, incluant cette thèse.

…Leïla, pour ta coordination et pour avoir accepté de présider mon jury.

…Nicolas, pour avoir été initiateur de ce projet et grandement facilitateur, aussi bien pour ce travail que pour d'autres que j'ai pu avoir en lien avec le DIM.

…Pr Roussel, pour avoir accepté d'apporter votre regard d'expert clinique sur ce travail.

…Mes co-internes, pour ces bouts de chemin effectués ensemble à l'occasion de nos stages, de nos formations ou de nos associations, tout autant à l'échelle de Nantes, du Grand Ouest ou du niveau national.

…Tous mes encadrants de stage, médecins ou non, qui m'ont permis un internat varié et riche en expériences, dans lequel j'aurai beaucoup avancé sur le niveau professionnel mais aussi personnel.

…Ma famille, sans qui je ne serais jamais arrivée jusque-là.

…Guillaume, sans qui je ne serais jamais arrivée jusque-là non plus !

# Table des matières

# Figures

# Tableaux

# I. Introduction

La médecine des XXème et XXIème siècles est caractérisée par une explosion de la quantité de connaissance découverte chaque année et de la finesse avec laquelle nous pouvons appréhender aujourd'hui le corps humain et ses pathologies. Si ces avancées scientifiques sont bien entendu souhaitables et bénéfiques pour la santé de manière générale, elles rendent de plus en plus complexe la prise en charge des patients : la masse de connaissances à intégrer n'est plus à la portée d'une seule personne et le niveau d'exigence augmente, aussi bien du côté des patients qui veulent être soignés de façon qualitative et personnalisée, que du côté des soignants du fait de leur professionnalisme et de leur satisfaction d'un travail de qualité. Il devient alors complexe de conjuguer précision scientifique et prise en charge holistique et individualisée. Les patients présentant des pathologies graves se retrouvent dans un parcours morcelé, partagé entre différents médecins spécialistes, voire différents établissements de santé.

Ce fractionnement est observable aussi bien à l'échelle de l'évolution d'une maladie chronique qu'à l'échelle d'un séjour hospitalier. Un patient hospitalisé pour sa pathologie cardiaque sera pris en charge par différentes spécialités (cardiologie, chirurgie cardiaque, anesthésie, réanimation…), par différentes équipes pluri-professionnelles qui devront se synchroniser aussi bien en interne qu'avec les autres équipes hospitalières et les acteurs de soins primaires qui le suivent habituellement (médecin traitant, cardiologue…). A ce parcours segmenté peuvent également s'appliquer des facteurs aléatoires le complexifiant (échecs thérapeutiques, complications imprévues, faible compliance du patient, errance médicale…), le rendant moins linéaire en vie réelle qu'il ne l'est tel qu'imaginé dans les protocoles. Ainsi, chacun des patients d'un service de chirurgie thoracique et cardio-vasculaire (CTCV) aura déjà son propre parcours antérieur, mais aussi son propre cheminement au cours et après l'hospitalisation.

Si ce morcellement du parcours du patient est ce qui va lui permettre d'être soigné avec une médecine de précision, il est aussi à l'origine d'une complexification de ce parcours, aussi bien par le patient qui peut avoir l'impression d'être perdu, que par les équipes en charge de la planification des séjours et de l'organisation d'un établissement de santé. Il n'est pas possible de considérer une prise en charge médicale ou chirurgicale de façon isolée, sans prendre en compte les relations entre services internes, mais aussi avec les ressources externes à l'établissement (médecine de ville, établissements de soins de

suite et réadaptation (SSR)). Dans le cas des grands établissements de santé, le pilotage est ainsi éclaté entre différents services et différentes directions, voire différents sites géographiques. Il est pourtant nécessaire d'avoir une vision transversale et flexible, et ce malgré un fonctionnement très compartimenté et des outils de pilotage peu adaptés à cette conception.

C'est par ce besoin que ce projet de thèse a démarré avec le service de CTCV du centre hospitalo-universitaire (CHU) de Nantes en 2018. L'organisation de ce secteur, déjà complexe de base, était en proie à des problèmes notamment liés à une forte augmentation de l'activité. Des difficultés de coordination et d'alignement des attentes et contraintes de chacun étaient signalées par les équipes, aussi bien au niveau de la prise en charge thérapeutique que sur les aspects administratifs.

Un premier bilan de l'activité clinique avait été élaboré en amont du travail présenté ici. Le taux d'occupation des 58 lits du service d'hospitalisation conventionnelle de CTCV était de 86,4 % en 2017, avec des durées moyennes de séjour (DMS) assez hautes par rapport à la moyenne nationale des établissements de même type. Par exemple, pour les 3 premiers mois de 2018, tous types de pathologies confondues, la DMS dans l'unité d'hospitalisation conventionnelle de CTCV était d'en moyenne 1,8 jours plus longue que la DMS inter-CHU. Les chirurgiens expliquaient cela principalement par des difficultés à trouver des lits d'aval (notamment en SSR) et à rééquilibrer les traitements par anti-vitamine K, de plus en plus prescrits par les cardiologues. Ils avaient également observé une complexification de la prise en charge des patients, à cause de multiples comorbidités et de l'augmentation des interventions chirurgicales complexes. Malgré les efforts, la durée de séjour de ces patients compliqués restait difficile à réduire. En parallèle, un nombre croissant de patients de CTCV étaient hébergés dans l'unité d'hospitalisation conventionnelle de cardiologie à défaut de place dans le service chirurgical dédié.

Les interventions chirurgicales programmées étaient quant à elle en nombre croissant, avec une augmentation de + 12,8 % (+ 260 interventions) estimée pour 2018. Au premier bilan de l'année 2018 réalisé fin avril, on dénombrait déjà 117 interventions reportées. La principale justification des reports (40 %) était la nécessité de céder la place à une urgence (hors greffe). Venaient ensuite les problèmes de manque de place en aval, en particulier en réanimation (14 %), la surcharge du programme opératoire (14 %), les contre-indications médicales à l'intervention (13 %), la place cédée à une greffe (8 %), la prolongation d'une intervention précédente (7 %) et le manque de ressources, notamment humaines (4 %). Quant aux délais d'attentes préopératoires, une forte augmentation était

également attendue. En effet, si la médiane se situait à 48 jours pour les interventions d'avril 2018, les rendez-vous posés au mois de mai concernaient des interventions prévues pour septembre et octobre 2018, soit plus de 4 mois plus tard, bien au-delà des recommandations et des moyennes nationales.

Face à cela, différentes options pouvant apporter une amélioration ont été émises par la direction de l'hôpital et étaient soumises à évaluation :

- L'ouverture de nouvelles vacations hebdomadaires au bloc opératoire,
- La mise en commun de lits d'hospitalisation simple ou de télémétrie avec le service de cardiologie,
- L'augmentation du nombre de lits de réanimation CTCV,
- L'ouverture d'une unité dédiée dans le service de médecine physique et de réadaptation.

Afin de choisir et affiner ces différentes propositions, une analyse d'impact avec étude de scénarios était en cours. Cependant, les indicateurs classiques (taux de remplissage des lits des unités, DMS, indices de performance hospitalière) avaient montré leurs limites : les chiffres étaient à des niveaux très compartimentés et agrégés, les réinterventions n'étaient pas prises en compte, les échanges entre services étaient difficilement mis en évidence et les hébergements dans des unités ne faisant habituellement pas partie du parcours n'étaient pas repérés.

Un outil a donc été élaboré par l'autrice du présent mémoire dans le cadre de son stage de Master 2 au sein de l'unité recherche du service de Santé Publique du CHU de Nantes. Cet outil devait permettre d'explorer les flux de patients et d'obtenir de nouveaux indicateurs en lien avec la notion de parcours. Il s'agissait d'une interface graphique interactive basée sur les informations de mouvements administratifs des patients opérés par les chirurgiens de CTCV en 2016 et 2017 (1). L'outil permettait d'obtenir des résultats à la fois graphiques et chiffrés répondant à des questions précises grâce à des algorithmes d'extraction de motifs séquentiels et de règles d'association.

Suite au développement de cet outil, nous avons voulu explorer les possibilités offertes par les méthodes modernes de traitement des données pour la représentation et la description séquentielle des parcours de patients au sein d'un établissement de santé. Nous nous sommes focalisés sur la chirurgie cardiaque, en excluant la chirurgie pulmonaire et les transplantations thoraciques, car il s'agit de la filière majoritaire de ce

service et qu'elle présentait une plus grande diversité, aussi bien en termes de tableaux cliniques que de parcours. Les idées de départ restaient cependant les mêmes :

- Trouver comment analyser et représenter les données pour mieux les comprendre et amener des arguments objectifs dans les débats,
- Valoriser des données administratives pour en faire ressortir leur intérêt clinique, sans avoir à mener un recueil de données médicales complémentaires.

Si la prise en charge individuelle d'un patient se veut aujourd'hui personnalisée, à grande échelle il existe toujours des profils-types permettant de prévoir la prise en charge de façon macroscopique, et ainsi d'anticiper l'activité d'un service. Nous avons donc orienté notre réflexion vers les techniques de *clustering*, également appelé classification ou partitionnement[1], qui permettent de regrouper les sujets semblables et d'individualiser des groupes avec des caractéristiques distinctes.

L'objectif du travail présenté ici était de faire le tour des techniques applicables à l'analyse du parcours de patients et d'évaluer la faisabilité et la pertinence d'un clustering basé sur des données administratives pour répartir les patients de chirurgie cardiaque en groupes ayant des parcours et des caractéristiques cliniques semblables.

[1] Ces trois termes pourront être utilisés de manière indifférenciée dans le manuscrit. De même, « cluster », « classe » et « groupe » seront considérés ici comme synonymes.

# II. Étude bibliographique

## A. Définition d'un « parcours »

Si la notion de parcours patient peut paraitre facile et claire au premier abord, elle se complexifie lorsque l'on s'intéresse aux détails et aux variations attendues ou inattendues. On peut par exemple s'intéresser aux parcours patient en termes de prise en charge par l'équipe de soin et de succession d'événements de santé (actes, prescriptions, diagnostics, etc.) au cours d'un séjour (2). Cette définition se rapproche du terme « parcours de soin », utilisé en particulier par l'HAS et le Ministère de la Santé pour parler de l'évolution du système de santé français. Cette approche se penche sur la question de la coordination des soins, notamment entre hôpital et ville, dans des pathologies particulières, comme les cancers ou l'insuffisance cardiaque chronique, pour permettre une prise en charge cohérente tout le long de l'évolution de la maladie et éviter les ruptures. Le terme trajectoire de soins peut également être rapproché de cette vision mais a été délaissé. Dans une vision plus large, le parcours de santé est défini tel que le « parcours de soins articulé en amont avec la prévention primaire et sociale, et en aval avec l'accompagnement médico-social et social, le maintien et le retour à domicile » (3). De son côté, la BDSP (Banque de Données en Santé Publique) définit le parcours de santé tel que la « Succession des différents points de soins que parcourt un malade tout au long de son traitement », synonyme de filière de soin et de chemin clinique. (4)

Si l'analyse du parcours des patients se faisait traditionnellement par regroupements d'« éléments de parcours » (5), la vision actuelle se veut plus globale. Un outil de visualisation du parcours de soins basé sur la combinaison des différents types d'informations fournies par la base PMSI nationale a d'ailleurs été étudié récemment (6). D'autres projets avaient également été menés auparavant sur cette thématique, mais dans des domaines plus précis, tels que  la prise en charge du cancer du sein (7) ou de la sclérose en plaque (8).

La notion de parcours d'un patient peut aussi être appréhendée dans un sens purement administratif, c'est-à-dire la succession de mutations entre unités au sein d'un établissement et/ou de transferts entre différents établissements. Le domaine des pathologies cardiovasculaires a d'ailleurs déjà fait l'objet d'une étude française par analyse de réseaux qui observait les mutations effectuées d'une unité d'hospitalisation à

l'autre (9). Cependant, ce travail a étudié ces mouvements de façon agglomérée et le parcours n'était pas analysé dans sa globalité. Nous avons également fait le choix de cette approche « administrative » car notre objectif était de valoriser des données déjà accessibles au Département d'Information Médicale (DIM), mais en conservant l'information sur les parcours de façon plus précise, dans toute sa longueur.

## B. Représentation théorique

Le premier défi rencontré, lorsque l'on souhaite analyser des séquences, est leur représentation. Cette question est posée dès le premier niveau de conceptualisation, c'est-à-dire la construction de la base de données. L'équipe de Ritschard en a fait une synthèse assez claire et en propose une ontologie qui distingue notamment séquences d'états et séquences d'événements (Figure 1) (10). Les premières sont des suites d'éléments caractérisés par une durée et un caractère constant qui les définit, alors que les secondes sont plutôt des enchainements de passage d'un caractère à l'autre à un moment précis.

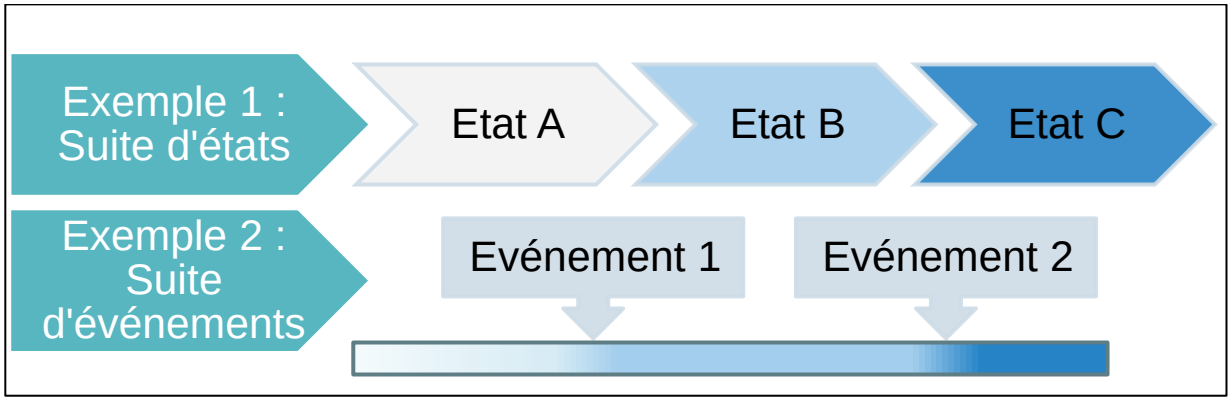



Figure 1 : Exemples de représentation théoriques de séquences : suite d'états ou suite d'évènements

Afin d'illustrer ces approches dans le domaine des parcours de soin, prenons l'exemple d'un patient qui arrive dans l'unité A, qui est opéré le lendemain, qui passe 2 jours dans l'UM B en sortant du bloc, puis qui retourne 5 jours dans l'unité A avant de sortir (Figure 2). Son parcours peut être représenté selon ces deux visions tel que dans le Tableau 1.

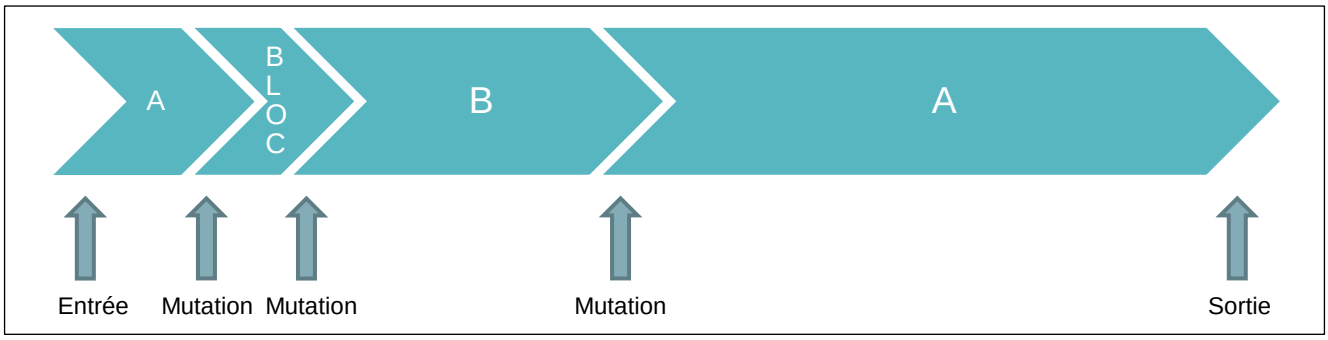



Figure 2 : Exemple fictif de parcours d'un patient

| **Suite d'états** | |
|---|---|
| **Durée** | **État** |
| 1 jour | Séjour en unité A |
| 4 heures | Passage au bloc |
| 2 jours | Séjour en unité B |
| 5 jours | Séjour en unité A |
| **Suite d'événements** | |
| **Date** | **Évènement** |
| J0 | Entrée en unité A |
| J1 | Intervention au bloc |
| J1 + 4h | Mutation en unité B |
| J5 | Mutation en unité A |
| J8 | Sortie |

Tableau 1 : Représentation d'un parcours de patient fictif sous forme d'une suite d'états et d'une suite d'événements

## C. Représentation graphique

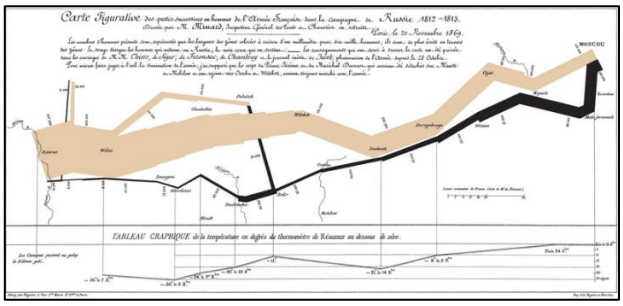
De même qu'il existe plusieurs options pour représenter numériquement des séquences, il existe de multiples approches pour les représenter graphiquement, ce qui constitue ici notre second niveau de représentation des données. Un large panel de visualisations adaptables aux flux de patients existe. Si nous commençons par les graphiques les plus simples, nous pouvons les catégoriser selon leur approche du problème (Tableau 2).

| Approche | Description |
|---|---|
| **Longitudinale** | Aperçu en longueur de la succession d'états, prenant en compte à la fois la longueur des états et leurs successions, sur le modèle d'une frise chronologique. |
| **Transversale** | Aperçu à chaque unité de temps de la distribution de la population, entrainant la perte de la visibilité de la durée des états individuels et de leur succession, mais permettant une vision plus synthétique (par exemple : histogramme, graphique en aires empilées (11)). Les courbes de survie peuvent également être incluses dans cette catégorie. |
| **Agrégée** | Perte de la notion de séquence, mais mise en avant de certaines caractéristiques étudiées spécifiquement : temps total passé dans chaque état, transitions entre états, état modal à chaque unité de temps, évolution d'indicateurs d'hétérogénéité, etc. (Le type de graphique sera alors adapté au format du résultat, en restant dans les « grands classiques » tels que les histogrammes ou les nuages de points.) |

Tableau 2 : Typologie des représentations graphiques simples des données séquentielles

Il existe également des graphiques plus complexes, dont nous qualifierions l'approche plutôt de mixte, entre longitudinale et transversale, avec un niveau d'agrégation plus ou moins important ; nous entrons ici dans le domaine des diagrammes de flux et des diagrammes de réseaux. Des exemples variés sont donnés dans la Figure 3. Le premier et célèbre exemple de visualisation complexe de données est d'ailleurs une représentation de flux : la Carte figurative des pertes successives en hommes de l'armée

française dans la campagne de Russie 1812-1813, publiée en 1869 par Charles Minard. Ces types de diagramme sont très variés et sont choisis en fonction des analyses qu'ils doivent représenter. Par exemple, les graphes orientés sont pertinents dans le cadre d'analyses effectuées sur la base de la théorie des treillis (12), la théorie des graphes (13) ou les réseaux bayésiens (14). Le diagramme de Sankey, quant à lui, est bien adapté à des flux qui se croisent et qui peuvent être quantifiés : initialement créé pour l'étude des transferts d'énergie (15), il se prête bien aux parcours de patients (6).




Carte figurative des pertes successives en hommes de l'armée française dans la campagne de Russie 1812-1813, Charles Minard, 1869

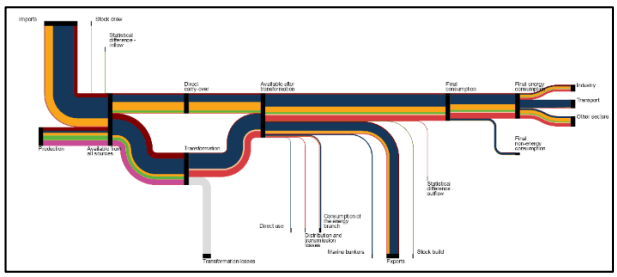


**Diagramme de Sankey**

Diagramme de flux d'énergie, Eurostat[2]

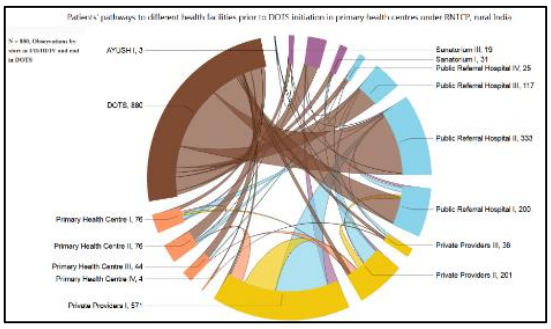



**Diagramme en cordes**

*Patients' pathways to different health facilities prior to DOTS initiation in primary health centers under RNTCP, India.* (16)

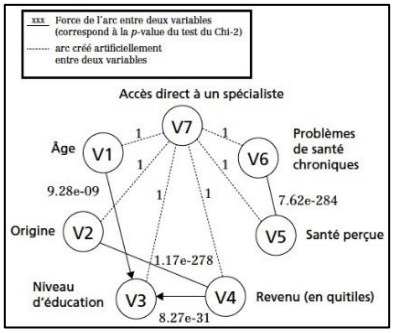



**Diagramme en réseau**

*Réseau bayésien rendant compte de l'accès direct au spécialiste pour les hommes pour toutes les variables sociales sélectionnées* (8)

Figure 3 : Exemples de représentations complexes de données séquentielles

[2] http://ec.europa.eu/eurostat/fr/web/energy/energy-flow-diagrams

## D. Analyses statistiques

Après avoir réussi à appréhender et à représenter des parcours, il est intéressant de pouvoir les analyser pour les caractériser. Diverses approches statistiques peuvent être utilisées :

- Les **analyses de séquences** sont les méthodes les plus fréquemment retrouvées. Ces analyses descriptives consistent en une analyse de similarité (17) afin de retrouver des ressemblances entre les séquences d'un corpus et permettre un partitionnement de celui-ci, voire une classification des séquences en types de parcours. (8)(18)(19)
- L'**extraction de motifs séquentiels** (20) est un deuxième type d'analyse de séquences qui repère les « morceaux » de séquences, appelés motifs, les plus fréquents. Cette technique ne permet pas de classer les parcours entiers, mais retrouve les éléments qui se suivent le plus fréquemment. De multiples algorithmes permettent d'affiner les résultats obtenus en intégrant par exemple des contraintes temporelles, des pondérations par éléments, de l'incertitude ou encore des analyses multidimensionnelles. (21)(22)
- Les **règles d'association** détectent les éléments fréquemment associés entre eux, quelle que soit leur fréquence. Elles s'appuient sur deux mesures : le support (force de l'association, c'est-à-dire le nombre de fois où la règle est trouvée) et la confiance (nombre de fois où la règle est respectée). Cette technique présente le défaut de ne pas conserver l'ordre des éléments de chaque côté de la règle (23) mais il existe des algorithmes spécialisés pour l'application de cette approche à des séquences : les règles d'association séquentielle (24).
- L'**analyse formelle de concepts** suit une toute autre logique que celles des séquences vues précédemment. Elle repose sur la théorie des treillis, sortes de grilles décrivant des relations entre des concepts qui regroupent des objets présentant des attributs communs. Son application au parcours des patients permet de visualiser les flux sous forme de réseaux. (12)(25)

Par exemple, si nous appliquions les 3 premières approches statistiques au corpus de séquences fictives donné dans la Figure 4, nous pourrions trouver que :

- Analyse de séquence : la séquence 1 ressemble plus à la séquence 3 qu'à la séquence 2,

- Extraction de motifs séquentiels : le motif B>C>D est le plus long et le plus fréquent que nous puissions trouver,
- Règles d'association séquentielle : B est suivi par D dans 100 % des séquences (support). Si B est présent, alors il est suivi par D dans 75 % (3/4) des cas (confiance).

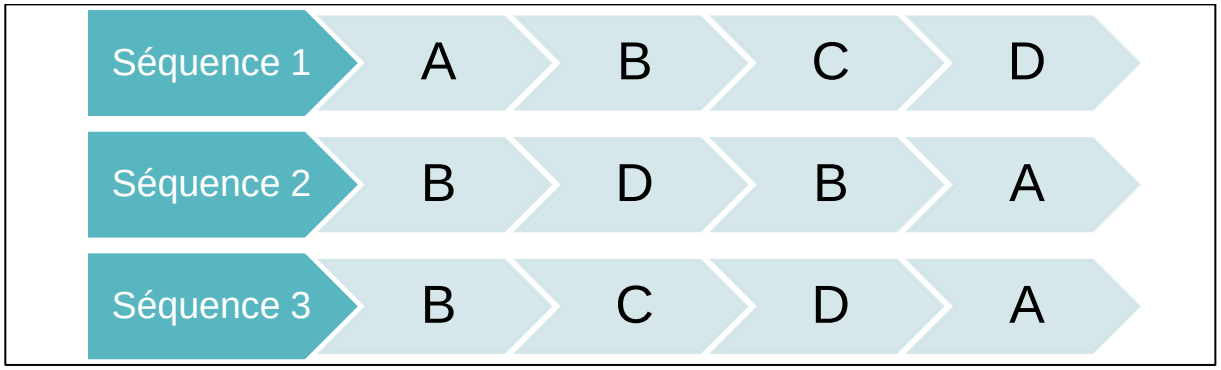



Figure 4 : Exemple de corpus de séquences fictives

Il existe encore une multitude d'autres approches qui n'ont pas encore, à notre connaissance, été transposées à l'exploration des parcours patient. En médecine, l'approche par séquence est utilisée pour le traitement des données omiques (26), pour l'analyse de signaux vitaux (27) ou la prédictions d'événements, comme des prescriptions (28) ou des infections nosocomiales (29). Mais nous pouvons nous inspirer d'autres domaines plus habitués à ces types d'analyse : les sciences sociales s'y intéressent depuis longtemps pour décrire des parcours de vie ; la recherche dans le monde du commerce et d'internet est particulièrement sensible à la compréhension des flux. Nous citerons par exemple les chaines de Markov, les réseaux bayésiens (14), les extractions de sous-graphes (13), *l'answer set programming* (programmation déclarative logique) (30), les modélisations de Poisson (31) ou les mesures de dissimilarité en grilles multidimensionnelles (32).

# III. Expérimentation d'une approche par le clustering

## A. Matériel et méthode

### 1. Population d'étude et sources de données

#### a. Sources de données et mise en forme des données de parcours

Notre échantillon était composé de patients concernés par une intervention de CTCV, et plus précisément de chirurgie cardiaque. Notre étude s'est basée sur la mise en commun de 2 bases de données :

- celle du logiciel de gestion du processus chirurgical : *QBloc* (*Evolucare*),
- et celle servant de base au PMSI.

En premier lieu, la base de QBloc nous a permis de sélectionner les patients opérés dans les salles réservées à la CTCV en 2016 et 2017. Nous avons pu en extraire les Identifiants Permanents du Patient (IPP), les dates et heures d'entrée et de sortie de bloc, et l'intervention réalisée (champ libre).

Partant de ces premières données, nous avons extrait des informations administratives et médicales depuis les bases PMSI du CHU en se basant sur l'IPP et l'année d'intervention.

Le PMSI (Programme de Médicalisation des Systèmes d'Information) est un dispositif national permettant de décrire de façon synthétique et standardisée l'activité médicale des établissements de santé. Depuis 2005, les données qu'il génère sont utilisées pour la Tarification à l'activité (T2A) des séjours hospitaliers en médecine, chirurgie et obstétrique (MCO). Lorsqu'une personne sort d'hospitalisation, un Résumé de Sortie Standardisé (RSS) est édité, permettant de classer le séjour dans un Groupe Homogène de Malades (GHM) et évaluer le tarif qui sera facturé. Le RSS est constitué d'un enchainement de Résumé d'Unité Médicale (RUM) dans lesquelles les équipes de l'hôpital auront codé des informations administratives et médicales (les diagnostics, les actes de soin, etc.) pour chacune des unités médicales par lesquelles le patient est passé. Les diagnostics sont codés selon la classification CIM 10 (annexe 2) et peuvent être de différents types :

- DP (diagnostic principal) : c'est le diagnostic majeur, à l'origine de l'hospitalisation
- DR (diagnostic relié) : lorsque le DP est plus un motif de recours aux soins qu'une réelle maladie, le DR est la pathologie qui a nécessité ce recours aux soins (par

exemple : type de cancer (DR) ayant nécessité une séance de chimiothérapie (DP))

- DAS (diagnostic associé significatif) : toutes les pathologies dont souffre le patient et qui auront eu un impact sur son hospitalisation (le plus souvent en rendant sa prise en charge plus lourde)

Les unités médicales (UM) sont les plus petites entités utilisées pour situer un patient au sein d'un établissement de santé. Il peut s'agir par exemple de l'unité d'hospitalisation conventionnelle en cardiologie (1410) ou de l'USI de cardiologie (3710). Elles sont ensuite regroupées en services, dans notre exemple le service de cardiologie. Une difficulté technique connue du PMSI est particulièrement attendue dans notre situation : celle des hospitalisations avec transferts d'un site géographique de l'établissement à un autre. Au CHU de Nantes, jusqu'en 2016, lorsqu'un patient passait d'un site à un autre au sein du CHU, ces mouvements étaient tous considérés comme des mutations (changement d'une UM à une autre au sein d'un séjour unique). Au 1er janvier 2016, afin d'harmoniser les pratiques entre les établissements multi-sites, les règles de facturation ont été modifiées. Les mouvements entre sites géographiques sont désormais considérés comme des transferts vers un autre établissement hospitalier et entrainent l'ouverture d'un second séjour (ainsi que la facturation du premier). Si cette décision a permis d'harmoniser la facturation à travers le territoire, elle entraine cependant une perte de continuité des informations lors de ces transferts. Ces hospitalisations multi-séjours ne bénéficient pas d'une identification unique et fragmente ainsi la base de données. Pour cette raison, nous utiliserons ici le terme « hospitalisation » plutôt que « séjour » qui est habituellement dédié, afin de ne pas mélanger ces 2 notions.

Pour pallier ce problème, nous avons fait le choix d'extraire les informations de la base PMSI de façon très large pour les filtrer ensuite :

- Dans un premier temps, en partant des IPP et des dates d'intervention de l'extraction QBloc, nous avons extrait tous les séjours d'un IPP sur l'année de l'intervention, mais également sur les années immédiatement précédentes et suivantes (3 ans au total).
- Ensuite, nous avons relié entre eux les séjours consécutifs lorsque la date de sortie de l'un était égale à la date d'entrée de l'autre, quelle que soit l'heure. Un identifiant unique d'hospitalisation était attribué à ces séries comme aux hospitalisations mono-séjour, permettant d'avoir un dénominateur plus proche de la réalité et cohérent avec notre approche.

- Enfin, nous avons éliminé toutes les hospitalisations pour lesquelles il n'y avait pas d'intervention en CTCV. Nous avons également éliminé les séjours se poursuivant en 2018 puisqu'à ce moment-là il était possible que le patient soit toujours hospitalisé et que son séjour ne soit pas encore codé, ou qu'il y ait du retard dans le codage de son séjour récent.

Nous avons également effectué des adaptations au niveau des RUM :

- Le bloc opératoire n'est pas considéré comme une unité d'hospitalisation, le passage du patient ne déclenche donc ni mouvement administratif, ni codage d'un RUM. Le patient est considéré comme hospitalisé dans son unité d'origine. Comme nous avions besoin d'insérer le bloc dans le parcours des patients au même titre que les UM, nous avons créé une UM factice et utilisé les heures d'entrée et de sortie de bloc données par Qbloc comme bornes de ce nouveau RUM. Si le patient retournait dans son UM d'origine après l'intervention (sortie de bloc antérieur à la sortie d'UM), nous avons scindé le RUM d'hospitalisation en 2 RUM. Si le patient était muté vers une autre UM après son intervention, le RUM d'origine était amputé de la durée de présence au bloc.
- Les RUM de moins d'une heure ont été supprimés car il s'agissait dans la grande majorité des cas d'erreurs de saisie corrigées dans la foulée (mais enregistrées par le logiciel de suivi administratif).

Les différentes modifications de codage des parcours (création des identifiants par hospitalisation, réorganisation des RUM) sont illustrées en Figure 5 sur un parcours fictif.

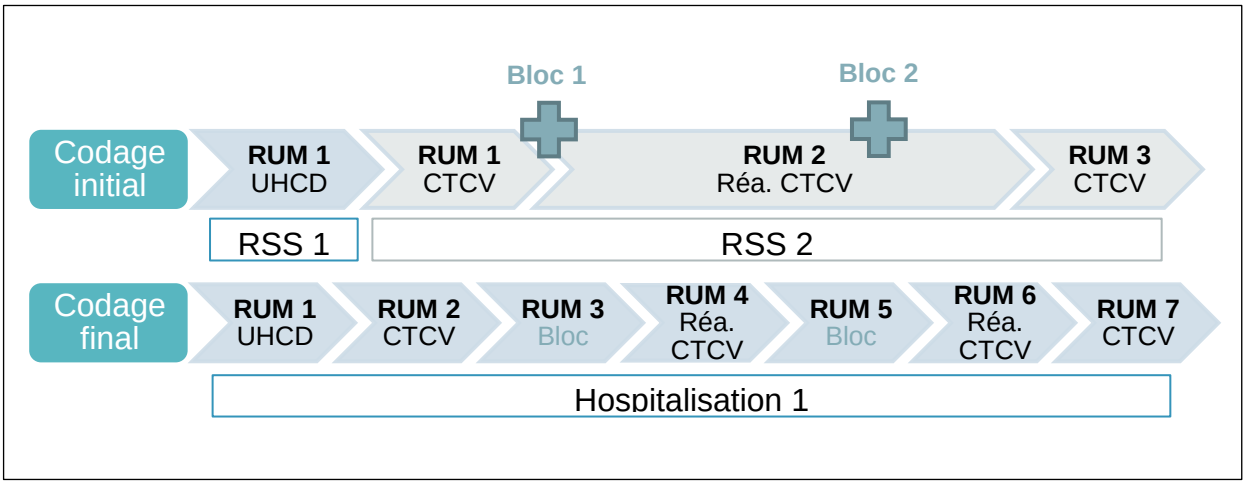



Figure 5 : Exemple de recodage des RUM sur un séjour fictif

### b. Sélection des hospitalisations

Le jeu de données obtenu après mise en forme des hospitalisations concernait toute la filière CTCV, incluant la chirurgie cardiaque, la chirurgie pulmonaire et thoracique, et les transplantations. Afin de nous focaliser sur la chirurgie cardiaque, un filtrage a été effectué sur la base du libellé des interventions chirurgicales (base QBloc). Ce filtrage a été réalisé par traitement automatique du langage (Tableau 3), classant les interventions en 4 catégories : transplantation thoracique, chirurgie cardiaque, chirurgie pulmonaire et thoracique, et autres interventions non classables.

| Catégorie de l'intervention | Expression régulière appliquée |
|---|---|
| Transplantation thoracique | "greff" |
| Chirurgie cardiaque | "coeur\|cardi\|pc\|valv\|péricard\|épicard\|ecmo\|cia\|aort\|tamponade\|heart\|tirone\|bentall\|pace-maker\|défibrillateur\|myxom\|coronar\|tsa\|rva\|ross\|maze\|vd/ap\|cec\|vg\|ablation sonde" |
| Chirurgie pulmonaire et thoracique | "pneum\|pulm\|poumon\|lobect\|xérèse atypique\|pariétectomie\|trach\|pleura\|thymectomie" |

Tableau 3 : Expressions régulières utilisées pour le filtrage des interventions chirurgicales

Les interventions ont été classées de façon séquentielle : tout d'abord les transplantations thoraciques, puis la chirurgie cardiaque et enfin la chirurgie pulmonaire et thoracique. Ainsi, celles qui étaient déjà classées comme « transplantations thoraciques » ne pouvaient pas être reclassées en chirurgie cardiaque ou pulmonaire, tout comme les interventions de chirurgie cardiaque ne pouvaient pas être reclassées en chirurgie pulmonaire (nous évitions ainsi de perdre tous les remplacements valvulaires pulmonaires par exemple). Les interventions dont l'intitulé ne présentait aucun des mots-clés étaient considérées comme non-classables et ont été étudiées afin de vérifier la perte d'information à ce stade.

La classification des séjours a été réalisée avec le même ordre de priorité que pour les interventions : transplantation thoracique, puis chirurgie cardiaque, puis chirurgie pulmonaire et thoracique, dès lors qu'une des interventions du séjour était classée dans

l'une de ces catégories. Ceci nous a permis de classer les hospitalisations dans un groupe précis, même lorsque plusieurs interventions chirurgicales avaient eu lieu (par exemple, des reprises chirurgicales cardiaque après une greffe). Finalement, seules les hospitalisations classées comme « chirurgie cardiaque » ont été retenues.

### c. Variables cliniques associées

#### *Choix des variables*

Comme précédemment énoncé, nous avons extrait de la base QBloc l'IPP, les dates et heures d'entrée et de sortie de bloc, et l'intervention réalisée pour chaque patient opéré en salle de CTCV. Ces informations nous ont principalement servi à réaliser un croisement avec la base PMSI et de filtrer les patients concernés par la chirurgie cardiaque.

C'est de cette base PMSI que nous avons pu récupérer des informations supplémentaires sur les patients. Même si elle est moins complète qu'un dossier médical, elle a l'avantage d'être facilement accessible, mieux organisée et plus homogène dans les informations transcrites (pas de champs de texte libre, il ne s'agit que de codes issus de classifications). Nous en avons donc extrait des informations :

- Administratives :
    - personnelles : âge, sexe, département de résidence, modes et dates d'entrée / de sortie
    - techniques : numéro de RUM et de RSS, indicateur de séjour principal, GHM, provenance et destination des mouvements
- Médicales : diagnostic principal, score de gravité

La codification des différents mouvements (entrée ou sortie) est synthétisée en annexe 3,

Pour ce qui est de la gravité, nous avons utilisé 2 types d'information :

- Le niveau de sévérité du séjour : c'est le dernier caractère du code de GHM et il est influencé principalement par l'âge, les diagnostics et l'éventuel décès du patient. Il est codé par un chiffre compris entre 1 et 4 représentant des niveaux de sévérité croissants. Il peut également s'agir d'une lettre pour des séjours courts et spécifiques (par exemple : J pour l'ambulatoire, T pour les séjours de très courte durée), voir l'annexe 4,

- Le score IGS2 : C'est un score de réanimation calculé sur différentes variables relevées au cours des 24 heures suivant l'admission dans l'unité de soins, telles que des mesures cardio-vasculaires ou biologiques.[3] Il permet de refléter l'état du patient de façon globale à son entrée dans un service de réanimation (éventuellement de soins intensifs).

Ces 2 scores permettaient d'avoir une idée de l'état général du patient et de la gravité de la situation, notions qui ne seraient pas évaluables par l'intermédiaire des données administratives ou des diagnostics seuls, mais qui le seraient relativement aisément une fois le patient en face d'un professionnel.

Afin de limiter le risque de confusion avec des complications graves en cours de séjour, nous avons restreint la récupération des scores de gravité à ce qui était présent à l'arrivée du patient. Le score IGS2 relevé correspondait à ce qui avait été coté soit dans le premier RUM, soit, s'il était absent ici, dans le deuxième si le premier durait moins de 7 jours. Le niveau de sévérité était celui du RUM principal[4] dans les cas où il s'agissait d'un séjour multi-RUM. Dans le cas d'une hospitalisation multi-séjours, si le niveau de sévérité du premier était peu informatif (codage par une lettre), nous prenions celui du séjour suivant. La même logique était suivie pour récupérer le DP, souvent codé de façon plus précise pour les séjours plus longs. Celui-ci étant obligatoirement présent à l'entrée du patient, les complications en cours de séjour ne l'impactaient pas.

Toutes les données administratives et médicales ont été enregistrées de façon simplifiées (durées à la place des dates-heures, département au lieu du code postal de résidence) et pseudonymisées (numéro d'identification propre à cette étude uniquement, absence de données personnelles clairement identifiantes).

### *Extraction et remodelage des données cliniques*

Les informations ont toutes été extraites dans le même format que celui qu'elles avaient dans les bases PMSI, c'est-à-dire avec une granularité au niveau des RUM. Or notre

[3] Grille de cotation détaillée sur https://www.atih.sante.fr/grille-de-cotation-du-score-igs2

[4] Le RUM principal est le premier RUM pour lequel il existe un DP assez précis pour une facturation. Les séjours de très courte durée ou les DP de symptômes sont évités tant que possible. Pour plus de détails sur l'algorithme de choix, se référer au 11° manuel des GHM (point 7,2 du volume 1, page 111) : https://www.atih.sante.fr/sites/default/files/public/content/2708/volume_1.pdf

étude avait une granularité moins précise : au niveau des hospitalisations (uni- ou multi-séjours selon les cas). Des adaptations et des simplifications ont donc été nécessaires.

Afin de récupérer l'information du motif d'entrée en hospitalisation du patient, nous avons extrait le DP du premier RUM. Lorsque le premier RUM était un séjour court (code de sévérité du GHM en « T » ou en « J »), c'est plutôt le DP du deuxième RUM qui a été conservé car celui-ci était généralement plus précis et informatif. Le codage du DP a été enregistré de façon intégrale (sous-catégorie de la CIM10) mais également de façon simplifiée ou thématique (catégorie seule, groupe et chapitre). Nous avons procédé avec la même logique pour le niveau de sévérité du séjour à l'arrivée du patient (premier RUM sauf si GHM en T ou en J).

L'information sur la gravité avait déjà été synthétisée à l'étape d'extraction des données, à savoir :

- Récupération du score IGS2 du premier RUM s'il y en avait un, sinon du deuxième RUM, mais uniquement si le premier RUM avait duré au maximum 7 jours
- Sévérité du séjour principal, sauf s'il s'agissait d'une hospitalisation multi-séjours et que le premier était très court, auquel cas c'était le RUM principal du second séjour qui était pris en compte

Les variables continues (âge et score IGS2) ont été recodées en classes. Pour le cas du score IGS2, ceci nous a également permis de créer une classe « NA » pour les patients n'ayant pas bénéficié de cette évaluation car n'ayant pas nécessité un passage en réanimation en début d'hospitalisation. Il était donc préférable de de considérer l'absence de score comme information à exploiter plutôt que comme donnée manquante. De plus, devant l'éventualité d'utiliser des modèles statistiques ne prenant pas en compte les données manquantes, l'exclusion de ces patients non graves à l'entrée aurait entrainé un biais de sélection dans nos analyses.

Les variables cliniques issues des bases PMSI sont récapitulées dans le Tableau 4.

| Catégorie | Variables |
|---|---|
| **Administratif** | sexe, âge (continu), âge_2 (catégoriel) |
| | département |
| | mode entrée PMSI, mode sortie PMSI |
| **Médical** | sévérité du séjour |
| | IGS2 (continu), IGS2_2 (catégoriel) |
| | DP, DP_cat (catégorie), DP_gp (groupe) |

Tableau 4 : Listes des variables concernant les données cliniques

## 2. Représentation d'un parcours

### a. Modèle de données

Nous avons fait le choix de modéliser nos séquences sous la forme de suites d'états. Appliqué à notre contexte, chaque UM correspondait à un état distinctif et chaque séjour dans cette UM constituait un élément de la séquence (équivalent d'un RUM en PMSI). Un parcours était donc défini comme l'enchainement de séjours successifs d'un patient dans différentes UM, d'une durée individuelle mesurée, de l'entrée dans l'établissement jusqu'à la sortie.

L'unité de temps choisie pour cette étude était l'heure. Cette décision alourdit nettement la base de données et complexifie les analyses mais, comme certains séjours pouvaient être très courts (en particulier les passages au bloc opératoire), ceux-ci auraient fortement altéré la qualité des données, quelle que soit la stratégie adoptée pour les traiter (suppression des séjours courts, arrondissement au jour entrainant un allongement artificiel des séjours…).[5]

### b. Représentation graphique

Tel que décrit dans la première partie, il existe de multiples façons de représenter graphiquement des séquences. Celles qui ont été utilisées dans ces travaux[6] (11) sont :

[5] Un jeu de données simplifié a été élaboré pour permettre la lecture des représentations graphiques. Dans ce jeu, l'unité de temps était le jour, les UM les moins fréquentées (moins de 5% de la population CTCV) étaient regroupées et les séjours étaient tronqués à 3 mois.

[6] Les résultats présentés en annexe 9 pourront servir d'exemple illustratif.

- « Sequence frequency plot » (catégorie longitudinale) : diagramme en barres horizontal représentant les séquences dans leur longueur. Les séquences identiques sont regroupées dans une barre unique dont la largeur est proportionnelle à la proportion de l'effectif total qu'elle représente. Les séquences les plus fréquentes sont en bas du graphique, le haut étant constitué des séquences les moins rencontrées, voire uniques. La totalité de la population est représentée et les barres s'interrompent lorsque les patients sont sortis.
- « State distribution plot » (catégorie transversale) : diagramme en barres vertical représentant la distribution de la population présente entre les différentes UM à chaque unité de temps. L'information est donnée de façon proportionnelle, il s'agit donc de pourcentages sur la population journalière ; une vigilance doit donc être maintenue pour l'interprétation des résultats les plus à droite, ceux-ci n'étant représentatifs que de la population encore hospitalisée à ce moment et non la population de départ.
- « Mean time plot » (catégorie agrégée) : histogramme du temps moyen passé dans chacune des UM par la totalité de la population.

## 3. Partitionnement des données

### a. Calcul des distances

*Algorithme de calcul*

Afin de rassembler les parcours similaires, il est nécessaire de calculer une mesure de dissimilarité. Il est possible pour cela de se baser sur les algorithmes élaborés par Hamming puis Levenshtein dans les années 50-60, permettant de comparer 2 chaines de caractères et de quantifier leurs différences. La distance de Levenshtein (ou distance d'édition) est un des plus connus de ces algorithmes. Elle consiste à compter le nombre de caractères qui doivent être supprimés, insérés ou substitués pour passer d'une chaine à l'autre ; la distance représente le coût minimal pour effectuer ce passage. L'ensemble des caractères utilisés est appelé un alphabet. Cet algorithme permet de calculer une matrice de distances comparant toutes les séquences 2 à 2 sur laquelle il est possible de se baser pour des analyses ultérieures.

Dans notre situation, on pouvait considérer un parcours comme une suite de caractères qui représentent de façon quotidienne l'UM où un patient était pris en charge, son hospitalisation constituant ainsi une chaine de caractères propre. L'alphabet était ainsi

l'ensemble des UM du centre hospitalier, et chaque chaine avait un nombre de caractères égal à la durée de séjour. Nous pouvions alors comparer les hospitalisations 2 à 2 en calculant des distances d'édition.

Reprenons le parcours patient fictif utilisé précédemment. Ce patient P1 est resté une journée dans l'UM A, est ensuite passé au bloc opératoire, puis deux jours dans l'UM B et enfin 5 jours à nouveau dans l'UM A. La séquence de Levenshtein de ce parcours est représentée dans la figure 6. Comparons maintenant ce parcours à un second parcours fictif, celui du patient P2, également présenté dans la figure 6, et calculons la distance de Levenshtein. Pour passer du premier au second parcours, il faut réaliser :

- Une substitution de A par C en position 1 (1 point)
- Une insertion de C en position 2 (1 point)
- Une suppression de A en position 10 (1 point)
    - ⇨ La distance entre les deux parcours est de 3 points.

Notons que nous également pu obtenir le second parcours en réalisant une substitution sur les positions 1, 2, 3 et 5 (4 points), mais la distance de Levenshtein est toujours la distance présentant un coût minimal. Le décalage d'une séquence par rapport à l'autre ne coûte pas plus que l'insertion ou la délétion de l'élément responsable de ce décalage.

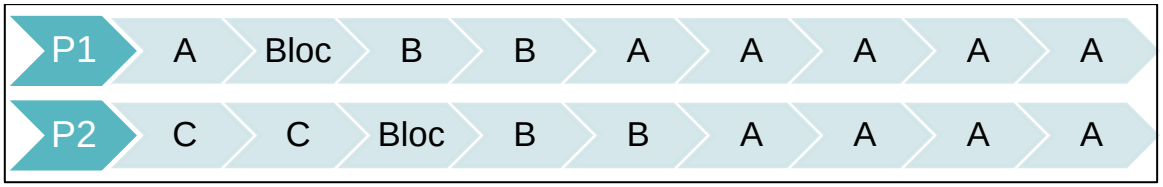



Figure 6 : Représentation du parcours de 2 patients fictifs adaptée à l'application de la méthode de Levenshtein

Ce calcul est ainsi réitéré pour comparer tous les patients entre eux et obtenir une matrice de distance telle que celle donnée dans la Figure 7. Celle-ci est toujours symétrique avec une diagonale à 0.

Des algorithmes de mesure de distance plus élaborés sont dérivés de cette mesure de Levenshtein (17), permettant notamment de pondérer plus finement ces transitions. Nous avons choisi d'utiliser l'une de ces variantes, construites initialement par les sciences sociales pour l'analyse des trajectoires de vie : l'appariement optimal (ou OM : *Optimal Matching*). Cette mesure repose toujours sur les 2 mêmes coûts de base : les indels

(insertion-délétions) et les substitutions. Dans cette approche, ceux-ci peuvent avoir des valeurs variables selon une matrice de coûts liée aux différents éléments de l'alphabet concerné : il est ainsi possible d'assigner un poids plus haut à la suppression d'un élément qu'à la suppression d'un autre, ou sur les transitions entre 2 éléments plutôt que 2 autres.

| | P1 | P2 | P3 | Pi |
|---|---|---|---|---|
| P1 | 0 | 3 | 1 | … |
| P2 | 3 | 0 | 4 | … |
| P3 | 1 | 4 | 0 | … |
| Pi | … | … | … | 0 |

Figure 7 : Exemple de matrice de distance obtenue par la méthode de Levenshtein à partir de patients fictifs

L'OM reste de manière générale très sensible aux durées et a montré ses limites pour la comparaison de séquences longues. Une nouvelle variante, l'OMSpell, modifie son comportement en ne considérant plus les séquences par unités de temps mais par « segments de séquence » dans un état distinct et de durée variable. Cette façon de penser et représenter les parcours est illustrée en reprenant notre exemple des 2 patients fictifs en Figure 8. Cette méthode permet de pondérer plus finement le calcul de distance en ajustant le coût des indels à la longueur des segments concernés et en limitant le coût de dilatation/compression de segments. Elle favorise la dilatation/compression à la substitution et minimise ainsi la distance entre des séquences qui suivent le même enchainement d'états mais dans une temporalité variable. Elle évite également le découpage de segments longs tel qu'a tendance à faire un OM classique.

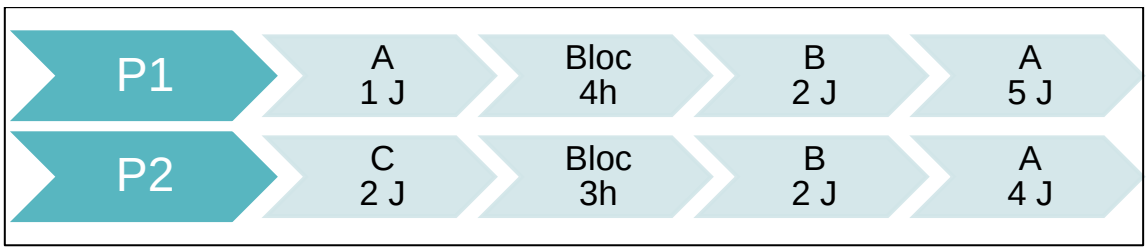



Figure 8 : Représentation du parcours de 2 patients fictifs adaptée à l'application de la méthode OMSpell

C’est ainsi ce dernier algorithme, OMSpell, qui a été choisi pour notre étude. Son comportement se prêtait bien à notre contexte : il était plus intéressant de savoir si un patient avait fait un détour par un autre service que de savoir s’il avait été hospitalisé quelques heures de plus que le second patient auquel il était comparé. La matrice de coûts de substitution nous permettait également de hiérarchiser les modifications de parcours selon leur intérêt clinique : il y avait plus de sens à savoir qu’un patient avait fait un séjour en réanimation alors qu’un autre non, plutôt que de savoir que l’un avait été hospitalisé dans l’aile A et l’autre dans l’aile B d’un même service d’hospitalisation conventionnelle. Le fonctionnement précis et les différents paramètres à régler sont détaillés en annexe 5, Comme nous n’avions aucune étude préexistante sur laquelle nous baser, plusieurs paramétrages ont été expérimentés en parallèle afin de tester leurs effets.

Le résultat de l’application de l’OMSpell reste le même que pour les algorithmes précités, à savoir une matrice de distance où tous les éléments sont comparés les uns aux autres. Il est ensuite possible de visualiser ces distances afin de comparer le comportement des algorithmes de clustering et les résultats obtenus.

### *Visualisation des distances*

Les matrices de distances pourraient être utilisées directement (en transposant ces distances théoriques en distances métriques) mais, dans le cadre de nos données de grandes dimensions, ceci aurait abouti à une représentation hautement multidimensionnelle, inexploitable par l’œil humain. Les approches vectorielles permettent de simplifier cette situation.

Cousine de la classique analyse en composantes principales (ACP), l’analyse en coordonnées principales (ACoP) permet de réduire la dimensionnalité de nos données en calculant des vecteurs qui synthétisent l’information. Il devient ainsi possible de tracer un nuage de points par projection en 2 dimensions, en utilisant 2 des premiers vecteurs comme axes du graphique.

Dans notre situation, l’objectif de cette projection était d’aider à la compréhension des partitionnements et à l’évaluation visuelle des principales tendances, de la « forme » de notre population, son homogénéité ou non, etc. Nous n’avons pas cherché à interpréter les résultats chiffrés de ces ACoP.

## b. Clustering

### *Algorithme de classification*

Après avoir calculé la similarité entre les parcours, l'étape suivante était de regrouper ceux qui étaient les plus proches. La première approche intuitive est généralement de considérer que plus deux parcours sont proches, plus ils sont susceptibles de faire partie du même groupe. C'est sur ce principe qu'est basé l'algorithme de classification ascendante hiérarchique (CAH).

De façon simplifiée, l'algorithme de CAH est le suivant :

- Initialement, chaque sujet constitue un groupe individuel,
- Puis, les groupes les plus proches (là où la mesure de dissimilarité est la plus basse) sont rassemblés en un nouveau groupe plus grand,
- Les groupes sont ainsi rassemblés jusqu'à l'obtention d'un seul et unique groupe.

À chaque étape, l'inertie[7] inter-classe et intra-classe est calculée. Le nombre optimal de groupe est déterminé en identifiant l'étape pour laquelle l'inertie intra-classe est minimisée et l'inertie inter-classe est maximisée. Afin que le résultat reste interprétable « humainement », nous avons ajouté un critère de choix : ne pas dépasser 15 groupes.

### *Sélection d'un clustering*

La qualité d'une classification peut être définie et évaluée de plusieurs façons. Il y a d'une part des indicateurs statistiques permettant de caractériser la forme (« Est-ce que les clusters sont bien homogènes ? Bien séparés des autres ? ») et d'autre part une approche plus empirique pour évaluer le fond (« Est-ce que la classification à un sens clinique ? Est-ce que diviser la population de cette façon a un intérêt ? »). Nous avons donc procédé à une sélection en 2 étapes :

- Une première pré-sélection sur des indicateurs statistiques pour choisir les clusterings les plus clairement définis (la forme),
- Puis une sélection finale par une évaluation qualitative pour choisir ceux qui ont la plus grande valeur médicale (le fond).

[7] L'inertie d'un nuage de points est la somme des carrés des distances des points au centre de gravité. Elle représente donc la « densité » du nuage de points, sa « concentration ». L'inertie inter-classe est donc le reflet de l'écartement entre les groupes et l'inertie intra-classe la proximité des individus au sein du groupe.

*Indicateurs de qualité*

De nombreux indicateurs existent pour décrire une classification. Nous en avons choisi 3 fréquemment utilisés, permettant de caractériser de façon complémentaire nos résultats :

- **Indice de Dunn :** moyenne du rapport entre la distance maximale qui sépare deux éléments classés ensemble et la distance minimale qui sépare deux éléments classés séparément. Il décrit la séparation entre les clusters.
- **Indice de Davies-Bouldin :** moyenne du rapport maximal entre la distance d'un point au centre de son cluster et la distance entre deux centres de cluster. Il décrit la similarité entre clusters voisins.
- **Score de silhouette :** moyenne de la différence entre la distance moyenne d'un point avec les points d'un même cluster et la distance moyenne avec les points des clusters voisins. Il décrit, pour un point, la similarité à son cluster attribué plutôt qu'aux autres clusters.

Une bonne classification aura un indice de Dunn haut (séparations nettes), un indice de Davies-Bouldin bas (clusters distants les uns des autres) et un score de silhouette haut (classification robuste).

Afin de synthétiser ces résultats, nous avons calculé un score final à partir des 3 scores précédemment cités en les normalisant par rapports aux résultats obtenus pour les autres classifications, puis en additionnant les scores à maximiser et retirant le score à minimiser. Nous avons obtenu ainsi un score global permettant de trier nos classifications sur un score unique. Le clustering obtenant le score global le plus haut parmi le groupe de ceux ayant le même nombre de clusters était sélectionné. Si le meilleur score d'un groupe était inférieur à 0, aucun clustering n'était conservé à cette étape.

Nous avons ensuite comparé ce score global entre classifications proches. En triant les clusterings par ordre croissant sur le nombre de clusters identifiés, il est mécanique que le score global diminue petit à petit (plus il y a de clusters dans un espace fini, plus la distance moyenne entre eux va diminuer). Des irrégularités dans cette progression peuvent toutefois survenir en cas de clustering particulièrement bien défini ou, au contraire, particulièrement bruyant. Nous avons donc écarté les clusterings pour lesquels le score global de qualité était moins bon que celui du clustering suivant.

### *Analyse qualitative des clusters*

Une fois que cette première étape de pré-sélection sur des bases statistiques a été effectuée, nous avons analysé le contenu de chaque cluster des différents clusterings retenus. Plusieurs critères entraient compte :

- Effectif du plus petit cluster : la CAH ne permettait pas de fixer un nombre minimal d'individus par groupe a priori, mais un groupe trop anecdotique aurait déséquilibré la suite de nos analyses et n'aurait pas eu d'intérêt clinique à l'échelle d'un centre hospitalier. Il était donc nécessaire d'exclure les classifications concernées.
- Lisibilité clinique des représentations graphiques des séquences de chaque cluster : il s'agissait ici d'une analyse subjective, basée sur des questions telles que : est-ce qu'un motif commun ressort ? Est-ce que ce motif est distinctif de celui des autres clusters ? Est-ce que ce cluster semble homogène ?

Pour l'évaluation du second critère, nous avons procédé à une évaluation visuelle de chaque clusters au sein des classifications retenues sur différents graphiques : la projection de la répartition des clusters sur une ACoP de la matrice de distance[8] et la représentation graphique des parcours sur le « Sequence frequency plot », le « State distribution plot » et le « Mean time plot »[9]. Plus les parcours au sein des groupes semblaient homogènes et avec des caractéristiques identifiables, plus les différences avec les parcours des autres clusters étaient claires, meilleure était l'évaluation. Une échelle ordinale à 5 modalités a été utilisée pour noter ce critère.

Il existe des indicateurs dérivant de l'analyse de séquence qui pouvaient approcher ces dernières questions (calcul d'entropie par exemple), mais leur calcul restait peu robuste devant la taille très importante de notre alphabet. De plus, nous avions la volonté de conserver un œil médical sur ces analyses exploratoires afin de les garder ancrées dans la pratique, d'où le choix de cette approche empirique à un stade où le nombre de possibilités devenait abordable pour un choix humain basé sur des critères de nature variée.

Nous avons fait le choix de conserver les 2 meilleures classifications afin de pouvoir comparer l'effet du paramétrage de calcul des distances et du nombre de groupes retenus.

[8] Description méthodologique dans la section précédente, page 27 : Visualisation des distances.

[9] Description méthodologique dans la section précédente page 23 : Représentation graphique.

## 4. Analyses descriptives

### a. Description des parcours

Dans la suite de l'évaluation qualitative des partitionnements, nous avons approfondi l'analyse descriptive des groupes de parcours un par un.

Concernant les indicateurs chiffrés, nous avons calculé au niveau de chaque cluster :

- Durée de séjour : moyenne et médiane
- Nombre de passage au bloc opératoire : moyenne et médiane
- Proportion du temps passé dans chaque UM

La dispersion était également calculée, à savoir l'écart-type pour les moyennes et l'espace interquartile pour les médianes.

### b. Description des données cliniques

De façon parallèle, nous avons décrit les données cliniques à la fois sur notre population entière et sur chacun des clusters de nos deux classifications.

Les variables qualitatives ont été représentées par des proportions et les variables quantitatives par une médiane et l'espace interquartile associé. Des histogrammes ont été élaborés pour représenter la dispersion en classes des variables quantitatives discrétisées (âge et score IGS2). Pour la description des DP, afin de simplifier la lecture des résultats, nous ne les avons décrits qu'au niveau du groupe de la CIM10,

Des tests bivariés non paramétriques ont été effectués sur les variables disponibles à l'entrée : âge, sexe, département de résidence, mode d'entrée, diagnostic principal, scores de gravité. Les variables continues ont été évaluées par un test de Kruskal-Wallis et les variables catégorielles par un test exact de Fisher. Le seuil de significativité était fixé à 0,05.

## 5. Matériel

Les logiciels suivants ont été utilisés pour mener ce travail :

- Extraction PMSI : SAS Enterprise Guide
- Mise en forme des données : Python 3,6 et R 3.5.1
- Analyses : R 3.5.1

Les principales bibliothèques R utilisés étaient :

- Analyses et graphiques de séquences : TraMineR
- Clustering : cluster, clustercrit
- Analyses vectorielles : ade4, FactoMineR

## B. Résultats

### 1. Bases de données

Un diagramme de flux de l'évolution de la population d'étude et du volume des jeux de données est disponible dans la Figure 9. L'extraction QBloc a retourné 3963 interventions (3264 patients) et l'extraction PMSI qui l'a suivie 20953 RUM, soit 10890 hospitalisations (après agrégation des hospitalisations multi-séjours). Après croisement sur les identifiants patients et premières modifications (recodage avec inclusion des passages au bloc comme RUM fictif, filtrage des aberrations), la base CTCV représentait 3332 hospitalisations.

Le filtre sur le libellé des interventions chirurgicales a permis de les répartir telles que :

- Transplantation thoracique : n = 95 interventions
- Chirurgie cardiaque : n = 3031 interventions
- Chirurgie pulmonaire et thoracique : n = 487 interventions
- Non classable : n = 326 interventions

Les interventions non-classées étaient majoritairement représentées par :

- Les reprises chirurgicales (n = 113) : « reprise hémostase » n = 51, « reprise cicatrice » n = 50, « reprise chirurgicale (fermeture sternale, décaillotage..) » n = 10, « reprise chirurgicale pour infection profonde du site opératoire » n = 2
- La mention « autres interventions » : n = 101
- « drainage » : n = 30
- Biopsie médiastin : n = 27
- « irrigation-lavage médiastinal » : n = 24

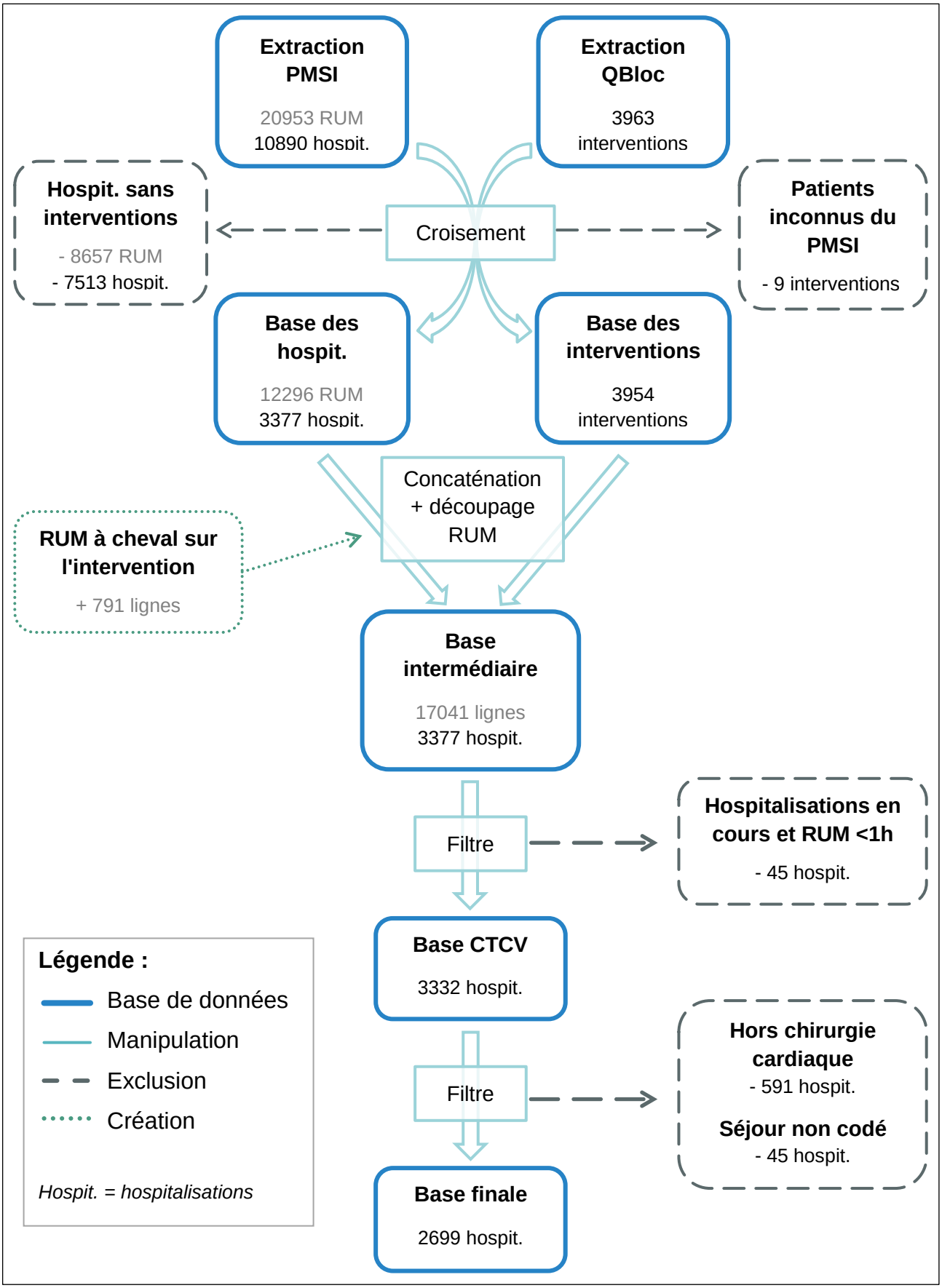



Figure 9 : Diagramme de flux de l'évolution de la population d'étude et des bases de données

Nous avons vérifié si les patients concernés par ces interventions non-classables étaient déjà inclus dans une catégorie d'intérêt grâce à une autre intervention, par exemple :

- « autres interventions » (n = 95 patients) : 14 en transplantation, 25 en chirurgie cardiaque, 6 en chirurgie pulmonaire et 50 non classés
- « reprise hémostase » (n = 45 patients) : 14 en transplantation, 26 en chirurgie cardiaque, 2 en chirurgie pulmonaire et 3 non classés
- « reprise cicatrice » (n = 43 patients) : 4 en transplantation, 29 en chirurgie cardiaque, 6 en chirurgie pulmonaire et 4 non classés
- « irrigation-lavage médiastinal » (n = 24 patients) : 1 en transplantation, 20 en chirurgie cardiaque, 1 en chirurgie pulmonaire et 2 non classés

Peu d'hospitalisations étaient donc perdues à cause de ces interventions non-classables. Quant à la catégorie « autres interventions » non-informative, il n'était de toute façon pas faisable de déduire la spécialité sur cette seule base, et il était possible qu'elle intègre des interventions non liées à la CTCV mais réalisées dans ces salles pour des raisons pratiques (par exemple : utilisation de la salle par une autre spécialité chirurgicale pour une urgence en l'absence d'une salle dédiée disponible).

A l'inverse, il restait également quelques interventions non exclues de la base finale ne relevant pas de la chirurgie cardiaque, telles que « trachéotomie » ou « drainage pleural », mais il s'agissait de patients ayant subi plusieurs interventions chirurgicales distinctes, dont au moins une concernait la chirurgie cardiaque, et ceux-ci représentaient une minorité des patients concernés par ces interventions (« trachéotomie » : 9/49 interventions, « drainage pleural » : 8/40 interventions). Ces patients ont été conservés dans notre population car concernés par la chirurgie cardiaque.

Au final, les 3031 interventions de chirurgie cardiaque concernaient 2689 patients. Après élimination de 44 patients pour lesquels une transplantation thoracique avait été réalisée lors d'une autre hospitalisation, nous n'avons conservé que 2645 patients sur les 3264 de la CTCV. Ceux-ci représentaient 2744 hospitalisations, parmi lesquelles 45 n'ont pas pu être associées à un séjour codé du côté PMSI. Nous avions donc au final 2602 patients et 2699 hospitalisations (2508 patients ont été hospitalisés une fois, 91 l'ont été 2 fois et 3 l'ont été 3 fois).

Un récapitulatif des 25 interventions les plus fréquentes dont ont bénéficié les patients intégrés dans la catégorie « chirurgie cardiaque » est disponible dans le Tableau 5 ; l'intégralité des résultats est en annexe 6.

Un récapitulatif de toutes les UM concernées par les hospitalisations sélectionnées est disponible en annexe 7 afin de rendre plus facile la lecture des futurs tableaux et graphiques. Les UM les plus souvent rencontrées sont récapitulées dans le Tableau 6.

| Intervention | Nombre |
|---|---|
| Remplacement valve aortique | 576 |
| PC 3 | 361 |
| Sup à PC 3 | 341 |
| Drainage / décaillotage péricardique | 121 |
| PC 2 | 113 |
| Plastie valve mitrale | 108 |
| Remplacement valve aortique + PC 1 | 82 |
| Tamponade | 73 |
| Remplacement valve mitrale | 56 |
| Remplacement valve aortique + PC 2 | 50 |
| Remplacement aorte ascendante | 48 |
| Drainage péricardique | 45 |
| Bentall | 44 |
| Valve aortique transapicale | 38 |
| ECMO / bio-médicus | 37 |
| Remplacement valve aortique + PC 3 | 35 |
| Reprise cicatrice | 35 |
| Reprise hémostase | 28 |
| Autres interventions | 26 |
| Drainage | 25 |
| Remplacement valve aortique + maze | 25 |
| Remplacement bi-valvulaire | 23 |
| Remplacement valve aortique + remplacement aorte ascendante | 23 |
| Ablation ECMO | 22 |
| Valve transcarotidienne | 22 |

Tableau 5 : Les 25 interventions les plus fréquentes dont ont bénéficié les patients du groupe "chirurgie cardiaque"

| Code UM | Libellé UM | Code serv. | Libellé service | Code pôle | Libellé pôle |
|---|---|---|---|---|---|
| 0000 | Bloc CTCV | | | | |
| 2670 | Chirurgie Cardiaque Congénitale Adulte (CCCA) | 2670 | Chirurgie Cardiaque Pédiatrique et Congénitale (CCPC) | 7090 | PHU 5 : Femme, Enfant, Adolescent[10] |
| 3740 | Pneumologie SI | 1310 | Pneumologie | 7610 | PHU 2 : Institut Du Thorax Et Du Système Nerveux |
| 1410 | Cardiologie | 1410 | Cardiologie | 7610 | PHU 2 |
| 3710 | Cardiologie SI | 1410 | Cardiologie | 7610 | PHU 2 |
| 3711 | Rythmologie SI | 1410 | Cardiologie | 7610 | PHU 2 |
| 2600 | CTCV | 2600 | CTCV | 7610 | PHU 2 |
| 3762 | USC Chirurgicale | 3850 | Anesthésie - réanimation chirurgicale HGRL | 7660 | PHU 3 : Médecines, Urgences et Soins Critiques |
| 3850 | Réanimation Chirurgicale Polyvalente HGRL | 3850 | Anesthésie - réanimation chirurgicale HGRL | 7660 | PHU 3 |
| 3870 | Réanimation CTCV | 3850 | Anesthésie - réanimation chirurgicale HGRL | 7660 | PHU 3 |

Tableau 6 : Cartographie des principales UM concernées par l'étude

[10] La chirurgie cardiaque congénitale est inclue dans le pôle Femme, Enfant, Adolescent mais n'est pas forcément liée à de la chirurgie pédiatrique. Les personnes concernées par ces pathologies sont suivies toute leur vie dans ce même service.

## 2. Clustering

Un diagramme de flux de l'évolution des bases de données est donné en Figure 10 pour toute l'étape de clustering, à la fois sur sa constitution et sur la sélection qui y a été appliquée par la suite.

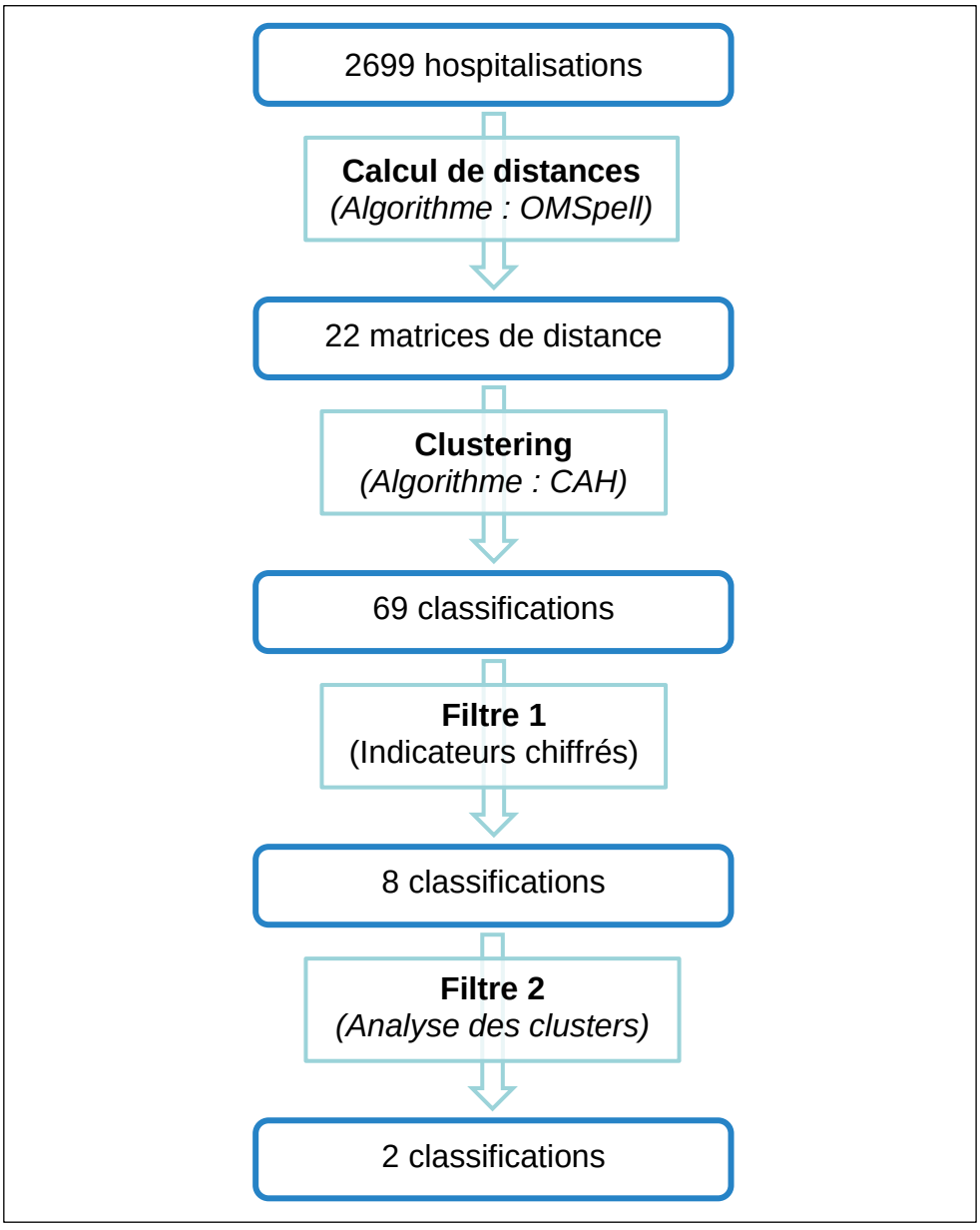



Figure 10 : Diagramme de flux des étapes de clustering

### a. Calcul des distances

Un total de 22 matrices de distances a été calculé, mais nous ne détaillerons pas ici les paramètres de chacune (le détail du fonctionnement de calcul est en annexe 5). Le

paramétrage des matrices de distance utilisées pour les 2 clusterings retenus est indiqué dans le Tableau 7. Le coût de substitution était uniforme pour la première matrice (fixé à 1, tel que dans le paramétrage par défaut) et était personnalisé pour la seconde matrice (matrice de coût individuel pour chaque unité vers une autre).

| Matrice de distance | Coût de substitution | Pondération temporelle | Coût de dilatation / compression |
|---|---|---|---|
| **MD1** | 1 | 0,8 | 0,2 |
| **MD2** | Matrice 3 | 0,5 | 0,04 |

Tableau 7 : Synthèse des paramètres utilisés pour les différents calculs de distance

Pour la matrice de coûts 3, les coûts étaient fixés par le produit des facteurs suivants : même service = 0,6, même type d'autorisation = 0,9 et UM hors CTCV = 0,7, Les coûts obtenus ainsi variaient de 1 à 0,378,

### b. Partitionnement

Partant de nos matrices de distances calculées à l'étape précédente, nous avons réalisé une CAH sur chacune d'entre elles. Plusieurs paliers ont été retenus pour chacune des CAH, aboutissant à 69 clusterings. Le calcul des indicateurs chiffrés de la qualité des classifications a été effectué sur chacun de ces partitionnements. Il est à noter que le calcul du score de Dunn a montré peu de variabilité mais que les autres étaient plus informatifs. Nous avons ensuite procédé au premier filtrage :

- Au-delà de 12 clusters, les clusterings avaient un score global inférieur à 0, aucun n'a donc été retenu,
- 11 clusterings ont été sélectionnés en prenant le meilleur score global de qualité pour chaque groupe de nombre de cluster,
- Parmi eux, 3 ont été écartés du fait d'un score global trop faible par rapport au clustering suivant dans le tri par ordre croissant de nombre de clusters.

Venait ensuite l'analyse à l'échelle des clusters. Parmi les 8 clusterings analysés, 2 avaient un cluster de moins de 10 individus et ont été directement écartés. La lecture des parcours de chaque cluster a enfin permis de retenir 2 classifications qui nous paraissaient intéressante.

Les résultats des scores et évaluations obtenus par nos 2 clusterings retenus sont récapitulés dans le Tableau 8. Le tableau complet de ces résultats pour les 8 classifications présélectionnées est en annexe 8, Les graphiques ayant servi de support à l'analyse de la lisibilité et aux analyses descriptives sont en annexe 9. Nous présentons ici seulement le state frequency plot car il est celui qui porte le plus d'informations (Figure 11 et Figure 12).

| Scores et indicateurs | Clustering C1 | Clustering C2 |
|---|---|---|
| **Nombre de clusters** | 4 | 6 |
| **Indice de Dunn** | 0,017 | 0,011 |
| **Indice de Davies-Bouldin** | 1,774 | 1,719 |
| **Coefficient de silhouette** | 0,231 | 0,231 |
| **Score global** | 2,24 | 1,75 |
| **Effectif du plus petit cluster** | 54 | 13 |
| **Lisibilité** | Bonne | Bonne |

Tableau 8 : Scores de qualité et indicateurs obtenus par les deux clusterings retenus

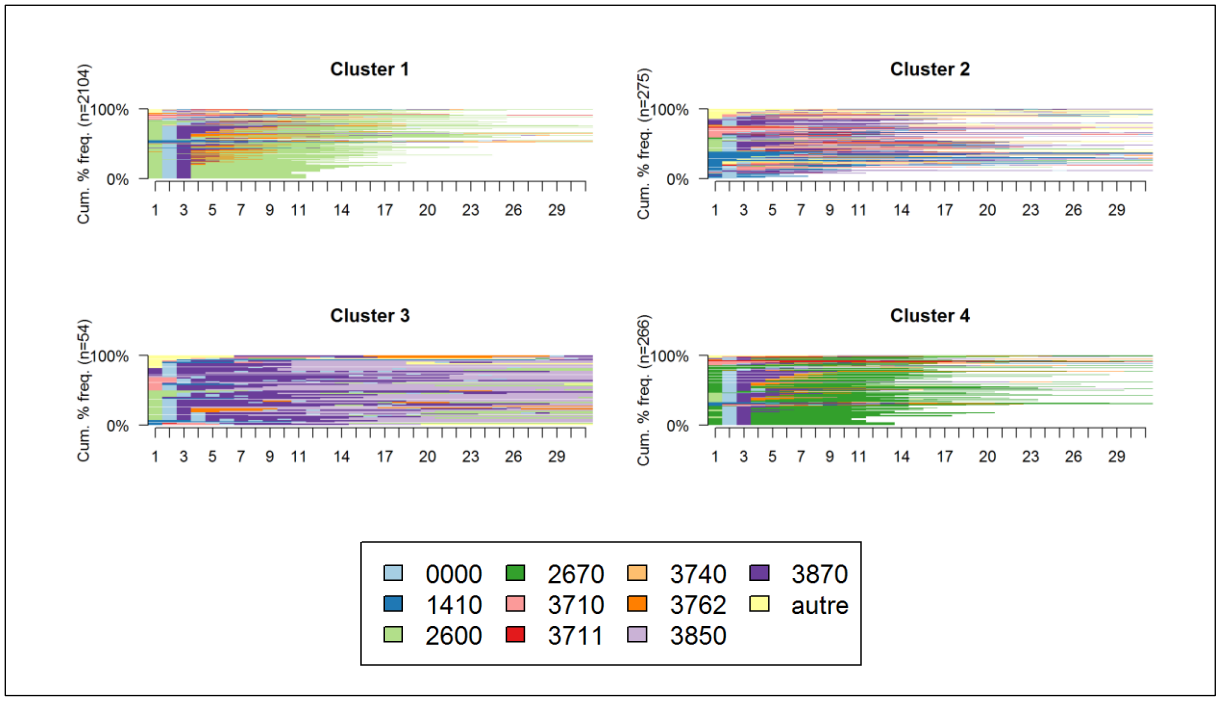



Figure 11 : State frequency plots associés aux clusters de la classification C1

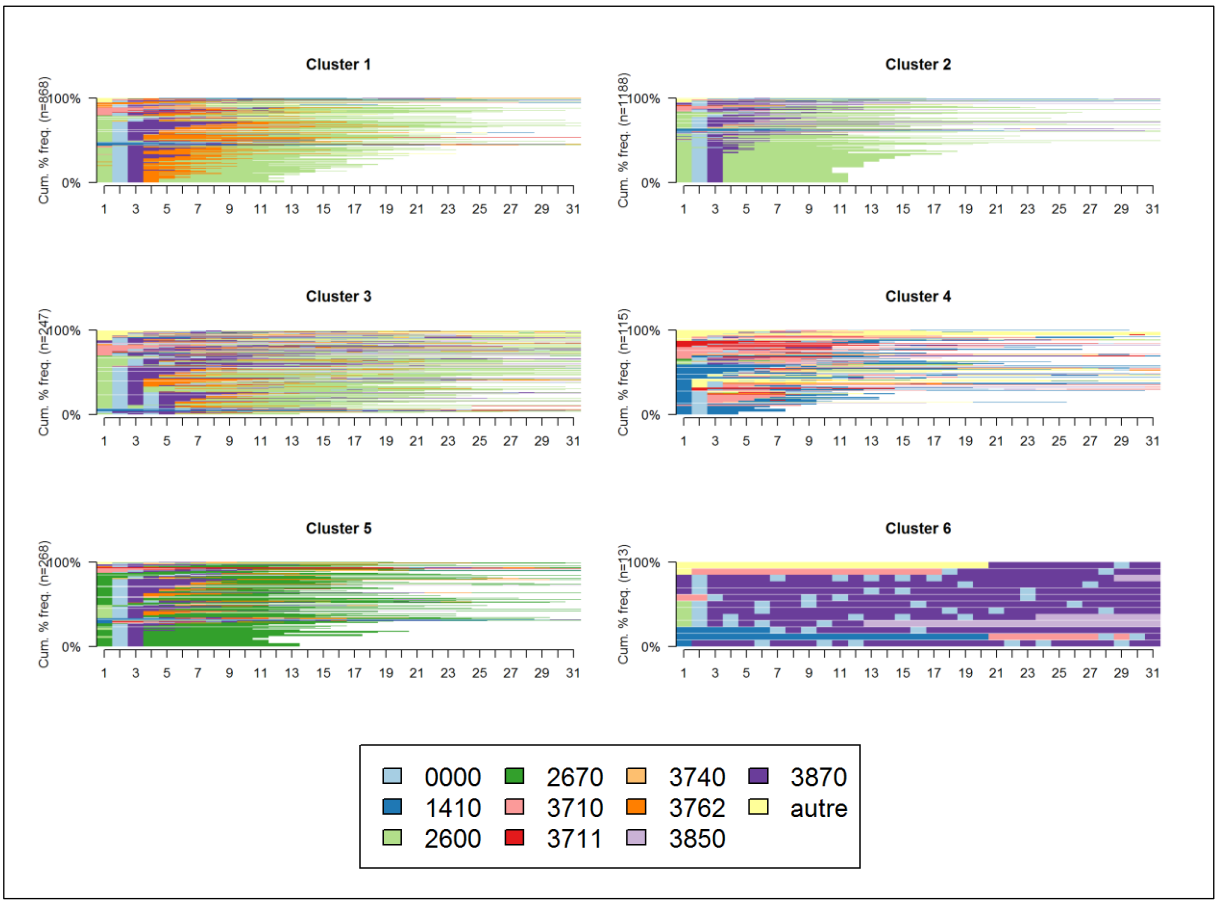



Figure 12 : State frequency plots associés aux clusters de la classification C2

Les classifications les plus grossières (à 2 ou 3 clusters) avaient le meilleur score global, mais n'ont pas été retenue sur cet argument seul car il apparaissait assez clairement que les clusters étaient encore composite. Si nous prenons l'exemple de la classification à 2 clusters illustrée par ses state frequency plots en Figure 13, il est visible que le premier cluster comporte à la fois des patients à la prise en charge simple et rapide (sortie en mois de 2 semaines) et d'autres qui séjournent plus longtemps en réanimation et / ou soins continus (UM 3870 et 3762). De même, dans le second cluster sont mélangés les patients suivis en CCCA (2670) et les patients venant de cardiologie (1410). Il n'a donc pas été jugé cliniquement opportun de poursuivre les analyses sur cette classification.

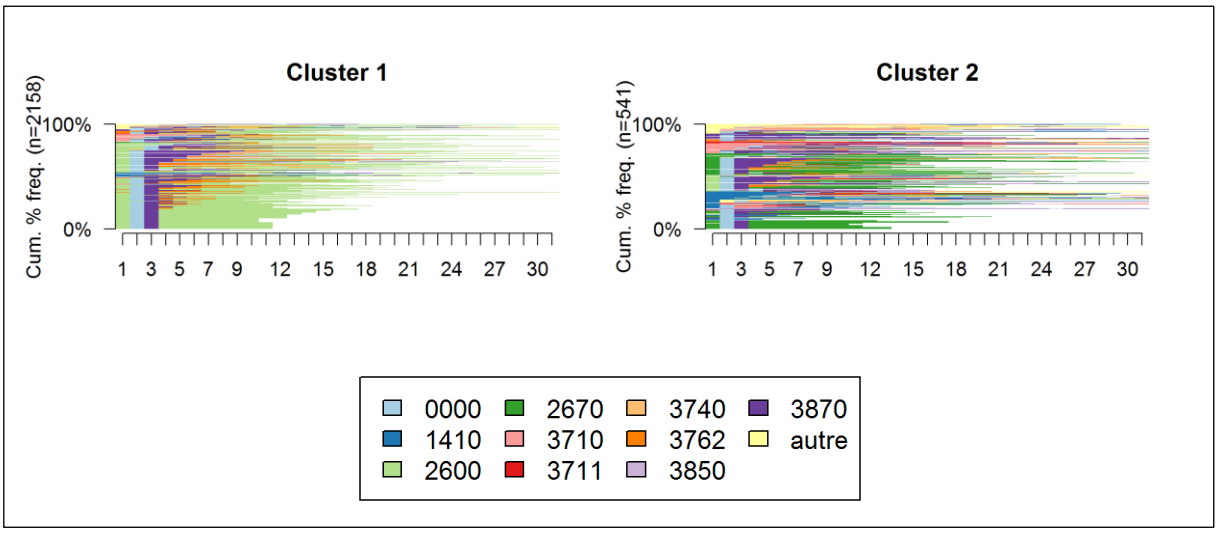



Figure 13 : State frequency plots associés aux 2 clusters de la classification retenue dans cette catégorie

Au-delà de leurs résultats aux évaluations, nous avons décidé de conserver les 2 classifications citées plus haut car elles présentaient des intérêts différents. D'un côté, C1 semblait avoir une ségrégation imparfaite avec des groupes encore composites mais laissait déjà entrevoir des clusters cohérents. De l'autre côté, C2 présentait des clusters de taille très variable, dont un particulièrement petit, risquant de rendre des analyses prédictives peu robustes, mais restait le meilleur choix parmi les classifications présentant plus de 4 groupes afin d'évaluer l'intérêt d'une ségrégation plus fine.

## 3. Analyses descriptives

### a. Parcours

La répartition du nombre de parcours de chaque cluster sur l'effectif global est représentée dans la Figure 14. La description des durées de séjour est donnée dans le Tableau 9 et le détail du temps passé dans chaque UM est en annexe 10, Le Tableau 10 récapitule les passages au bloc opératoire (c'est-à-dire le nombre de RUM dans l'UM 0000) pour les patients de chacun des clusters.

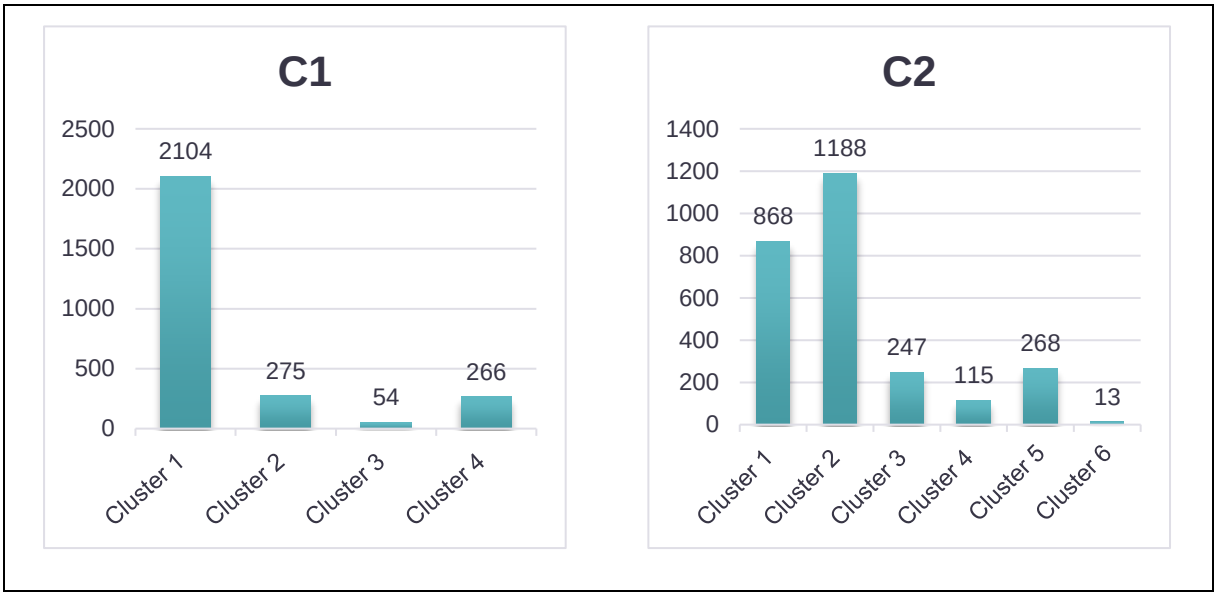



Figure 14 : Répartition des effectifs entre les différents clusters

| Durée de séjour (jours) : | | Moyenne (écart-type) | Médiane [EIQ] |
|---|---|---|---|
| **Clustering C1** | Cluster 1 | 15,3 (8,3) | 13,1 [10,0 ; 17,9] |
| | Cluster 2 | 21,8 (23,2) | 14,1 [7,3 ; 27,9] |
| | Cluster 3 | 58,2 (28,4) | 50,1 [40,0 ; 68,7] |
| | Cluster 4 | 16,1 (9,4) | 13,9 [10,9 ; 18,1] |
| **Clustering C2** | Cluster 1 | 16,5 (8,7) | 14,8 [11,1 ; 19,1] |
| | Cluster 2 | 13,7 (9,2) | 11,1 [9,0 ; 15,2] |
| | Cluster 3 | 30,5 (21,3) | 24,2 [17,5 ; 37,3] |
| | Cluster 4 | 18,8 (18,1) | 12,9 [7,4 ; 24,1] |
| | Cluster 5 | 15,9 (8,9) | 13,9 [10,8 ; 18,1] |
| | Cluster 6 | 78,9 (37,5) | 66,0 [50,4 ; 99,0] |

Tableau 9 : Moyenne et médiane de la durée de séjour totale pour chaque cluster des 2 classifications retenues

| Nombre de passages au bloc : | | Moyenne (écart-type) | Médiane [EIQ] |
|---|---|---|---|
| **Clustering C1** | Cluster 1 | 1,09 (0,30) | 1,0 [1,0 ; 1,0] |
| | Cluster 2 | 1,36 (0,87) | 1,0 [1,0 ; 1,0] |
| | Cluster 3 | 2,20 (1,48) | 2,0 [1,0 ; 3,0] |
| | Cluster 4 | 1,11 (0,37) | 1,0 [1,0 ; 1,0] |
| **Clustering C2** | Cluster 1 | 1,01 (0,10) | 1,0 [1,0 ; 1,0] |
| | Cluster 2 | 1,03 (0,16) | 1,0 [1,0 ; 1,0] |
| | Cluster 3 | 2,01 (0,78) | 2,0 [2,0 ; 2,0] |
| | Cluster 4 | 1,06 (0,24) | 1,0 [1,0 ; 1,0] |
| | Cluster 5 | 1,10 (0,32) | 1,0 [1,0 ; 1,0] |
| | Cluster 6 | 4,92 (0,76) | 5,0 [4,0 ; 5,0] |

Tableau 10 : Moyenne et médiane du nombre de passages au bloc pour chaque cluster des 2 classifications retenues

### b. Données cliniques

Pour des questions de lisibilité, nous ne présenterons ici qu'une partie des variables analysées, dans les Tableau 11 et Tableau 12. Les résultats complets sur l'échantillon entier et sur chacun des clusters des 2 classifications retenues sont en annexes 11-13,

| Variable | | Proportion (effectif) ou médiane [EIQ] |
|---|---|---|
| **Échantillon entier (n = 2699)** | | |
| **Sexe** | Homme | 73,5 (1984) |
| | Femme | 26,5 (715) |
| **Age** | | 69,0 [61,0 ; 76,0] |
| **IGS2** | | 24,0 [18,0 ; 30,0] |
| **Entrée PMSI** | Domicile (80) | 87,5 (2361) |
| | Via SAU (85) | 2,8 (76) |
| **Décès (Sortie PMSI = 90)** | | 4,0 (108) |

Tableau 11 : Analyses quantitatives des données cliniques sur l'échantillon entier

| Variable | | Proportion (effectif) ou médiane [EIQ] | | | | | | p |
|---|---|---|---|---|---|---|---|---|
| **Clustering C1** | | | | | | | | |
| | | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | | | |
| n = | | 2104 | 275 | 54 | 266 | | | |
| **Sexe** | Homme | 72,8 (1532) | 76,7 (211) | 77,8 (42) | 74,8 (199) | | | 0,436 |
| | Femme | 27,2 (572) | 23,3 (64) | 22,2 (12) | 25,2 (67) | | | |
| **Age** | | 69,0 [62,0 ; 76,0] | 68,0 [55,5 ; 77,0] | 65,0 [60,3 ; 73,8] | 69,5 [62,0 ; 77,0] | | | 0,035 |
| **IGS2** | | 24,0 [18,0 ; 29,0] | 25,0 [0,0 ; 50,0] | 39,0 [22,0 ; 51,0] | 24,0 [18,0 ; 29,0] | | | < 0,001 |
| **Entrée PMSI** | Domicile | 89,3 (1879) | 73,5 (202) | 83,3 (45) | 88,3 (235) | | | |
| | Via SAU | 2,3 (48) | 6,9 (19) | 7,4 (4) | 1,9 (5) | | | |
| **Décès** | | 0,5 (10) | 30,9 (85) | 20,4 (11) | 0,8 (2) | | | |
| **Clustering C2** | | | | | | | | |
| | | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** | **Cluster 6** | |
| n = | | 868 | 1188 | 247 | 115 | 268 | 13 | |
| **Sexe** | Homme | 72,2 (627) | 72,3 (859) | 78,9 (195) | 80,9 (93) | 73,5 (197) | 100,0 (13) | 0,021 |
| | Femme | 27,8 (241) | 27,7 (329) | 21,1 (52) | 19,1 (22) | 26,5 (71) | 0,0 (0) | |
| **Age** | | 70,0 [62,0 ; 77,0] | 68,0 [61,0 ; 75,0] | 68,0 [60,0 ; 76,0] | 67,0 [55,0 ; 78,0] | 70,0 [62,0 ; 77,0] | 64,0 [59,0 ; 68,0] | < 0,001 |
| **IGS2** | | 25,0 [19,0 ; 31,0] | 23,0 [18,0 ; 29,0] | 26,0 [18,0 ; 37,0] | 0,0 [0,0 ; 0,0] | 24,0 [18,0 ; 30,0] | 55,0 [36,3 ; 64,8] | < 0,001 |
| **Entrée PMSI** | Domicile | 87,2 (757) | 89,8 (1067) | 82,2 (203) | 75,7 (87) | 88,8 (238) | 69,2 (9) | |
| | Via SAU | 2,3 (20) | 2,3 (27) | 6,9 (17) | 6,1 (7) | 1,1 (3) | 15,4 (2) | |
| **Décès** | | 0,0 (0) | 6,3 (75) | 8,9 (22) | 3,5 (4) | 1,5 (4) | 23,1 (3) | |

Tableau 12 : Analyses quantitatives des données cliniques sur les clusters identifiés

# IV. Synthèse et discussion

## A. Synthèses narratives des clusters

Devant la multitude et la variété des résultats, nous avons jugé pertinent d'en faire une synthèse narrative. Celle-ci avait pour objectif d'apporter plus de clarté et de refaire le lien entre technicité des résultats et sens clinique à réaliser ces classifications automatiques.

### 1. Clustering C1

#### a. Cluster 1 (n = 2104) :

- Parcours : ce groupe représentait la grande majorité des parcours (78 %) et correspondait au parcours attendu pour une chirurgie programmée : une hospitalisation en CTCV la veille de l'intervention, un séjour court (2,1 jours en moyenne) en réanimation CTCV à la sortie de bloc avant de retourner dans le service de CTCV pour quelques jours, en général une semaine, avec un retour à domicile. Certains patients bénéficiaient d'un passage en USC chirurgical avec de retourner en service d'hospitalisation conventionnelle.
- Clinique : Sans trop de surprise, les caractéristiques de ce groupe étaient proches des caractéristiques de la population entière, avec 72,8 % d'individus masculins et une moyenne d'âge de 69 ans. Notons tout de même qu'il s'agissait de patients relativement simples, arrivant en grande majorité du domicile (89,3 %) et y retournant directement (74,6 %), avec une petite proportion tout de même envoyée en rééducation (16,3 %). Le niveau de gravité était également assez bas, aussi bien du côté du niveau de sévérité du GHM (60,3 % niveau 2 et 27,7 niveau 3) que du score IGS2 (63,1 % entre 15 et 30 et 17,7 % entre 30 et 45). Les DP était largement dominés par les groupes I20-I25 (36,0 %, Cardiopathies ischémiques) et I30-I52 (55,0 %, Autres formes de cardiopathies).

#### b. Cluster 2 (n = 275) :

- Parcours : Ce groupe semblait moins homogène, mais il concernait principalement des patients qui étaient pris en charge par des UM du service de cardiologie à un moment de leur parcours (cardiologie conventionnelle, USI cardiologie, USI

rythmologie) et qui ne passaient quasiment pas dans le service de CTCV (2,5 % de la durée de séjour). Les séjours pouvaient d'être d'une durée très variable (EIQ = [7,3 ; 27,9]) mais on observait une proportion de temps passé au bloc nettement plus importante que dans les autres clusters (4,20 %).

- Clinique : Les caractéristiques démographiques de ce groupe étaient assez proches de celles du premier, en notant néanmoins des origines géographiques un peu plus étendues (seulement 48,4 % de patients originaires du département 44 contre 58,0 % dans le cluster 1). Une proportion plus importante de patients provenait d'autres établissements de santé (79,8 %), mais c'est surtout le mode de sortie qui était marquant pour ce groupe, avec 30,9 % de décès. Seuls 40,0 % retournaient à domicile, les autres étant renvoyés en établissement de santé MCO (16,4 %) ou SSR (11,6 %). La gravité de ces patients était nettement plus élevée que dans le premier groupe avec une sévérité haute (25,1 % niveau 3 et 38,9 % niveau 4). L'IGS2 était un peu moins cohérent avec ce tableau, avec notamment 18,2 % de données manquantes. Ce dernier point ceci pouvait être expliqué par le fait que les parcours étaient moins stéréotypés, les interventions chirurgicales n'étaient pas réalisées dès le début de l'hospitalisation, donc le passage en réanimation pouvait être au-delà de la limite de 7 jours que nous avions fixée et le score IGS2 n'était pas enregistré. Les DP étaient également moins stéréotypés ; toujours en majorité dans le groupe I30-I52 (52,0 %) mais plus variés et peu précis ensuite, avec par exemple 5,5 % en T80-T88 (Complications de soins chirurgicaux et médicaux) et la grande majorité des diagnostics en R (Symptômes, signes et résultats anormaux d'examens cliniques) ou Z (Facteurs influant sur l'état de santé) de l'échantillon (10,3 % au total pour ces 2 chapitres).

### c. Cluster 3 (n = 54) :

- Parcours : Les patients de ce cluster étaient ceux dont la situation et la prise en charge semblaient très compliquées : séjours principalement en réanimation (réanimation CTCV ou réanimation chirurgie polyvalente), réinterventions fréquentes (2,2 passages au bloc en moyenne) et durée d'hospitalisation très longue (médiane 50,1 jours avec EIQ = [40,0 ; 68,7])
- Clinique : Ce cluster restait assez proche des 2 premiers du point de vue démographique et mode d'entrée. On notait tout de même 7,4 % de passage par le SAU, proche du 6,9 % du cluster 2, mais restant loin derrière l'arrivée directe du

domicile (83,3 %). En revanche, comme attendu au vu du parcours long avec beaucoup de temps passé en réanimation, la situation finale de ces patients était la plus compliquée de tous les groupes. Il y avait seulement 22,2 % de retour à domicile contre 20,4 % de décès et 55,6 % de transfert vers un autre établissement. La gravité était haute avec un niveau de sévérité 4 pour 94,4 % de la population et un IGS2 assez variable mais dont la moyenne dépassait largement celle de la population totale (39,0 versus 24,0). Au niveau du DP, nous avons retrouvé une proportion plus importante que pour les autres clusters dans le groupe I70-I79 (20,4 % Maladies des artères, artérioles et capillaires) mais le groupe majoritaire restait toujours I30-I52 (57,4 %).

### d. Cluster 4 (n = 266) :

- Parcours : Il était question ici de patients ayant un parcours-type semblable à celui du cluster 1, mais faisant partie de la filière de chirurgie cardiaque congénitale adulte (CCCA). Comme le cluster 1, les hospitalisations avaient une durée médiane de 13,9 jours avec un parcours assez stéréotypé (CCCA / bloc / réanimation CTCV / CCCA) mais certaines pouvaient être plus longues si la situation le nécessitait (3ème quartile à 18,1 jours).
- Clinique : ce groupe se différenciait peu du cluster 1 et de la population générale, aussi bien au niveau de la description clinique que des DP. Comme nous l'avions vu lors de la description des parcours, il s'agissait des patients ayant un parcours relativement classique mais au sein du service de CCCA, ils ont donc été regroupés parce qu'ils appartenaient à cette filière très spécifique et non mélangée à celle des patients au suivi cardiologique plus tardif. Le niveau de regroupement du code CIM-10 des DP ne nous permettait pas de faire de différence sur les pathologies ayant motivé l'hospitalisation et l'intervention chirurgicale. Les DP en Q20-Q28 (Malformations congénitales de l'appareil circulatoire) y étaient plus représentés que dans les autres clusters mais restaient minoritaires (2,6 % de l'effectif) et tous les patients concernés par ce groupe de DP n'y étaient pas rassemblés (7 sur 19 patients, dont 11 étaient dans le cluster 1).

## 2. Clustering C2

### a. Cluster 1 (n = 868)

- Parcours : Ce groupe était le deuxième le plus fréquent et correspondait aux patients au parcours « classique » (tel que le cluster 1 de la classification C1) mais ayant nécessité un passage en USC chirurgical entre leur passage en réanimation et leur retour en service conventionnel. La durée de séjour était légèrement augmentée mais restait dans les normes de ce qui était attendu (environ un jour de plus sur la durée totale, la moyenne en réanimation restait à 2,2 jours mais les patients passaient en moyenne 3,6 jours en USC).
- Clinique : Comme attendu, les caractéristiques étaient très proches de celles de la population totale. La seule chose qui pouvait être relevée était que la sévérité du séjour était plus centrée autour des niveaux 2 et 3 (55,3 % et 35,4 % respectivement) que la répartition moyenne. Rien de particulier n'était à noter par rapport à la distribution des DP dans l'échantillon complet, les groupes I20-I25 (33,3 %) et I30-I52 (58,9 %) étaient largement majoritaires.

### b. Cluster 2 (n = 1188) :

- Parcours : Ce groupe majoritaire correspond à l'autre moitié des patients qui auraient pu être dans le Cluster 1 de la classification C1. La principale différence était que ces patients n'avaient pas nécessité de passage en USC. Ils avaient la médiane de durée de séjour la plus basse de tous les clusters (11,1 jours).
- Clinique : Ce cluster dominant restait très proche des scores de notre échantillon global comme le précédent. Nous pouvions tout de même noter une sévérité légèrement plus basse, toujours centrée sur les niveaux 2 et 3 (63,0 % et 19,4 % respectivement). Du côté des DP, les groupes I20-I25 (37,8 %) et I30-I52 (50,3 %) étaient toujours les plus représentés.

### c. Cluster 3 (n = 247) :

- Parcours : Ce cluster représentait des parcours un peu moins homogènes que les deux premiers, rassemblant majoritairement des UM de la filière CTCV et mais aussi d'autres sans lien avec la spécialité. On pouvait néanmoins remarquer qu'il s'agissait de patients ayant nécessité une réintervention précoce, le plus souvent dans la première semaine d'hospitalisation (EIQ du nombre d'intervention à [2,0 ;

2,0]). Les durées de séjours étaient de ce fait prolongées (médiane à 24,2 jours, avec en moyenne 26,2 % du temps dans un service de réanimation), la situation du patient ayant été plus complexe.

- Clinique : Les caractéristiques démographiques de ce cluster étaient proches de ceux de la population d'étude. Le mode de sortie était un peu plus variable, avec seulement 57,9 % de retour à domicile pour une mortalité un peu plus élevée (8,9 %) et plus de transfert vers d'autres établissements de court séjour (8,9 %) ou SSR (23,5 %). La sévérité était également plus élevée, plutôt centrée sur des niveaux 3 et 4 (34,4 % et 39,7 % respectivement), et l'IGS2 était également légèrement plus haut que la moyenne générale, sans l'être significativement. Les DP principaux restaient toujours les 2 mêmes groupes (23,1 % en I20-I25 et 58,7 % en I30-I52) mais on pouvait observer une plus forte proportion que la moyenne dans le groupe I70-I79 (10,9 % dans ce groupe contre 4,3 % en moyenne). Ce profil clinique est donc cohérent avec le profil de parcours « classique mais un peu plus grave avec une convalescence plus longue ».

### d. Cluster 4 (n = 115) :

- Parcours : Nous retrouvions ici le groupe des patients passant par les UM du service de cardiologie du cluster 2 de C1. Il présentait un effectif plus réduit mais était plus spécifique de la cardiologie car certaines hospitalisations avaient été ventilées sur d'autres clusters, notamment le n° 3 avec les réinterventions précoces (retour du nombre moyen d'intervention à 1,06 versus 1,36). La durée de séjour restait très variable (EIQ à [7,4 ; 24,1]).
- Clinique : Comme dans C1, le profil de ces patients était un peu plus flou, avec des origines géographiques un peu plus variées, des âges plus étalés (EIQ de [55,0 ; 78,0]), une proportion non négligeable de patients venant d'autres établissements (18,3 %) et y sortant (22,6 % en MCO et 7,8 % en SSR). Le niveau de sévérité était assez panaché, sans niveau représenté de façon majoritaire. Cependant, il était original d'observer un IGS2 moyen à 0,0, correspondant à 59,1 % de la population avec un score de 0 et 26,1 % sans cotation du score. Les DP étaient également plus variés, avec 10,4 % en T80-T88 et 12,2 % en Z. Nous retrouvions donc bien ce profil de patients non-orientés vers la CTCV initialement, mais dont l'état s'est dégradé en cours de séjour, de façon encore plus distincte que dans la classification C1.

### e. Cluster 5 (n = 268) :

- Parcours : Nous retrouvions ici un cluster très semblable au cluster 4 de C1 : les patients de la filière CCCA. La taille et la description de ce groupe était inchangée entre les deux classifications, dénotant une filière très spécifique et facilement identifiable, au parcours bien borné.
- Clinique : Toujours comme le cluster 4 de C1, il y avait peu d'éléments à faire ressortir par rapport à la population totale. Les DP en Q20-Q28 y sont plus représentés mais restent minoritaires (2,6 %).

### f. Cluster 6 (n = 13) :

- Parcours : Ce dernier groupe rassemblait les parcours les plus complexes avec une moyenne du nombre d'interventions chirurgicales à 4,92 et une médiane de séjour à 66,0 jours. Il correspondait aux patients les plus graves du cluster 3 de C1, une fois les patients avec réintervention précoce réattribués au cluster 3 de C2. Les patients passaient en moyenne 77,0 % du temps en réanimation et 9,1 % en USI.
- Clinique : Comparé au reste de l'échantillon, nous pouvions noter une plus grande proportion d'entrées via le SAU (15,4 %), de mortalité (23,1 %), et un retour à domicile quasiment anecdotique (15,4 %) en comparaison avec les transferts (15,4 % en MCO et 46,2 % en SSR). Au niveau de la gravité de la situation, 100 % des patients avaient un niveau de sévérité 4 et l'IGS2 moyen était de 55,0, soit plus du double de la moyenne générale. Au niveau des DP, les groupes I20-I25 (38,5 %) et I30-I52 (23,1 %) étaient toujours dominants, mais on notait également 23,1 % en R50-R69 (Symptômes et signes généraux), catégorie qui inclut les états de chocs (R57).

## B. Discussion des résultats

Malgré le peu de ressources sur lesquelles se baser et les difficultés rencontrées, détaillées ci-après, nous sommes parvenus à atteindre notre objectif principal, à savoir diviser la population de chirurgie cardiaque en groupes ayant chacun un sens clinique, uniquement à partir de données administratives.

Sur la question du clustering, il est apparu que de minimes changements, intervenant aussi bien au stade du calcul des distances que de la classification, pouvaient grandement influencer les résultats finaux. Les groupes montraient des similitudes entre les 2 partitionnements mais n'étaient pas toujours strictement superposables. Tous les résultats n'ont pas été reportés ici, mais certaines classifications pouvaient faire ressortir d'autres typologies, par exemple un groupe de séjours débutant tous en USIC ou la séparation du groupe CCCA en 2 clusters. Pour autant, malgré la difficulté de lecture dans les clusterings plus fragmentés que ceux que nous avons analysés ici, il y avait toujours des points communs et des différences à faire ressortir au niveau de la description des parcours. Ceux que nous avons sélectionnés ici présentaient des caractéristiques assez claires mais un bruit de fond assez prononcé à cause de ces profils proches mais non-identiques, ainsi que des parcours atypiques, difficilement classables.

Si la question du nombre optimal de clusters à rechercher s'est posée en cours d'analyse, nous n'avons pas vraiment pu y apporter de réponse claire. Cependant, chaque degré de détail présentait ses avantages, entre précision et facilité de lecture. Il ne semble donc pas y avoir de seuil optimal ; celui-ci n'aurait de toute façon eu de valeur que dans le cadre spécifique de la chirurgie cardiaque en CHU, puisqu'il n'aurait pas été applicable à une autre spécialité médicale ou à la chirurgie cardiaque dans un autre type d'établissement de santé, où le profil des patients n'aurait pas été le même.

Le rapprochement entre données de parcours et données médicales a permis de vérifier l'idée a priori qu'un parallèle pouvait être fait entre ces informations et que les profils-types de patients restaient cohérents avec ce qui était attendu. Cependant, pour la description de la clinique, et notamment des DP, nous avons été confrontés aux limites que porte le choix de limiter l'analyse au groupe de la CIM10. Cette option nous permettait d'obtenir des résultats agrégés, et de limiter le bruit de fond et les effectifs trop petits pour pouvoir être visibles et analysables. Malheureusement, pour les groupes en lien direct avec la chirurgie cardiaque, ce niveau d'agrégation nous a fait perdre beaucoup d'information. En particulier, le groupe I30-I52 « Autres formes de cardiopathies » regroupe des profils très hétérogènes de patients : péricardite, endocardite, pathologies valvulaires non rhumatismales, troubles de conduction, arythmies, insuffisance cardiaque, arrêt cardiaque etc… Or tous ces profils ne vont pas correspondre à un même niveau de gravité et à une même charge de soins. Ce groupe est retrouvé dans tous les clusters, on peut supposer que des différences auraient émergé si un niveau de précision supérieur (catégorie CIM10 par exemple) avait été utilisé. Nous avons pu remettre un sens clinique

sur des données purement administratives et peu utilisées dans un contexte d'étude médicale. Nous avons ainsi exploré les questions techniques de cette association entre différents types de données et fait un premier tour des difficultés que cette approche pouvait amener.

## C. Difficultés rencontrées

Tout d'abord, aucune étude de ce type n'avait déjà été menée. Les données de mouvements sont principalement utilisées à des fins administratives et financières, éventuellement analysées pour des objectifs de pilotage d'établissement, mais pas pour leur signification clinique. Nous n'avions donc que très peu d'information sur la façon d'aborder et de manipuler ces données dans ce contexte. Les algorithmes de clustering sont également peu utilisés dans la recherche médicale, et jamais sur ce type de données. S'il est relativement aisé de travailler sur des données de PMSI car leur intérêts et limites sont bien connus, les bases administratives bénéficient d'un bien moindre intérêt dans le monde scientifique et la bibliographie à leur sujet est très pauvre. Nous avions peu d'informations sur les pièges à éviter avec ces bases de données, ils ont été éliminés un à un après avoir été repérés lorsque nous pouvions. Nous n'avions pas non plus d'idées sur les algorithmes qui se prêtaient le mieux à ce type de données et avons avancé de façon itérative.

Nous nous sommes tout de même appuyés sur des approches statistiques de la sociologie et des études les utilisant pour le clustering, nous permettant de partir d'une première base, même si celle-ci n'était pas directement applicable à notre contexte, notamment à cause de la taille de nos données. Par exemple, dans ce qui pouvait être illustré à propos des parcours de vie, ces derniers étaient représentés par moins d'une dizaine d'états différents, alors qu'ici notre alphabet da base (sans les durées associées) comportait 43 états. De la même façon pour la question de la durée d'observation, pour une espérance de vie moyenne de 83 ans en France (33) nous avions une durée de séjour moyenne de 16,9 jours, donc 405 heures, avec une bien plus grande variabilité. La durée en nombre d'unités était donc bien plus grande que ce qui est utilisé en sciences sociales. Nous n'avions pas de valeurs de références auxquelles nous comparer, ni de paramétrage par défaut adapté à notre situation. Ceci nous a obligés à multiplier les essais en avançant de manière empirique et par essai-erreur. Le nombre de tentatives

augmentait de façon exponentielle entre l'exploration des possibilités de paramétrage et les choix de codage des variables, aussi bien au moment du clustering que de la modélisation. Le seuil du nombre de groupes à diviser par la CAH n'était pas non plus évident à fixer (à cause de cassures dans les courbes présentes à plusieurs endroits) et entre 2 et 4 possibilités ont été retenues pour chaque hiérarchisation (exemples en annexe 14). Chacune pouvait avoir un sens et nous avons fait le choix d'en conserver plusieurs pour les comparer et éliminer les moins performantes ensuite. Nous avons ainsi calculé 22 matrices de distances qui ont servi de base à l'élaboration de 69 clusterings, parmi lesquels 8 sont passés à l'étape de sélection qualitative après la première étape d'étude des indicateurs chiffrés.

Nous pouvons poursuivre notre réflexion sur une autre difficulté proche et qui en découle : l'absence d'automatisation dans la sélection des résultats obtenus. Tout comme nous n'avions pas de valeurs par défaut pour le paramétrage, nous n'avions pas de bornes pour évaluer la qualité des résultats de manières générales, ni de critère utilisé de façon classique pour faire les choix finaux. Nous aurions pu faire le choix d'un indicateur nous paraissant représenter le mieux notre vision du sujet, mais pour le clustering par exemple, nous avons bien vu que performances chiffrées et intérêt clinique n'étaient pas forcément parallèles et qu'il fallait jouer sur ces deux aspects pour essayer de trouver le meilleur compromis. Nous aurions pu dérouler tout le protocole pour chacun des clusterings obtenus à la toute première étape, mais la quantité de résultats n'aurait pas été gérable et il n'aurait plus été possible de savoir comment les trier, entre la qualité du clustering, les performances du modèle ou le sens clinique que nous pouvions y mettre, ces différents critères n'évoluant pas toujours de façon parallèle ou synchrone.

À côté de ces problèmes en lien avec les techniques d'analyse, nous étions également confrontés à des difficultés liées directement aux données elles-mêmes. Notre base peut être considérée comme très homogène et très hétérogène à la fois. Homogène car, comme nous l'avons vu à l'étape de modélisation, les parcours se ressemblaient beaucoup pour une grande partie, entrainant un fort déséquilibre entre le ou les quelques clusters majoritaires (jusqu'à 79 % de l'échantillon) et les petits clusters satellites qui peinaient à être correctement caractérisés. C'est la principale source d'erreurs à l'étape de la prédiction, rendant caduque à la fois le résultat de classement du modèle et l'interprétation de ses coefficients. Mais la base de données était également très hétérogène, avec un bruit de fond très fort, des parcours très atypiques par rapport à ce qui est habituellement programmé, et des variables pouvant prendre un très grand

nombre de modalités. Même les petits clusters dont le parcours était relativement stéréotypé avaient une grande variété de présentation clinique. Ces deux aspects mélangés nous ont donné cette base que l'on pourrait représenter par un nuage avec un centre très dense et une périphérie très hétérogène et mouvante selon le point de vue. C'est donc qui nous a posé problème pour trouver comment calculer des similitudes, comment trancher parmi des zones de densités différentes et comment les caractériser alors que les attributs étaient multidimensionnels, sans rapprochement évident. Il s'agit ici du reflet de l'aspect « vie réelle » de nos données, mais il s'est avéré être bien plus fort et déséquilibré que ce nous avions estimé en imaginant cette étude.

Cette difficulté liée à une hétérogénéité de densité se rapproche des problèmes de choix dans la précision des DP tel que déjà décrit plus haut : une masse majoritaire (88,2 % pour les groupes I20-et I30-I52) aurait nécessité une description au niveau de la catégorie et une multitude de groupes périphérique pouvait se contenter du niveau du groupe.

Nous avons cependant échappé ou su contourner certaines difficultés attendues :

- Le codage des séjours ne correspondant pas à la vie réelle, entre hospitalisations multi-séjours, erreurs de codage et biais d'information face aux questions de facturation. (non-codage de ce qui n'est pas valorisé).
- Un très petit nombre de matrices de distances présentaient des anomalies empêchant leur utilisation pour la suite (par exemple : non-respect de l'inégalité triangulaire).
- Les représentations graphiques étaient assez simplifiées pour permettre leur lecture par un œil averti.

## D. Intérêts de l'étude et perspectives

Notre étude n'a pas su nous permettre de pousser l'utilisation des données de mouvement jusqu'à un objectif de prédiction, mais elle reste la première à essayer de rapprocher données administratives et informations cliniques. Notre bibliographie était pauvre sur ce thème, mais nous présentons ici justement un premier tour du sujet. Nous proposons une approche qui n'avait jamais été essayée dans ce domaine et entamons les réflexions, aussi bien sur l'utilisation des données de mouvements que sur l'analyse des parcours en médecine dans un sens plus large. Nous avons pu pointer les difficultés que ce sujet et ce type de données amènent. Ainsi, ce travail peut servir de base à une suite s'appuyant

sur ce que nous avons déjà repéré, en rectifiant ce qui a mal fonctionné et en ajustant ce qui semblait intéressant. N'oublions pas de noter que ces données sont plutôt facilement accessibles, puisqu'enregistrées de façon automatique pour tous les patients et codifiées de façon homogène au sein d'un établissement, et qu'elles ne nécessitent pas de recueil spécifique à l'étude comme peuvent le requérir les données cliniques.

Nous avons pu entrevoir des résultats encourageants sur la question du clustering. Malgré ce fort bruit de fond, les clusters ont pu être explicités et être rapprochés de l'analyse qu'aurait pu faire un clinicien. Nous avons pu donner un sens à nos résultats et y voir une logique que l'on pouvait rapprocher de l'expérience clinique. Ces débuts sont encourageants et il semble réaliste de continuer à imaginer, une fois un algorithme plus performant et des données plus adéquates en place, l'élaboration d'un système d'aide à la décision pour la planification des interventions programmées (actuellement seulement soumise à l'appréciation d'un chirurgien référent). Il est tout à fait possible de rassembler des parcours semblables, mais le problème qu'il reste ici est de savoir comment le prédire pour un nouveau patient.

Différentes pistes peuvent être explorées, mais à l'heure actuelle, il semblerait judicieux de commencer par créer un protocole de base avec des paramétrages par défaut. Il pourrait par exemple s'agir de clustering sur une base de données simulée ou sur une spécialité médicale plus stéréotypée. Nous pourrions également nous baser sur un jeu de données pour lequel il y aurait eu un travail « à la main » fait en amont pour regrouper les parcours selon une typologie décidée à l'avance et entrainer un algorithme sur ces données pré-étiquetées. Ceci permettrait d'affiner un premier socle pour lequel le clustering serait clair et la prédiction serait réellement performante, puis de le réappliquer à des données de vie réelle où la procédure serait plus résistante au bruit de fond et nécessiterait moins d'essais et moins de décisions empiriques.

Il serait également intéressant d'essayer d'autres algorithmes de classification plus « souples » que la CAH. Nous avons par exemple essayé d'appliquer un clustering par DBSCAN (Density-Based Spatial Clustering of Applications with Noise), mais cet essai s'est soldé par un échec pour plusieurs raisons. Tout d'abord, cet algorithme se base sur la densité des nuages de points mais, de ce fait, est en difficulté face à des échantillons présentant des densités très hétérogènes, ce qui était notre situation. Ensuite, et en lien avec ce premier écueil, nous avons été obligés de forcer le paramétrage pour obtenir autre chose qu'un unique méga-cluster central, au prix de la mise à l'écart d'une grande proportion de notre population, considérée comme valeurs aberrantes. Les algorithmes

basés sur la densité paraissaient prometteurs mais ne sont finalement pas apparus comme les plus adaptés à cette situation puisqu'elle les met face à leur principale faille. Ils restent néanmoins intéressants à utiliser car ils sont plus performants pour les clusters dont la représentation graphique est de forme irrégulière ou imbriquée avec d'autres clusters. Il serait par exemple possible d'imaginer une clustering en plusieurs étapes, avec le recours à un algorithme basé sur la densité pour identifier les clusters, puis une réattribution des individus exclus dans le cluster le plus proche pour limiter ainsi la perte d'informations.

Une autre piste que nous avions envisagée, que nous avons abandonnée, mais que nous considérons toujours intéressantes à explorer, est celle de la modélisation statistique. Notre idée était de pouvoir prédire le type de parcours qu'un nouveau patient allait avoir, en se basant sur son état clinique à l'entrée et sa pathologie cardiaque, avec l'hypothèse d'une application dans le domaine de la planification des séjours et des interventions. Nos résultats furent médiocres, en grande partie à cause du déséquilibre de taille entre nos différents clusters, et ce malgré le recours à une régression logistique, réputée robuste face à ce défaut. D'autres tentatives seraient intéressantes à mener, par exemple en utilisant un autre modèle adapté aux échantillons déséquilibrés ou en se basant sur d'autres informations, notamment l'intervention chirurgicale (programmée ou réalisée en urgence). Nos données cliniques étaient basées sur le PMSI qui est connu pour être incomplet et biaisé par l'aspect économique de ce recueil. Cette approche basée sur l'intervention chirurgicale garderait néanmoins ce handicap face à des groupes très déséquilibrés et à une variable explicative avec un très grand nombre de facteurs, en lien avec la grande diversité de la classification CCAM. L'appui d'un expert clinicien permettrait sans doute de simplifier cette classification mais demanderait un réel travail de création d'ontologie, sans certitude sur sa pertinence.

En poussant encore plus loin l'exploration sur cette thématique, il serait tout à fait intéressant de se pencher sur la question de l'analyse de parcours de patients au-delà des mouvements intra-hospitaliers. Dans une définition plus proche de celle du parcours de santé, nous pourrions imaginer réutiliser ces approches pour des analyses de suivi en soins de ville, de parcours en amont et en aval d'un séjour hospitalier, etc. De multiples applications peuvent être imaginées et ont tout intérêt à être envisagée afin de mieux comprendre ces notions de parcours souvent mal appréhendées et difficiles d'accès pour les analyses. En ville comme à l'hôpital, les données de mouvement ou d'événement sont plus précises et plus facile d'accès que les données médicales pures, nécessitant un

recueil complémentaire dans la grande majorité des cas, et devraient pouvoir être exploitées à hauteur de leur potentiel.

Pour réaliser ces études à plus grande échelle nous pouvons facilement imaginer mettre à profit les entrepôts de données créés en collaboration par plusieurs établissements ou les bases PMSI gérées en commun au niveau de Groupements Hospitaliers de Territoire (GHT), lorsqu'ils existent. Ces bases de données multi-établissements permettraient d'évaluer les parcours de patients ayant une prise en charge partagée et difficile à tracer. Il est possible, par exemple, pour un patient en cancérologie, de bénéficier d'une chirurgie dans un établissement de santé et d'un traitement par chimiothérapie ou radiothérapie dans une seconde structure. Il serait même envisageable d'étudier des parcours mêlant hospitalisation en établissement de santé et prise en charge en Hospitalisation A Domicile (HAD). Le PMSI permettant une harmonisation du codage entre les établissements et les données de mouvements étant assez rudimentaires, une réplication de ce travail de clustering ne semblerait pas beaucoup plus compliquée à l'échelle territoriale qu'à l'échelle d'un établissement, dès lors qu'une base commune existe.

S'il n'y a pas d'application clinique directe prévue pour ces résultats, ce travail aura servi de première incursion dans le sujet des données de mouvements administratifs et des parcours intra-hospitaliers des patients. Il pourra servir de point de départ pour d'autres travaux en ayant commencé à explorer les possibilités et les difficultés que ce domaine apporte. Nous avons pu démontrer qu'il était tout à fait possible de tirer des informations médicalement utiles de ces données habituellement perçues comme vides de sens (clinique) mais facilement accessibles, ce qui est encourageant pour donner suite à ce type d'exploration. Les idées d'amélioration ne manquent pas après ce premier tour exploratoire et d'autres études sur ces données sont à envisager.

# Bibliographie


1. Trutt L, Mauduit N, Leclère B. Development of a Graphical Interface to Visualize and Analyze the Pathways of Patients During Their Hospital Stay for Thoracic Surgery. Stud Health Technol Inform. 21 août 2019;264:1882‑3.

2. Fitzgerald A, Curry J. Patient journey modelling : a patient centric approach to heal (f)ailing healthcare systems. Present 2011 Int Perspect Health Care Logist Symp May 27 2011 Univ Twente Neth [Internet]. 2011 [cité 10 nov 2017]; Disponible sur: http://researchdirect.westernsydney.edu.au/islandora/object/uws%3A22357/

3. SGMCAS. Lexique des parcours de A à Z [Internet]. 2016. Disponible sur: https://solidarites-sante.gouv.fr/IMG/pdf/2016-01-11_lexique_vf.pdf

4. Banque de données en santé publique. Filière soins (Parcours de santé). In 2019. Disponible sur: https://bdsp-ehesp.inist.fr/vibad/index.php?action=thesaurusDetail&lang=fr&code=http%3A%2F%2Fdata.loterre.fr%2Fark%3A%2F67375%2FTSP-MP3BV55D-C&esid=5

5. Riou F, Jarno P. Représentation et modélisation des trajectoires de soins. ITBM-RBM. oct 2000;21(5):313‑7.

6. Cossin S, Heve D, Jamet O. Datavisualisation des parcours de soins. Rev DÉpidémiologie Santé Publique. mars 2018;66:S38‑9.

7. Rollet A, Defossez G, Dameron O, CoRIM Poitou-Charentes, CRISAP Poitou-Charentes, Ingrand P. Développement et évaluation d'un algorithme de représentation des parcours de soins de patientes atteintes de cancer du sein à partir des données d'un système d'information régional. Rev DÉpidémiologie Santé Publique. mars 2013;61:S17.

8. Roux J, Le Meur N, Grimaud O, Leray E. Utilisation des bases médico-administratives pour l'étude des parcours de soins des patients atteints de sclérose en plaques en France de 2007 à 2013. Rev DÉpidémiologie Santé Publique. mars 2016;64:S18‑9.

9. Dellinger A, Bolard P. Étude de relations de mutation entre unités fonctionnelles d'un centre hospitalier par l'analyse structurale de réseaux: Prat Organ Soins. 1 déc 2010;Vol. 41(4):341‑8.

10. Ritschard G, Gabadinho A, Studer M, Müller NS. Converting between Various Sequence Representations. In: Ras ZW, Dardzinska A, éditeurs. Advances in Data Management [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009 [cité 16 févr 2018]. p. 155‑75. Disponible sur: http://link.springer.com/10.1007/978-3-642-02190-9_8

11. Gabadinho A, Ritschard G, Müller NS, Studer M. Analyzing and Visualizing State Sequences in *R* with TraMineR. J Stat Softw [Internet]. 2011 [cité 12 févr 2018];40(4). Disponible sur: http://www.jstatsoft.org/v40/i04/

12. Jay N, Napoli A, Kohler F. Cancer patient flows discovery in DRG databases. Stud Health Technol Inform. 2006;124:725‑30.

13. Buzmakov, Aleksey and Kuznetsov, Sergei O. and Napoli, Amedeo. Efficient Mining of Subsample-Stable Graph Patterns. In New Orleans, United States; 2017. p. 1‑6. Disponible sur: https://hal.inria.fr/hal-01668663

14. Dimeglio C, Delpierre C, Chauvin P, Lefèvre T. Utilisation des réseaux bayésiens comme technique de fouille de données massives – application à des données de recours aux soins. Rev Fr Aff Soc. 2017;(4):27‑55.

15. Eurostat. Sankey diagrams for energy balance. In: Statistics explained [Internet]. 2018. Disponible sur: http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Sankey_diagrams_for_energy_balance

16. Veesa KS, John KR, Moonan PK, Kaliappan SP, Manjunath K, Sagili KD, et al. Diagnostic pathways and direct medical costs incurred by new adult pulmonary tuberculosis patients prior to anti-tuberculosis treatment – Tamil Nadu, India. Dowdy DW, éditeur. PLOS ONE. 7 févr 2018;13(2):e0191591.

17. Studer Matthias, Ritschard Gilbert. A comparative review of sequence dissimilarity measures [Internet]. LIVES Working Papers; 2014. Disponible sur: https://www.lives-nccr.ch/sites/default/files/pdf/publication/33_lives_wp_studer_sequencedissmeasures.pdf

18. Martin Prodel. Process discovery, analysis and simulation of clinical pathways using health-care data [Internet]. Université de Lyon; 2017. Disponible sur: https://tel.archives-ouvertes.fr/tel-01665163/document

19. Prodel M, Blein C, Fernandez J, Chouaid C. Diversité des parcours de soins des patients atteints de cancer du poumon entre quatre régions françaises : définition de mesures de dissimilarité. Rev DÉpidémiologie Santé Publique. mars 2018;66:S24.

20. Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., Thomas, R. A Survey of Sequential Pattern Mining. Data Sci Pattern Recognit DSPR Vol 11 Pp 54-77 [Internet]. 2017; Disponible sur: https://pdfs.semanticscholar.org/81b8/0dc27f2ee556bb6f7b5d9fb5f227c931cf28.pdf

21. Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med. sept 2012;56(1):35‑50.

22. Egho E, Jay N, Raïssi C, Ienco D, Poncelet P, Teisseire M, et al. A contribution to the discovery of multidimensional patterns in healthcare trajectories. J Intell Inf Syst. avr 2014;42(2):283‑305.

23. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R. Mining Healthcare Data with Temporal Association Rules: Improvements and Assessment for a Practical Use. In: Combi C, Shahar Y, Abu-Hanna A, éditeurs. Artificial Intelligence in Medicine [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009 [cité 5 juin 2018]. p. 16‑25. Disponible sur: http://link.springer.com/10.1007/978-3-642-02976-9_3

24. Fournier-Viger P, Nkambou R, Tseng VSM. RuleGrowth: mining sequential rules common to several sequences by pattern-growth. In ACM Press; 2011 [cité 26 janv 2018]. p. 956. Disponible sur: http://portal.acm.org/citation.cfm?doid=1982185.1982394

25. Aleksey Buzmakov, Elias Egho, Nicolas Jay, Sergei O. Kuznetsov, Amedeo Napoli, et al. FCA and pattern structures for mining care trajectories. In: What FCA can do for artificial intelligence? [Internet]. Beijing, Chine; 2013. Disponible sur: https://hal.inria.fr/hal-00910290

26. Brejova, Dimarco, Vinar, Romero-Hidalgo, Holguin, Patten. Finding Patterns in Biological Sequences [Internet]. 2001. (Technical Report CS-2000-22, University of Waterloo). Disponible sur: http://engr.case.edu/li_jing/papers/00798gpattern.pdf

27. Patnaik D, Butler P, Ramakrishnan N, Parida L, Keller BJ, Hanauer DA. Experiences with mining temporal event sequences from electronic medical records: initial successes and some challenges. In ACM Press; 2011 [cité 31 janv 2018]. p. 360. Disponible sur: http://people.cs.vt.edu/naren/papers/ind0147-patnaik.pdf

28. Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential pattern mining to predict next prescribed medications. J Biomed Inform. févr 2015;53:73‑80.

29. Ma L, Tsui FC, Hogan WR, Wagner MM, Ma H. A framework for infection control surveillance using association rules. AMIA Annu Symp Proc AMIA Symp. 2003;410‑4.

30. Thomas Guyet, Yves Moinard, René Quiniou. Using Answer Set Programming for pattern mining. In Angers, France; 2014. Disponible sur: https://hal.inria.fr/hal-01069092/document

31. Witten DM. Classification and clustering of sequencing data using a Poisson model. Ann Appl Stat. déc 2011;5(4):2493‑518.

32. Romain Guigourès, Dominique Gay, Marc Boullé, Fabrice Clérot. Clustering de séquences d'évènements temporels. In: Revues des nouvelles technologies de l'information [Internet]. Rennes, France; 2014. Disponible sur: http://www.marc-boulle.fr/publications/GuigouresEtAlEGC14.pdf

33. Insee. Espérance de vie – Mortalité [Internet]. Insee Références. 2020. Disponible sur: https://www.insee.fr/fr/statistiques/4277640

# Annexes

## Annexe 1 : Abréviations

| Abréviation | Signification |
|---|---|
| **ACM** | Analyse des Correspondances Multiples |
| **ACoP** | Analyse en Coordonnées Principales |
| **ACP** | Analyse en Composantes Principales |
| **BDSP** | Banque de Données en Santé Publique |
| **CAH** | Classification Ascendante Hiérarchique |
| **CCCA** | Chirurgie Cardiaque Congénitale Adulte |
| **CEC** | Circulation ExtraCorporelle |
| **CHU** | Centre Hospitalo-Universitaire |
| **CIA** | Communication InterAuriculaire |
| **CIM-10** | Classification internationale des maladies - 10ème édition |
| **CTCV** | Chirurgie Thoracique et Cardio-Vasculaire |
| **DAS** | Diagnostic Associé Significatif |
| **DBSCAN** | Density-Based Spatial Clustering of Applications with Noise |
| **DIM** | Département d'Information Médicale |
| **DMS** | Durée Moyenne de Séjour |
| **DP** | Diagnostic Principal |
| **DR** | Diagnostic Relié |
| **ECMO** | ExtraCorporeal Membrane Oxygenation |
| **ETO** | Echocardiographie Trans-Oesophagienne |
| **GHM** | Groupe Homogène de Malades |
| **GHT** | Groupement Hospitalier de Territoire |
| **HAD** | Hospitalisation A Domicile |
| **IGS2** | Indice de Gravité Simplifié 2 |
| **IPP** | Identifiant Permanent du Patient |
| **MCO** | Médecine, Chirurgie, Obstétrique |
| **OM** | Optimal Matching |
| **PC** | Pontage coronarien |
| **PHU** | Pôles Hospitalo-Universitaire |
| **PMSI** | Programme de Médicalisation des Systèmes d'Information |
| **RSS** | Résumé de Sortie Standardisé |
| **RUM** | Résumé d'Unité Médicale |
| **RV(A)** | Remplacement Valvulaire (Aortique) |
| **SAU** | Service d'Accueil des Urgences |
| **SSR** | Soins de suite et réadaptation |
| **T2A** | Tarification A l'Activité |
| **TNI** | Taux de Non-Information |
| **TSA** | Troncs Supra-Aortiques |
| **UF** | Unité Fonctionnelle |
| **UH(T)CD** | Unité d'Hospitalisation de (Très) Courte Durée |
| **UM** | Unité Médicale |
| **(U)SC** | (Unité de) Soins Continus |
| **(U)SI** | (Unité de) Soins Intensifs |
| **VD/G** | Ventricule Droit / Gauche |

## Annexe 2 : Tableau récapitulatif de tous les chapitres de la CIM10 et des groupes de codes utilisés dans ce document

| Chapitre | Codes | Nom du chapitre | Groupe | Nom du groupe |
|---|---|---|---|---|
| **I** | A00-B99 | Certaines maladies infectieuses et parasitaires | A30-A49 | Autres maladies bactériennes |
| **II** | C00-D48 | Tumeurs | C00-C97 | Tumeurs malignes |
| | | | D10-D36 | Tumeurs bénignes |
| | | | D37-D48 | Tumeurs à évolution imprévisible ou inconnue |
| **III** | D50-D89 | Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire | D55-D59 | Anémies hémolytiques |
| **IX** | I00-I99 | Maladies de l'appareil circulatoire | I05-I09 | Cardiopathies rhumatismales chroniques |
| | | | I20-I25 | Cardiopathies ischémiques |
| | | | I26-I28 | Affections cardiopulmonaires et maladies de la circulation pulmonaire |
| | | | I30-I52 | Autres formes de cardiopathies |
| | | | I60-I69 | Maladies cérébrovasculaires |
| | | | I70-I79 | Maladies des artères, artérioles et capillaires |
| | | | I95-I99 | Troubles autres et non précisés de l'appareil circulatoire |
| **X** | J00-J99 | Maladies de l'appareil respiratoire | J09-J18 | Grippe et pneumopathie |
| | | | J40-J47 | Maladies chroniques des voies respiratoires inférieures |

| Chapitre | Codes | Nom du chapitre | Groupe | Nom du groupe |
|---|---|---|---|---|
| | | | J80-J84 | Autres maladies respiratoires touchant principalement le tissu interstitiel |
| | | | J85-J86 | Maladies suppurées et nécrotiques des voies respiratoires inférieures |
| | | | J90-J94 | Autres affections de la plèvre |
| | | | J95-J99 | Autres maladies de l'appareil respiratoire |
| **XI** | K00-K93 | Maladies de l'appareil digestif | K20-K31 | Maladies de l'œsophage, de l'estomac et du duodénum |
| | | | K90-K93 | Autres maladies de l'appareil digestif |
| **XII** | L00-L99 | Maladies de la peau et du tissu cellulaire sous-cutané | L00-L08 | Infections de la peau et du tissu cellulaire sous-cutané |
| | | | L80-L99 | Autres affections de la peau et du tissu cellulaire sous-cutané |
| **XIII** | M00-M99 | Maladies du système ostéo-articulaire, des muscles et du tissu conjonctif | M00-M25 | Arthropathies |
| | | | M60-M79 | Affections des tissus mous |
| | | | M80-M94 | Ostéopathies et chondropathies |
| | | | M95-M99 | Autres maladies du système ostéo-articulaire, des muscles et du tissu conjonctif |
| **XV** | O00-O99 | Grossesse, accouchement et puerpéralité | O00-O08 | Grossesse se terminant par un avortement |
| **XVII** | Q00-Q99 | Malformations congénitales et anomalies chromosomiques | Q20-Q28 | Malformations congénitales de l'appareil circulatoire |

| Chapitre | Codes | Nom du chapitre | Groupe | Nom du groupe |
|---|---|---|---|---|
| **XVIII** | R00-R99 | Symptômes, signes et résultats anormaux d'examens cliniques et de laboratoire, non classés ailleurs | R00-R09 | Symptômes et signes relatifs aux appareils circulatoire et respiratoire |
| | | | R50-R69 | Symptômes et signes généraux |
| **XIX** | S00-T98 | Lésions traumatiques, empoisonnements et certaines autres conséquences de causes externes | S00-S09 | Lésions traumatiques de la tête |
| | | | S20-S29 | Lésions traumatiques du thorax |
| | | | T36-T50 | Intoxications par des médicaments et des substances biologiques |
| | | | T80-T88 | Complications de soins chirurgicaux et médicaux, non classées ailleurs |
| **XXI** | Z00-Z99 | Facteurs influant sur l'état de santé et motifs de recours aux services de santé | Z00-Z13 | Sujets en contact avec les services de santé pour des examens divers |
| | | | Z40-Z54 | Sujets ayant recours aux services de santé pour des actes médicaux et des soins spécifiques |
| | | | Z80-Z99 | Sujets dont la santé peut être menacée en raison d'antécédents personnels et familiaux et de certaines affections |

## Annexe 3 : Codes mouvements PMSI MCO (à partir de 2009)

| Code d'entrée | | Signification |
|---|---|---|
| **Chiffre 1** | Chiffre 2 | |
| **6** | 0 | Mutation… (même établissement) |
| | 1 | … depuis une unité de soins de courte durée |
| | 2 | … depuis une unité de soins de suite ou de réadaptation |
| | 3 | … depuis une unité de soins de longue durée |
| | 4 | … depuis une unité de psychiatrie |
| | 6 | … depuis une unité d'hospitalisation à domicile |
| **7** | 0 | Transfert… (autre établissement) |
| | 1 | … depuis une unité de soin de courte durée |
| | 2 | … depuis une unité de soins de suite ou de réadaptation |
| | 3 | … depuis une unité de soins de longue durée |
| | 4 | … depuis une unité de psychiatrie |
| | 6 | … depuis une unité d'hospitalisation à domicile |
| **8** | 0 | Du domicile… |
| | 5 | … avec passage par le service d'accueil des urgences |
| | 7 | D'une structure d'hébergement médico-sociale |

| CCode de sortie | | Signification |
|---|---|---|
| **Chiffre 1** | Chiffre 2 | |
| **6** | 0 | Mutation… (même établissement) |
| | 1 | … vers une unité de soin de courte durée |
| | 2 | … vers une unité de soins de suite ou de réadaptation |
| | 3 | … vers une unité de soins de longue durée |
| | 4 | … vers une unité de psychiatrie |
| | 6 | … vers une unité d'hospitalisation à domicile |
| **7** | 0 | Transfert… (autre établissement) |
| | 1 | … vers une unité de soins de courte durée |
| | 2 | … vers une unité de soins de suite ou de réadaptation |
| | 3 | … vers une unité de soins de longue durée |
| | 4 | … vers une unité de psychiatrie |
| | 6 | … vers une unité d'hospitalisation à domicile |
| **8** | 0 | Vers le domicile |
| | 7 | Vers une structure d'hébergement médico-sociale |
| **9** | 0 | Par décès |

## Annexe 4 : Codage du niveau de sévérité d'un GHM

| Code | Signification |
|---|---|
| **1 - 4** | Niveaux de sévérité de 1 à 4 |
| **T** | Très courte durée (T1 : 0 ou 1 nuit, T2 : de 0 à 2 nuits) |
| **J** | Ambulatoire (0 nuit) |
| **E** | Décès |
| **Z** | Non concerné par une question de sévérité ou de courte durée |

## Annexe 5 : Détails du calcul des distances selon la méthode OMspell

### A. Méthodologie détaillée

Rappel des termes utilisés :

- Alphabet : liste de tous les éléments pouvant être dans une séquence
- Coût d'indel : coût attribué à la modification d'une séquence par l'ajour ou la suppression d'un élément
- Coût de substitution : coût attribué à la modification d'une séquence par le remplacement d'un élément par un autre

Les méthodes OM et OMspell ont le même fonctionnement de base que le calcul de distance de Levenshtein mais apportent plus de finesse en permettant d'affiner les coûts en fonction des situations, là où l'algorithme originel attribuait le score de 1 à chaque modification. Les coûts d'indel et de substitution peuvent avoir une autre valeur que 1, mais également avoir une valeur personnalisée pour chaque élément de l'alphabet. Ce deuxième cas de figure impose de construire une matrice des coûts de transitions reliant tous les éléments de l'alphabet entre eux. Celle-ci peut être déterminée selon des bases théoriques (en lien avec une connaissance a priori du terrain) ou de façon automatisée (probabilité inverse de passer d'un élément à l'autre, calculée sur un échantillon de séquences). Elle doit respecter 3 règles : être symétrique (A → B équivaut à B → A), respecter l'inégalité triangulaire (A → B est plus court que A → C → B) et être nulle pour la substitution d'un élément par lui-même.

L'équilibre entre les coûts d'indel et de substitution permet de faire varier le comportement de l'algorithme : un coût d'indel haut le rend plus sensible aux différences de durées et un coût d'indel bas le rend plus sensible aux décalages dans la séquence.

La méthode OMspell apporte une autre grande différence (par rapport à l'OM et la distance de Levenshtein) qui est qu'au lieu d'avoir un alphabet où chaque élément est défini par un état unique et une durée de 1 unité de temps, les éléments sont ici de durées variables et plusieurs peuvent donc être plusieurs à correspondre à un même état. La taille de l'alphabet est ainsi démultipliée mais peut être ensuite simplifiée grâce à l'introduction de nouveaux paramètres de calcul :

- $\sigma$ : pondération d'un segment, proportionnel à sa durée (facteur exponentiel)

- $\delta$ : dilatation / compression d'un segment, modification de la durée d'un segment d'une unité de temps (≥ 0)

L'alphabet utilisé pour l'application de l'algorithme peut ainsi être à nouveau réduit à des éléments définis par un état unique et une durée de 1 unité de temps. Les coûts d'indel et de substitution peuvent être paramétrés de façon unitaire, sans avoir à les calculer pour chaque élément du premier alphabet :

- $c_I$ : indel unitaire, délétion / insertion d'un segment de longueur de 1 unité de temps (valeur unique commune ou valeur individualisée pour chaque état)
- $\gamma$ : substitution unitaire, matrice de coût pour la substitution d'un segment dans un état par un segment dans un autre état (durées identiques, rapportées à 1 unité de temps)

Rapportée au fonctionnement de l'algorithme d'OM « classique », la distance entre un segment $x$ dans un état $a$ et de durée $t_1$, et un segment $y$ dans un état $b$ et de durée $t_2$ est calculée avec les coûts de base :

- Indel ($x$) : $\boxed{c_I^S(a_{t_1}) = c_I(a) + \delta({t_1}^\sigma - 1)}$
- Substitution ( $x$ - $y$ ) : $\boxed{\gamma^S(a_{t_1}, b_{t_2}) = \gamma(a,b) + \delta({t_1}^\sigma + {t_2}^\sigma - 2)}$

  Dans le cas particulier où $a = b$, il ne s'agit plus que de comparer les durées des deux segments $x$ et $y$, donc le coût de substitution peut être simplifié à : $\boxed{\gamma^S(a_{t_1}, b_{t_2}) = \delta|{t_1}^\sigma - {t_2}^\sigma|}$

## B. Application dans l'étude

Puisqu'il n'existait pas d'étude semblable antérieure nous permettant de préjuger de comment évaluer les coûts pour rapprocher le résultat de l'algorithme à ce qui est médicalement intéressant, nous avons effectué plusieurs essais pour tester différents réglages et voir ce qui avait le plus de sens. Les coûts pour le calcul des distances ont été fixés tels que :

- **Pondération d'un segment** : l'unité de temps choisie ayant été l'heure pour plus de précision, ce paramètre exponentiel a été légèrement minimisé (classiquement fixé à 1).

- **Dilatation/compression** : de même que le coût précédent, il a été minimisé afin de rattraper la petite unité de temps par rapport aux durées totales des segments et des parcours (classiquement fixé à 0,5).
- **Indel unitaire** : fixé à 1
- **Substitution unitaire** : plusieurs matrices de coût ont été élaborées pour tester différentes pondérations plus ou moins sévères. En se basant sur le fichier de structure du CHU et en partant d'un coût de substitution de base à 1, nous avons testé plusieurs indices permettant de réduire ce coût lorsque les UM avaient des points communs :
    - même service : hospitalisation dans des UM à la thématique proche, avec le même niveau d'expertise de l'équipe de soin
    - même type d'autorisation : code en lien avec la tarification des séjours, il permet de repérer les UM spécialisées et aux prises en charge lourdes telles que les réanimations, les USI et les USC. Deux UM avec le même type d'autorisation auront une activité et une spécialisation proche, une substitution de l'une par l'autre aura donc moins de valeur que par rapport à une UM d'un autre type.
    - non-appartenance à la filière CTCV : l'objectif étant de suivre les patients au sein de la filière, il n'était plus intéressant de suivre leur parcours avec précision une fois qu'ils en étaient sortis et devenaient des parcours atypiques. (La filière incluait aussi bien des UM conventionnelles que des UM spécialisées, tous services confondus)

  Le coût de substitution final était le produit de ces 3 pondérations avec le coût de base (à 1). Le coût minimum était obtenu pour les substitutions qui n'avaient que très peu d'intérêt dans notre contexte (par exemple : 2 UM d'hospitalisation conventionnelle d'une spécialité médicale n'étant pas en lien avec la CTCV). Au contraire, des poids importants signifiaient des différences importantes dans le parcours et la prise en charge médicale nécessitée par les patients (par exemple : sortie de bloc en réanimation CTCV ou dans un service de médecine dont était issu le patient pour une pathologie sans lien avec la CTCV).

  Afin d'évaluer l'intérêt du calcul de ce type de matrice, nous avons également testé de laisser ce paramètre de substitution unitaire fixé à 1 de manière uniforme.

## Annexe 6 : Interventions dont ont bénéficié les patients du groupe "chirurgie cardiaque", triées par fréquence

**Abréviations :**

- CEC : Circulation ExtraCorporelle
- CIA : Communication InterAuriculaire
- ECMO : ExtraCorporeal Membrane Oxygenation
- ETO : Echocardiographie Trans-Oesophagienne
- PC : Pontage coronarien
- RV(A) : Remplacement Valvulaire (Aortique)
- TSA : Troncs Supra-Aortiques
- VD/G : Ventricule Droit / Gauche

| Intervention | n |
|---|---|
| remplacement valve aortique | 576 |
| pc 3 | 361 |
| sup à pc 3 | 341 |
| drainage / décaillotage péricardique | 121 |
| pc 2 | 113 |
| plastie valve mitrale | 108 |
| remplacement valve aortique + pc 1 | 82 |
| tamponade | 73 |
| remplacement valve mitrale | 56 |
| remplacement valve aortique + pc 2 | 50 |
| remplacement aorte ascendante | 48 |
| drainage péricardique | 45 |
| bentall | 44 |
| valve aortique transapicale | 38 |
| ecmo / bio-médicus | 37 |
| remplacement valve aortique + pc 3 | 35 |
| reprise cicatrice | 35 |
| reprise hémostase | 28 |
| autres interventions | 26 |
| drainage | 25 |
| remplacement valve aortique + maze | 25 |
| remplacement bi-valvulaire | 23 |
| remplacement valve aortique + remplacement aorte ascendante | 23 |
| ablation ecmo | 22 |
| valve transcarotidienne | 22 |
| tirone david | 21 |
| irrigation-lavage médiastinal | 20 |
| pc 1 | 20 |
| plastie valve mitrale + plastie valve tricuspide | 19 |
| remplacement valve mitrale + plastie valve tricuspide | 19 |

| Intervention | n |
|---|---|
| remplacement valve aortique + pc 1 | 18 |
| pose de sonde épicardique | 17 |
| plastie bi-valvulaire | 16 |
| pose pace-maker | 16 |
| ablation pace-maker | 15 |
| remplacement valve aortique + plastie valve tricuspide | 15 |
| valve transaortique | 15 |
| ablation défibrillateur | 14 |
| ecmo | 14 |
| hearth ware | 14 |
| remplacement valve aortique + pc 2 | 11 |
| cec urgente | 10 |
| myxome | 10 |
| pc 3 + remplacement valve aortique | 10 |
| reprise chirurgicale (fermeture sternale, décaillotage..) | 10 |
| plastie valve mitrale + maze | 9 |
| sup à pc 3 + remplacement valve aortique | 9 |
| trachéotomie | 9 |
| rva + pc2 | 8 |
| ablation pace-maker + pose pace-maker | 7 |
| drainage / décaillotage pleural | 7 |
| plastie valve mitrale + pc 1 | 7 |
| remplacement bi-valvulaire + plastie | 7 |
| remplacement valve aortique + remplacement valve mitrale | 7 |
| cia | 6 |
| fermeture d'une cia simple avec cec | 6 |
| pc 3 + plastie valve mitrale | 6 |
| remplacement aorte horizontale | 6 |
| remplacement valve aortique + sup à pc 3 | 6 |
| remplacement valve mitrale + pc 1 | 6 |
| remplacement valve mitrale + pc 2 | 6 |
| ablation défibrillateur + pose défibrillateur | 5 |
| pose défibrillateur | 5 |
| ross | 5 |
| rva + pc1 | 5 |
| tumeur du coeur | 5 |
| bentall + maze | 4 |
| bentall + pc 1 | 4 |
| debranching des tsa | 4 |
| pc 3 + maze | 4 |
| plastie valve mitrale + maze + plastie valve tricuspide | 4 |
| plastie valve mitrale + pc 2 | 4 |
| remplacement aorte ascendante + remplacement aorte horizontale | 4 |
| remplacement valve aorti + remplacement valve mitra + plastie valve tricuspide | 4 |
| remplacement valve mitrale + maze | 4 |
| remplacement valve tricuspide | 4 |

| Intervention | n |
|---|---|
| rva + pc3 | 4 |
| sup à pc 3 + maze | 4 |
| ablation de fils d'acier | 3 |
| ablation pace-maker + ablation sondes | 3 |
| autre geste (cia, bigelow ...) | 3 |
| bentall + sup à pc 3 | 3 |
| biopsie (coeur) | 3 |
| pc 2 + remplacement valve aortique | 3 |
| péricardectomie pour constriction | 3 |
| plastie bi-valvulaire + maze | 3 |
| plastie valve mitrale + plastie valve tricuspide + maze | 3 |
| plastie valve tricuspide | 3 |
| remplacement aorte ascendante + remplacement valve aortique | 3 |
| remplacement bi-valvulaire + pc 1 | 3 |
| remplacement valve aortique + autre geste (cia, bigelow ...) | 3 |
| remplacement valve aortique + remplacement aorte ascendante | 3 |
| remplacement valve mitrale + plastie valve tricuspide | 3 |
| remplacement valvulaire pulmonaire | 3 |
| ablation de fils sternaux | 2 |
| ablation défibrillateur + ablation sondes | 2 |
| ablation pace-maker + pose de sonde épicardique | 2 |
| ablation pace-maker sous scopie | 2 |
| anévrisme ou rupture vg | 2 |
| bentall + pose de sonde épicardique | 2 |
| carotide | 2 |
| myxome + pc 1 | 2 |
| pc 1 + remplacement valve aortique | 2 |
| pc 2 + maze | 2 |
| plastie bi-valvulaire + pc 1 | 2 |
| plastie bi-valvulaire + pc 3 | 2 |
| plastie valve mitrale + plastie valve tricuspide + pc 1 | 2 |
| plastie valve mitrale + valve tricuspide | 2 |
| pose pace-maker + ablation pace-maker | 2 |
| poumon | 2 |
| remplacement bi-valvulaire + pc 2 | 2 |
| remplacement bi-valvulaire + pose de sonde épicardique | 2 |
| remplacement valve aorti + pc 2 + maze | 2 |
| remplacement valve aorti + remplacement aorte ascen + pc 1 | 2 |
| remplacement valve aorti + remplacement valve mitra + maze | 2 |
| remplacement valve aorti + remplacement valve mitra + remplacement valve tricu | 2 |
| remplacement valve aortique + carotide | 2 |
| remplacement valve aortique + pc 1 + maze | 2 |
| remplacement valve aortique + pc 3 + maze | 2 |
| remplacement valve aortique + plastie valve mitrale | 2 |
| remplacement valve aortique + remplacement aorte ascendante + pc 1 | 2 |
| remplacement valve aortique + pc 3 | 2 |

| Intervention | n |
|---|---|
| remplacement valve aortique + remplacement valve mitrale | 2 |
| remplacement valve mitrale + plastie valve tricuspide + maze | 2 |
| remplacement valve mitrale + remplacement valve aortique | 2 |
| remplacement valvulaire pulmonaire + plastie valve tricuspide | 2 |
| reprise chirurgicale pour infection profonde du site opératoire | 2 |
| rv mitrale + pc2 | 2 |
| rva + plastie valve tricuspide | 2 |
| sténose sous-valvulaire aortique (morrow, morrow étendue, konno modifié...) | 2 |
| sup à pc 3 + plastie valve mitrale | 2 |
| tirone david + remplacement aorte horizontale | 2 |
| tube vd/ap | 2 |
| 6 pontages coronariens | 1 |
| ablation défibrillateur + ablation sondes + pose défibrillateur + pose de sonde épicardique | 1 |
| ablation défibrillateur sous scopie | 1 |
| ablation pace-maker + pose pace-maker + drainage / décaillotage péricardique | 1 |
| ablation pace-maker sous scopie + ablation sondes sous scopie | 1 |
| ablation pace-maker sous scopie + pose pace-maker | 1 |
| ablation sondes | 1 |
| ablation sondes + pose de sonde épicardique | 1 |
| ablation sondes + pose défibrillateur | 1 |
| ablation sondes sous scopie | 1 |
| anévrisme ou rupture vg + pc 2 | 1 |
| autre intervention digestive ou endocrinienne | 1 |
| autres interventions + poumon | 1 |
| bentall + pc 2 | 1 |
| bentall + plastie valve tricuspide | 1 |
| bentall + plastie valve tricuspide + maze | 1 |
| bentall + remplacement aorte horizontale | 1 |
| bentall + remplacement valve aortique | 1 |
| cec urgente + remplacement valve aortique | 1 |
| conversion de fontan + maze | 1 |
| coronarographie sans vg (rac, etc) | 1 |
| drainage / décaillotage péricardique + autre geste (cia, bigelow ...) | 1 |
| drainage / décaillotage péricardique + pose de sonde épicardique sous scopie | 1 |
| drainage pleural droit | 1 |
| fermeture d'une cia avec rvpap avec cec + pc 1 | 1 |
| fistule oesophagienne | 1 |
| lobectomie supérieure droite | 1 |
| lobectomie supérieure gauche | 1 |
| myxome + remplacement valve aortique | 1 |
| pc 2 + bentall | 1 |
| pc 2 + carotide | 1 |
| pc 2 + carotide + remplacement valve aortique | 1 |
| pc 2 + pc 1 | 1 |

| Intervention | n |
|---|---|
| pc 2 + plastie valve mitrale + ecmo / bio-médicus | 1 |
| pc 2 + remplacement valve mitrale | 1 |
| pc 3 + autre intervention | 1 |
| pc 3 + carotide | 1 |
| pc 3 + remplacement aorte ascendante | 1 |
| pc 3 + remplacement valve aortique + maze | 1 |
| pc 3 + remplacement valve aortique + plastie valve tricuspide | 1 |
| pc2 | 1 |
| plastie bi-valvulaire + plastie valve tricuspide | 1 |
| plastie valve aortique | 1 |
| plastie valve mit + plastie valve tri + remplacement valv + maze | 1 |
| plastie valve mitrale + maze + pc 1 | 1 |
| plastie valve mitrale + maze + remplacement valve aortique + plastie valve tricuspide | 1 |
| plastie valve mitrale + pc 1 + maze | 1 |
| plastie valve mitrale + pc 2 + maze | 1 |
| plastie valve mitrale + pc 3 | 1 |
| plastie valve mitrale + plastie valve tricuspide + maze | 1 |
| plastie valve mitrale + plastie valve tricuspide + pc 2 | 1 |
| plastie valve mitrale + plastie valve tricuspide + pc 3 | 1 |
| plastie valve mitrale + remplacement valve aortique | 1 |
| plastie valve mitrale + remplacement valve mitra + plastie valve tricuspide | 1 |
| plastie valve mitrale + sup à pc 3 | 1 |
| plastie valve mitrale + tirone david | 1 |
| plastie valve mitrale + valve tricuspide + maze | 1 |
| pneumonectomie dr + tumeur du coeur + autres interventi + exérèse atypique | 1 |
| pneumonectomie droite | 1 |
| pose pace-maker sous scopie | 1 |
| remplacement aorte ascen + ecmo / bio-médicus + pc 2 | 1 |
| remplacement aorte ascen + remplacement valve aorti + pc 3 | 1 |
| remplacement aorte ascendante + autre geste (cia, bigelow ...) | 1 |
| remplacement aorte ascendante + bentall | 1 |
| remplacement aorte ascendante + maze | 1 |
| remplacement aorte ascendante + pc 3 | 1 |
| remplacement aorte ascendante + plastie valve tricuspide | 1 |
| remplacement aorte ascendante + tirone david | 1 |
| remplacement bi-valvulai + plastie valve tricuspide + pc 1 | 1 |
| remplacement bi-valvulaire + maze | 1 |
| remplacement bi-valvulaire + maze + pc 1 + plastie valve tricuspide | 1 |
| remplacement bi-valvulaire + plastie valve tricuspide | 1 |
| remplacement bi-valvulaire + remplacement valve tricuspide | 1 |
| remplacement crosse avec technique "trompe éléphant" | 1 |
| remplacement valv + pc 1 + maze + remplacement aort | 1 |
| remplacement valv + plastie valve mit + plastie valve tri + maze | 1 |
| remplacement valv + remplacement valv + remplacement aort + plastie valve tri | 1 |
| remplacement valve aorti + plastie valve tricuspide | 1 |

| Intervention | n |
|---|---|
| remplacement valve aorti + autre geste (cia, bigelo + drainage / décaillotage | 1 |
| remplacement valve aorti + maze + pc 1 | 1 |
| remplacement valve aorti + pc 1 + maze | 1 |
| remplacement valve aorti + pc 1 + plastie valve tricuspide | 1 |
| remplacement valve aorti + pc 1 + remplacement aorte ascen | 1 |
| remplacement valve aorti + pcx1 | 1 |
| remplacement valve aorti + plastie valve tricuspide + maze | 1 |
| remplacement valve aorti + plastie valve tricuspide + pc 1 | 1 |
| remplacement valve aorti + plastie valve tricuspide + remplacement aorte ascen | 1 |
| remplacement valve aorti + remplacement valve mitra + pc 2 | 1 |
| remplacement valve aorti + remplacement valve tricu + pc 1 | 1 |
| remplacement valve aortique + biopsie (médiastin) | 1 |
| remplacement valve aortique + ecmo / bio-médicus | 1 |
| remplacement valve aortique + maze + pc 1 | 1 |
| remplacement valve aortique + pc 1 + plastie valve tricuspide | 1 |
| remplacement valve aortique + pcx3 | 1 |
| remplacement valve aortique + plastie bi-valvulaire | 1 |
| remplacement valve aortique + plastie valve mitrale + plastie valve tricuspide | 1 |
| remplacement valve aortique + plastie valve mitrale + sup à pc 3 | 1 |
| remplacement valve aortique + plastie valve tricuspide + autre geste (cia, bigelow ...) + remplacement valve mitrale | 1 |
| remplacement valve aortique + pose de sonde épicardique | 1 |
| remplacement valve aortique + pose défibrillateur | 1 |
| remplacement valve aortique + remplacement aorte ascendante + sup à pc 3 | 1 |
| remplacement valve aortique + remplacement valve mitrale + plastie valve tricuspide | 1 |
| remplacement valve aortique + remplacement valve mitrale + plastie valve tricuspide | 1 |
| remplacement valve aortique + chirurgie de l'arche aortique avec cec | 1 |
| remplacement valve aortique + fermeture d'une cia simple avec cec | 1 |
| remplacement valve aortique + maze | 1 |
| remplacement valve aortique + pc 1 + remplacement aorte ascendante | 1 |
| remplacement valve aortique + pc1 | 1 |
| remplacement valve aortique + plastie valve mitrale | 1 |
| remplacement valve aortique + plastie valve tricuspide | 1 |
| remplacement valve aortique + plastie valve tricuspide + remplacement aorte ascendante | 1 |
| remplacement valve aortique + pose de sonde épicardique | 1 |
| remplacement valve aortique + remplacement aorte ascendante + plastie valve tricuspide | 1 |
| remplacement valve aortique + sténose sous-valvulaire aortique (résection de membrane ss-valvulaire aortique) + remplacement aorte ascendante | 1 |
| remplacement valve aortique + sup à pc 3 | 1 |
| remplacement valve mitra + pc 1 + plastie valve tricuspide | 1 |
| remplacement valve mitra + plastie valve tricuspide + maze | 1 |

| Intervention | n |
|---|---|
| remplacement valve mitra + plastie valve tricuspide + pc 1 | 1 |
| remplacement valve mitra + remplacement valve aorti + plastie valve tricuspide | 1 |
| remplacement valve mitra + remplacement valve tricu + plastie valve mitrale | 1 |
| remplacement valve mitrale + eto sous ag | 1 |
| remplacement valve mitrale + maze + plastie valve tricuspide | 1 |
| remplacement valve mitrale + pc 3 | 1 |
| remplacement valve mitrale + plastie valve tricuspide + pose de sonde épicardique | 1 |
| remplacement valve mitrale + remplacement valve aortique + pc 2 | 1 |
| remplacement valve mitrale + remplacement valve tricuspide | 1 |
| remplacement valve mitrale + pc 1 | 1 |
| remplacement valve mitrale + remplacement valve aortique + pc 1 + plastie valve tricuspide | 1 |
| remplacement valve tricu + pose pace-maker + ablation pace-maker | 1 |
| remplacement valve tricuspide + pc 1 | 1 |
| remplacement valve tricuspide + pose de sonde épicardique | 1 |
| remplacement valve tricuspide + pose de sonde épicardique | 1 |
| rv mitrale + pc1 + plastie valve tricuspide | 1 |
| rv mitrale + plastie valve tricuspide | 1 |
| rv mitrale + plastie valve tricuspide + pc1 | 1 |
| rva + maze | 1 |
| rva + maze + remplacement aorte ascendante | 1 |
| rva + pc1 + plastie valve tricuspide | 1 |
| rva + pc3 + plastie valve tricuspide | 1 |
| rva + remplacement aorte ascendante + pc1 | 1 |
| rva + rv mitrale | 1 |
| rva + rv mitrale + pc 1 | 1 |
| sup à pc 3 + carotide | 1 |
| sup à pc 3 + plastie valve mitrale + maze | 1 |
| sup à pc 3 + plastie valve tri + plastie valve mit + maze | 1 |
| sup à pc 3 + plastie valve tricuspide + plastie valve mitrale | 1 |
| sup à pc 3 + remplacement valve mitrale | 1 |
| tamponade + remplacement valve aortique | 1 |
| tirone david + maze | 1 |
| tirone david + pc 1 | 1 |
| tirone david + plastie valve tricuspide | 1 |
| valve transcarotidienne d | 1 |
| valve transfémorale | 1 |

## Annexe 7 : Structure et organisation administrative des UM concernées par l'étude

| UM | Libellé UM | Ser. | Libellé Service | pôle | Libellé Pôle | Aut. | Libellé Autorisation |
|---|---|---|---|---|---|---|---|
| **0000** | **Bloc CTCV** | | | | | | |
| **1043** | UM MAG HGRL | 6020 | Serv. Med. Aigue Geriat. | 7070 | Gérontologie Clinique | 27 | Médecine gériatrique |
| **1635** | Soins Cont. Clin. Med. Ped | 1610 | Pédiatrie Générale | 7090 | Femme-Enfant-Adolescent | 1401 | Soins surveillance continue pédiatrique hors grands brûlés |
| **2670** | **CCCA** | **2670** | **Chir. Card. Ped. Et Congenitales (CCPC)** | **7090** | **Femme-Enfant-Adolescent** | **50** | **Chirurgie cardiaque** |
| **1020** | UM Endocrinologie | 1020 | Endocrinologie | 7610 | Inst. Thorax Et S. Nerveux | 29 | Unité de médecine indifférenciée |
| **1210** | UM Neurologie | 1210 | Neurologie | 7610 | Inst. Thorax Et S. Nerveux | 29 | Unité de médecine indifférenciée |
| **1215** | UM Neuro-Vasculaire SI | 1210 | Neurologie | 7610 | Inst. Thorax Et S. Nerveux | 18 | Soins intensifs en unité neurovasculaire |
| **1310** | UM Pneumologie | 1310 | Pneumologie | 7610 | Inst. Thorax Et S. Nerveux | 29 | Unité de médecine indifférenciée |
| **1322** | Pneumo. Surv. Continue UMTR | 1310 | Pneumologie | 7610 | Inst. Thorax Et S. Nerveux | 301 | Soins surveillance continue adulte hors grands brûlés |
| **3740** | **Pneumologie Soins Intensifs** | **1310** | **Pneumologie** | **7610** | **Inst. Thorax Et S. Nerveux** | **202** | **Autres soins intensifs** |
| **1410** | **UM Cardiologie** | **1410** | **Cardiologie** | **7610** | **Inst. Thorax Et S. Nerveux** | **29** | **Unité de médecine indifférenciée** |
| **3710** | **Cardiologie Soins Intensifs** | **1410** | **Cardiologie** | **7610** | **Inst. Thorax Et S. Nerveux** | **201** | **Soins intensifs en cardiologie = USIC** |
| **3711** | **UM Rythmologie SI** | **1410** | **Cardiologie** | **7610** | **Inst. Thorax Et S. Nerveux** | **201** | **Soins intensifs en cardiologie = USIC** |

| UM | Libellé UM | Ser. | Libellé Service | pôle | Libellé Pôle | Aut. | Libellé Autorisation |
|---|---|---|---|---|---|---|---|
| **2600** | **CTCV** | **2600** | **C.T.C.V.** | **7610** | **Inst. Thorax Et S. Nerveux** | **53** | **Chirurgie indifférenciée adulte** |
| **2620** | UM U.T.T. | 2620 | U.T.T. | 7610 | Inst. Thorax Et S. Nerveux | 202 | Autres soins intensifs |
| **2630** | UM Chirurgie Vasculaire | 2630 | Chirurgie Vasculaire | 7610 | Inst. Thorax Et S. Nerveux | 53 | Chirurgie indifférenciée adulte |
| **1160** | UM Rhumatologie HD | 1160 | Rhumatologie HD | 7620 | OTONN | 29 | Unité de médecine indifférenciée |
| **2410** | UM Orthopedie HD | 2410 | Clin. Chir. Ortho Et Trauma | 7620 | OTONN | 53 | Chirurgie indifférenciée adulte |
| **2421** | UM Septique Osteo | 2410 | Clin. Chir. Ortho Et Trauma | 7620 | OTONN | 53 | Chirurgie indifférenciée adulte |
| **3410** | UM Neurochirurgie | 3430 | Neurochir-Neurotraumato | 7620 | OTONN | 51 | Neurochirurgie |
| **3415** | Cetd | 3430 | Neurochir-Neurotraumato | 7620 | OTONN | 61 | Unité de prise en charge de la douleur chronique |
| **3420** | UM Neurotraumatologie | 3430 | Neurochir-Neurotraumato | 7620 | OTONN | 53 | Chirurgie indifférenciée adulte |
| **1070** | UM Nephrologie | 1070 | Nephrologie-Immunologie | 7630 | Itun-Imad-Dermato-Hemato | 29 | Unité de médecine indifférenciée |
| **1111** | Hep. Gast. Entero. Ass. Nutrit | 1120 | Hep. Gast. Entero. Ass. Nutrit | 7630 | Itun-Imad-Dermato-Hemato | 29 | Unité de médecine indifférenciée |
| **1120** | UM Hep. Gast. Entero. Ass. Nutrit | 1120 | Hep. Gast. Entero. Ass. Nutrit | 7630 | Itun-Imad-Dermato-Hemato | 29 | Unité de médecine indifférenciée |
| **3790** | UM HGE Soins Intensifs 2ès | 1120 | Hep. Gast. Entero. Ass. Nutrit | 7630 | Itun-Imad-Dermato-Hemato | 202 | Autres soins intensifs |
| **1511** | Dermatologie Générale | 1510 | Dermatologie | 7630 | Itun-Imad-Dermato-Hemato | 29 | Unité de médecine indifférenciée |
| **2010** | UM Clin. Chir. Digest. Endocr. | 2010 | Clin.Chir.Dig.Endocrin. | 7630 | Itun-Imad-Dermato-Hemato | 53 | Chirurgie indifférenciée adulte |
| **3781** | UM CCDE Soins Intensifs 2ème | 2010 | Clin.Chir.Dig.Endocrin. | 7630 | Itun-Imad-Dermato-Hemato | 202 | Autres soins intensifs |

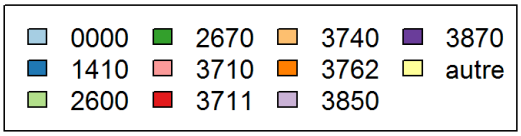
| UM | Libellé UM | Ser. | Libellé Service | pôle | Libellé Pôle | Aut. | Libellé Autorisation |
|---|---|---|---|---|---|---|---|
| **1051** | MEDECINE INTERNE 7ee - U2 | 1040 | Médecine Interne | 7660 | Med. Urg. Soins Critiques | 29 | Unité de médecine indifférenciée |
| **1064** | MEDECINE INTERNE 7ème - U1 | 1040 | Médecine Interne | 7660 | Med. Urg. Soins Critiques | 29 | Unité de médecine indifférenciée |
| **1150** | Mal. Infections Tropicales | 1150 | Mal. Infections. Tropicales | 7660 | Med. Urg. Soins Critiques | 29 | Unité de médecine indifférenciée |
| **2086** | UHTCD | 2080 | Urgences-Accueil | 7660 | Med. Urg. Soins Critiques | 29 | Unité de médecine indifférenciée |
| **2091** | ZSCD Urgences | 2081 | UHCD | 7660 | Med. Urg. Soins Critiques | 701 | UHCD structures des urgences générales |
| **2088** | Med. Polyv. Urgence 6ème | 2082 | Urgence-Medecine Polyvalente | 7660 | Med. Urg. Soins Critiques | 29 | Unité de médecine indifférenciée |
| **2096** | Med. Polyv. Urgence 5sud | 2082 | Urgence-Medecine Polyvalente | 7660 | Med. Urg. Soins Critiques | 29 | Unité de médecine indifférenciée |
| **3810** | UM Rea. Med. Polyv. Jm | 3810 | Med. Intensive-Rea HD | 7660 | Med. Urg. Soins Critiques | 101 | Réanimation adulte hors grands brûlés |
| **3811** | USC Médicale | 3810 | Med. Intensive-Rea HD | 7660 | Med. Urg. Soins Critiques | 301 | Soins surveillance continue adulte hors grands brûlés |
| **3830** | UM Rea.Chirurgicale Jm | 3830 | Anesth. Rea. Chirurgicale HD | 7660 | Med. Urg. Soins Critiques | 101 | Réanimation adulte hors grands brûlés |
| **3762** | **Unite Surv. Continue Chir.** | **3850** | **Anesth. Rea. Chirurgicale HD** | **7660** | **Med. Urg. Soins Critiques** | **301** | **Soins surveillance continue adulte hors grands brûlés** |
| **3850** | **Um Rea. Chir. Polyvalente HGRL** | **3850** | **Anesth. Rea. Chirurgicale HD** | **7660** | **Med. Urg. Soins Critiques** | **101** | **Réanimation adulte hors grands brûlés** |
| **3851** | USC Chir. Poly | 3850 | Anesth. Rea. Chirurgicale HD | 7660 | Med. Urg. Soins Critiques | 301 | Soins surveillance continue adulte hors grands brûlés |
| **3870** | **Rea CTCV** | **3850** | **Anesth. Rea. Chirurgicale HD** | **7660** | **Med. Urg. Soins Critiques** | **101** | **Réanimation adulte hors grands brûlés** |

## Annexe 8 : Tableau récapitulatif des indicateurs servant de base à l'évaluation des clusters après la première sélection de clusterings

| Clustering | Nombre de clusters | Effectif minimal | Score global | Lisibilité | Conclusion |
|---|---|---|---|---|---|
| **gp_compar2.1** | 2 | 541 | 3,68 | Moyenne | |
| **gp_compar1.1** | 3 | 270 | 2,33 | Assez bonne | |
| **C1 (gp_compar2.2)** | 4 | 54 | 2,24 | Bonne | **Retenu** |
| **C2 (gp12.1)** | 6 | 13 | 1,75 | Bonne | **Retenu** |
| **gp8.2** | 8 | 17 | 1,56 | Moyenne | |
| **gp_compar7.2** | 9 | 9 | 1,29 | Moyenne | |
| **gp_compar7.3** | 11 | 9 | 0,92 | Mauvaise | |
| **gp_compar6.4** | 12 | 11 | 0,73 | Mauvaise | |

## Annexe 9 : Représentations graphiques des parcours des 2 classifications sélectionnées

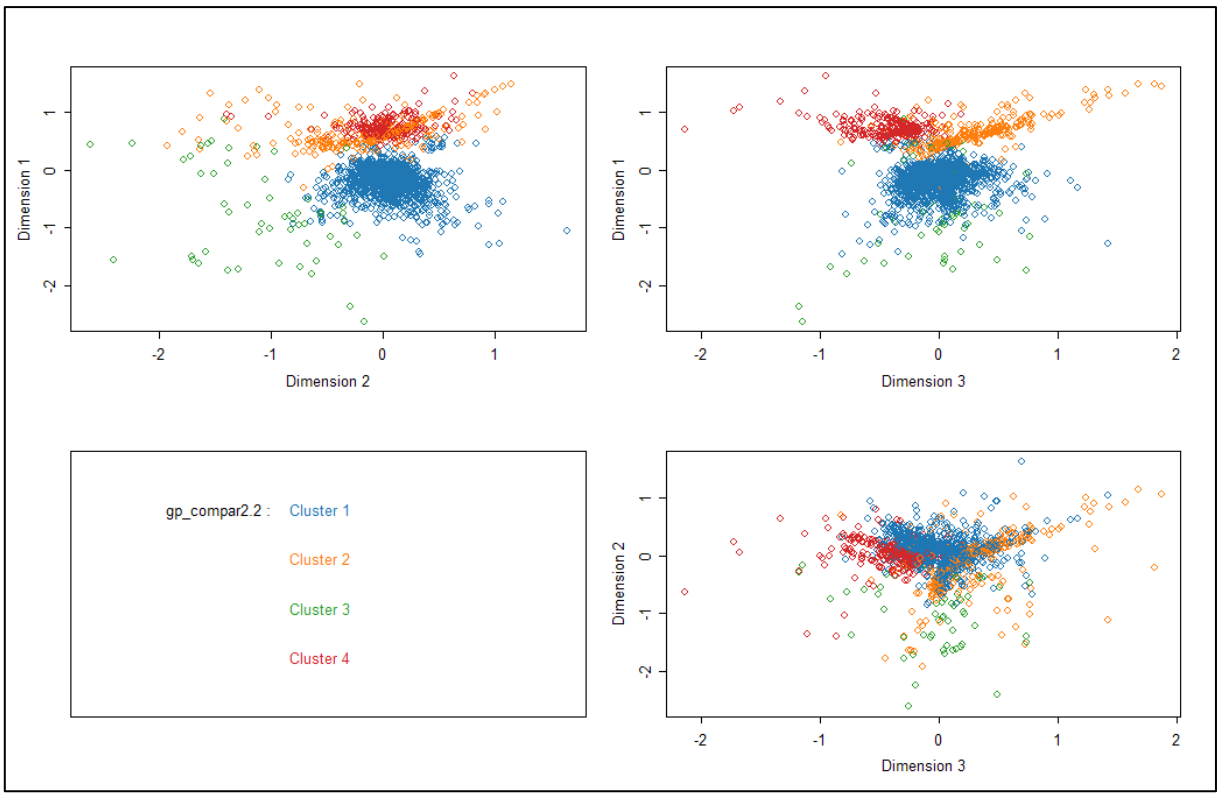
### A. Légende commune



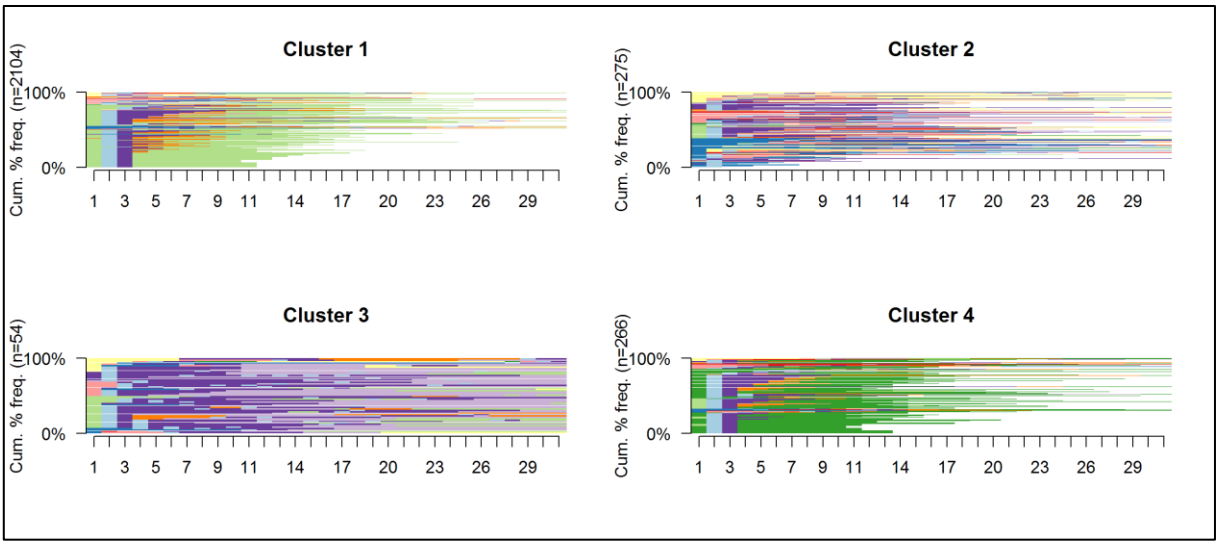

Les couleurs sont valables pour tous les graphiques de séquence de cette annexe (hors ACoP : légendes dans les projections correspondantes). Les UM sont celles qui ont été mises en gras dans le tableau en Annexe 7.

Les durées sont indiquées en jours et les parcours ont été tronqués à 3 mois (91 jours)

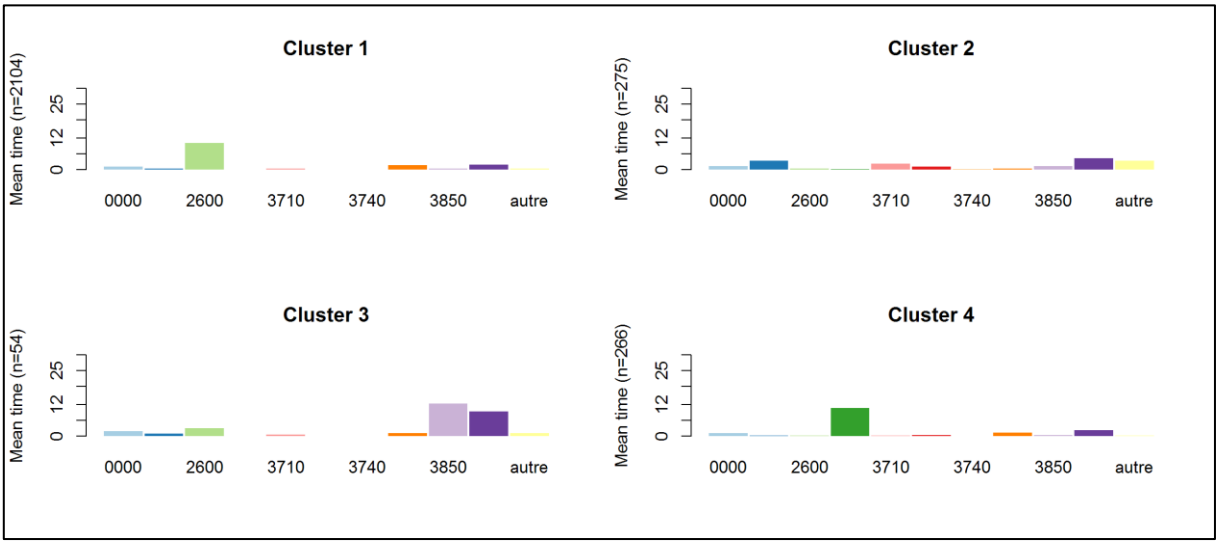
### B. Classification C1



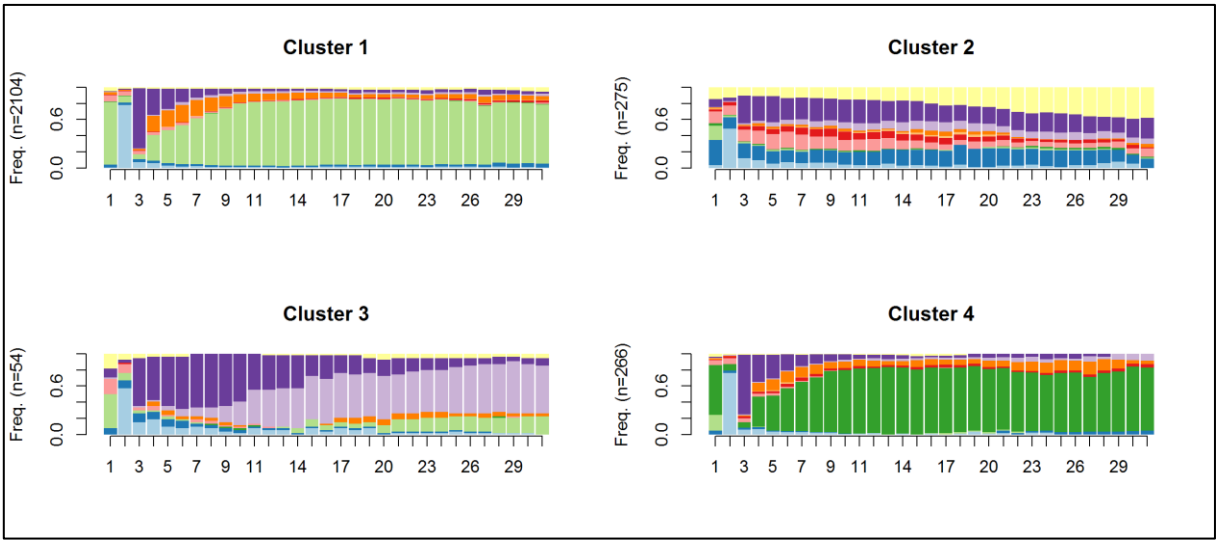

Projection des clusters sur une ACoP de la matrice de distance (C1)




Sequence frequency plot (C1)

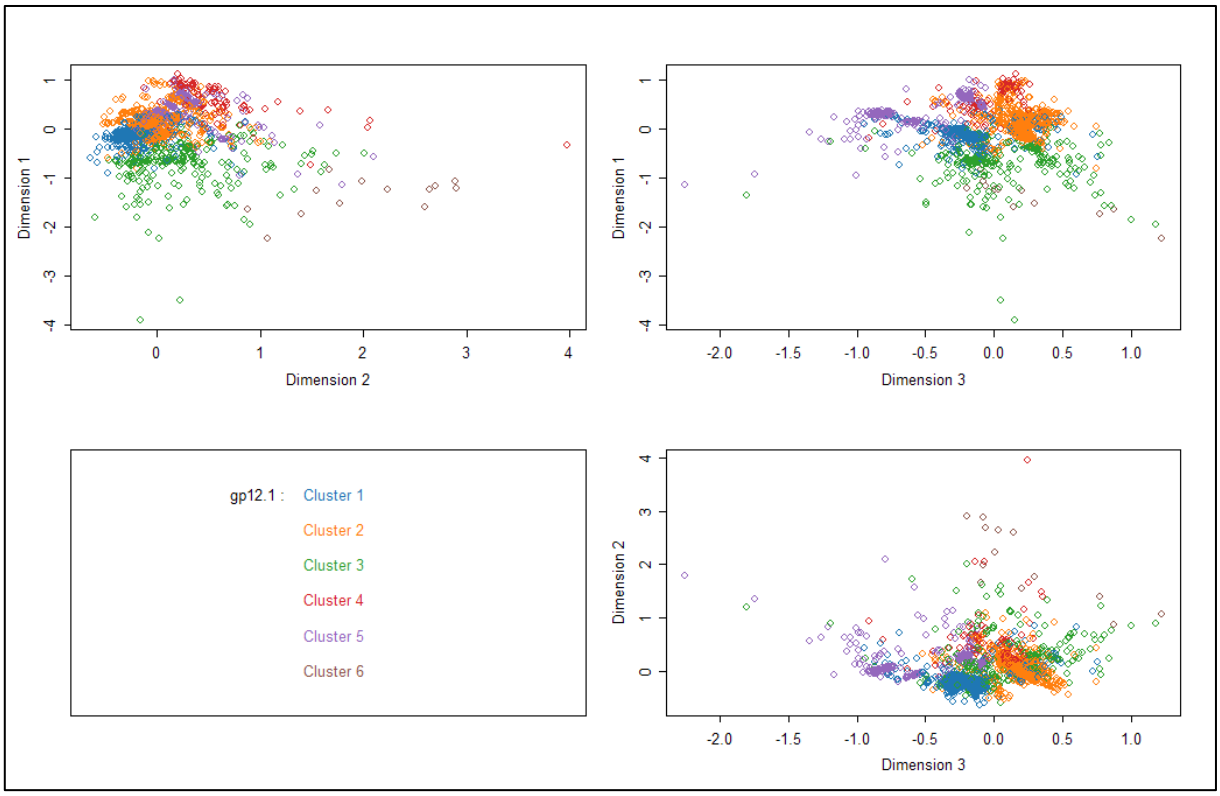



Mean time plot (C1)

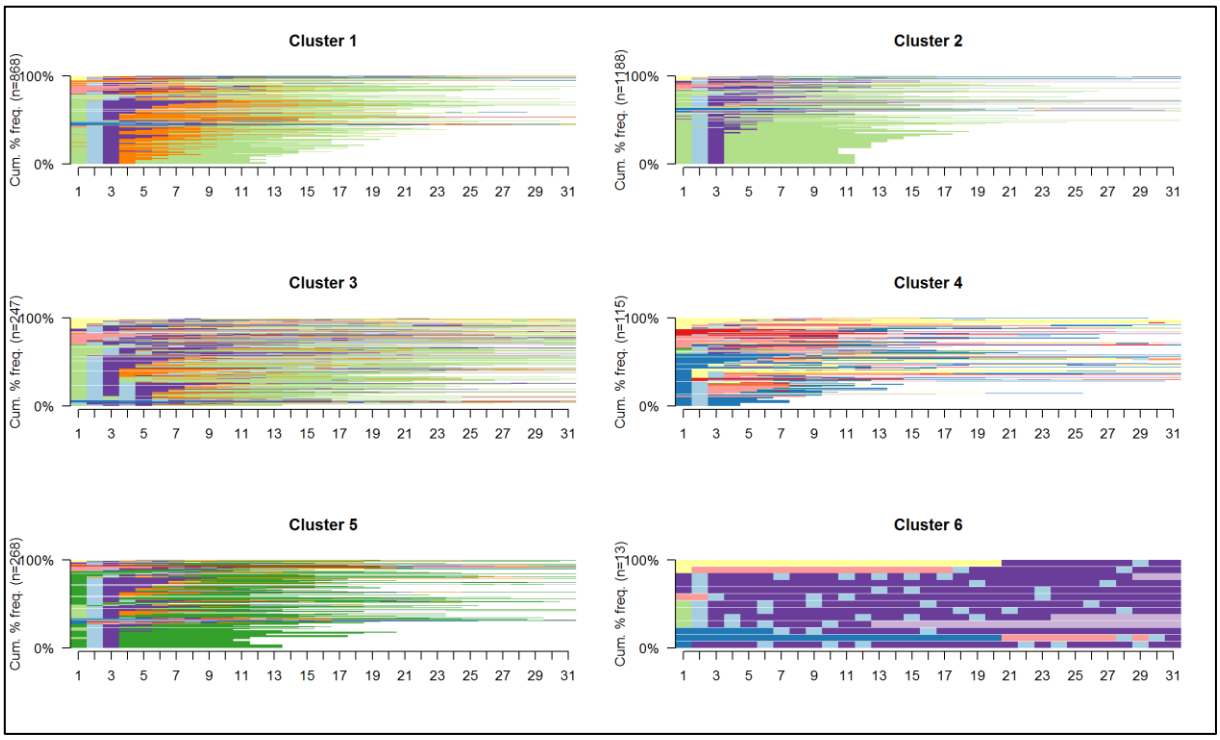



State distribution plot (C1)

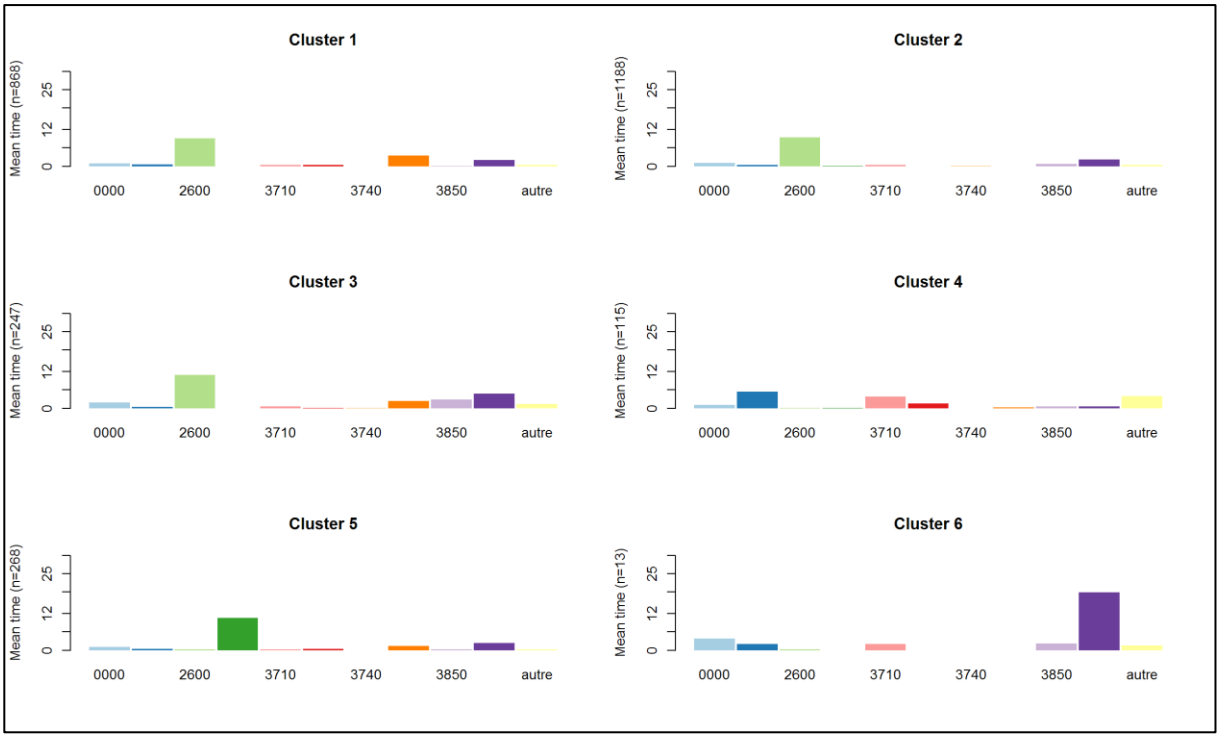
## C. Classification C2



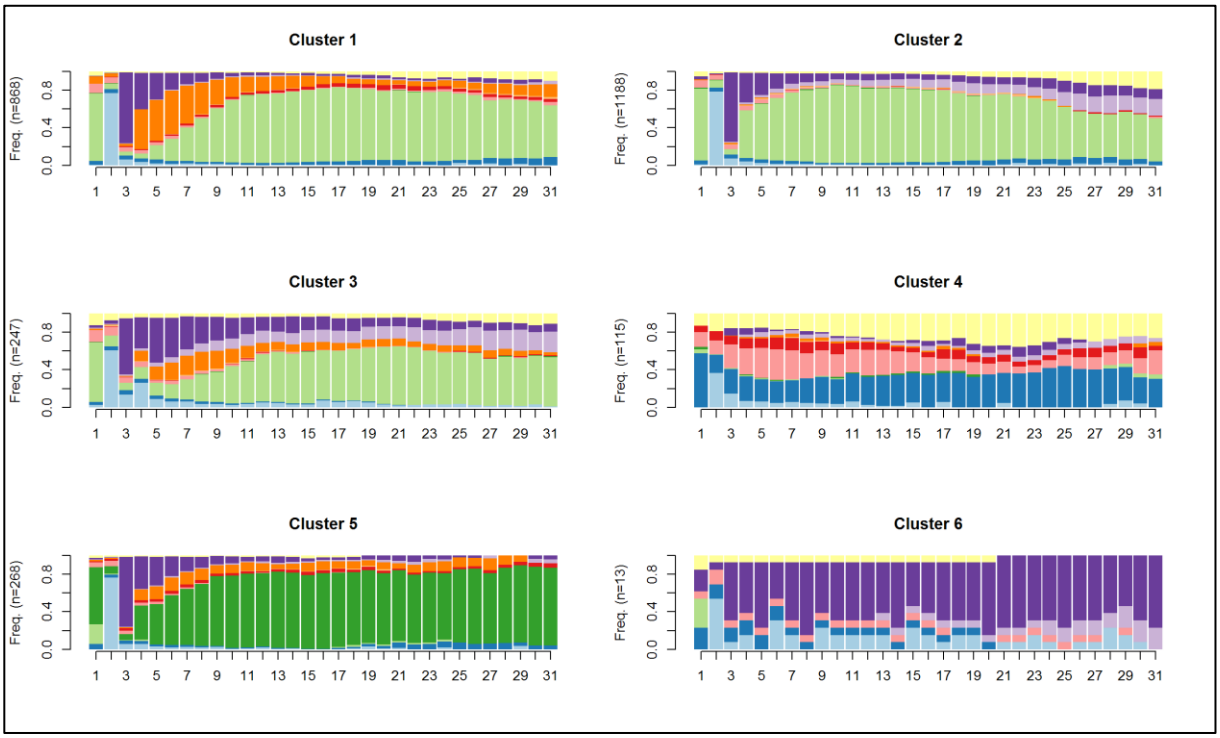

Projection des clusters sur une ACoP de la matrice de distance (C2)



Sequence frequency plot (C2)




Mean time plot (C2)




State distribution plot (C2)

## Annexe 10 : Proportion du temps passé par UM durant les hospitalisations de chaque cluster des 2 classifications retenues

Les UM correspondant aux codes sont détaillés en annexe 7.

Les résultats en gras sont les temps > 5 %.

### A. Classification C1

| UM | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| **Bloc** | 1,80 % | 4,20 % | 0,80 % | 1,60 % |
| **1020** | 0,10 % | 0 % | 0 % | 0,10 % |
| **1043** | 0,01 % | 0,10 % | 0 % | 0 % |
| **1051** | 0,02 % | 0 % | 0 % | 0 % |
| **1064** | 0 % | 0,10 % | 0 % | 0 % |
| **1070** | 0,03 % | 0,01 % | 0 % | 0 % |
| **1111** | 0 % | 0,10 % | 0 % | 0 % |
| **1120** | 0,03 % | 0 % | 0 % | 0 % |
| **1150** | 0,02 % | 0,60 % | 0,60 % | 0,10 % |
| **1160** | 0,02 % | 0 % | 0 % | 0 % |
| **1210** | 0,10 % | 0 % | 0 % | 0 % |
| **1215** | 0,03 % | 0 % | 0 % | 0 % |
| **1310** | 0,04 % | 1,00 % | 0 % | 0 % |
| **1322** | 0,10 % | 0,40 % | 0,20 % | 0,04 % |
| **1410** | 1,60 % | **19,30 %** | 1,50 % | 0,90 % |
| **1511** | 0 % | 0,20 % | 0 % | 0 % |
| **1635** | 0 % | 0,30 % | 0 % | 0 % |
| **2010** | 0,02 % | 0,10 % | 0 % | 0 % |
| **2086** | 0,03 % | 0,30 % | 0,04 % | 0,01 % |
| **2088** | 0,00 % | 0,20 % | 0 % | 0 % |
| **2091** | 0,00 % | 0 % | 0,05 % | 0 % |
| **2096** | 0 % | 0,02 % | 0,10 % | 0 % |
| **2410** | 0 % | 0,03 % | 0 % | 0 % |
| **2421** | 0 % | 0,03 % | 0 % | 0 % |
| **2600** | **68,70 %** | 2,50 % | **29,50 %** | 1,80 % |
| **2620** | 0 % | 3,80 % | 0 % | 0 % |
| **2630** | 0,03 % | 1,50 % | 0 % | 0 % |
| **2670** | 0,10 % | 1,40 % | 0,10 % | **69,30 %** |
| **3410** | 0,03 % | 0,10 % | 0 % | 0 % |
| **3415** | 0,01 % | 0 % | 0 % | 0 % |
| **3420** | 0 % | 0,20 % | 0 % | 0 % |
| **3710** | 2,00 % | **13,30 %** | 1,10 % | 1,70 % |
| **3711** | 0,80 % | **13,40 %** | 0,01 % | 1,60 % |
| **3740** | 0,30 % | 1,10 % | 0,20 % | 0,30 % |
| **3762** | **10,10 %** | 2,50 % | 3,30 % | **7,60 %** |

| UM | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| **3781** | 0,01 % | 0 % | 0 % | 0 % |
| **3790** | 0,00 % | 0,01 % | 0 % | 0,10 % |
| **3810** | 0,03 % | 0,50 % | 0,30 % | 0 % |
| **3811** | 0,00 % | 0,03 % | 0,04 % | 0 % |
| **3830** | 0,00 % | 0,10 % | 0 % | 0 % |
| **3850** | 1,60 % | **6,00 %** | **44,50 %** | 1,00 % |
| **3851** | 0,10 % | 0,20 % | 0,20 % | 0,30 % |
| **3870** | **12,30 %** | **26,30 %** | **17,50 %** | **13,50 %** |

## B. Classification C2

| UM | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
|---|---|---|---|---|---|---|
| **Bloc** | 1,70 % | 2,60 % | 1,60 % | 1,10 % | 1,60 % | 1,20 % |
| **1020** | 0 % | 0,10 % | 0,30 % | 0 % | 0,10 % | 0 % |
| **1043** | 0 % | 0,01 % | 0 % | 0,20 % | 0 % | 0 % |
| **1051** | 0 % | 0,03 % | 0,03 % | 0 % | 0 % | 0 % |
| **1064** | 0 % | 0,02 % | 0 % | 0,10 % | 0 % | 0 % |
| **1070** | 0 % | 0,10 % | 0 % | 0,02 % | 0 % | 0 % |
| **1111** | 0,03 % | 0 % | 0 % | 0 % | 0 % | 0 % |
| **1120** | 0,10 % | 0 % | 0 % | 0 % | 0 % | 0 % |
| **1150** | 0,04 % | 0 % | 0,20 % | 1,50 % | 0,10 % | 0 % |
| **1160** | 0 % | 0,03 % | 0 % | 0 % | 0 % | 0 % |
| **1210** | 0,10 % | 0,05 % | 0 % | 0 % | 0 % | 0 % |
| **1215** | 0,01 % | 0,04 % | 0,10 % | 0 % | 0 % | 0 % |
| **1310** | 0,02 % | 0,20 % | 0,20 % | 1,00 % | 0 % | 0 % |
| **1322** | 0,10 % | 0,03 % | 0,50 % | 0 % | 0,04 % | 1,00 % |
| **1410** | 2,10 % | 1,80 % | 1,30 % | **37,20 %** | 1,50 % | 4,40 % |
| **1511** | 0 % | 0 % | 0 % | 0,50 % | 0 % | 0 % |
| **1635** | 0 % | 0 % | 0 % | 0,80 % | 0 % | 0 % |
| **2010** | 0,02 % | 0,02 % | 0 % | 0,20 % | 0 % | 0 % |
| **2086** | 0,03 % | 0,10 % | 0,10 % | 0,10 % | 0,01 % | 0,10 % |
| **2088** | 0,01 % | 0 % | 0,20 % | 0 % | 0 % | 0 % |
| **2091** | 0,00 % | 0 % | 0,01 % | 0 % | 0 % | 0 % |
| **2096** | 0 % | 0,01 % | 0,03 % | 0 % | 0 % | 0 % |
| **2410** | 0 % | 0 % | 0 % | 0,10 % | 0 % | 0 % |
| **2421** | 0 % | 0 % | 0 % | 0,10 % | 0 % | 0 % |
| **2600** | **55,90 %** | **71,40 %** | **52,80 %** | 0,80 % | 2,20 % | **7,10 %** |
| **2620** | 0 % | 0,50 % | 1,30 % | 0,40 % | 0,20 % | **6,30 %** |
| **2630** | 0,02 % | 0,10 % | 1,30 % | 0,04 % | 0 % | 0 % |
| **2670** | 0,10 % | 0,50 % | 0,50 % | 0,90 % | **67,50 %** | 0 % |
| **3410** | 0,02 % | 0,04 % | 0 % | 0 % | 0 % | 0 % |
| **3415** | 0 % | 0,02 % | 0 % | 0 % | 0 % | 0 % |
| **3420** | 0 % | 0 % | 0 % | 0,50 % | 0 % | 0 % |
| **3710** | 2,00 % | 2,40 % | 1,80 % | **25,30 %** | 1,50 % | 2,80 % |
| **3711** | 3,00 % | 0,20 % | 0,50 % | **22,10 %** | 1,80 % | 0 % |
| **3740** | 0,10 % | 0,60 % | 0,30 % | 0,20 % | 0,10 % | 0 % |
| **3762** | **21,50 %** | 0,50 % | **10,20 %** | 1,30 % | **8,10 %** | 0 % |
| **3781** | 0,02 % | 0 % | 0 % | 0 % | 0 % | 0 % |
| **3790** | 0,00 % | 0 % | 0,03 % | 0 % | 0,10 % | 0 % |
| **3810** | 0,10 % | 0,02 % | 0,30 % | 0,20 % | 0 % | 2,90 % |
| **3811** | 0,00 % | 0,01 % | 0 % | 0,02 % | 0 % | 0 % |
| **3830** | 0 % | 0,03 % | 0,02 % | 0 % | 0 % | 0 % |
| **3850** | 0,60 % | 3,30 % | **10,50 %** | 2,40 % | 0,70 % | **16,30 %** |
| **3851** | 0,10 % | 0,10 % | 0,30 % | 0,30 % | 0,30 % | 0 % |
| **3870** | **12,50 %** | **15,20 %** | **15,80 %** | 2,80 % | **14,10 %** | **57,80 %** |

## Annexe 11 : Analyse descriptive des données cliniques (hors DP)

### A. Population générale

| Variable | Valeur | Proportion (effectif) ou médiane [EIQ] | Données manquantes |
|---|---|---|---|
| **Sexe :** | 1 | 73,5 (1984) | 0,0 % |
| | 2 | 26,5 (715) | |
| **Age** | | 69,0 [61,0 ; 76,0] | 0,0 % |
| **Département :** | 35 | 0,3 (8) | 0,0 % |
| | 44 | 56,7 (1529) | |
| | 49 | 6,4 (172) | |
| | 53 | 0,1 (2) | |
| | 56 | 7,0 (190) | |
| | 79 | 2,0 (54) | |
| | 85 | 24,1 (651) | |
| | Autre | 3,4 (93) | |
| **Mode Entrée PMSI :** | 71 | 9,3 (251) | 0,0 % |
| | 72 | 0,4 (11) | |
| | 80 | 87,5 (2361) | |
| | 85 | 2,8 (76) | |
| **Mode Sortie PMSI :** | 62 | 0,3 (9) | 0,0 % |
| | 63 | 0,0 (1) | |
| | 66 | 0,0 (1) | |
| | 71 | 9,6 (260) | |
| | 72 | 16,0 (431) | |
| | 73 | 0,1 (2) | |
| | 74 | 0,1 (2) | |
| | 80 | 69,8 (1885) | |
| | 90 | 4,0 (108) | |
| **Sévérité :** | 1 | 3,8 (103) | 0,0 % |
| | 2 | 55,8 (1505) | |
| | 3 | 26,5 (714) | |
| | 4 | 13,4 (361) | |
| | J | 0,0 (1) | |
| | T | 0,4 (12) | |
| | Z | 0,1 (3) | |
| **IGS2** | | 24,0 [18,0 ; 30,0] | 5,7 % |
| **IGS2 (classes) :** | ]60 ; 110] | 1,8 (49) | 0,0 % |
| | ]45 ; 60] | 3,7 (99) | |
| | ]30 ; 45] | 17,4 (469) | |
| | ]15 ; 30] | 57,1 (1541) | |
| | ]1 ; 15] | 6,0 (163) | |
| | [0 ; 1] | 8,3 (223) | |
| | NA | 5,7 (155) | |

## B. Classification C1

| Variable | Valeur | Proportion (effectif) ou médiane [EIQ] | | | | p |
|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | |
| n | | 2104 | 275 | 54 | 266 | |
| Sexe | 1 | 72,8 (1532) | 76,7 (211) | 77,8 (42) | 74,8 (199) | 0,436 |
| | 2 | 27,2 (572) | 23,3 (64) | 22,2 (12) | 25,2 (67) | |
| Age | | 69,0 [62,0 ; 76,0] | 68,0 [55,5 ; 77,0] | 65,0 [60,3 ; 73,8] | 69,5 [62,0 ; 77,0] | 0,035 |
| Dépar-tement | 35 | 0,2 (5) | 0,4 (1) | 1,9 (1) | 0,4 (1) | < 0,001 |
| | 44 | 58,0 (1221) | 48,4 (133) | 59,3 (32) | 53,8 (143) | |
| | 49 | 6,3 (133) | 7,3 (20) | 1,9 (1) | 6,8 (18) | |
| | 53 | 0,1 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| | 56 | 6,4 (135) | 7,3 (20) | 5,6 (3) | 12,0 (32) | |
| | 79 | 1,9 (41) | 4,0 (11) | 1,9 (1) | 0,4 (1) | |
| | 85 | 24,3 (511) | 25,8 (71) | 24,1 (13) | 21,1 (56) | |
| | Autre | 2,7 (56) | 6,9 (19) | 5,6 (3) | 5,6 (15) | |
| Mode Entrée PMSI | 71 | 8,0 (169) | 18,9 (52) | 9,3 (5) | 9,4 (25) | < 0,001 |
| | 72 | 0,4 (8) | 0,7 (2) | 0,0 (0) | 0,4 (1) | |
| | 80 | 89,3 (1879) | 73,5 (202) | 83,3 (45) | 88,3 (235) | |
| | 85 | 2,3 (48) | 6,9 (19) | 7,4 (4) | 1,9 (5) | |
| Mode Sortie PMSI | 62 | 0,3 (6) | 0,7 (2) | 0,0 (0) | 0,4 (1) | < 0,001 |
| | 63 | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| | 66 | 0,0 (0) | 0,0 (0) | 1,9 (1) | 0,0 (0) | |
| | 71 | 8,5 (179) | 16,4 (45) | 16,7 (9) | 10,2 (27) | |
| | 72 | 16,0 (337) | 11,6 (32) | 38,9 (21) | 15,4 (41) | |
| | 73 | 0,1 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| | 74 | 0,0 (1) | 0,0 (0) | 0,0 (0) | 0,4 (1) | |
| | 80 | 74,6 (1569) | 40,0 (110) | 22,2 (12) | 72,9 (194) | |
| | 90 | 0,5 (10) | 30,9 (85) | 20,4 (11) | 0,8 (2) | |

| Variable | Valeur | Proportion (effectif) ou médiane [EIQ] | | | | p |
|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | |
| Sévéri-té | 1 | 3,4 (72) | 9,1 (25) | 1,9 (1) | 1,9 (5) | < 0,001 |
| | 2 | 60,3 (1269) | 21,8 (60) | 0,0 (0) | 66,2 (176) | |
| | 3 | 27,7 (582) | 25,1 (69) | 3,7 (2) | 22,9 (61) | |
| | 4 | 8,6 (180) | 38,9 (107) | 94,4 (51) | 8,6 (23) | |
| | J | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| | T | 0,0 (1) | 3,6 (10) | 0,0 (0) | 0,4 (1) | |
| | Z | 0,0 (0) | 1,1 (3) | 0,0 (0) | 0,0 (0) | |
| IGS2 | | 24,0 [18,0 ; 29,0] | 25,0 [0,0 ; 50,0] | 39,0 [22,0 ; 51,0] | 24,0 [18,0 ; 29,0] | < 0,001 |
| IGS2_2 | ]60 ; 110] | 0,4 (9) | 11,3 (31) | 13,0 (7) | 0,8 (2) | < 0,001 |
| | ]45 ; 60] | 2,4 (51) | 10,9 (30) | 22,2 (12) | 2,3 (6) | |
| | ]30 ; 45] | 17,7 (373) | 11,6 (32) | 27,8 (15) | 18,4 (49) | |
| | ]15 ; 30] | 63,1 (1327) | 16,0 (44) | 18,5 (10) | 60,2 (160) | |
| | ]1 ; 15] | 6,7 (141) | 2,5 (7) | 0,0 (0) | 5,6 (15) | |
| | [0 ; 1] | 5,3 (112) | 29,5 (81) | 16,7 (9) | 7,9 (21) | |
| | NA | 4,3 (91) | 18,2 (50) | 1,9 (1) | 4,9 (13) | |

## C. Classification C2

| Variable | Valeur | Proportion (effectif) ou médiane [EIQ] | | | | | | p |
|---|---|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | |
| n | | 868 | 1188 | 247 | 115 | 268 | 13 | |
| Sexe | 1 | 72,2 (627) | 72,3 (859) | 78,9 (195) | 80,9 (93) | 73,5 (197) | 100,0 (13) | 0,021 |
| | 2 | 27,8 (241) | 27,7 (329) | 21,1 (52) | 19,1 (22) | 26,5 (71) | 0,0 (0) | |
| Age | | 70,0 [62,0 ; 77,0] | 68,0 [61,0 ; 75,0] | 68,0 [60,0 ; 76,0] | 67,0 [55,0 ; 78,0] | 70,0 [62,0 ; 77,0] | 64,0 [59,0 ; 68,0] | < 0,001 |
| Départ-tement | 35 | 0,5 (4) | 0,1 (1) | 0,8 (2) | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,018 |
| | 44 | 58,1 (504) | 58,0 (689) | 55,5 (137) | 44,3 (51) | 52,6 (141) | 53,8 (7) | |
| | 49 | 5,6 (49) | 6,6 (78) | 6,1 (15) | 10,4 (12) | 6,7 (18) | 0,0 (0) | |
| | 53 | 0,1 (1) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| | 56 | 6,5 (56) | 5,9 (70) | 7,3 (18) | 12,2 (14) | 11,6 (31) | 7,7 (1) | |
| | 79 | 2,3 (20) | 2,1 (25) | 0,8 (2) | 5,2 (6) | 0,4 (1) | 0,0 (0) | |
| | 85 | 24,1 (209) | 24,2 (287) | 26,3 (65) | 23,5 (27) | 22,4 (60) | 23,1 (3) | |
| | Autre | 2,9 (25) | 3,1 (37) | 3,2 (8) | 4,3 (5) | 6,0 (16) | 15,4 (2) | |
| Mode Entrée PMSI | 71 | 10,4 (90) | 7,4 (88) | 10,5 (26) | 17,4 (20) | 9,3 (25) | 15,4 (2) | < 0,001 |
| | 72 | 0,1 (1) | 0,5 (6) | 0,4 (1) | 0,9 (1) | 0,7 (2) | 0,0 (0) | |
| | 80 | 87,2 (757) | 89,8 (1067) | 82,2 (203) | 75,7 (87) | 88,8 (238) | 69,2 (9) | |
| | 85 | 2,3 (20) | 2,3 (27) | 6,9 (17) | 6,1 (7) | 1,1 (3) | 15,4 (2) | |

| Variable | Valeur | Proportion (effectif) ou médiane [EIQ] | | | | | | p |
|---|---|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | |
| Mode Sortie PMSI | 62 | 0,3 (3) | 0,3 (3) | 0,4 (1) | 0,9 (1) | 0,4 (1) | 0,0 (0) | < 0,001 |
| | 63 | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |
| | 66 | 0,0 (0) | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| | 71 | 10,8 (94) | 7,4 (88) | 8,9 (22) | 22,6 (26) | 10,4 (28) | 15,4 (2) | |
| | 72 | 16,8 (146) | 14,4 (171) | 23,5 (58) | 7,8 (9) | 15,3 (41) | 46,2 (6) | |
| | 73 | 0,0 (0) | 0,2 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| | 74 | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,4 (1) | 0,0 (0) | |
| | 80 | 72,0 (625) | 71,4 (848) | 57,9 (143) | 64,3 (74) | 72,0 (193) | 15,4 (2) | |
| | 90 | 0,0 (0) | 6,3 (75) | 8,9 (22) | 3,5 (4) | 1,5 (4) | 23,1 (3) | |
| Sévéri-té | 1 | 1,3 (11) | 5,8 (69) | 0,4 (1) | 14,8 (17) | 1,9 (5) | 0,0 (0) | < 0,001 |
| | 2 | 55,3 (480) | 63,0 (749) | 25,1 (62) | 32,2 (37) | 66,0 (177) | 0,0 (0) | |
| | 3 | 35,4 (307) | 19,4 (231) | 34,4 (85) | 26,1 (30) | 22,8 (61) | 0,0 (0) | |
| | 4 | 8,1 (70) | 10,6 (126) | 39,7 (98) | 26,1 (30) | 9,0 (24) | 100,0 (13) | |
| | J | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| | T | 0,0 (0) | 0,9 (11) | 0,0 (0) | 0,0 (0) | 0,4 (1) | 0,0 (0) | |
| | Z | 0,0 (0) | 0,1 (1) | 0,4 (1) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |
| IGS2 | | 25,0 [19,0 ; 31,0] | 23,0 [18,0 ; 29,0] | 26,0 [18,0 ; 37,0] | 0,0 [0,0 ; 0,0] | 24,0 [18,0 ; 30,0] | 55,0 [36,3 ; 64,8] | < 0,001 |
| IGS2_2 | ]60 ; 110] | 0,3 (3) | 2,4 (28) | 4,5 (11) | 0,0 (0) | 0,7 (2) | 38,5 (5) | < 0,001 |
| | ]45 ; 60] | 3,0 (26) | 3,9 (46) | 7,7 (19) | 0,9 (1) | 1,9 (5) | 15,4 (2) | |
| | ]30 ; 45] | 22,0 (191) | 13,5 (160) | 22,7 (56) | 6,1 (7) | 19,8 (53) | 15,4 (2) | |
| | ]15 ; 30] | 59,8 (519) | 61,9 (735) | 47,4 (117) | 7,0 (8) | 59,7 (160) | 15,4 (2) | |
| | ]1 ; 15] | 6,1 (53) | 6,9 (82) | 5,3 (13) | 0,9 (1) | 5,2 (14) | 0,0 (0) | |
| | [0 ; 1] | 6,1 (53) | 5,6 (66) | 6,1 (15) | 59,1 (68) | 7,5 (20) | 7,7 (1) | |
| | NA | 2,6 (23) | 6,0 (71) | 6,5 (16) | 26,1 (30) | 5,2 (14) | 7,7 (1) | |

## Annexe 12 : Analyse descriptive de la répartition des DP

Légende des codes de la CIM10 en Annexe 2.

Les résultats en gras sont dont la proportion est > 5 % de l'effectif du cluster concerné.

### A. Population entière

| | Groupe | Nom du groupe | Proportion (effectif) |
|---|---|---|---|
| **DP_gp** | A30-A49 | Autres maladies bactériennes | 0,1 (3) |
| | C00-C97 | Tumeurs malignes | 0,3 (7) |
| | D10-D36 | Tumeurs bénignes | 0,6 (15) |
| | D37-D48 | Tumeurs à évolution imprévisible ou inconnue | 0,1 (2) |
| | D55-D59 | Anémies hémolytiques | 0,0 (1) |
| | I05-I09 | Cardiopathies rhumatismales chroniques | 1,2 (32) |
| | I20-I25 | Cardiopathies ischémiques | 33,8 (913) |
| | I26-I28 | Affections cardiopulmonaires et maladies de la circulation pulmonaire | 0,1 (4) |
| | I30-I52 | Autres formes de cardiopathies | 54,4 (1467) |
| | I60-I69 | Maladies cérébrovasculaires | 0,2 (5) |
| | I70-I79 | Maladies des artères, artérioles et capillaires | 4,3 (115) |
| | I95-I99 | Troubles autres et non précisés de l'appareil circulatoire | 0,0 (1) |
| | J09-J18 | Grippe et pneumopathie | 0,1 (2) |
| | J40-J47 | Maladies chroniques des voies respiratoires inférieures | 0,0 (1) |
| | J80-J84 | Autres maladies respiratoires touchant principalement le tissu interstitiel | 0,0 (1) |
| | J85-J86 | Maladies suppurées et nécrotiques des voies respiratoires inférieures | 0,5 (13) |
| | J90-J94 | Autres affections de la plèvre | 0,1 (2) |
| | J95-J99 | Autres maladies de l'appareil respiratoire | 0,3 (8) |
| | K20-K31 | Maladies de l'œsophage, de l'estomac et du duodénum | 0,0 (1) |
| | K90-K93 | Autres maladies de l'appareil digestif | 0,1 (2) |

| | Groupe | Nom du groupe | Proportion (effectif) |
|---|---|---|---|
| | L00-L08 | Infections de la peau et du tissu cellulaire sous-cutané | 0,2 (5) |
| | L80-L99 | Autres affections de la peau et du tissu cellulaire sous-cutané | 0,0 (1) |
| | M00-M25 | Arthropathies | 0,0 (1) |
| | M60-M79 | Affections des tissus mous | 0,0 (1) |
| | M80-M94 | Ostéopathies et chondropathies | 0,1 (2) |
| | M95-M99 | Autres maladies du système ostéo-articulaire, des muscles et du TC | 0,0 (1) |
| | O00-O08 | Grossesse se terminant par un avortement | 0,0 (1) |
| | Q20-Q28 | Malformations congénitales de l'appareil circulatoire | 0,7 (19) |
| | R00-R09 | Symptômes et signes relatifs aux appareils circulatoire et respiratoire | 0,2 (5) |
| | R50-R69 | Symptômes et signes généraux | 0,6 (15) |
| | S00-S09 | Lésions traumatiques de la tête | 0,0 (1) |
| | S20-S29 | Lésions traumatiques du thorax | 0,2 (5) |
| | T36-T50 | Intoxications par des médicaments et des substances biologiques | 0,0 (1) |
| | T80-T88 | Complications de soins chirurgicaux et médicaux, non classées ailleurs | 1,0 (28) |
| | Z00-Z13 | Sujets en contact avec les services de santé pour des examens divers | 0,0 (1) |
| | Z40-Z54 | Sujets ayant recours aux services de santé pour des soins spécifiques | 0,6 (15) |
| | Z80-Z99 | Sujets dont la santé peut être menacée en raison d'antécédents personnels et familiaux et de certaines affections | 0,1 (2) |

## B. Classification C1

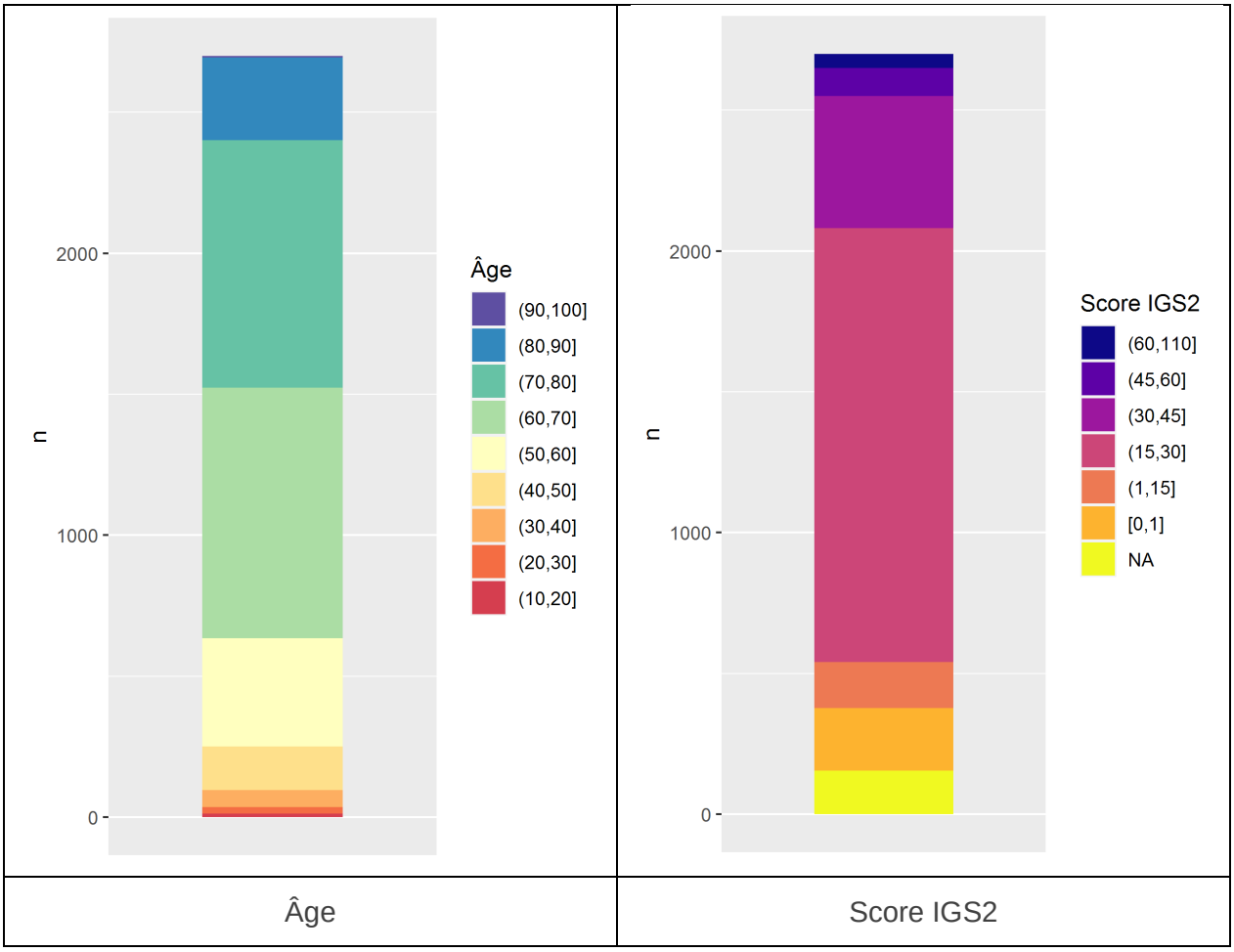
| Groupe | Proportion (effectif) | | | | p |
|---|---|---|---|---|---|
| | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | |
| n = | 2104 | 275 | 54 | 266 | |
| **A30-A49** | 0,0 (0) | 1,1 (3) | 0,0 (0) | 0,0 (0) | |
| **C00-C97** | 0,3 (6) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **D10-D36** | 0,6 (13) | 0,4 (1) | 0,0 (0) | 0,4 (1) | |
| **D37-D48** | 0,1 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **D55-D59** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **I05-I09** | 1,3 (27) | 0,7 (2) | 3,7 (2) | 0,4 (1) | |
| **I20-I25** | **36,0 (757)** | **14,9 (41)** | **16,7 (9)** | **39,8 (106)** | |
| **I26-I28** | 0,1 (2) | 0,7 (2) | 0,0 (0) | 0,0 (0) | |
| **I30-I52** | **55,0 (1157)** | **52,0 (143)** | **57,4 (31)** | **51,1 (136)** | |
| **I60-I69** | 0,2 (5) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **I70-I79** | 3,4 (72) | **8,7 (24)** | **20,4 (11)** | 3,0 (8) | |
| **I95-I99** | 0,0 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **J09-J18** | 0,0 (0) | 0,7 (2) | 0,0 (0) | 0,0 (0) | |
| **J40-J47** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **J80-J84** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **J85-J86** | 0,5 (10) | 0,4 (1) | 0,0 (0) | 0,8 (2) | |
| **J90-J94** | 0,1 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **J95-J99** | 0,3 (7) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **K20-K31** | 0,0 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | < 0,001 |
| **K90-K93** | 0,0 (1) | 0,0 (0) | 0,0 (0) | 0,4 (1) | |
| **L00-L08** | 0,1 (3) | 0,7 (2) | 0,0 (0) | 0,0 (0) | |
| **L80-L99** | 0,0 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **M00-M25** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **M60-M79** | 0,0 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **M80-M94** | 0,1 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **M95-M99** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **O00-O08** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **Q20-Q28** | 0,5 (11) | 0,4 (1) | 0,0 (0) | 2,6 (7) | |
| **R00-R09** | 0,1 (3) | 0,4 (1) | 0,0 (0) | 0,4 (1) | |
| **R50-R69** | 0,1 (2) | 4,0 (11) | 1,9 (1) | 0,4 (1) | |
| **S00-S09** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **S20-S29** | 0,2 (4) | 0,0 (0) | 0,0 (0) | 0,4 (1) | |
| **T36-T50** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **T80-T88** | 0,6 (13) | **5,5 (15)** | 0,0 (0) | 0,0 (0) | |
| **Z00-Z13** | 0,0 (0) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |
| **Z40-Z54** | 0,0 (0) | **5,1 (14)** | 0,0 (0) | 0,4 (1) | |
| **Z80-Z99** | 0,0 (1) | 0,4 (1) | 0,0 (0) | 0,0 (0) | |

## C. Classification C2

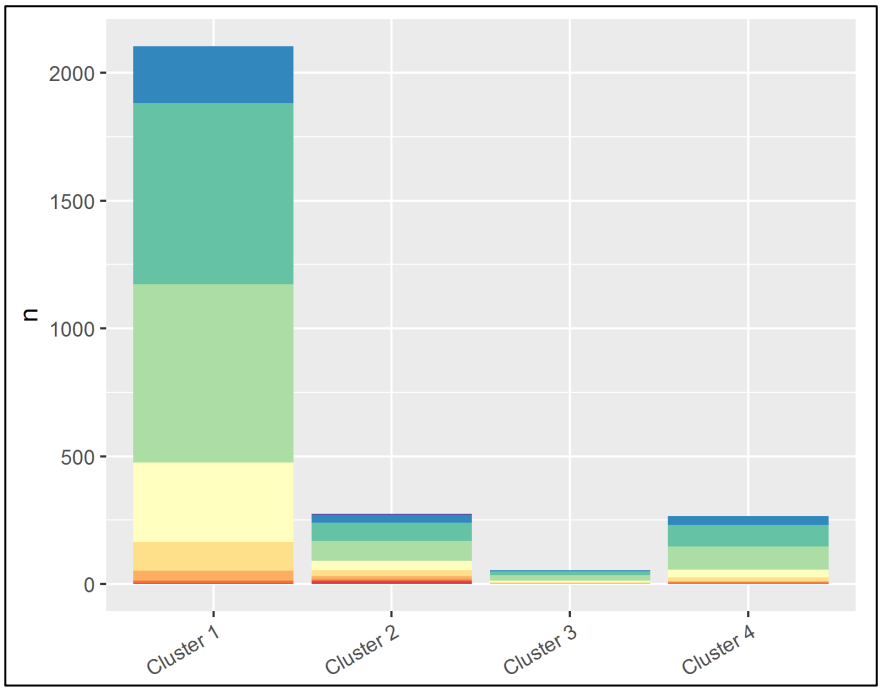
| Groupe | Proportion (effectif) | | | | | | p |
|---|---|---|---|---|---|---|---|
| | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | |
| n = | 868 | 1188 | 247 | 115 | 268 | 13 | |
| **A30-A49** | 0,0 (0) | 0,0 (0) | 0,4 (1) | 1,7 (2) | 0,0 (0) | 0,0 (0) | < 0,001 |
| **C00-C97** | 0,1 (1) | 0,4 (5) | 0,0 (0) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |
| **D10-D36** | 0,5 (4) | 0,8 (10) | 0,0 (0) | 0,0 (0) | 0,4 (1) | 0,0 (0) | |
| **D37-D48** | 0,1 (1) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **D55-D59** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **I05-I09** | 1,6 (14) | 1,2 (14) | 1,2 (3) | 0,0 (0) | 0,4 (1) | 0,0 (0) | |
| **I20-I25** | **33,3 (289)** | **37,8 (449)** | **23,1 (57)** | **7,8 (9)** | **38,8 (104)** | **38,5 (5)** | |
| **I26-I28** | 0,0 (0) | 0,2 (2) | 0,8 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **I30-I52** | **58,9 (511)** | **50,3 (598)** | **58,7 (145)** | **60,9 (70)** | **52,2 (140)** | **23,1 (3)** | |
| **I60-I69** | 0,1 (1) | 0,2 (2) | 0,8 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **I70-I79** | 4,0 (35) | 3,7 (44) | **10,9 (27)** | 0,9 (1) | 3,0 (8) | 0,0 (0) | |
| **I95-I99** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **J09-J18** | 0,0 (0) | 0,2 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **J40-J47** | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |
| **J80-J84** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **J85-J86** | 0,2 (2) | 0,6 (7) | 0,8 (2) | 0,0 (0) | 0,7 (2) | 0,0 (0) | |
| **J90-J94** | 0,0 (0) | 0,1 (1) | 0,4 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **J95-J99** | 0,0 (0) | 0,5 (6) | 0,8 (2) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **K20-K31** | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **K90-K93** | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,4 (1) | 0,0 (0) | |
| **L00-L08** | 0,0 (0) | 0,3 (3) | 0,0 (0) | 1,7 (2) | 0,0 (0) | 0,0 (0) | |
| **L80-L99** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **M00-M25** | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |
| **M60-M79** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **M80-M94** | 0,0 (0) | 0,1 (1) | 0,4 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **M95-M99** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **O00-O08** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **Q20-Q28** | 0,2 (2) | 0,7 (8) | 0,4 (1) | 0,0 (0) | 2,6 (7) | 7,7 (1) | |
| **R00-R09** | 0,0 (0) | 0,3 (4) | 0,0 (0) | 0,0 (0) | 0,4 (1) | 0,0 (0) | |
| **R50-R69** | 0,1 (1) | 0,5 (6) | 1,2 (3) | 0,9 (1) | 0,4 (1) | **23,1 (3)** | |
| **S00-S09** | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |
| **S20-S29** | 0,0 (0) | 0,3 (4) | 0,0 (0) | 0,0 (0) | 0,4 (1) | 0,0 (0) | |
| **T36-T50** | 0,0 (0) | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,0 (0) | |
| **T80-T88** | 0,5 (4) | 1,0 (12) | 0,0 (0) | **10,4 (12)** | 0,0 (0) | 0,0 (0) | |
| **Z00-Z13** | 0,0 (0) | 0,0 (0) | 0,0 (0) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |
| **Z40-Z54** | 0,0 (0) | 0,1 (1) | 0,0 (0) | **10,4 (12)** | 0,4 (1) | **7,7 (1)** | |
| **Z80-Z99** | 0,1 (1) | 0,0 (0) | 0,0 (0) | 0,9 (1) | 0,0 (0) | 0,0 (0) | |

## Annexe 13 : Répartition des variables continues recodées en classes (âge et IGS2)

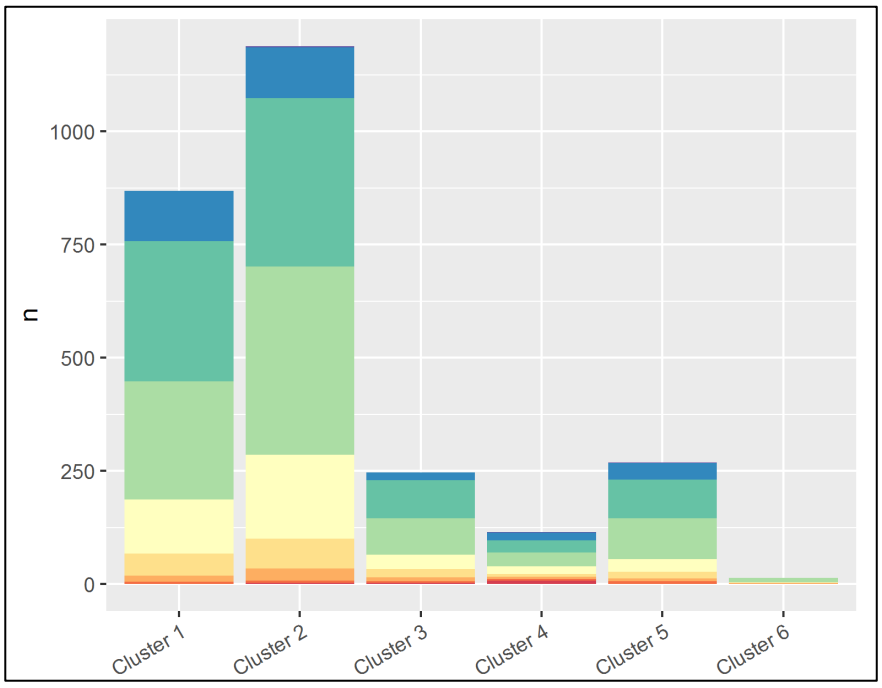
### A. Échantillon entier et légende



## B. Répartition des classes d’âge dans les classifications C1 et C2

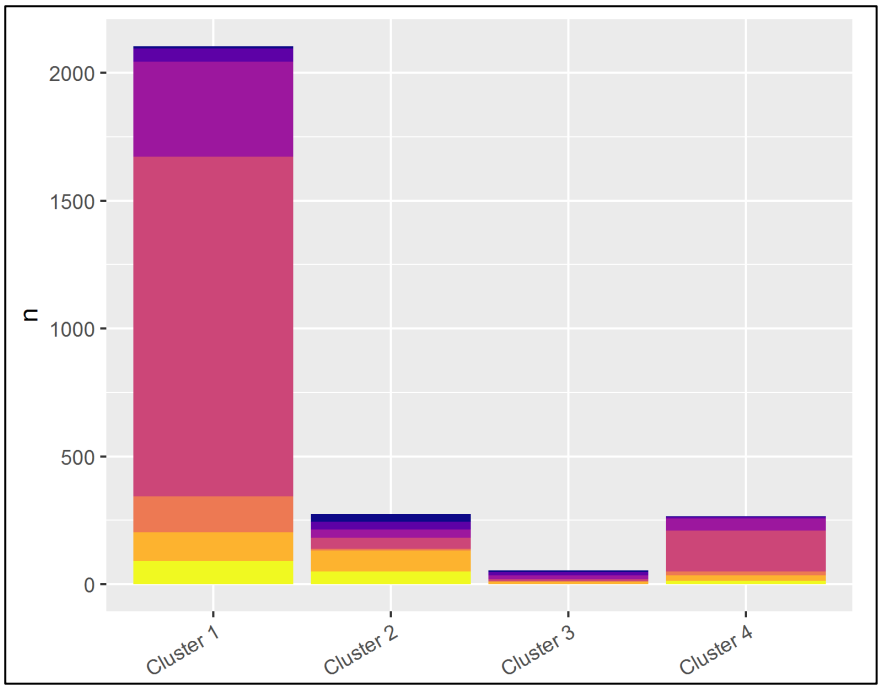



Classification C1

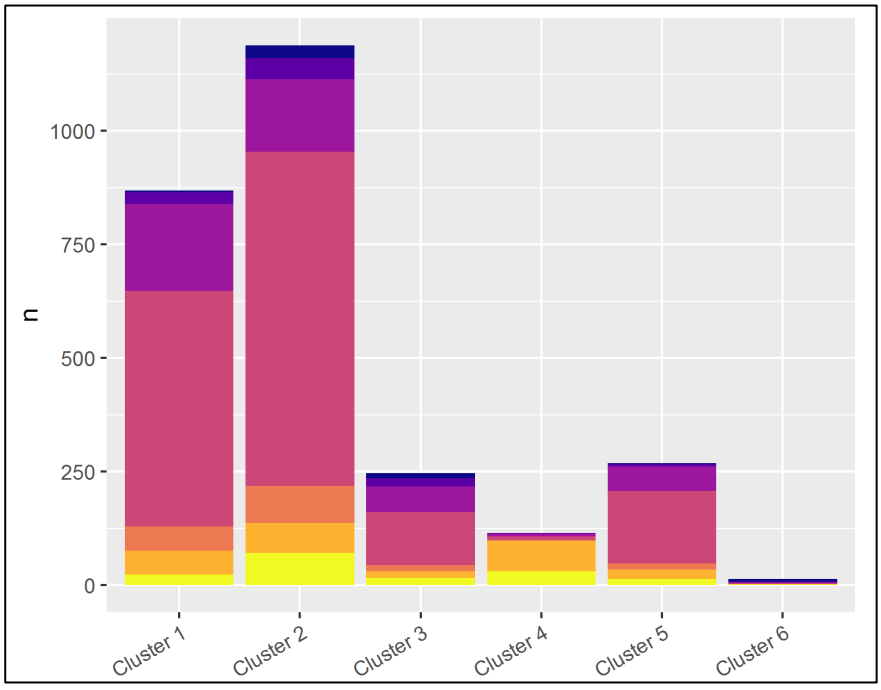



Classification C2

## C. Répartition des classes de score IGS2 dans les classifications C1 et C2



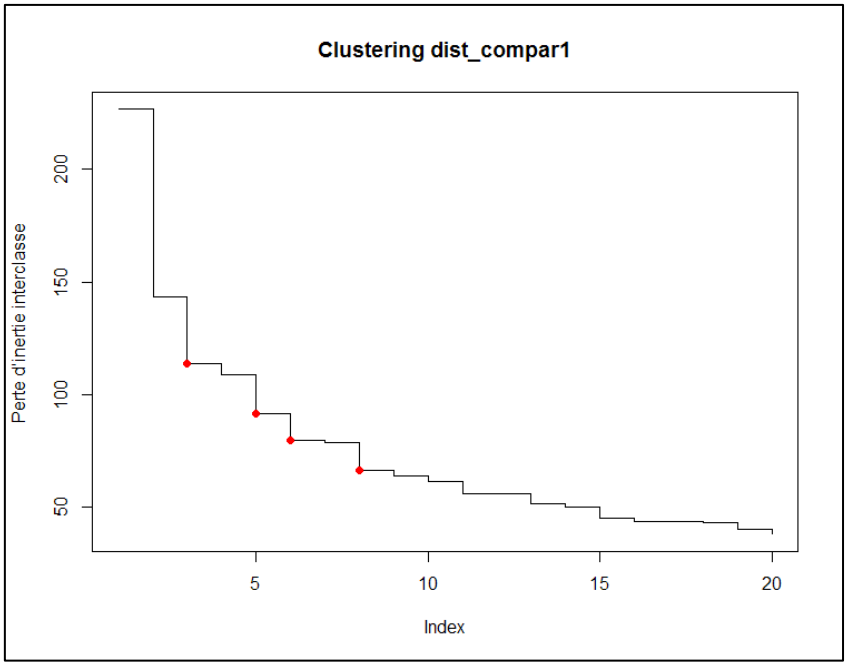

Classification C1



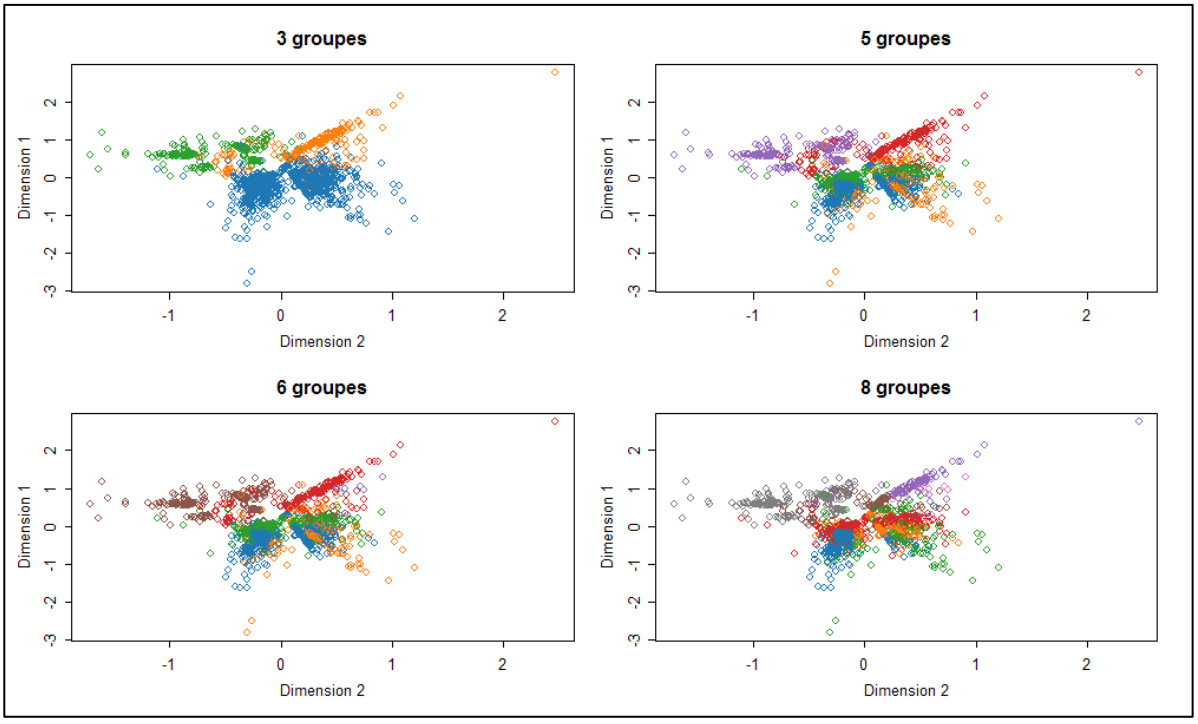

Classification C2

## Annexe 14 : Exemple de choix de seuils pour le nombre de groupes sur une CAH

### A. Perte d'inertie interclasse en fonction des regroupements lors de la CAH effectuée sur la matrice dist_compar1




### B. Représentation graphique de différents choix du nombre de groupes pour une CAH sur la matrice dist_compar1, via une projection par ACoP



NOM : TRUTT PRENOM : Lucile

**Titre de Thèse :** Représentation et description des parcours patients en chirurgie cardiaque : une approche exploratoire par le clustering

---

RESUME

La prise en charge médicale des patients ne peut plus être étudiée de façon épisodique et ponctuelle, mais doit s'inscrire dans la durée sous forme de parcours. Cette représentation nécessite l'utilisation d'outils et d'analyse adaptés aux données séquentielles. En prenant l'exemple des patients ayant bénéficié d'une chirurgie cardiaque, nous avons étudié la faisabilité d'un clustering de leurs parcours hospitaliers en nous basant sur une méthode d'appariement optimal (*optimal matching*). Malgré un grand déséquilibre dans la répartition de nos clusters, ceux-ci ont pu se révéler intéressants, aussi bien au niveau de la description des données de parcours que des données cliniques. Cette première expérimentation ouvre des pistes de réflexion et un premier tour des difficultés à prévoir pour d'autres études basée sur cette approche.

---

**MOTS-CLES**

PROGRAMME CLINIQUE, PLANIFICATION HOSPITALIÈRE, GESTION DES RESSOURCES EN ÉQUIPE EN SOINS DE SANTÉ, ARCHIVES ADMINISTRATIVES HOSPITALIÈRES, ÉVALUATION DE PROCESSUS EN SOINS DE SANTÉ, STATISTIQUES COMME SUJET, DONNÉES PRÉLIMINAIRES, ÉTUDES DE FAISABILITÉ, ANALYSE DE REGROUPEMENTS, CLASSIFICATION, CHRONOLOGIE, REPRÉSENTATION GRAPHIQUE, INFORMATIQUE MÉDICALE