Prédiction des parcours sous l'angle médico-économique : Une approche basée sur l'intelligence artificielle

Alice Martin

N° d'ordre NNT : 2023ISAL0050

THESE de DOCTORAT DE L'INSA LYON, membre de l'Université de Lyon

Ecole Doctorale N° 512
(INFORMATIQUE ET MATHEMATIQUES)

Spécialité/ discipline de doctorat :
Informatique et applications

Soutenue publiquement le 12/07/2023, par :
Alice MARTIN

Prédiction des parcours sous l'angle médico-économique : une approche basée sur l'intelligence artificielle

---

Devant le jury composé de :

DI MASCOLO, Maria Directrice de recherche CNRS Présidente

LAHRICHI, Nadia Professeure Université Laval du Québec Rapporteure
SOUALMIA, Lina Professeure Université de Rouen Rapporteure
AUGUSTO, Vincent Professeur Ecole des Mines de St Etienne Examinateur
DI MASCOLO, Maria Directrice de recherche CNRS Examinatrice
ROBARDET, Céline Professeure INSA Lyon Examinatrice
GUINET, Alain Professeur émérite INSA Lyon Directeur de thèse
FONDREVELLE, Julien Maître de Conférences INSA Lyon Co-encadrant de thèse
GUILLAUME, Jean-Baptiste Ingénieur IAC Partners Invité
RAKOTONDRANAIVO Auguste Médecin gériatre, Maître de Conférences Université de Lorraine Invité

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **https://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **https://edeea.universite-lyon.fr** Sec. : Stéphanie CAUVIN Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 **philippe.delachartre@insa-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**Mme Sandrine CHARLES** Université Claude Bernard Lyon 1 UFR Biosciences Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69622 Villeurbanne CEDEX **sandrine.charles@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://ediss.universite-lyon.fr** Sec. : Bénédicte LANZA Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 **stephane.benayoun@ec-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **https://edsciencessociales.universite-lyon.fr** Sec. : Mélina FAVETON INSA : J.Y. TOUSSAINT Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Bruno MILLY** Université Lumière Lyon 2 86 Rue Pasteur 69365 Lyon CEDEX 07 **bruno.milly@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

PREDICTION *des* PARCOURS *sous l'*ANGLE MEDICO-ECONOMIQUE *: une approche basée sur l'*INTELLIGENCE ARTIFICIELLE

PROJET PROACTIVE

Alice Martin

Mots-clefs

In English.

Patient journey; Care trajectory; Neurocognitive disorders; Homecare Hospital; Medical-economic simulation; Unsupervised classification; Supervised classification; Operational research; Markov chain.

En français.

Parcours patient ; Parcours de soins ; Troubles neurocognitifs ; Hospitalisation à Domicile (HAD) ; Simulation médico-économique ; Classification non supervisée ; Classification supervisée ; Recherche opérationnelle ; Chaîne de Markov.

Synthèse

In English. Health care providers are experiencing difficulties in the organization of care and the management of their patients, particularly chronic ones. These block roads are multiple, including the increasing prevalence of chronic diseases and aging of the population, territorial divide in access to care, pressure on costs and efficiency - and can have a strong impact on the health prospects of populations. Healthcare organizations, particularly hospitals, are trying to overcome these difficulties by optimizing patient and care pathways as a whole. Within these trajectories, one of the levers for efficiency is to be able to understand which aspects of a patient's profile are correlated with the events that have an impact on the use and consumption of care, so that they can be anticipated.

Recent technological advances in AI make it possible to study a wide variety of care pathways and to analyze a broad range of variables. In this work, we want to represent and analyze the pathways in several clinical contexts, using billing and medico-economic data as a proxy to reconstruct the individual trajectory of a patient. The ultimate goal is to relieve the operational pressures on hospital resources while improving the comfort of patient care.

We present two case studies: health journey prediction from a real-life cohort of patients with neurocognitive disorders and care prediction from a homecare hospital in France. We portrayed the trajectories and investigated the main drivers of variation using patient clinical characteristics including disease progression. We then used the same drivers to predict patient journey variation. Our methodology was built on two steps: identifying relevant medico-economical groups of patients, through clustering for example – then predicting the variation of care required through time. Our model allowed to predict patient journey variation with an accuracy ranging from of 60.5% to 90% depending on scenarios.

En français. Les structures de santé rencontrent des difficultés structurelles dans l'organisation des soins et la prise en charge de leurs patients, notamment chroniques. Ces blocages sont multiples - prévalence croissante des maladies chroniques et vieillissement de la population, fracture territoriale dans l'accès aux soins, pression sur les coûts et l'efficience - et peuvent avoir un fort impact sur les perspectives de santé des populations. Les organisations, en particulier les hôpitaux, tentent de surmonter ces barrières en optimisant les parcours patients et de soins. Au sein de ces trajectoires, l'un des leviers d'efficience est de pouvoir comprendre quels aspects du profil d'un patient sont corrélés aux événements ayant un impact sur le recours et la consommation de soins, afin de pouvoir les anticiper.

Les récentes avancées technologiques en matière d'IA permettent d'étudier une grande variété de parcours et d'analyser un large panel de variables. Dans ces travaux, nous souhaitons représenter et analyser les parcours dans plusieurs contextes cliniques, en utilisant les données médico-économiques et de facturation comme proxy pour reconstruire la trajectoire individuelle d'un patient. L'objectif final est d'alléger les pressions opérationnelles sur les ressources hospitalières tout en améliorant le confort des soins et la qualité de la prise en charge.

Nous présentons deux études de cas : la prédiction du parcours d'une cohorte en vie réelle de patients atteints de troubles neurocognitifs et la prédiction des soins d'un hôpital à domicile en France. Nous avons décrit les trajectoires et étudié les principaux facteurs de variation en utilisant les caractéristiques cliniques des patients, y compris l'évolution de la maladie. Nous avons ensuite utilisé ces mêmes facteurs pour prédire la variation du parcours des patients. Notre méthodologie s'articule autour de deux étapes : l'identification de groupes médico-économiques de patients pertinents, par le biais d'un regroupement par exemple, puis la prédiction des soins requis au fil du temps. Notre modèle a permis de prédire les variations des parcours avec une précision allant de 60,5 % à 90 % selon les scénarios.

Remerciements

Ce projet de recherche n'aurait pas pu voir le jour sans le soutien de Jean-Baptiste Guillaume, Nicolas Grangier et Olivier Saint-Esprit, associés d'IAC Partners. Je les remercie chaleureusement pour toutes les opportunités offertes depuis mon arrivée chez IAC, et notamment de m'avoir accompagnée pour transformer quelques folles idées en projets concrets.

Cette thèse a abouti grâce aux conseils, toujours avisés, de mes encadrants, Alyn Guinet et Julien Fondrevelle. Je suis ravie d'être la dernière doctorante d'Alyn, à qui je ne peux que souhaiter une excellente retraite, amplement méritée. Un grand merci également à vous Julien, qui avez suivi mon parcours de jeune étudiante à docteure (et corrigé ce manuscrit à la virgule près).

Je remercie aussi tous les contributeurs de ces travaux, en premier lieu Victor Manach, les équipes de Memora et de Soins et Santé, et les rapporteurs et examinateurs de ce manuscrit. Tous les soutiens reçus sont par ailleurs des contributions indirectes, et je ne peux que saluer mes collègues, amis et proches qui ont supporté mes doléances.

Je ne serais, bien entendu, jamais allée si loin sans le soutien inconditionnel de mes parents, Anne et Thierry, et de ma sœur Fanny. Ce manuscrit est aussi le vôtre. A défaut de le lire entièrement (je ne vous en voudrai pas), il trouvera certainement une place de choix sur une étagère de la maison.

"To err is human but to really foul up requires a computer"

Paul Ehrlich

Sommaire

Glossaire

Sigle	Définition
ALD	Affection Longue Durée
AP-HP	Assistance Publique Hôpitaux de Paris
APL	Accessibilité Potentielle Localisée
ARS	Agences Régionales de Santé
ATIH	Agence Technique de l'Information sur l'Hospitalisation
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
BNA	Banque Nationale Alzheimer
CIM-10	Classification Internationale Maladies, 10e révision
Classification ATC	Classification Anatomique, Thérapeutique et Chimique
CMD	Catégorie Majeure de Diagnostic
CMRR	Centre Mémoire Ressource Recherche
CNAM	Caisse Nationale de l'Assurance Maladie
CPOM	Contrat Pluriannuel d'Objectifs et de Moyens
CPU	Central Processing Unit
DGOS	Direction Générale de l'Offre de Soins
DIM	Département d'Informations Médicales
DREES	Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques
EHR	Electronic Health Record
ERP	Enterprise Resource Planning
ESPIC	Etablissement de Santé Privé d'Intérêt Collectif
FNEHAD	Fédération Nationale des Etablissements d'Hospitalisation à Domicile
GHM	Groupe Homogène de Maladie
GHPC	Groupe Homogène de Prise en Charge
GHS	Groupe Homogène de Séjour
GPU	Graphics Processing Unit
HAD	Hospitalisation à Domicile
HCFiPS	Haut Conseil du Financement de la Protection sociale
HCL	Hospices Civils de Lyon
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
IA	Intelligence Artificielle
IADL	Instrumental Activities of Daily Living
IDE	Infirmier Diplômé d'Etat
IDEC	Infirmier Diplomé d'Etat de Coordination
IDEL	Infirmier Diplômé d'Etat Libéral
IK	Indice de Karnofsky
INSEE	Institut National des Statistiques et des Etudes Economiques
IRDES	Institut de Recherche et Documentation en Economie de la Santé
MCO	Médecine, Chirurgie, Obstétrique et Odontologie
MMSE	Mini Mental State Examination
MPA	Mode de Prise en Charge Associé
MPP	Mode de Prise en Charge Principal
NPI	Inventaire NeuroPsychiatrique
ONDAM	Objectif National des Dépenses d'Assurance Maladie
PEC	Prise En Charge

Sigle	Définition
PMSI	Programme de Médicalisatin des Systèmes d'Information
PSY	Psychiatrie
SMR	Soins Médicaux et Réadaptation, anciennement Soins de Suite et Réadaptation (SSR)
SSIAD	Service de Soins Infirmiers à Domicile
SVM	Support Vector Machine
T2A	Tarification A l'Activité
TNC	Trouble NeuroCognitif
t-SNE	t-distributed Stochastic Neighbor Embedding
VIF	Variance Influence Factor

1

Introduction et mise en perspective du parcours patient

Ce premier chapitre décrit les problématiques contemporaines d'organisation de la santé en France et les vecteurs d'innovation qui sont développés pour y répondre.

Contenu

1.1 Articulation de la problématique générale

SYNOPSIS Où nous explicitons les phénomènes qui ont abouti à la problématique générale traitée par ces travaux de recherche.

Plusieurs mécanismes mettent à l'épreuve l'écosystème de la santé en France, notamment :

1. Des mouvements démographiques : à l'instar de ses voisins européens, la France connaît un vieillissement de sa population et une augmentation de l'espérance de vie. On observe en conséquence depuis plusieurs décennies une croissance des maladies chroniques et liées au grand âge (à titre d'exemple, +50% de prévalence du diabète entre 2010 et 2020) ; un accroissement qui pèse sur le système de soins.

2. Des inégalités d'accès aux soins : Les inégalités territoriales, de revenus et socio-culturelles impactent négativement l'accès aux soins des individus. Elles engendrent un retard dans l'identification des pathologies et entravent les recours aux soins aux étapes clés du parcours patient.

3. Une remise en cause du système de financement : la tarification à l'activité a été déployée en 2003 pour créer une emphase sur l'efficience des actes de soins. Ce système est aujourd'hui de plus en plus décrié, notamment car il ne favorise pas le développement de l'innovation. Or les hôpitaux font face à des contraintes économiques et opérationnelles grandissantes, et tendent à se transformer en plateaux techniques prenant en charge des patients sur des séjours courts pour réaliser des actes spécifiques.

Ces phénomènes induisent sur le système de santé plusieurs changements de paradigmes, plus particulièrement une nécessité de considérer les prises en charge des patients de manière globale et coordonnée, et donc de se recentrer autour des parcours patients et de

soins[1] en s'éloignant de la vue découpée à l'acte. On retrouve par ailleurs un effort commun des acteurs de la santé pour accroître l'efficience opérationnelle et financière du système tout en démontrant la valeur apportée par le soin dans la prise en charge.

4. Des avancées technologiques : parallèlement, on assiste à une disponibilité croissante des données de santé et à des progrès considérables dans l'applicabilité de l'intelligence artificielle à des bases de données en vie réelle y compris les données de parcours patients.

Ces changements systémiques bousculent l'écosystème de la santé, qui se voit contraint de transformer son offre de soins autour du parcours patient avec un impératif d'efficience globale, notamment de coût. Pour que tous les acteurs soient impliqués dans des parcours individuels et optimisés, il est nécessaire d'être en mesure de les caractériser puis de les anticiper. L'intelligence artificielle est un outil prometteur pour répondre à ces enjeux. Nous souhaitons explorer son impact dans ces travaux.

[1] Différence entre parcours de soins et parcours patient. Le parcours patient se réfère à l'ensemble des étapes que le patient va traverser dans son expérience d'une condition donnée, de l'apparition des premiers symptômes à la guérison. Ces étapes sont animées par de multiples effecteurs : médecine de ville, hôpital, pharmacie, etc. Le parcours de soins, quant à lui, se concentre sur la coordination de plusieurs professionnels de santé pour effectuer un ensemble de soins destiné à traiter une condition. Dans ces travaux, nous nous concentrerons sur les parcours de soins intra-hospitalier c'est-à-dire relevant d'une seule structure de santé.

1.2 Le contexte de la santé en France

SYNOPSIS Où nous décrivons les grands phénomènes qui se manifestent au sein de l'écosystème de la santé en France et comment ceux-ci génèrent des problématiques qui fédèrent les acteurs du soin autour du parcours patient.

1.2.1 LA POPULATION VIEILLISSANTE ET LA PROGRESSION DES MALADIES CHRONIQUES PESENT SUR LE SYSTEME DE SANTE

Cette section présente l'évolution récente de la démographie française et son impact sur la prévalence des maladies chroniques et donc sur le système de santé.

Caractérisation de la démographie française. Au 1er janvier 2020, la France compte plus de 67 millions d'individus contre 60,5 millions en 2000[2].

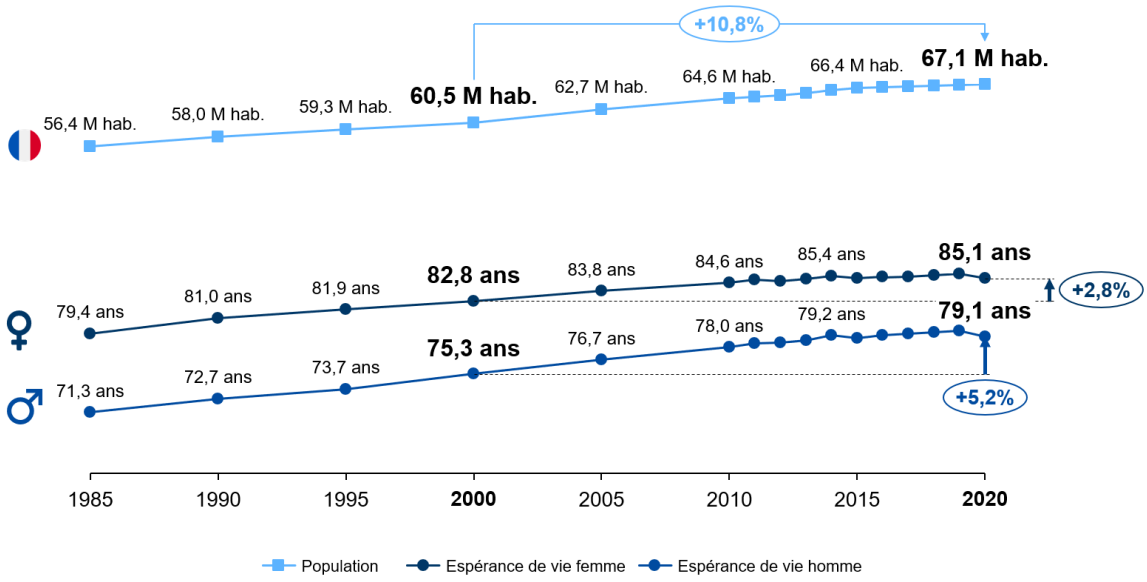


Figure 1 : Progression de la population française et de l'espérance de vie des hommes et des femmes entre 1985 et 2020 - Source INSEE

[2] Selon la pyramide des âges au 1er janvier 2020 – estimations de population de l'INSEE

Parallèlement, l'espérance de vie à la naissance augmente de manière stable : 85,13 ans pour les femmes et 79,13 ans pour les hommes en 2020 soit respectivement 2,35 et 3,88 ans de plus qu'il y a 20 ans[3]. Les Français sont donc plus nombreux et vivent plus longtemps (cf. figure 1 ci-dessus), une progression qui est susceptible de persister sur les prochaines décennies. La répartition de l'âge de la population française au 1er janvier 2020 donne une forme cylindrique à la pyramide des âges (cf. figure 2 ci-dessous, en gris), qui est caractéristique d'une société en cours de vieillissement[4]. La France n'est pas le seul pays à connaître ces tendances démographiques, que l'on observe à l'identique dans la plupart des pays occidentaux.

Influence de la démographie sur la prévalence des affections chroniques. Ces dynamiques sont à l'origine d'une augmentation de la prévalence des maladies chroniques – ou affections de longue durée (ALD) – au sein de la population. A partir de 45 ans, la part de patients atteints d'une ALD s'accroît sensiblement, pour toucher en moyenne plus du tiers des individus (cf. figure 2). Au 1er janvier 2020, cette portion parmi les plus de 45 ans représentait près de 10 millions de patients en France[5].

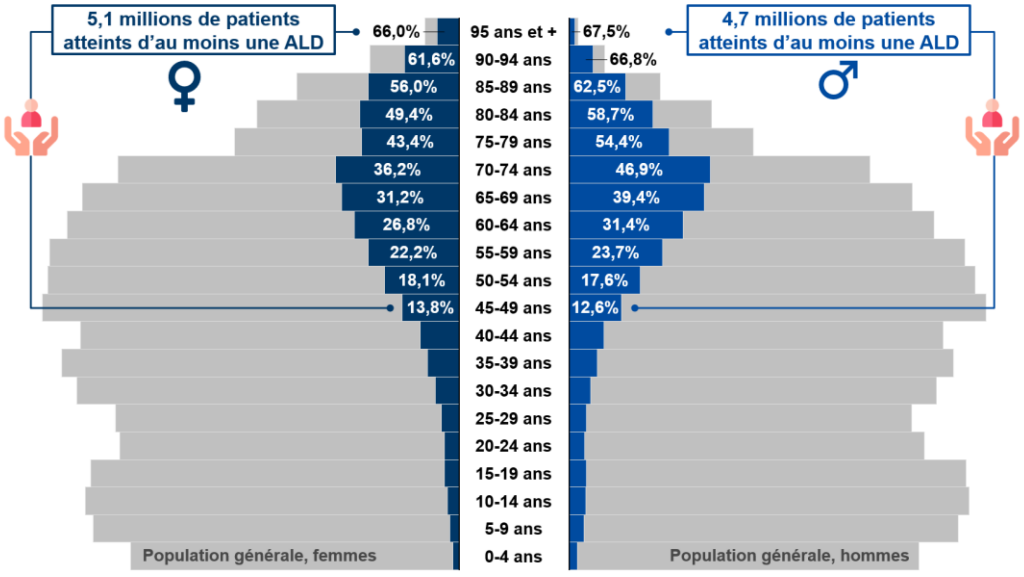


Figure 2 : Pyramide des âges de la prévalence des affections de longue durée chez les femmes et les hommes parmi la population française au 1er janvier 2020 - source CNAM et INSEE

[3] Selon les tables de mortalité de 1994 à 2021 – statistiques de l'état civil et estimations de population de l'INSEE
[4] Selon « L'état de santé de la population en France – Rapport 2017 » de la DREES
[5] Selon « Effectif, prévalence et caractéristiques des bénéficiaires d'une ALD en 2019 » - tables d'effectifs de patients pris en charge pour ALD 30-31-32 de l'Assurance Maladie

Cette dégradation est fortement portée par les cinq affections les plus courantes : diabète de type I et II, tumeurs malignes, affections psychiatriques de longue durée, maladie coronaire et insuffisance cardiaque. Les mouvements démographiques ne sont certes pas les seuls responsables de la forte hausse de leur prévalence et incidence, mais elles sont passées de 70% des ALD en 2010 à 80% en 2020.

De meilleures techniques diagnostiques mais surtout la persistance de facteurs de risques individuels (tabagisme, sédentarité, etc.), professionnels et environnementaux viennent influencer ces mesures.

Quelles qu'en soient les causes, ces accroissements pèsent de plus en plus lourd sur le système de soins. Entre 2010 et 2020, l'augmentation moyenne de l'effectif et de la prévalence pour ces cinq pathologies est de respectivement +56% (soient 3 millions de patients) et +42% (cf. figure 3 ci-dessous)[6]. Cette progression concerne également les moins de 65 ans.

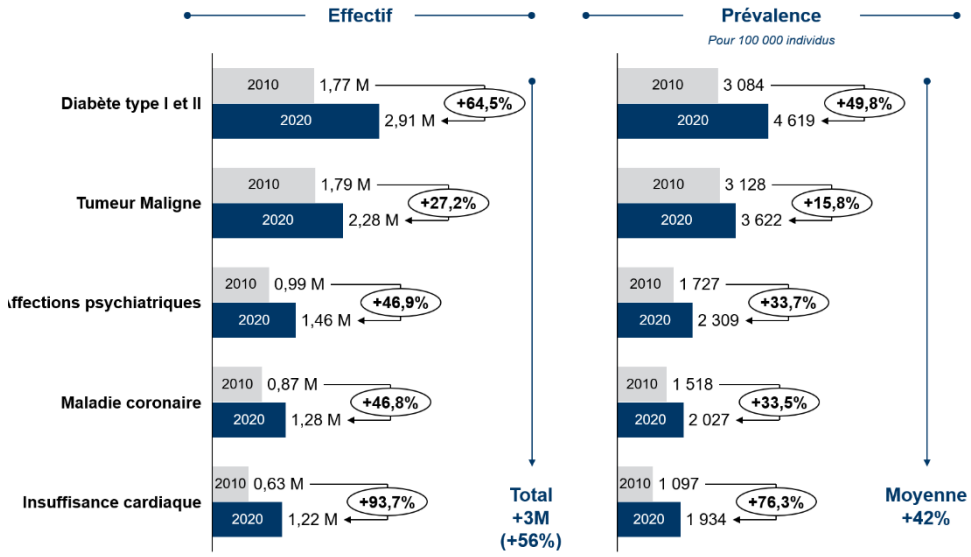


Figure 3 : Progression de l'incidence et la prévalence des 5 affections de longue durée les plus courantes en France entre 2010 et 2020 - Source CNAM

[6] Nota Bene. Il existe 3 familles d'affections prises en compte par l'Assurance Maladie :
1. Les ALD 30 : liste d'affections comportant un traitement prolongé et une thérapeutique particulièrement couteuse. Elle est établie par décret.
2. Les ALD 31 « hors liste » sont des formes graves de maladies qui ne figurent pas sur la liste des ALD30 mais qui ouvrent droit à une exonération du ticket modérateur.
3. Les ALD 32 « polypathologies » sont des affections qui combinées, entraînent un état pathologique invalidant.
Il existe également des ALD non exonérantes, c'est-à-dire qui n'ouvrent pas droit à l'exonération du ticket modérateur. Celles-ci ne sont pas prises en compte dans ces chiffres.
Définitions extraites du site de l'Assurance Maladie.

Ces pathologies ont un poids économique et opérationnel pour le système de santé. Les patients atteints d'une ALD vont recourir plus régulièrement à une variété d'acteurs du soin plus élevée que le reste de la population et leur prise en charge requiert une coordination renforcée notamment dans le lien ville-hôpital.

En effet, ces ALD pèsent de manière significative sur le système de santé et représentent en moyenne 60% des remboursements de l'assurance maladie, soit, à titre d'exemple près de 90 milliards d'euros en 2016[7]. Une enveloppe importante qui devrait continuer à croître : l'augmentation du nombre d'aidants informels sera bien moindre que celle du nombre de personnes dépendantes dans les années à venir[8], une charge supplémentaire qui sera reportée sur l'écosystème de la santé.

1.2.2 L'ACCES AUX SOINS SOUFFRE DE FRACTURES MULTIPLES

Cette section présente les indicateurs de l'écart d'accessibilité aux soins, et comment leur mesure permet de mettre en valeur une disparité sur le territoire qui rend les communes inégales face aux flux de patients.

Notions d'accessibilité aux soins. Si l'on s'accorde sur l'existence d'écarts dans l'accès aux soins, ceux-ci sont mal mesurés. L'accès aux soins n'étant pas uniforme sur le territoire, les structures de santé doivent absorber une charge variable selon leur localisation, mais aussi le type de public qu'ils adressent. Si les déserts médicaux sont régulièrement évoqués au cours des débats sur l'offre de santé en France, ce ne sont pas les seuls freins pour les patients. On recense trois notions autour de l'accessibilité aux soins[9] :

1. L'accessibilité géographique
   L'offre de soins doit être disponible en termes de temps d'accès mais également en nombre de consultations vacantes.
2. L'accessibilité financière
   Il s'agit de mesurer les freins économiques qui entravent la capacité d'un patient à consulter un praticien, en prenant en compte le coût

[7] Selon « Rapport Sécurité Sociale 2016 » de la Cour des Comptes
[8] Projections démographiques de l'INSEE, citées par (Ankri et al., 2013)
[9] Citées par (Chambaud, 2018)

du service, mais aussi celui de son accès (transport, garde d'enfants, etc.).

3. L'accessibilité socio-culturelle
   De nombreux facteurs sociaux peuvent fortement teinter la reconnaissance d'une pathologie et la décision de prendre contact avec des services de santé : par exemple l'âge, le genre, ou le niveau d'éducation.

Evaluation de l'écart d'accessibilité aux soins. Depuis 2012, l'effectif global de médecins en France stagne, mais cette évolution varie selon les spécialités. En 2021, on compte 94 500 médecins généralistes et 120 000 spécialistes, soit respectivement une chute de 5,6% et une hausse de 6,4% par rapport à 2012[10]. Cette baisse est notamment expliquée par la création en 2017 des spécialités de « gériatrie » et « médecine d'urgence », qui ne sont plus comptabilisées dans la catégorie « généralistes » et se sont transvasées vers les « spécialistes ». La répartition territoriale est pour autant extrêmement variable avec des écarts de densités de médecins notables entre régions pourtant voisines : l'Île-de-France compte par exemple 354 médecins pour 100 000 habitants contre 243 pour la Picardie[11].

De nombreux indicateurs sont utilisés pour évaluer la fracture territoriale dans l'accès aux soins. Parmi les plus courants on peut citer : le temps d'accès au médecin le plus proche[12], et le nombre de consultations ou d'Equivalent Temps Plein (ETP) disponibles par habitants. Le suivi de ces mesures montre effectivement une disparité géographique dans la répartition des services de santé selon les professionnels recherchés (cf. figures 4 et 5 ci-dessous)[13].

[10] Selon (Anguis et al., 2021)
[11] Source : Médecins actifs de moins de 70 ans, ayant au moins une activité en France métropolitaine ou dans les DROM en 2021. RPPS, INSEE, traitement DREES.
[12] A titre d'exemple, dans la littérature | La densité, localisation spatiale de l'offre de soins et le temps nécessaire pour y accéder sont utilisées comme variables de modélisation de l'accessibilité des soins, avec un cas d'application à l'espace transfrontalier des Alpes du Sud. Les auteurs proposent de quantifier l'accessibilité aux soins à l'aide d'un modèle issu de la théorie des graphes. L'étude fait notamment le lien entre les espaces de densité démographique faible et un indice de vieillesse de la population élevé. Ces territoires, qui souffrent d'une offre sanitaire absente, sont particulièrement fragiles. A retrouver dans (Decoupigny et al., 2007)
[13] D'après « La France et ses territoires, édition 2021 » de l'INSEE. Fiche 4.4 « Accessibilité aux professionnels de santé ».

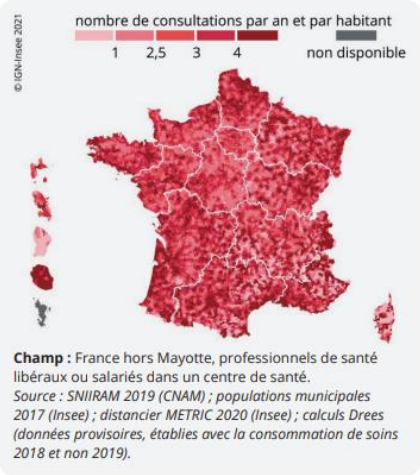


Figure 4 : Cartographie du nombre de consultations de médecins généralistes par an et par habitant en 2018 – Source DREES[14]

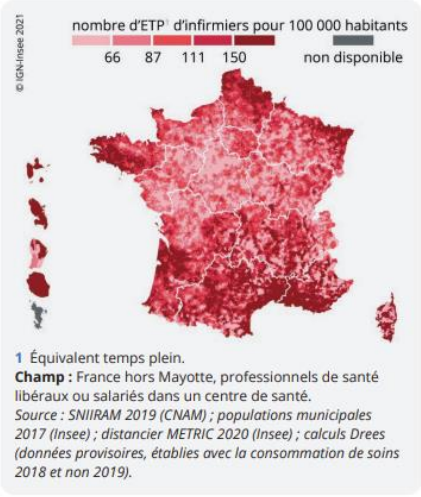


Figure 5 : Cartographie du nombre d'ETP d'infirmiers pour 100 000 habitants en 2018 – source DREES

Si la densité des services de santé permet de tracer un parallèle entre la répartition territoriale des praticiens et des individus, elle n'intègre pas les spécificités des populations et leur impact sur le recours aux soins. L'âge est par exemple un facteur déterminant dans la différenciation des besoins en consultations (cf. figure 6 ci-dessous). La pondération des consommations de santé en fonction de l'âge montre que les patients âgés de 45 ans et plus ont un recours aux soins en moyenne 30% plus important que le reste de la population. Ce ratio est largement tiré par les consultations en médecine spécialisée (ex : cardiologie), puis en médecine générale. Les soins en odontologie ont tendance à augmenter de manière stable au cours de la vie, puis décroissent à partir de 65 ans. Les consultations prééminentes parmi les 20 à 40 ans sont celles des sage-femmes et gynécologues, pointant également une disparité des recours aux soins selon le genre[15].

[14] La Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES) est une direction de l'administration des ministères sanitaires et sociaux qui produit des statistiques, synthèses et études pour doter ses ministères de tutelles d'une capacité d'observation et d'évaluation de leurs actions et environnement. Définition issue du site de la DREES. Drees.solidarites-sante.gouv.fr

[15] Selon (Anguis et al., 2021).

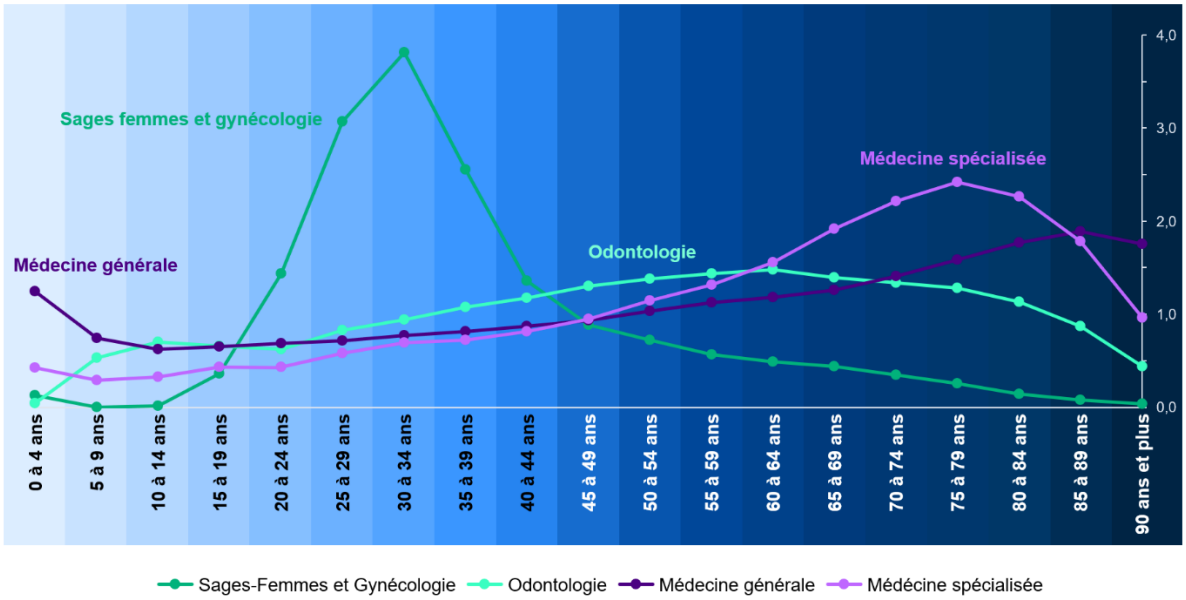


Figure 6 : Pondération des consommations de soins observées par tranche d'âge en France en 2018 – source « Echantillon général des bénéficiaires » de l'Assurance Maladie. En ordonnée : pondération relative du recours (ex : Les 30 – 34 ans ont 4 fois plus recours aux sage-femmes et gynécologues qu'à la médecine générale).

Afin de combiner la densité et la disponibilité des professionnels aux besoins spécifiques de la population locale, la DREES et l'IRDES[16] introduisent dans leurs publications à partir de 2012 l'Accessibilité Potentielle Localisée (APL)[17]. Il s'agit d'une mesure du nombre de consultations disponibles par an pour un patient sur un territoire donné.

L'APL comme mesure améliorée de l'accessibilité. La DREES publie en 2017 une étude dont l'objectif est d'évaluer la part de la population française qui cumule une ou plusieurs difficultés d'accès à des services de santé[18]. Il en résulte que 5,3 millions d'individus (8,1% de la population) en France ont une Accessibilité Potentielle Localisée strictement inférieure à 2,5 consultations chez le généraliste par an. Les territoires les plus touchés sont d'abord les DROM où en moyenne ¼ de la population vit sous ce seuil, à l'exception notable de la Réunion. En France métropolitaine, les régions les plus touchées sont le Centre-Val-de-Loire, la Bourgogne Franche-Comté, l'Auvergne Rhône-Alpes et la Corse (cf figure 8 ci-dessous).

[16] L'Institut de Recherche et Documentation en Economie de la Santé (IRDES) est un organisme de recherche chargé de produire des données statistiques sur le système de santé français. Définition issue du site de l'IRDES. Irdes.fr
[17] A retrouver chez (Barlet et al., 2012).
[18] Selon (Vergier & Chaput, 2017)

Encadré 1

Calcul de l'indicateur d'Accessibilité Potentielle Localisée

L'APL se construit en deux étapes :

1. Calcul de la zone de patientèle pour toutes les communes
   On détermine un ratio R qui représente le nombre de patients pouvant être desservis par les médecins d'une commune.

   Le ratio R_j de la commune *j* est défini par

$$R_j = \frac{m_j}{\sum_{D_{ij}<D_0} N_i . w_{ij}}$$

   Avec

- m_j : l'activité des médecins de la zone, calculée à partir du nombre d'actes réalisés sur la période à l'étude (base Assurance Maladie)
- D_{ij} : distance de la commune *i* à *j*
- D_0 : seuil maximal en dessous duquel on considère qu'une commune est accessible
- N_i : nombre d'habitants de la commune *i* pondérés par leur taux de recours moyen en fonction de leur âge (base INSEE et Assurance Maladie pour les besoins en soins, cf. figure 6)
- w_{ij} : pondération de la commune *i* en fonction de la distance de *i* à *j* (plus une commune *i* est éloignée de la commune *j*, moins elle est accessible et plus w_{ij} est faible)

2. Calcul de la zone de recours pour une commune donnée
   Pour chaque commune on précise sa zone de recours, c'est-à-dire le nombre de communes pouvant être accessibles pour un service de santé. On somme ensuite les ratios de toutes les communes présentes dans la zone de recours. Cette somme est l'APL.

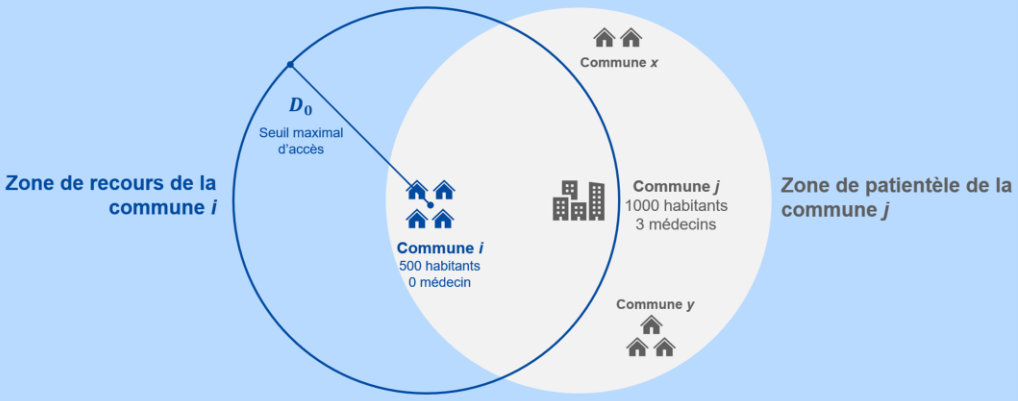


Figure 7 : Illustration des zones de recours et de patientèle pour deux communes i et j

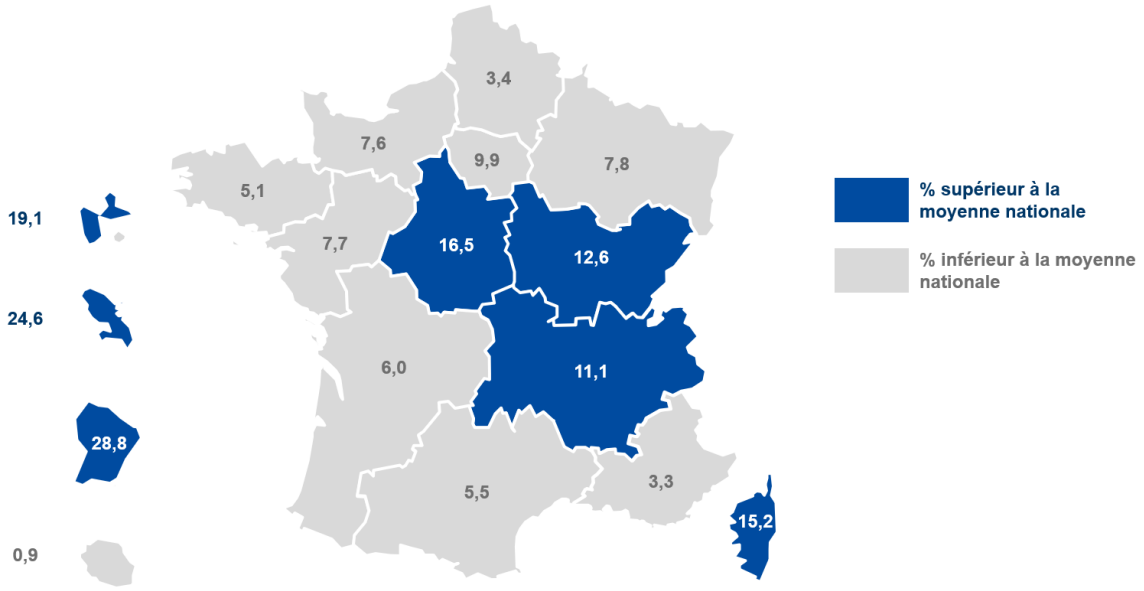


Figure 8 : Part de la population régionale (en %) vivant sous le seuil d'APL de 2,5 consultations / an / habitant en 2017 - source DREES

L'APL est une mesure plus fine de l'accessibilité que la densité, il est cependant important de souligner que cet indicateur dépend fortement des hypothèses de calcul qui sont considérées, notamment le seuil d'accessibilité maximal D_0[19] et l'estimation des besoins en soins des individus.

De nombreuses études utilisent les consommations réelles de soins comme proxy aux besoins, ce qui introduit un biais via deux phénomènes[20]. Premièrement, le niveau de consommation local dépend du niveau d'offre du territoire[21]. D'autre part, le non-recours aux soins – pour des raisons financières par exemple – peut induire une sous-estimation de la consommation, plus particulièrement pour les médecins spécialistes et alors même que ces patients peuvent avoir des besoins supérieurs.

[19] Ces valeurs seuils évoluent dans la littérature. L'étude de la DREES datant de 2017 fixe un temps de trajet maximal de 20 minutes pour fixer le seuil d'accessibilité d'un médecin généraliste. Elle propose également une définition des pondérations des distances entre communes w_{ij} par palier, en considérant une accessibilité parfaite, réduite ou très réduite si le temps de trajet est respectivement strictement inférieur à 10 minutes, 15 minutes et 20 minutes. A noter que ces valeurs ont été relâchées depuis la première étude présentant l'APL sur recommandation d'un groupe de travail issu de la DREES et de l'IRDES.

[20] Cités par (Lucas-Gabrielli et al., 2022)

[21] Une première étude économétrique montre en 2000 que dans les territoires où la densité de médecins augmente, on constate parallèlement une hausse de la consommation en soins. Cette conclusion est confirmée par une étude de la DREES et l'INSEE, publiée en 2019. (Delattre & Dormont, 1999) et (Choné et al., 2019).

Des travaux récents[22] ont montré qu'être pauvre en conditions de vie[23] multiplie le risque de renoncer à des soins par 3,2.

Pour que l'ensemble de la population bénéficie d'une offre de soins structurée et équitable, il s'agit de calibrer les services en fonction des besoins spécifiques des individus. Les indicateurs de mesure de l'accessibilité des soins peinent à intégrer une vision fine des besoins, il sera donc nécessaire de s'appuyer sur des outils complémentaires pour mieux cerner les spécificités des individus. Cet enjeu est d'autant plus prégnant que les projections démographiques du nombre de médecins généralistes en France sont peu dynamiques[24] (seulement +35% entre 2021 et 2050) et laissent entrevoir une pénurie face au vieillissement de la population, particulièrement en territoire rural.

1.2.3 UN ECOSYSTEME DE LA SANTE AU FINANCEMENT COMPLEXE

Cette section présente la structuration du système hospitalier en France, son mode de financement et comment celui-ci a pu favoriser une vision découpée à l'acte de soins, accentuée par une forte pression économique et opérationnelle.

Le secteur hospitalier en France. L'activité des établissements hospitaliers est décomposée en trois disciplines :

1. Médecine, Chirurgie, Obstétrique et Odontologie (MCO)
   La MCO désigne les activités aigües de courte durée réalisées dans les établissements de santé[25]. Ce sont les activités principales des hôpitaux.

2. Psychiatrie (PSY)
   Patients souffrant de troubles psychiatriques et suivis en unité d'hospitalisation ou dans des centres d'accueil.

[22] Selon (Lapinte et al., 2021).
[23] Définition de l'INSEE : « La pauvreté en conditions de vie mesure conventionnellement la proportion de ménages qui déclarent au moins huit restrictions matérielles parmi une liste de 27 difficultés, regroupées en quatre grandes dimensions : insuffisance de ressources, retards de paiement, restrictions de consommation et difficultés de logement. »
[24] Projections citées par (Anguis et al., 2021).
[25] Glossaire Financement des établissements de santé, en ligne sur le site internet du ministère de la santé, solidarités-santé.gouv.fr

3. Soins Médicaux et Réadaptation (SMR)
   Les SMR sont un ensemble de prises en charge spécialisées qui ont pour objectif le recouvrement maximal des conditions de vie d'un patient à la suite d'un séjour hospitalier.[26]

La MCO est le secteur d'activité prépondérant avec plus de 200 000 lits en hospitalisation complète en 2019, avec une tendance de baisse des places disponibles en hospitalisation complète comme partielle (- 14 529 lits au total en 10 ans, entre 2009 et 2019, voir figure 9 ci-dessous). Cette chute s'explique notamment par une amélioration des techniques médicales et médicamenteuses notamment anesthésiques, qui permettent à un nombre croissant de procédures d'être effectuées en dehors du cadre hospitalier traditionnel[27].

La SMR est la seule discipline à connaître une croissance d'activité au global (+12 797 lits sur la même période). En 2019, elle offre 120 000 lits dédiés aux moyens séjours pour 1 million de patients comptabilisant 37 millions de journées d'hospitalisation. Ces séjours représentent 28% de l'activité hospitalière en France et continuent à se développer pour l'accueil des patients autonomes ou légèrement dépendants[28].

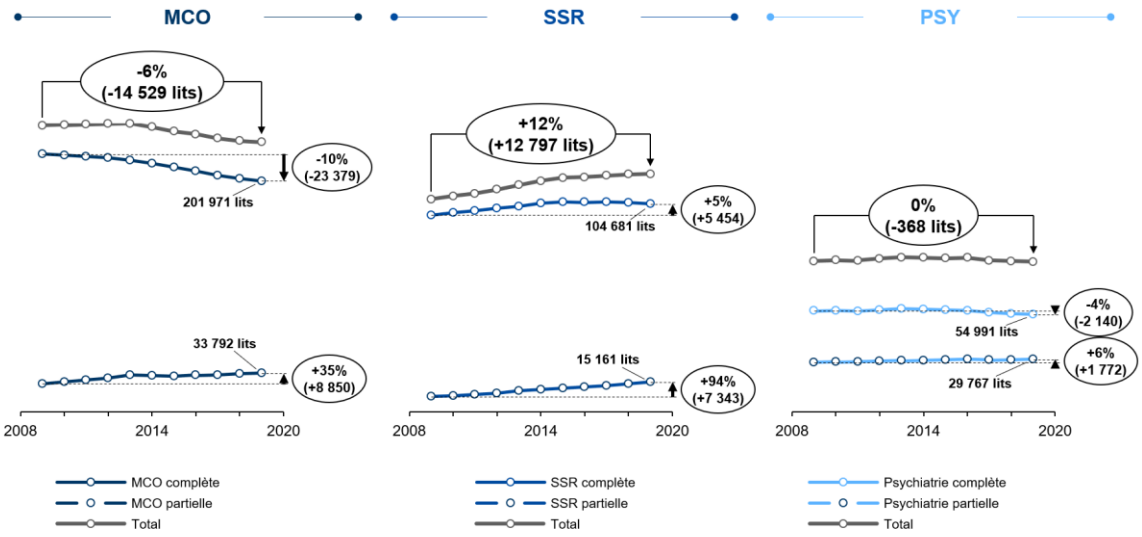


Figure 9 : Evolution du nombre de lits par discipline et modalité d'hospitalisation entre 2009 et 2019 – source SAE et DREES

[26] Définition issue de la page « Les soins médicaux et de réadaptation (SMR) » du ministère de la santé, solidarités-santé.gouv.fr
[27] D'après « Les établissements de santé ». Panoramas de la DREES, édition 2021
[28] Selon (Charavel et al., 2018).

L'offre de soins en psychiatrie connaît une certaine stabilité, la fermeture de lits en hospitalisation complète étant compensée par des ouvertures en hospitalisation partielle. La PSY présente des caractéristiques de prises en charge spécifiques, avec une faible proportion d'actes techniques dans les soins mais des séjours récurrents et une place importante de l'ambulatoire.

Chaque année, environ 13 millions de patients sont hospitalisés dans l'une de ces disciplines[29]. Ces dernières, réparties sur les secteurs privés, privés non lucratifs et publics, ont chacune leurs modalités de prises en charge (hospitalisation complète, ambulatoire, etc.). Parallèlement au virage ambulatoire, la France a fortement développé la prise en charge en Hospitalisation à Domicile (HAD). Ces établissements hospitaliers médicalisent le domicile du patient et coordonnent des soins complexes pour éviter ou raccourcir une hospitalisation conventionnelle et/ou en soins de suite[30]. Chacune de ces activités connaît des spécificités dans ses modes de financement, malgré un modèle qui se veut convergent depuis plusieurs années.

Historique du financement et philosophie de la T2A. Le financement des établissements de santé en France a connu plusieurs vagues de réformes successives, dont la dernière a été lancée en 2020 et est toujours en cours. Le mécanisme de recettes des établissements repose sur 2 blocs fondateurs[31] :

1. Financements liés à l'activité : composés de la tarification à l'activité (T2A) et de recettes supplémentaires liées au remboursement intégral de certains médicaments onéreux (ex : chimiothérapie, facteurs de coagulation, médicaments orphelins[32]) et dispositifs médicaux.

[29] Selon « Les dépenses de santé en 2019. Résultats des comptes de la santé ». Panoramas de la DREES, édition 2020.
[30] Le développement et le fonctionnement de l'HAD en France sera abordé plus en détail au cours du chapitre 3.
[31] Selon « La tarification à l'activité. Réforme de l'allocation de ressources des établissements de santé ». Présentation de la Mission Tarification à l'Activité sous la coordination de Christophe Andréoletti pour le Ministère de la santé (Mai 2007)
[32] Un médicament est dit « orphelin » lorsqu'il est destiné au traitement de maladies rares (dites « orphelines »). Le développement et la mise sur le marché de ces médicaments sont stimulés par l'Union Européenne via des mesures d'incitation (ex : crédits, assistance à l'élaboration de protocoles et exclusivité commerciale de 10 ans à la mise sur le marché). Selon la page « Les médicaments orphelins » du site internet du ministère de la santé, solidarités-santé.gouv.fr

2. Dotations complémentaires : certaines activités peuvent faire l'objet d'un budget additionnel, notamment les missions d'intérêt général (SAMU, actions de prises en charge en milieu carcéral, missions d'enseignement et de recherche) ou des activités d'expertise.

Historiquement, les établissements de santé publics bénéficiaient d'une enveloppe budgétaire annuelle reconduite selon les dépenses de l'année passée, et donc déconnectée de l'évolution de l'activité. Le secteur privé facturait des prestations à l'assurance maladie sur la base de tarifs négociés au cas par cas avec les Agences Régionales de Santé (ARS)[33].

Depuis 2003, la T2A constitue le mode de financement principal lié à l'activité de médecine, chirurgie et obstétrique des établissements publics et privés[34]. La philosophie de la T2A est d'équilibrer les recettes des établissements de santé en fonction de l'activité produite pour permettre l'engagement de moyens qui stimuleront à leur tour l'activité[35] (cf. Figure 10 ci-dessous).

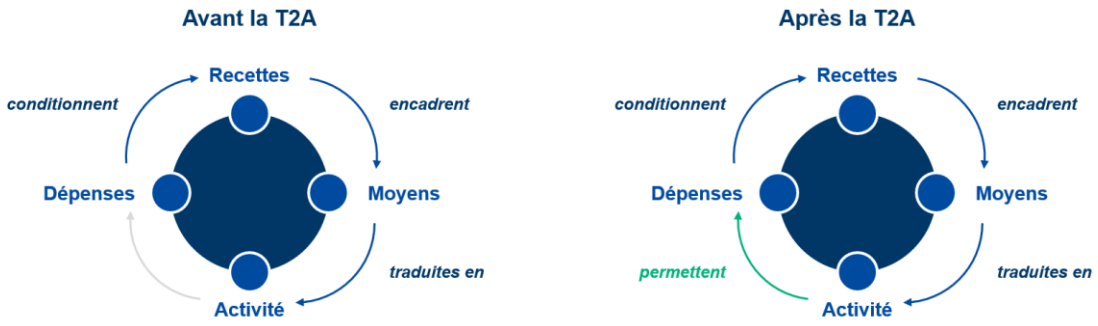


Figure 10 : Philosophie de la T2A – avant/après pour les établissements de santé publics – d'après Rapport de la Mission Tarification à l'Activité pour le Ministère de la Santé

L'objectif affiché est d'harmoniser les modes de financement pour permettre une convergence sectorielle, tout en respectant les spécificités de chaque activité.

[33] Décrits dans la section 'Les modes de financements antérieurs de la page « Financement des établissements de santé » du site internet du ministère de la santé, solidarités-santé.gouv.fr
[34] Inscrit dans la Loi n°2003-1199 du 18 décembre 2003 de financement de la sécurité sociale pour 2004.
[35] Selon « La tarification des établissements de santé. Rappel des enjeux, des modalités, des schémas cibles et transitoires ». Rapport de la Mission Tarification à l'activité sous la coordination de Christophe Andréoletti pour le Ministère de la Santé (2007).

Mode de fonctionnement de la T2A. Les recettes des établissements sont attribuées proportionnellement à l'activité enregistrée. Le prix de chaque activité est fixé par le ministère de la santé en fonction de deux paramètres : le Groupe Homogène de Malades (GHM) et le Groupe Homogène de Séjour (GHS) du patient[36]. Pour chaque patient entamant un séjour d'hospitalisation, ses informations administratives (âge, genre et durée du séjour) ainsi que médicales (diagnostic principal, diagnostics associés et actes) sont combinées dans le Programme de Médicalisation des Systèmes d'Information (PMSI) pour attribuer au séjour un couple GHM – GHS. Ces groupes conditionnent le tarif de prise en charge par les régimes d'assurance maladie[37] et constituent la base du paiement au séjour.

Des déclinaisons dans ces groupes tarifaires existent selon le secteur d'activité (MCO, PSY, HAD ou SMR), mais le mécanisme global reste identique.

Une remise en cause de la T2A. Malgré une liste GHM – GHS commune aux secteurs publics et privés, le calcul des tarifs diffère en pratique :

- Secteur public : tarifs évalués à partir d'une étude nationale de coûts réalisée sur un échantillon d'établissements.
- Secteur privé : tarifs déterminés par une valorisation des données historiques de facturation à l'assurance maladie.

Cependant, des études ont montré que l'application de la réforme entraînait des effets secondaires. De par son fonctionnement, la T2A encourage les établissements de santé à se concentrer sur les activités les mieux rémunérées. Par ailleurs, tous les hôpitaux ne se ressemblent pas et peuvent subir des variations dans leurs structures de coûts, par exemple dues à une géolocalisation qui affecte les salaires ou encore une économie d'échelle réalisée sur des consommables médicaux et liée à la taille du groupement hospitalier. Chaque individu pris en charge peut par la suite apporter sa propre variabilité en coûts de par son historique de soins et sa condition de santé.

[36] Exemple – arrêté tarifaire du 27 février 2007 : GHM 14Z02A « Accouchement par voie basse sans complication » et GHS 5451 donneront un tarif de 2 236.29 €. Des suppléments sont appliqués pour certains séjours, à titre d'exemple 838.16 € par journée de présence en unité de réanimation.

[37] Le Programme de Médicalisation des Systèmes d'Information (PMSI) permet de classer le séjour de chaque patient au sein d'un Groupe Homogène de Malades (GHM), auquel est associé un ou plusieurs Groupes Homogènes de Séjour (GHS). Selon la page « Financement des établissements de santé » du site internet du ministère de la santé, solidarités-santé.gouv.fr

Ces différences par rapport au tarif moyen national peuvent créer des incitations à dégrader la qualité des soins, à sélectionner des patients, ou à optimiser le codage pour déclarer le GHM le plus lourd et rémunérateur – dont l'appréciation exacte est difficilement vérifiable[38]. Avec le passage à une nomenclature plus fine, la probabilité d'être codé dans le degré de sévérité le plus faible a baissé de 2,1% alors que la probabilité d'un surcodage à une sévérité plus élevée a crû de 8%[39].

Une forte pression sur les coûts et l'efficience opérationnelle. Conjointement aux réformes successives d'un mode de financement complexe, le système de santé subit une forte pression sur les coûts qui le force à mettre l'accent sur l'efficience opérationnelle des parcours de soins. L'accélération du déficit de l'assurance maladie, aggravé par la crise du Covid en est notamment l'une des sources. En 2022, celui-ci devrait atteindre 21,9 Md€. Bien que le gouvernement ait annoncé une croissance de l'Objectif National des Dépenses d'Assurance Maladie (ONDAM)[40] de 2,5% par an à partir de 2024[41], le Haut Conseil du Financement de la Protection Sociale (HCFiPS) projette que la hausse des prix en 2023 aura des effets importants sur les comptes sociaux qu'il est difficile de quantifier à ce stade.

Depuis plusieurs années, les hôpitaux tendent ainsi à devenir des plateaux techniques qui réalisent les actes complexes puis transfèrent les patients chroniques et polypathologiques vers des systèmes de soins « relais » de type médecine de ville, prise en charge à domicile ou en soins de suite.

1.2.4 Une disponibilite croissante des donnees de sante et une meilleure applicabilite de l'IA

Cette section présente un historique de l'IA dans la santé, les opportunités actuelles apportées par cette technologie et comment elles permettent de résoudre des problématiques structurantes de la santé malgré des freins.

[38] Selon (Frédéric Pierru, 2020).
[39] Cité par (Mougeot & Naegelen, 2014).
[40] L'ONDAM est un objectif annuel de dépenses totales de l'assurance maladie en France. Il est fixé chaque année par le gouvernement français et a pour but de contrôler le budget global de la santé. Définition issue de la page sante.gouv.fr
[41] Dans une note publiée le 14 décembre 2022 par le Haut Conseil du Financement de la Protection Sociale. « Etat des lieux du financement de la protection sociale. Un redressement des comptes sociaux incontestable, inégal et fragile ».

Un bref historique de l'intelligence artificielle dans la santé. La genèse de l'intelligence artificielle (IA) est souvent attribuée à Alan Turing, qui décrit en 1950 le concept de *« machine intelligente »*[42] dans son texte fondateur du célèbre test de Turing[43]. Le terme d' « *intelligence artificielle »* a quant à lui été employé pour la première fois six ans plus tard, par le chercheur John McCarthy, qui l'a défini comme « *la science et l'ingénierie de la conception de machines intelligentes* »[44]. Les chercheurs et industriels se sont depuis longtemps intéressés aux applications de l'intelligence artificielle dans la santé. Plusieurs systèmes ont été mis en avant[45] comme des exemples précurseurs de l'application de l'intelligence artificielle à la santé, parmi lesquels :

- MYCIN (1975)[46] est un système logique qui, à partir d'une série d'informations patient, liste les bactéries pathogènes dont il est potentiellement atteint et les traitements antibiotiques à suivre ainsi que leur dosage. Cette application repose sur une base de connaissances clinique de 600 règles 'SI (…) => ALORS (…)'.

- CASNET (1978)[47] est un programme simulant une consultation chez l'ophtalmologue pour les patients atteints d'un glaucome. Le système apporte un soutien au praticien pour diagnostiquer, traiter et suivre le patient ainsi qu'un pool d'avis d'experts pour certains aspects cliniques faisant débat.

- DXplain (1986)[48] est un système d'aide à la décision qui produit des diagnostics différentiels à partir de la description du syndrome[49] d'un patient. Ce logiciel, développé par le Massachussetts General Hospital pour couvrir 500 pathologies à son lancement, est toujours actif et en comporte aujourd'hui 2 600.

[42] Selon (Turing, 1950).
[43] « Le test de Turing est un test visant à déterminer si un ordinateur peut faire preuve d'un comportement intelligent similaire à celui d'un être humain, et qui consiste à faire évaluer par une tierce personne une conversation entre une machine et un humain. On considère que la machine réussit le test si l'évaluateur ne peut pas la distinguer de son interlocuteur humain ». Définition issue du Cambridge Dictionary (dictionary.cambridge.org)
[44] Décrit par John McCarthy sur la page « What is Artificial Intelligence ? », stanford.edu
[45] A retrouver dans (Ramesh et al., 2004) et (Kaul et al., 2020).
[46] Décrit par (Shortliffe et al., 1975).
[47] Décrit par (Weiss et al., 1978).
[48] Selon (Barnett et al., 1987).
[49] En médecine, le syndrome est l'ensemble de plusieurs symptômes ou signes en rapport avec un état pathologique donné et permettant, par leur groupement, d'orienter le diagnostic. Définition du Larousse.

Une récente étude conduite au Saint Mary's Hospital (Minnesota, USA) a estimé que l'utilisation systématique de DXplain lors des diagnostics les plus complexes pourrait permettre d'économiser 1 281 dollars par patient soit 2 millions de dollars par an en coûts de service[50].

Opportunités en santé de l'intelligence artificielle contemporaine. Malgré des initiatives fondatrices et apportant un réel appui à la fonction médicale, il a fallu attendre les années 2010 pour constater des avancées significatives dans l'adoption de l'intelligence artificielle dans la pratique clinique. Plusieurs dynamiques concomitantes ont permis cette envolée : une explosion du volume de données disponibles sur un large périmètre de la prise en charge patient, des progrès sur les techniques d'analyses des données mais surtout sur les capacités computationnelles et de stockage.

Une étude de 2018 estime par ailleurs que 30% des données stockées dans le monde sont des données de santé et que le patient moyen génère 80 mégabits de données en imagerie et dossier électronique médical par an[51]. Ce volume d'information continue ne cesse de croître et ouvre la porte à de nombreuses innovations. Les données générées sont de multiples origines : dossiers médicaux, imagerie, résultats de laboratoire, objets connectés, smartphones, données génomiques, etc.

La quantité de données n'est pas le seul moteur, avec des progrès notables dans l'accessibilité et le prix des technologies de pointe. A titre d'exemple, le coût du séquençage d'un génome humain a baissé d'un facteur un million en quinze ans[52]. Au début des années 2000, des premières études[53] se sont penchées sur l'apport des GPU[54] par rapport aux CPU pour accélérer l'implémentation et l'exécution des algorithmes d'intelligence artificielle. En 2008, une équipe de chercheurs de Stanford démontre que le temps nécessaire pour entraîner un réseau de neurones de 4 couches et 100 millions de paramètres est divisé par 70 en utilisant

[50] Voir (Elkin et al., 2010).
[51] Cité par (Suter-Crazzolara, 2018).
[52] Selon le livre (E. Topol, 2016).
[53] Une des premières études a été publiée par (Oh & Jung, 2004).
[54] « Les unités de traitement graphiques (*Graphics Processing Unit*, GPU) et unités centrales de traitement (*Central Processing Unit*, CPU) sont des moteurs de calculs informatiques. Le CPU est souvent appelé le « cerveau » de l'ordinateur car cette unité exécute les commandes et processus nécessaires au système d'exploitation. Elle détermine également la vitesse d'exécution des programmes. La GPU, est elle un processeur composé de nombreux cœurs plus petits et performants qui peuvent traiter plusieurs tâches en parallèle. » Définition issue du site d'Intel (intel.fr)

un GPU, soit réduit de plusieurs semaines à une seule journée[55]. Ces avancées technologiques ont favorisé l'accélération de la recherche et de l'innovation pour un certain nombre de tâches qui sont résolues au moyen de modèles intensifs en calcul et entraînés sur des jeux de données massifs, notamment la reconnaissance d'images et de sons, le traitement et la compréhension du langage, et la simulation[56]. En 2021, près de quatre mille articles ont été publiés par mois dans la catégorie « AI » d'arXiv[57], un site populaire de publications scientifiques. Un rythme de publication qui double tous les 23 mois environ.

Parallèlement aux travaux de recherche, c'est la mise à disposition de plateformes de déploiement open-source – dont la plus populaire est probablement TensorFlow, publiée en 2015 par Google[58] – qui a permis une expansion sans précédent du nombre d'applications d'intelligence artificielle en production. Ces plateformes de « bout en bout » viennent appuyer le processus de développement au complet : préparation des données, création des algorithmes, déploiement du modèle sur l'infrastructure choisie (smartphone, site internet, objet connecté, etc.) puis maintien en production.

Un outil pour résoudre les problématiques structurantes du monde de la santé. Un nombre croissant d'articles se penchent sur les bénéfices de l'implémentation de l'intelligence artificielle dans un contexte clinique, même si peu d'études sont en mesure de démontrer un impact positif quantifié[59]. Les apports varient suivant l'entité concernée, essentiellement :

1. Les cliniciens, pour assister la pratique et accroître la productivité : en particulier avec une visée de gain de temps, comme avec la synthèse de notes cliniques issues du suivi patient[60] ou encore l'analyse automatique d'images de radiologie.

2. Les patients, pour améliorer leur prise en charge : notamment réduire le temps nécessaire au diagnostic[61] et les erreurs

[55] Expérience décrite par (Raina et al., 2009).
[56] Selon (Dean, 2022).
[57] Données et graphique publiés par (Krenn et al., 2022).
[58](Abadi et al., 2015). Software available from tensorflow.org.
[59] Notamment économique, voir (Wolff et al., 2020).
[60] Comme décrit par (Kreuzthaler et al., 2017).
[61] Selon (Dilsizian & Siegel, 2013).

thérapeutiques qui sont inhérentes à la pratique clinique[62] tout en proposant des traitements et un suivi adaptés et personnalisés, c'est-à-dire prenant en compte toutes les spécificités du patient[63].

3. Les structures de santé, pour optimiser les ressources opérationnelles et financières : par exemple les stocks de médicaments et consommables[64], l'ordonnancement des chirurgies aux blocs opératoires[65], ou encore le nombre de tests superflus prescrits aux patients[66].

Face aux difficultés structurantes que connaît l'écosystème de la santé, l'intelligence artificielle apparaît comme un excellent levier pour réduire la pression sur les ressources humaines, économiques et opérationnelles.

Des freins spécifiques à la santé subsistent. Les avancées de l'intelligence artificielle n'ont jamais été aussi prometteuses dans tous les domaines cliniques, d'autant que l'écart de performance entre les algorithmes et les experts humains se réduit[67]. Des facteurs techniques[68] et organisationnels contribuent cependant à freiner l'implémentation en vie réelle de nombreux cas d'usages.

Facteurs techniques.

1. L'effet boîte noire. Ce terme désigne une catégorie d'algorithmes qui produisent des résultats difficilement interprétables, généralement car les variables explicatives du modèle sont combinées par des fonctions mathématiques de manière complexe et non compréhensible par un être humain[69]. Le manque d'interprétabilité engendre une défiance des utilisateurs, particulièrement dans le secteur de la santé où les cliniciens prennent des décisions qui peuvent avoir un impact crucial sur la vie des patients[70].

2. Des données complexes. La complexité et la performance des algorithmes dépendent pour beaucoup du volume d'informations à

[62] Selon (Neill, 2013).
[63] Selon (Abul-Husn & Kenny, 2019).
[64] Démontré par (Abu Zwaida et al., 2021).
[65] Implémenté par (Bellini et al., 2019).
[66] Selon (Gönel, 2020).
[67] Selon (E. J. Topol, 2019).
[68] A retrouver chez (Ghassemi et al., 2020).
[69] Comme défini par (Rudin & Radin, 2019).
[70] Selon (Petch et al., 2022).

traiter et de leur format. On retrouve trois caractéristiques principales dans les données de santé qui ont trait au *Big Data*[71]*, et* qui ont longtemps freiné l'impact de l'IA :

- Le volume généré par le suivi patient (comme décrit plus haut).
- La rapidité de création de nouvelles données. A titre d'exemple, un système informatisé de dossiers médicaux (EHR pour *Electronic Health Record* en anglais) est mis à jour plusieurs dizaines de fois par jour suivant la taille de la structure de santé[72].
- L'hétérogénéité. Les sources de données sont multiples et silotées : séjour à l'hôpital, données de facturation de l'assurance maladie, consultation chez le médecin généraliste, etc. Si les sources diffèrent, les informations sont également retranscrites dans une grande variété de formats – dates, images, notes écrites[73].

3. Des relations causes-effets difficiles à capter. La plupart des questions qui animent la recherche en santé sont causales par nature (e.g. « *qu'arrivera-t-il si le patient X prend le traitement Y ?* »). Tous les algorithmes ne sont pas en mesure d'adresser cette classe de problèmes. Par ailleurs, la plupart des données en santé sont observationnelles, c'est-à-dire qu'elles décrivent des événements qui ont pu être influencés par des facteurs extérieurs et non capturés dans les variables[74]. Cet état de fait complexifie la lecture des prédictions issues d'une IA. Le paradoxe de Simpson[75] est un parfait exemple de difficulté d'interprétation causale (voir encadré 2 ci-dessous)[76].

[71] Le Big Data, littéralement « données volumineuses » en anglais, caractérise un ensemble d'informations volumineuses, rapides et/ou variées qui exigent des techniques de traitement des données efficaces et innovantes pour en améliorer l'analyse et soutenir la prise de décision. « Definition of Big Data—Gartner Information Technology Glossary ». Gartner.
[72] Selon (White, 2014).
[73] Comme décrit par (Abul-Husn & Kenny, 2019).
[74] Selon (Pearl, 2009).
[75] Décrit par (Simpson, 1951).
[76] L'étude citée dans l'encadré a été menée par (Appleton et al., 1996).

Encadré 2
Le paradoxe de Simpson

Le paradoxe de Simpson est un parfait exemple du danger d'ignorer une variable significative dans un modèle statistique. Décrit par Edward Simpson en 1951, il stipule qu'un phénomène peut être observé pour différents groupes (d'individus, d'observations) mais s'inverser lorsque ces mêmes groupes sont rassemblés. L'exemple suivant, issu d'une étude épidémiologique réelle, illustre ce paradoxe :

- Fumer c'est bon pour la santé

En 1977, un groupe de chercheurs anglais investigue les facteurs susceptibles de conduire à des problèmes de santé tels que les maladies cardiovasculaires ou thyroïdiennes, au moyen d'un questionnaire sur les habitudes de vie incluant la consommation quotidienne de tabac. La population sélectionnée pour l'étude a ensuite été suivie pendant deux décennies.

Sur 1314 femmes suivies, le taux de mortalité après 20 ans chez les fumeuses est de 24% vs. 31% chez les non-fumeuses, un résultat contrintuitif. Or si on segmente ce même taux de mortalité par classe d'âge, on retrouve bien le facteur aggravant du tabagisme sur l'espérance de vie.

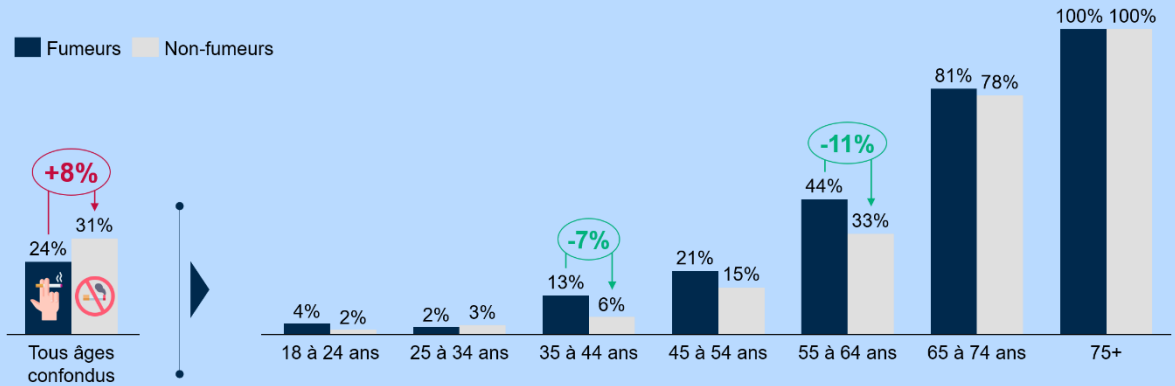


Figure 11: Taux de mortalité après 20 ans des fumeurs et non-fumeurs sur la population générale de l'étude (graphique de gauche) et groupé par classe d'âge (graphique de droite).

- D'où vient le paradoxe ?

Dans la population initiale de l'étude, les femmes âgées sont sur-représentées chez les non-fumeuses : 54% d'entre elles ont plus de 55 ans contre 28% pour les fumeuses. Dans chaque tranche d'âge, les non-fumeuses ont un taux de mortalité inférieur, mais cet avantage est compensé par l'effet « âge », qui fait que les personnes de plus de 55 ans ont une probabilité de décès à 20 ans plus élevée.
Le paradoxe de Simpson est observé lorsque l'échantillon étudié n'est pas distribué de manière homogène et qu'il existe une variable explicative non décrite qui influe sur la variable dépendante (le résultat). On parle alors de biais de sélection et de facteur de confusion.

4. Des données imparfaites. Les données en vie réelle sont susceptibles de reproduire les mêmes biais que les êtres humains, notamment liés au genre et aux origines. En se nourrissant de cette source d'information, les algorithmes reproduisent ces biais, ce qui représente un danger conséquent pour les groupes sous-représentés ou discriminés dans la population d'étude initiale[77]. Quelques-unes des innovations les plus impactantes dans la santé restent aujourd'hui centrées en grande partie sur les populations occidentales de descendance européenne, notamment en génomique[78]. En intelligence artificielle, cet obstacle est reconnu depuis quelques années[79] et a contribué à l'émergence d'un nouveau champ de recherche appelé IA équitable (*Machine Learning Fairness* en anglais)[80].

Ces freins techniques ont constitué pendant longtemps des barrières significatives à l'adoption de l'intelligence artificielle par les acteurs du soin. Bien que ces difficultés soient réelles et d'actualité, plusieurs études estiment qu'elles ne sont responsables qu'entre 5 à 20% des échecs d'implémentation de ces technologies[81] - et qu'il est donc crucial de se pencher également sur les blocages organisationnels à l'œuvre.

Facteurs organisationnels.

1. Une responsabilité floue. Les algorithmes, au même titre que les experts médicaux, sont susceptibles de commettre des erreurs. La question de la responsabilité quant à l'interprétation des résultats est épineuse, alors même que la législation et les recommandations éthiques tardent à verrouiller ces aspects[82].

2. Des investissements souvent sous-estimés. L'implémentation en vie réelle de cas d'usages requiert des investissements financiers et opérationnels importants au-delà du design de l'algorithme même. A titre d'exemple, les données administratives et issues des dossiers médicaux ont besoin d'être préparées et formatées,

[77] Plusieurs types de biais sont présentés par (Rajkomar et al., 2018).
[78] Selon Jessica Wapner. « The Search for a Cancer Cure Has Ignored African DNA ». Newsweek. (2018, juillet 18).
[79] Exemple d'évaluation de l'impact en pratique du biais socio-économique, de genre et d'origine ethnique pour un modèle prédictif de la réadmission à 30 jours dans une unité de soins intensive par (I. Y. Chen et al., 2019).
[80] Récent panorama décrit par (Fosch-Villaronga et al., 2022).
[81] Cité par (Lebcir et al., 2021).
[82] Selon (Padovan et al., 2023).

l'infrastructure informatique doit être adaptée pour permettre la mise en production de l'application, qui doit ensuite être maintenue[83]. L'implémentation de ce type de technologie, même fournie par un prestataire extérieur, nécessite un investissement de ressources humaines fort pour accompagner la transformation des processus impactés. Dans un contexte de ressources contraintes, les organisations de la santé ont tendance à sous-évaluer ces coûts qui sortent de son périmètre d'expertise[84].

3. Des rôles et compétences des praticiens fortement impactés. Dépendamment du niveau d'automatisation que la technologie propose, la prise en charge patient peut être modifiée à plusieurs échelles. Ces changements doivent être pris en compte dès la phase de conception pour s'adapter au mieux aux réalités du terrain. Les cliniciens opèrent habituellement avec un haut niveau d'autonomie, qui doit être conservé[85]. A noter que la formation des professionnels de santé est un facteur positif cité dans de nombreuses études[86] qui facilite l'implémentation de tout type d'applications numériques.

4. Des données sensibles et vulnérables. La question de la cybersécurité est celle qui inquiète en premier lieu les acteurs de la santé à l'implémentation de ces technologies. Dans ce secteur, la vulnérabilité des systèmes est concentrée au niveau de la fuite de données sensibles via des connexions et des appareils non sécurisés, des lacunes en matière d'authentification des utilisateurs mais également des autorisations d'accès accordées qui s'avèrent excessives[87]. Un rapport publié par le FBI révèle que l'organisation a reçu 148 signalements de cyberattaques visant des structures de santé, ce qui en fait le secteur le plus ciblé, devant la finance[88]. Le volume de patients concernés est important, à titre d'exemple une étude stipule qu'en 2015, plus de 110 millions d'entre eux aux Etats-Unis seuls ont vu leurs données compromises[89].

[83] Selon (Alami, Lehoux, Auclair, et al., 2020).
[84] Selon (Alami, Lehoux, Denis, et al., 2020).
[85] Comme décrit par (Levenson et al., 2008).
[86] Par exemple par (Cresswell & Sheikh, 2009).
[87] Cité par (Paul et al., 2023).
[88] Publié par Federal Bureau of Investigation. « Internet Crime Report 2021 ». https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf
[89] Cité par (Martin et al., 2017).

Malgré des obstacles techniques et organisationnels réels, l'intelligence artificielle présente de réelles perspectives de progrès dans la santé. Depuis 2016, les applications dans ce secteur représentent maintenant la plus grande part d'investissements de la recherche en intelligence artificielle[90]. En 2020, la taille du marché européen de l'IA dans la santé a dépassé les 700 millions de dollars (vs. 1,15 milliard de dollars pour l'Amérique du Nord) et devrait connaitre un taux de croissance de 43,9% entre 2021 et 2027[91]. L'impact que ces technologies pourront délivrer aux structures de soins, aux praticiens et aux patients dépendra toutefois de la prise en compte de ces freins par les autorités et les décisionnaires en santé. L'articulation de la problématique de recherche à partir de ces constats est décrite dans la section suivante.

[90] Cité par (Amisha et al., 2019).
[91] Source : Graphical Research Report ; cité par (Dicuonzo et al., 2023).

1.3 Question de recherche

SYNOPSIS Où nous décrivons la problématique principale qui a animé le travail de recherche, ainsi que les verrous scientifiques à lever pour y répondre.

1.3.1 DESCRIPTION DU PROBLEME ET DES VERROUS SCIENTIFIQUES

Les acteurs de la santé connaissent des difficultés structurelles dans l'organisation des soins et la gestion de leurs patients, notamment chroniques. Ces difficultés sont multiples – comme décrit précédemment, accroissement de la prévalence des maladies chroniques et vieillissement de la population, fracture territoriale dans l'accès aux soins, pression sur les coûts et sur l'efficience – et peuvent impacter fortement les perspectives de santé des populations. Les organisations de la santé, et notamment hospitalières, tentent de surmonter ces difficultés en optimisant les parcours patients et de soins dans leur ensemble : l'enjeu étant de garantir le niveau de qualité de la prise en charge tout en améliorant la gestion des ressources opérationnelles et financières.

Au sein de ces parcours, l'un des leviers d'efficience est de pouvoir comprendre à quels aspects du profil d'un patient sont corrélés les événements impactant le recours et la consommation de soins, afin de pouvoir les anticiper. L'objectif final est de soulager les pressions opérationnelles des ressources hospitalières tout en améliorant le confort des prises en charge des patients.

Les récentes avancées technologiques et la disponibilité croissante des données de santé permettent d'étudier une vaste diversité de parcours et d'analyser un large panel de variables, en ajoutant une dimension quantitative aux études qualitatives habituelles. Les données de facturation sont par ailleurs un excellent moyen de recomposer la trajectoire individuelle d'un patient. Dans ce travail, nous souhaitons représenter et analyser les parcours dans plusieurs contextes cliniques, au travers de la problématique suivante :

« COMMENT CARACTERISER ET ANTICIPER LES PARCOURS PATIENT ET DE SOINS SOUS L'ANGLE DES COUTS ? »

Cette question de recherche devra traiter plusieurs verrous scientifiques pour ouvrir des pistes de solutions qui répondent aux besoins du terrain des acteurs de la santé, à savoir :

1. Comment intégrer la multitude de formats de données qui composent le parcours patient ?
2. Comment représenter de manière synthétique la diversité des parcours de soins sans effacer des informations cliniques pertinentes ?
3. Quels outils implémenter pour identifier et corriger les biais présents dans les données ?
4. Comment produire des résultats interprétables à partir d'algorithmes complexes ?

Ces verrous scientifiques sont aujourd'hui partiellement couverts par la littérature scientifique, un point qui sera détaillé dans le chapitre suivant.

1.3.2 Structure du manuscrit

Le chapitre 1 a eu pour vocation de décrire les problématiques contemporaines d'organisation de la santé en France et les vecteurs d'innovation qui sont développés pour y répondre. L'objectif a été de replacer la question de recherche en perspectives de ces grandes problématiques.

Le chapitre 2 présente les notions de parcours patient et de soins et leurs problématiques respectives, décrit l'approche proposée et la littérature scientifique associée, et introduit les cas d'application qui seront développés au cours de ces travaux.

Le chapitre 3 décrit notre premier cas d'application, la prédiction de parcours patients atteints de troubles neurocognitifs grâce aux croisements de données cliniques issues de la base Memora et des données médico-économiques de la CPAM.

Le chapitre 4 décrit un cas d'application dans un contexte d'hospitalisation à domicile dans la région Auvergne Rhône-Alpes.

Le chapitre 5 présente la conclusion en miroir des objectifs et verrous identifiés en introduction.

2

Description de l'approche proposée

Ce deuxième chapitre présente les notions de parcours patient et de parcours de soins et leurs problématiques respectives, décrit l'approche proposée et la littérature scientifique de support, et introduit les cas d'application qui seront développés au cours de ces travaux.

Contenu

2.1 Parcours patient et parcours de soins

SYNOPSIS Où nous introduisons les définitions et concepts clefs autour des parcours patient et de soins, ainsi que les principaux enjeux associés et comment ceux-ci se déclinent en problématiques spécifiques selon les types de parcours.

2.1.1 DEFINITIONS ET CONCEPTS CLEFS

Cette section introduit les définitions et concepts clefs autour des parcours patients et des parcours de soins.

Le parcours patient et le parcours de soins sont deux termes fréquemment utilisés pour décrire l'ensemble des étapes que les patients traversent lors de leur prise en charge médicale ainsi que les interactions avec le système de santé que ces étapes vont générer. On parle également de trajectoire. Si ces termes sont parfois employés de manière interchangeable dans la littérature, ils couvrent pourtant des aspects différents de la prise en charge du patient. Nous proposons de théoriser ces termes à partir des cas d'application présentés dans la littérature. L'objectif est de présenter les spécificités de chacun de ces concepts et notamment de synthétiser les problématiques qui sont respectivement traitées.

Selon l'Assurance Maladie[92], le parcours de soins est l'ensemble des étapes qui permettent à un patient de bénéficier d'une prise en charge pour une pathologie donnée. Cette prise en charge comprend le diagnostic, le traitement, le suivi et la coordination des soins entre les différents professionnels de santé qui sont impliqués[93]. L'objectif du parcours de soins est de garantir une prise en charge optimale et cohérente pour une maladie ou un trouble médical spécifique. Il repose souvent sur de bonnes pratiques professionnelles et la définition de

[92] Site de l'Assurance Maladie. https://www.ameli.fr/assure/remboursements/etre-bien-rembourse/medecin-traitant-parcours-soins-coordonnes
[93] Définition issue du site du ministère de la santé. https://sante.gouv.fr/systeme-de-sante/parcours-de-sante-vos-droits/liberte-de-choix-et-acces-aux-soins/article/qu-est-ce-que-le-parcours-de-soins

protocole de soins, et peut être interne à une structure ou encore impliquer plusieurs effecteurs.

Le parcours patient[94] quant à lui, est défini par l'ensemble des étapes et des interactions qu'un patient donné va développer avec le système de santé au cours de sa prise en charge. Le périmètre du parcours patient s'étend de la prévention, jusqu'au diagnostic, puis la fin du traitement et de la réadaptation, ou la gestion des symptômes s'il s'agit d'une maladie incurable. Le parcours patient est souvent associé au concept de parcours de soins, mais il peut inclure également des aspects non médicaux, tels que le soutien psychologique, l'aide à la gestion des aspects administratifs et financiers dans la prise en charge de la maladie, la réinsertion sociale, etc. Le parcours patient doit avoir une visée d'amélioration de l'expérience du patient et de la qualité des soins.

Concept	Définition générale	Objectif	Niveau d'intervention	Evénements déclencheurs
Parcours patient	Ensemble des étapes et interactions avec le système de santé	Améliorer l'expérience et la qualité de la prise en charge du point de vue du patient.	Global sur le système de santé	Exposition à un facteur de risque ou reconnaissance d'un ensemble de symptômes
Parcours de soins	Etapes spécifiques d'un traitement médical donné	Garantir une prise en charge optimale et cohérente pour une maladie ou un trouble spécifique.	Local à l'échelle de la prise en charge. Peut être interne à une structure de santé	Diagnostic d'une pathologie ou d'un trouble. Démarche d'accès aux soins initiée par le patient.

Table 1 : Synthèse des définitions et objectifs des concepts de parcours patient et parcours de soins.

Quel que soit le niveau d'intervention de ces parcours, l'objectif commun reste l'amélioration de la qualité de la prise en charge clinique et thérapeutique qui doit permettre, pour le patient, le juste enchaînement au bon moment des différentes compétences professionnelles liées aux soins. Cette démarche qualité doit par ailleurs s'exprimer sur toutes les dimensions suivantes : pertinence, sécurité, efficacité clinique, accessibilité, continuité et « point de vue patient »[95].

[94] Définition issue du site du ministère de la santé. https://sante.gouv.fr/systeme-de-sante/parcours-des-patients-et-des-usagers/article/parcours-de-sante-de-soins-et-de-vie

[95] D'après les questions / réponses sur les parcours de soins de la Haute Autorité de la Santé (HAS). Consultable en ligne : https://www.has-sante.fr/upload/docs/application/pdf/2012-05/quest-rep_parcours_de_soins.pdf

2.1.2 Problematiques des parcours patient et de soins

Cette section présente les principaux enjeux des parcours patients et de soins et comment ceux-ci se déclinent ensuite en problématiques à chaque niveau, en s'appuyant sur l'exemple concret de la maladie de Parkinson.

Enjeux de ces parcours et expression des problématiques. Les parcours patients et de soins répondent par ailleurs à des enjeux communs[96], notamment celui de promouvoir une gestion coordonnée de la prise en charge, de se reposer sur les bonnes pratiques médicales tout en restant personnalisés et adaptés au patient, et de faciliter l'implication de celui-ci dans sa prise en charge. Ces enjeux se déclinent ensuite à chacun des niveaux d'intervention des parcours, quelques exemples figurent dans la Table 1Table 2 ci-dessous.

Enjeu	Problématique parcours patient	Problématique parcours de soins
Promouvoir une gestion coordonnée de la prise en charge	Comment organiser la coopération ville-hôpital et des passerelles entre la MCO et le médico-social ? Comment garantir la transmission des données et des informations ?	Comment coordonner des soins correspondant à plusieurs affections à la fois et impliquant un grand nombre de professionnels de santé ?
Garantir l'application des bonnes pratiques médicales tout en s'adaptant au patient	Comment réduire l'errance diagnostique[97] et diminuer le temps moyen au diagnostic ?	Quelles sont les difficultés rencontrées par les patients dans la prise d'un traitement donné ?
Faciliter l'implication du patient dans sa prise en charge	Comment atteindre les personnes exposées à un facteur de risque pathologique ?	Comment impliquer les aidants informels du patient dans la démarche thérapeutique ?

Table 2 : Exemples de déclinaisons des problématiques par type d'enjeu, sur le parcours patient et le parcours de soins.

L'expression de chaque problématique diffère bien selon le périmètre considéré. Des solutions communes peuvent être envisagées mais il s'agit de prendre en compte les spécificités de chaque cas d'application et d'adapter les méthodes à déployer.

[96] Selon (Rodde-Dunet & Mounic, 2016)
[97] L'errance diagnostique est définie comme la période allant de l'apparition des premiers symptômes à la date à laquelle un diagnostic précis est posé. Définition issue du site MaRIH : https://marih.fr/banque-nationale-de-donnees-maladies-rares/errance-impasse-diagnostiques/

Exemple de la maladie de Parkinson. Nous illustrons les concepts de parcours patients et parcours de soins via l'exemple de la maladie de Parkinson[98]. Cette pathologie est une maladie neurologique dégénérative qui se caractérise par un déclin progressif des neurones qui produisent la dopamine, un neurotransmetteur essentiel à la bonne régulation des mouvements. Cette dégénérescence entraîne des symptômes moteurs tels que la dyskinésie (difficulté ou anomalie dans l'exécution d'un mouvement), la bradykinésie (ralentissement des mouvements), des tremblements au repos ou encore de la rigidité musculaire.

Elle est la seconde maladie neurodégénérative la plus fréquente après la maladie d'Alzheimer et la seconde cause de handicap d'origine moteur chez le sujet âgé, après les accidents vasculaires cérébraux[99]. Son origine est encore méconnue, mais il est généralement admis qu'elle résulte d'une combinaison de facteurs génétiques et environnementaux. Il n'existe aujourd'hui pas de traitement curatif, mais plusieurs options de traitement sont disponibles pour permettre au patient de diminuer l'impact des symptômes sur son quotidien, notamment des médicaments dopaminergiques, une thérapie physique et des adaptations dans le mode de vie. La stratégie thérapeutique à appliquer dépend du projet personnalisé du patient dans le parcours de soins et de l'avancement de la maladie (cf. Figure 12 ci-dessous).

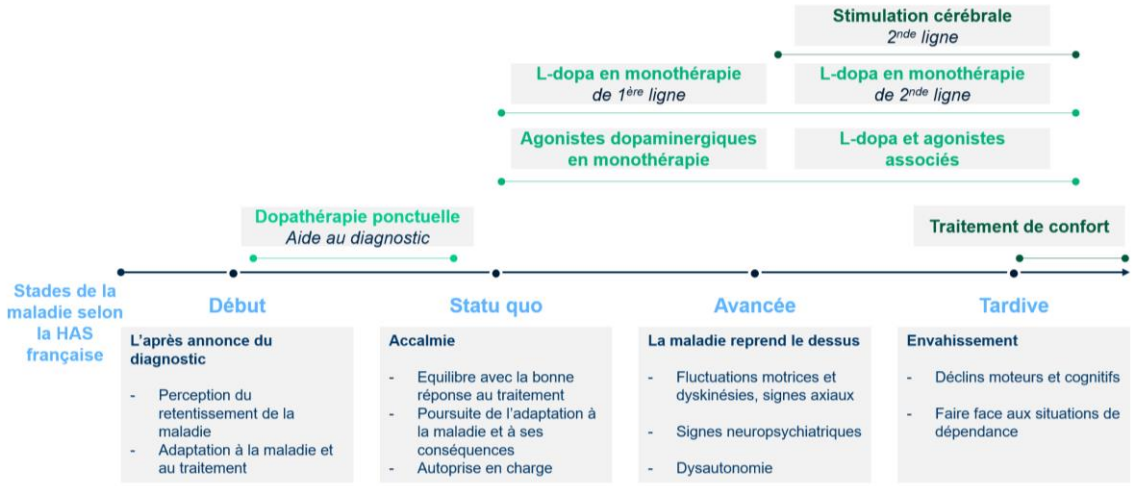


Figure 12 : Stratégies thérapeutiques médicamenteuses du parcours de soins de la maladie de Parkinson, selon les stades de la maladie tels que définis par la HAS. Schéma construit avec l'aide d'une coordinatrice « parcours patient » du Centre Expert Parkinson de l'AP-HP.

[98] Cet exemple s'inscrit dans la typologie des parcours qui seront étudiés dans les chapitres 3 et 4 de ce manuscrit.
[99] Selon le guide parcours de soins de la maladie de Parkinson, publié par la HAS et consultable en ligne. (*Guide parcours de soins maladie de Parkinson*, s. d.)

Comme toute pathologie complexe, la maladie de Parkinson s'exprime différemment chez chaque patient et les symptômes peuvent fortement varier d'une personne à l'autre, rendant difficile le diagnostic différentiel. Elle peut également s'accompagner de troubles associés tels que la dépression, l'anxiété, etc. Dans son guide traitant de la maladie de Parkinson, la HAS liste quatre points critiques de la prise en charge que nous proposons de classer selon leur niveau d'expression sur le parcours (cf. Table 3 ci-dessous).

Point critique du parcours (selon la HAS)	Niveau d'expression
La place des médecins généralistes dans le repérage de la maladie de Parkinson et l'orientation du patient vers un neurologue au moindre doute.	Parcours patient
La nécessité pour les patients d'être écoutés et de participer activement aux décisions thérapeutiques.	Parcours de soins
L'importance des traitements non médicamenteux dans le cadre d'une prise en charge pluridisciplinaire coordonnée (infirmiers, kinésithérapeutes, orthophonistes, psychologues).	Parcours de soins
L'intérêt de privilégier une prise en charge en ville (de proximité ou à domicile), en lien avec l'hôpital, et en prenant en compte les aspects médicaux et sociaux.	Parcours patient

Table 3 : Synthèse des points critiques du parcours de la maladie de Parkinson et classification selon le niveau d'expression sur le parcours.

Malgré une certaine proximité dans leurs définitions, objectifs et enjeux, les parcours patients et de soins sont des concepts bien distincts. Si la problématique de recherche que nous investiguons reste commune, il faudra néanmoins adapter la démarche à déployer pour intégrer les spécificités de chaque contexte d'application. Il s'agit maintenant de décomposer la problématique générale posée dans le chapitre précédent en sous-questions qui nécessiteront chacune un type de méthodologie pour y répondre.

2.2 Présentation de l'approche proposée

SYNOPSIS Où nous décomposons la problématique de recherche introduite dans le chapitre précédent, explicitons les méthodes présentées dans la littérature pour y répondre et détaillons les perspectives de recherche ouvertes par les articles référencés.

2.2.1 APPROCHE ET LITTERATURE EXISTANTE

Cette section présente les différentes étapes de l'approche à déployer et la littérature scientifique existante sur ces thématiques.

Décomposition de la problématique de recherche en approche. La problématique que nous investiguons, présentée dans le chapitre précédent est la suivante :

« COMMENT CARACTERISER ET ANTICIPER LES PARCOURS PATIENT ET DE SOINS SOUS L'ANGLE DES COUTS ? »

Nous choisissons de décomposer cette problématique en plusieurs étapes de résolution :

Il s'agit premièrement de recomposer le coût et la trajectoire individuelle de chaque patient, puis d'identifier les patients qui ont une trajectoire identique, c'est-à-dire qui appartiennent à des sous-groupes médico-économiques similaires et donc ont des consommations de soins proches en termes de nature mais aussi de coût. Par la suite, il faudra modéliser et représenter visuellement cette trajectoire, puis en prédire les évolutions. On notera par ailleurs que les données de facturation constituent un excellent proxy pour le suivi des soins dans le temps et c'est pourquoi nous sélectionnons les deux angles coûts / soins dans cette étude[100].

[100] D'après (Vogt et al., 2018).

Estimer le coût d'un parcours. On retrouve dans la littérature plusieurs méthodologies d'analyse des coûts dans l'objectif de reconstituer l'impact économique d'un parcours ou plusieurs parcours :

1. L'analyse des micro-coûts (plus connue sous le terme *Micro Costing Analysis* en anglais) est une méthode d'évaluation des coûts dite « de bas en haut (*bottom up*) ». Il s'agit d'énumérer et d'évaluer le coût de chaque ressource consommée par le traitement d'un patient donné, méthode qui est largement appliquée dans le domaine de la santé. Elle requiert un effort assez important pour décomposer les étapes du processus à évaluer et pour identifier de manière exhaustive les composantes individuelles de coût. Aujourd'hui il n'existe pas de méthode standardisée pour conduire ce type d'analyse, certains articles recommandent de compléter ce type d'analyse en y intégrant l'inflation et une analyse de sensibilité[101].

2. L'analyse des macro-coûts (*Macro Costing Analysis* ou encore *Gross Costing* en anglais) est une méthode d'estimation des coûts utiliser pour mesurer et quantifier l'impact économique global d'un programme, service ou d'une prise en charge. Ce processus implique d'identifier les dépenses totales directes et indirectes, effectuées pour toutes les composantes de l'analyse, puis éventuellement de les ventiler par patient ou par lit par exemple[102]. Cette méthode est particulièrement utile lorsqu'il s'agit d'estimer l'impact économique d'un service au global ou encore lorsque l'information détaillée pour chaque patient n'est pas disponible.

3. L'estimation des coûts basée sur les activités (*activity-based costing* en anglais) est une méthode de comptabilité de gestion d'entreprise qui permet de recomposer le coût d'un service ou d'un produit à partir des activités qui soutiennent la prestation ou la production. Elle est habituellement appliquée pour identifier les activités les plus coûteuses et les optimiser[103]. L'estimation des coûts basée sur les activités et les temps (*time-driven activity-based costing* en anglais) est une variante de cette analyse qui permet d'ajouter la notion de temps passé à réaliser une activité donnée[104].

[101] Selon (Xu et al., 2021).
[102] Selon (Raftery, 2000).
[103] D'après (Tchamdja et al., 2015).
[104] A retrouver chez (Keel et al., 2017).

4. Dans la recomposition des coûts d'une pathologie donnée ou d'une intervention particulière, les méthodes basées sur des cas-témoins (*case control* en anglais) sont particulièrement répandues[105]. L'objectif est de reconstituer deux sous-groupes de patients, les cas qui ont été exposés à la pathologie et les témoins qui ne l'ont pas été. Il s'agira ensuite de comparer les coûts associés aux deux groupes pour déterminer l'impact de l'exposition par différence.

Des exemples d'applications dans la littérature de chacune de ces méthodes sont présentés dans la table ci-dessous.

Méthodologie d'estimation des coûts	Exemple d'application	Contexte clinique
Macro-costing approach	(Geitona et al., 2010)	Unité de soins intensifs
Micro-costing approach	(Karabatsou et al., 2016)	Unité de soins intensifs
Time-driven activity-based costing	(Keel et al., 2017)	Variés : services de chirurgie, ambulatoire, etc.
Activity-based costing	(Tchamdja et al., 2015)	Unité de soins néonataux intensifs
Case control	(Chang et al., 2004)	Base de données administratives sur le cancer

Table 4 : Présentation des principales approches d'estimation des coûts des parcours et exemples d'applications dans la littérature scientifique.

Classifier les parcours en sous-groupes médico-économiques. Dans la littérature, plusieurs études ont montré l'apport des méthodes de classification non supervisée (*clustering* en anglais) dans l'analyse des trajectoires de patients pour des maladies chroniques, sous l'angle de l'évolution de la maladie[106], des coûts[107] ou des traitements reçus par le patient[108]. La classification non supervisée est particulièrement utile dans un contexte où les trajectoires ne sont pas labelisées a priori dans une classe qui permette d'identifier leur appartenance à un sous-groupe de patients.

[105] Un exemple chez (Chang et al., 2004).
[106] Voir (Anh Luong & Chandola, 2017).
[107] Exemple chez (Hajat et al., 2021).
[108] D'après (Najjar et al., 2018).

Effectuée sur la base d'un critère dérivé à partir de la consommation en soins du patient, les sous-groupes obtenus contiendront des parcours similaires en termes de soins reçus et donc de profil clinique. Plusieurs approches de segmentation sont régulièrement testées, parmi les plus fréquentes : celles basées sur les centroïdes, basées sur la densité ou encore hiérarchiques (cf. la table ci-dessous). Le choix des méthodologies est souvent arbitraire et peu documenté[109], nous implémenterons et comparerons donc plusieurs de ces approches.

Approche de classification	Algorithme	Exemple d'application	Contexte clinique
Basée sur les centroïdes	K-moyennes	(Anh Luong & Chandola, 2017)	Segmentation de la progression de la maladie chronique du rein
		(Hajat et al., 2021)	Classification des coûts en fonction des comorbidités chroniques
Basée sur les centroïdes	K-médoïdes	(Vogt et al., 2018)	Identification de séquences de prises en charge ambulatoires pour insuffisance cardiaque
Basée sur la densité	DBSCAN[110]	(Yan et al., 2019)	Segmentation des patients onéreux en fonction de leur pathologie
Hiérarchique	Arbre de décision	(Mei et al., 2015)	Classification des patients pour la suggestion de traitements

Table 5 : Présentation des principales approches de classification de groupes et exemples d'applications dans la littérature scientifique.

Les méthodes basées sur les centroïdes sont des approches qui consistent à regrouper un ensemble de données en clusters en minimisant la distance entre les données et les centres de chaque cluster. Pour les algorithmes K-moyennes (*K-means* en anglais) et K-médoïdes (*K-medoids* en anglais), les centres de chaque cluster sont respectivement définis comme la moyenne des points ou le point central du cluster. Contrairement aux algorithmes basés sur les centroïdes, les méthodes basées sur la densité se concentrent sur la densité des points de données et leur connectivité pour identifier des clusters. En particulier DBSCAN utilise un rayon pour capter les points situés dans un voisinage. Il différencie ensuite les points centraux des points de bruit pour former des

[109] Selon (Menger et al., 2019).
[110] Density-Based Spatial Clustering of Applications with Noise (DBSCAN) ou, en français, Classification non supervisée spatiale des applications avec bruit basée sur la densité.

clusters. Les clusters seront donc des zones de l'espace où la densité d'observations est importante[111].

Les méthodes de classification hiérarchique sont également une catégorie d'algorithmes répandue pour les tâches de segmentation. Généralement basées sur des arbres, elles sont soit agglomératives, soit divisives, selon qu'elles appliquent une approche *bottom-up* ou *top-down*[112] dans la création des clusters. De nombreuses références dans la littérature citent l'algorithme BIRCH comme une approche permettant d'obtenir des performances supérieures aux arbres dans de nombreux contextes, y compris cliniques[113].

Les avantages et limites de chacun de ces algorithmes seront détaillés dans le chapitre 3.

Modéliser et prédire le parcours. De nombreux articles s'intéressent à la modélisation des parcours patients et de soins. Ces trajectoires sont généralement représentées comme un enchaînement d'états qui correspondent à la progression d'une pathologie, à une série de soins et de traitement reçus, ou toutes autres données longitudinales cliniques permettant de rendre compte de la dimension temps du parcours. De par cet aspect séquentiel, c'est une typologie de problèmes auxquels sont particulièrement adaptées certaines approches spécifiques, qui intègrent les flux et notamment[114] :

1. Les méthodes basées sur l'apprentissage profond (*deep learning* en anglais) et notamment les réseaux récurrents de neurones (RNN pour *Recurrent Neural Networks* en anglais), tels que l'algorithme DeepCare[115].

2. Les approches combinées de *process mining* et de simulation à événements discrets[116].

[111] Publié par (Ester et al., 1996)
[112] Dans l'approche *bottom-up*, on parle de singletons que l'on regroupe progressivement en sous-ensembles jusqu'à ce qu'un point d'arrêt soit rencontré. Dans l'approche *top-down*, à l'inverse, on part de l'ensemble contenant toutes les observations que l'on divise progressivement jusqu'à ce qu'on point d'arrêt soit rencontré.
[113] Notamment (Ramadhani et al., 2020; Zhang et al., 1997).
[114] Revue de ces solutions proposée par (Silva & Matos, 2021).
[115] Présenté par (Pham et al., 2017)
[116] (Augusto et al., 2016).

3. Les méthodes de modélisation stochastiques telles que les chaînes de Markov, couplées ou non à des méthodes d'apprentissage supervisé et non supervisé.

4. Les méthodes d'apprentissage automatique (*machine learning* en anglais) dites classiques : arbres de décision, régressions, algorithmes ensemblistes, etc.

L'apprentissage profond est une technique algorithmique très puissante et de plus en plus répandue pour l'analyse des trajectoires dans un contexte médical, en témoigne le nombre de publications scientifiques dans ce domaine de la recherche qui est fort actif. Ces types d'approches ne figurent cependant pas toujours parmi les choix les plus appropriés et notamment lorsque les jeux de données sont relativement petits – ce qui peut engendrer des difficultés à entraîner un modèle qui généralise bien[117]. Par ailleurs, et pour chacun des cas d'application qui seront présentés dans la section suivante, nous souhaitons conserver un haut niveau d'interprétabilité – ce qui est toujours préférable dans un contexte d'utilisation clinique.

En ce qui concerne les méthodes de *process mining*, et malgré des perspectives de recherche particulièrement encourageantes, nous ne disposons pas du format de données adéquat[118].

Les chaînes de Markov sont particulièrement adaptées pour décrire et modéliser un système qui évolue au cours du temps et passe d'un état à un autre de manière stationnaire – c'est-à-dire que la probabilité de transition entre deux états est indépendante du temps écoulé depuis la dernière transition. Les applications à la modélisation des parcours sont nombreuses : prédiction de la mortalité au sein d'une unité de soins intensifs[119], modélisation dynamique de la progression de la sepsis pédiatrique (infection généralisée)[120], modélisation et prédiction de la trajectoire de traitement en réadaptation[121], ou encore analyse de la

[117] (C. Chen, 2004) discute, dans cette étude, des avantages des Random Forests par rapport aux réseaux de neurones, en particulier dans le contexte d'entraînement avec de petits ensembles de données.
[118] Les données utilisées pour le process mining sont généralement structurées sous forme de journaux d'événements, c'est-à-dire une séquence chronologique d'événements associés à une instance de processus. Ces journaux contiennent généralement a minima les informations suivantes : Identifiant unique | Activité | Date et heure à laquelle l'événement s'est produit (Aalst, 2016).
[119] Présenté par (Vairavan et al., 2012).
[120] Selon (Kausch et al., 2021).
[121] A retrouver chez (Kapadia et al., 1985).

progression du diabète de type 2[122]. Par ailleurs, la chaîne de Markov permet de visualiser de manière simple l'enchaînement des états et donc l'ensemble des trajectoires qui sont représentées. La simplicité d'implémentation de ce type d'algorithme en fait une approche de choix, mais elle repose sur l'hypothèse que la probabilité de transition d'un état à un autre ne dépend pas de l'historique de transitions. Dans certains contextes cliniques où l'historique peut avoir une très forte influence sur la trajectoire, il n'est pas toujours recommandé de s'appuyer sur ce type de méthode[123].

Les méthodes d'apprentissage automatique classiques sont généralement plus faciles à comprendre et à interpréter que des techniques plus complexes tel que le *deep learning*[124]. Par ailleurs, elles peuvent bien fonctionner avec des jeux de données relativement petits[125] et conviennent particulièrement à des contextes où les ressources de calcul à disposition sont contraintes. On notera pour autant que ces méthodes peuvent avoir du mal à modéliser et restituer des relations entre variables complexes et non linéaires[126]. Certains algorithmes en particulier, comme les arbres de décision, peuvent être sujets au surajustement, il est généralement recommandé d'atténuer cet aspect par l'utilisation d'algorithmes ensemblistes comme les forêts aléatoires ou le *boosting*[127].

Notes sur la visualisation des parcours. Quelques articles abordent des méthodes de visualisation des parcours patient et de soins, généralement à partir de graphes qui représentent particulièrement bien l'enchaînement des différentes étapes et les interactions avec le système de santé[128]. Nous choisirons de représenter les parcours au moyen d'un diagramme de Sankey, qui est l'approche retenue par (Huang et al., 2015). Ces diagrammes sont à la fois simples à implémenter et à lire, il existe par

[122] (Derevitskii & Kovalchuk, 2019).
[123] (Jackson et al., 2003) notent que les modèles de Markov traditionnels peuvent ne pas être suffisamment flexibles pour décrire correctement la progression de certaines maladies, en particulier lorsque l'hypothèse de Markov n'est pas respectée.
[124] (Rudin, 2019) argumente en faveur de l'utilisation de modèles d'apprentissage automatique interprétables, notamment pour les décisions à fort enjeu.
[125] (C. Chen, 2004).
[126] (Hastie et al., 2009) discute, dans cet ouvrage, des limites des méthodes d'apprentissage automatique classiques pour modéliser des relations complexes et non linéaires.
[127] (Breiman, 2001).
[128] (Dabek et al., 2015), (Widanagamaachchi et al., 2018) et (Huang et al., 2015).

ailleurs de nombreuses bibliothèques et logiciels en ligne qui permettent de les tracer automatiquement.

2.2.2 Perspectives ouvertes par la litterature referencee

Cette section présente les limitations discutées dans la littérature référencée et sur laquelle s'appuie l'approche choisie.

Estimer le coût d'un parcours. (Geitona et al., 2010) précisent que les données utilisées dans leur étude de coûts sont focalisées sur un service de soins intensifs d'un hôpital donné et ne peuvent pas être généralisées au système de santé. Par ailleurs, les auteurs n'ont pas pu disposer de données individuelles patients et ont appliqué une approche de Macro Costing. Cette approche a l'avantage de la simplicité mais ne peut pas prendre en compte des facteurs cliniques qui peuvent expliquer des variations individuelles pour certaines prises en charge. L'approche de Micro Costing appliquée par (Karabatsou et al., 2016) permet d'adresser une partie de ces limitations, bien que les chercheurs soulignent également un biais dans la collecte des données, car la totalité des dépenses n'a pas pu être couverte du fait de l'inexistence d'un dossier médicalisé centralisé. (Keel et al., 2017) confirment par ailleurs que ce coût doit être recomposé à partir des données informatisées pour assurer la fiabilité et l'exhaustivité de l'analyse. Ils recommandent également de séparer les coûts directs et indirects. Enfin, (Chang et al., 2004) rapportent l'importance d'intégrer des variables cliniques qui peuvent expliquer les différences dans les inducteurs de coûts des parcours.

Classifier les parcours en sous-groupes médico-économiques. (Anh Luong & Chandola, 2017) évoquent les difficultés d'optimisation des hyperparamètres de leur modèle (*Probabilistic Subtyping Model* pour segmenter les sous-groupes pathologiques). Pour pouvoir appliquer directement l'algorithme K-moyennes à des données longitudinales, ils suggèrent également de regrouper les observations en séries temporelles de même dimension et d'évaluer les données manquantes. Sur une méthodologie similaire, (Hajat et al., 2021) commentent également qu'il est difficile de dresser des liens de causalité entre les différentes comorbidités et les coûts associés. (Vogt et al., 2018), quant à eux, soulignent la nécessité de tester la généralisation du modèle à un échantillon de patients affichant des comorbidités et des degrés de sévérité dans la progression de la maladie plus larges. La granularité de

temporalité utilisée au cours de l'étude est d'un an et devrait être réduite pour améliorer la pertinence des clusters de trajectoires identifiés.

En concluant sur l'efficacité comparée de plusieurs approches de classification non supervisée, (Yan et al., 2019) mettent en avant la nécessité d'utiliser une méthode qui inclut les points de bruit, c'est-à-dire des observations non attribuées à un cluster, un phénomène fréquent dans des données cliniques en vie réelle. La proportion d'observations entre les classes devra également être un point d'attention, particulièrement si cette répartition est déséquilibrée et donc susceptible d'influencer fortement les résultats de l'algorithme.

Modéliser et prédire le parcours. (Kausch et al., 2021) ont développé un modèle Markovien pour estimer la progression de la maladie après un diagnostic de sepsis pour des enfants admis en soins intensifs. Les auteurs soulignent la nécessité d'une validation externe de leur méthode et notamment de tester la généricité sur une population différente que les patients de soins intensifs en pédiatrie. Par ailleurs, la définition de la gravité de la sepsis ne repose que sur un seul score clinique, la précision serait donc probablement améliorée en incorporant des indicateurs cliniques supplémentaires qui caractérisent l'avancée de la maladie. Une piste ouverte est également d'intégrer les interventions cliniques qui sont effectuées en cours de parcours pour en mesurer l'impact sur la trajectoire.

(Silva & Matos, 2021) évoquent la possibilité d'inclure des données non structurées dans l'analyse telles que les notes cliniques – mais également la nécessité de développer des modèles robustes de sélection des variables disponibles pour réduire la haute dimensionnalité des modèles, tout en conservant les caractéristiques patient d'intérêt. Une meilleure interprétabilité des trajectoires et des modèles transparents sont attendus pour favoriser l'adoption de ces technologies dans des contextes cliniques.

Visualiser le parcours. La visualisation des parcours est une étape nécessaire pour identifier des tendances communes aux parcours et simplifier l'interprétation des trajectoires. (Widanagamaachchi et al., 2018) présentent une méthode de visualisation et d'analyse qui permet d'explorer la progression de la trajectoire des patients au cours du temps. L'identification et le regroupement des trajectoires est basée sur une note de similarité.

Ce score est calculé en fonction de l'état actuel du patient et de son profil administratif dressé à l'admission à l'hôpital. L'algorithme a par ailleurs été entraîné sur une base de données de santé publique. Les auteurs identifient deux principales perspectives futures : l'intégration de l'historique de santé dans le groupement des trajectoires et l'application de la méthodologie à des données de santé en vie réelle.

La méthode présentée par (Huang et al., 2015) permet de simplifier la visualisation des parcours en filtrant les variables cliniques d'intérêt. L'article évoque plusieurs pistes d'améliorations parmi lesquelles une meilleure interprétation des associations intra et inter-clusters, et une revue des trajectoires avec une équipe clinique d'utilisateurs. Ce dernier point est aussi évoqué par (Dabek et al., 2015).

En conclusion, nous retenons trois catégories d'axes d'améliorations relevés dans la littérature, que nous nous proposons d'adresser dans ces travaux :

1. Une attention sur la sélection des variables d'intérêt : notamment intégrer plusieurs indicateurs cliniques qui permettent de caractériser la progression de l'état du patient et d'interpréter des variations dans les coûts de prise en charge ou les trajectoires ; inclure les interventions et soins réalisés dans l'analyse des parcours ;

2. Des exigences à respecter au niveau de la méthodologie : et particulièrement combiner des approches d'estimation des coûts macro et micro pour aboutir à une évaluation exhaustive ; veiller à choisir une méthode de classification qui prenne à la fois en compte une répartition déséquilibrée des observations entre les classes et puisse labelliser les observations « bruit » comme telles ;

3. Une démarche d'interprétation et d'évaluation des résultats robuste : parmi les recommandations, on compte notamment l'implémentation sur des données en vie réelle ; l'emploi d'une méthode robuste d'interprétation des clusters et le fait de faire d'évaluer la cohérence des trajectoires par une équipe de soignants et d'experts.

2.3 Présentation des cas d'application

SYNOPSIS Où nous présentons les cas d'application qui seront investigués au cours de ces travaux, ainsi que l'adaptation de l'approche déployée.

2.3.1 BREVE INTRODUCTION DES CAS D'APPLICATION

Prédire le parcours patient, exemple des troubles neurocognitifs. Les Centres Mémoires Ressource Recherche (CMRR) sont des pôles de soins et de recherche pour des patients souffrant de troubles neurocognitifs. En 2012, une équipe de médecins et chercheurs du CMRR des Hospices Civils de Lyon a lancé la collecte d'une cohorte prospective de données de santé en vie réelle des patients reçus en consultation Mémoire à l'hôpital des Charpennes. Cette base de données combine des informations sur le profil du patient, l'évolution de son trouble mesurée par plusieurs indicateurs cliniques et l'ensemble des soins consommés et des consultations réalisées, retracés à partir des données de coûts de l'Assurance Maladie. Ces données de coûts sont agrégées par semestre. Le début du parcours patient est marqué par la première consultation en centre mémoire et peut durer jusque quatre ans. Dans ce cas d'application, on cherche à classifier les patients en sous-groupes économiques cohérents, puis à modéliser la trajectoire du patient et à évaluer la probabilité de transitionner d'un sous-groupe à un autre en cours de parcours. Ces travaux seront détaillés dans le chapitre 3.

Prédire le parcours de soins, exemple de l'hospitalisation à domicile. Soins et Santé est une structure d'Hospitalisation à Domicile (HAD) qui accueille des patients en sortie ou en substitution à une hospitalisation conventionnelle. A l'admission, on attribue un mode de prise en charge principal, un mode associé et un indice de Karnofsky à chaque patient. Cette association, appelée séquence de soins, définit le sous-groupe économique du séjour. Pour chaque parcours de soins, nous disposons aussi du profil administratif et clinique du patient. Au cours de son séjour, le patient va recevoir une à plusieurs visites par jour de soignants libéraux, par exemple infirmiers ou kinésithérapeutes. Dans ce cas d'application, on cherche à prédire le nombre de visites requis par jour dans plusieurs scénarios de données – avec ou sans historique de visites. Ces travaux seront détaillés dans le chapitre 4.

2.3.2 Des differences de contexte justifient des differences d'approche

La table ci-dessous résume les caractéristiques et spécificités des cas d'application présentés dans la section précédente. Nous argumentons également l'adaptation de l'approche déployée en fonction du contexte clinique des données.

		Memora	Soins et Santé
Caractéristiques générales	Niveau d'intervention	Parcours patient	Parcours de soins
	Profils de patient	Troubles neurocognitifs	Hospitalisation à domicile
	Granularité temporelle	Semestre	Journée
	Effecteurs	Multi-structures de santé	Propre à une structure de santé
	Classification médico-économique	A construire	Connue a priori via la séquence de soins
Approche proposée en fonction du contexte	**Estimer le coût d'un parcours**	Approche micro exhaustive	Approche macro et micro partielle sur les visites seulement
	Classifier les parcours en groupes médico-économiques	Plusieurs méthodes de clustering seront testées	Les groupes sont déjà connus
	Modéliser le parcours	La granularité par semestre permet d'envisager une chaîne de Markov car l'état actuel comportera un « historique » de six mois	La granularité par jour ne permet pas d'envisager une chaîne de Markov. On testera plusieurs modèles d'apprentissage automatique « classiques » sur des scénarios avec/sans historique
	Prédire le parcours	A partir de la matrice de transition de la chaîne de Markov	A partir de la prédiction du nombre de visites sur les semaines du séjour

Table 6 : Synthèse des caractéristiques de chaque cas d'application et déclinaison de l'approche proposée en fonction du contexte.

Les chapitres 3 et 4 font suite à cette section et développent les travaux produits sur chacun de ces cas d'application tout en intégrant les axes d'amélioration mentionnés dans la littérature et résumés plus haut.

3

Prédire le parcours patient, exemple des troubles neurocognitifs

Ce troisième chapitre décrit notre premier cas d'application, la prédiction de parcours patients atteints de troubles neurocognitifs grâce aux croisements de données cliniques issues de la base Memora et des données médico-économiques de la CPAM.

Contenu

3.1 Description du contexte

SYNOPSIS Où nous présentons brièvement les troubles neurocognitifs et leurs spécificités de prise en charge, la base de données issue des consultations Mémoire qui nous permet de mener ces travaux de recherche, ainsi que la problématique explorée.

3.1.1 INTRODUCTION ET EXPLICATION DU CONTEXTE

Cette section pose la définition des troubles neurocognitifs ainsi que les modes de prise en charge et leur impact économique sur le patient et les aidants.

Définition des troubles neurocognitifs. Un trouble neurocognitif (TNC) est une « réduction acquise, significative et évolutive des capacités dans un ou plusieurs domaines cognitifs, souvent associée à un changement de comportement et de personnalité, et qui ne peut être expliqué par une dépression ou des troubles psychotiques[129] ». Il s'agit d'un syndrome qui peut être dû à des étiologies multiples telles que la maladie d'Alzheimer, la maladie de Parkinson, la maladie à corps de Lewy, une infection par le VIH ou encore l'utilisation d'un médicament. L'atteinte neurocognitive peut être plurielle et toucher de nombreux domaines, parmi lesquels :

1. Attention complexe : avoir des difficultés à se concentrer et être facilement distrait par des stimuli extérieurs ; les opérations mentales prennent plus de temps.
2. Fonctions exécutives : difficultés à planifier, prendre des décisions, à faire fonctionner la mémoire de travail.
3. Apprentissage et mémoire : incapacité à mémoriser des informations récentes et à faire appel à la mémoire immédiate.
4. Langage : fluence verbale diminuée, écholalie[130], mutisme.

[129] Définition posée dans le DSM-5, le manuel diagnostique et statistique des troubles mentaux publié par American Psychiatric Association. (Crocq & Guelfi, 2015)
[130] Trouble du langage qui consiste à répéter de manière systématique les derniers mots entendus. Définition issue du Larousse.

5. Activités perceptivomotrices : baisse de la dextérité dans les activités précédemment familières (conduite, maniement d'outils), confusion face à une baisse de la luminosité ou des ombres.

6. Cognition sociale : comportement en dehors des normes sociales, moindre habileté à reconnaitre les expression faciales et diminution de l'empathie.

L'atteinte sur les fonctions cognitives varie selon la sévérité du trouble et peut être mesurée au moyen de différentes échelles de tests[131]. Ces évaluations peuvent s'accompagner d'examens de laboratoire et de neuroimagerie qui permettront de poser un diagnostic différentiel[132].

Les TNC se développent habituellement conjointement au vieillissement. On estime qu'entre 2 et 10% des cas se déclarent avant l'âge de 65 ans, puis que la prévalence du trouble double tous les cinq ans à partir de ce seuil[133]. En 2019, on estimait que 57,4 millions de personnes vivaient avec des troubles neurocognitifs dans le monde, et la prévalence devrait atteindre 152,8 millions de cas d'ici 2050[134]. Dans l'extrême majorité des cas, il n'existe pas aujourd'hui de traitement curatif aux TNC. Le projet thérapeutique consiste essentiellement en l'accompagnement des patients atteints pour limiter l'impact des symptômes et leur permettre une vie la plus sereine possible[135].

Prise en charge des troubles neurocognitifs. La prise en charge de ces troubles peut nécessiter une approche médicamenteuse. Il s'agit surtout de coordonner une stratégie de soins pluridisciplinaire : neurologue, gériatre, orthophoniste, kinésithérapeute, ergothérapeute, psychologue, auxiliaires de vie et aides-soignants pour en citer quelques-uns. L'évolution des troubles neurocognitifs étant complexe et spécifique à chaque patient, il existe une grande diversité de parcours. Ces trajectoires peuvent être influencées par de multiples facteurs, parfois difficiles à mesurer, notamment : la progression du déclin cognitif, l'apparition de comorbidités impactantes, la présence d'aidants et la possibilité de rester

[131] Par exemple le Mini Mental State Examination (MMSE) qui sera présenté plus en détail dans la section suivante.
[132] Selon (Büla et al., 2007)
[133] Chiffres cités par (World Health Organization, 2012).
[134] Estimations présentées par (Nichols et al., 2022).
[135] Selon (Sol et al., 2018).

à domicile dans un environnement adapté. L'évolution des soins requis est donc difficile à anticiper.

Les conséquences des symptômes se répercutent sur le patient, mais aussi sur ses proches, avec une charge psychologique et économique liée aux soins et à la dépendance. En France seule, les coûts médicaux et paramédicaux des soins portés par le système de santé pour diagnostiquer, traiter et prendre en charge les patients atteints de la maladie d'Alzheimer sont évalués à 5,3 milliards d'euros par an. Le coût total associé aux soins informels, tels que l'aide quotidienne pour le ménage, l'habillage, les déplacements, pris en charge par les proches, est évalué à 14 milliards d'euros par an, dont une grande partie est payée par les aidants eux-mêmes[136]. En parallèle, l'Institut National des Statistiques et des Etudes Economiques (INSEE) estime que l'augmentation du nombre de patients âgés nécessitant une aide quotidienne sera bien supérieure au nombre d'aidants informels disponibles[137]. Pouvoir anticiper l'entrée en dépendance et le coût de la prise en charge devient donc un enjeu de santé publique et une préoccupation majeure.

3.1.2 Presentation du centre de la memoire

Présentation des consultations Mémoire. Le Centre Mémoire Ressources Recherche (CMRR) est un pôle de soins et de recherche implanté au sein des Hospices Civils de Lyon, dans plusieurs hôpitaux. Il accueille des patients souffrant d'un trouble cognitif suspecté ou avéré et a pour objectif de poser un diagnostic initial, d'assurer le suivi des malades et de coordonner les parcours de soins en lien avec les structures médico-sociales locales[138]. Le centre est également un catalyseur de la recherche et pilote pour ses patients des essais cliniques médicamenteux et non-médicamenteux.

Présentation du projet Memora. En 2012, une équipe de médecins et chercheurs initie la création d'une cohorte prospective de données de santé en vie réelle des patients reçus en consultation Mémoire.

[136] D'après (Bérard et al., 2015)
[137] Estimatifs cités par (Bérard et al., 2015)
[138] D'après le site internet du CHU de Lyon. https://www.chu-lyon.fr/centre-memoire-ressources-recherche

Initialement implanté au sein de l'hôpital des Charpennes, le dispositif s'est développé parmi les consultations Mémoire partenaires de la région Auvergne-Rhône-Alpes et notamment aux CHU de Grenoble et Saint-Etienne. Cette base, qui bénéficie du financement de la fondation MSD Avenir, a également été appairée avec les données de la caisse d'Assurance Maladie de la région et présente des opportunités uniques de recherche qui se reflètent dans une série de travaux déjà menés[139] :

- Evaluer l'impact économique des TNC à partir des coûts médicaux directs de l'Assurance Maladie[140],
- Etudier les trajectoires du déclin cognitif chez les patients atteints de la maladie d'Alzheimer au stage léger et leur prédicteur[141],
- Suivre l'impact déclaré de la maladie d'Alzheimer et des démences apparentées sur les aidants à domicile[142].

Par ailleurs, cette base de données combine de précieuses informations sur le profil du patient, l'évolution mesurée de son trouble et les soins consommés en dehors des consultations Mémoire[143]. Elle constitue donc une excellente référence de l'étude des parcours patients autour des troubles neurocognitifs.

3.1.3 Problematique a l'etude

Les objectifs de l'étude sont de décrire les parcours des patients souffrant de troubles neurocognitifs ou de plaintes cognitives subjectives[144], et d'étudier les principaux facteurs de variation en utilisant les caractéristiques cliniques des patients, y compris l'évolution de la maladie. Nous utilisons ensuite les mêmes facteurs pour prédire la variation du parcours des patients. Nous pensons que ce travail aidera les acteurs de la santé publique à comprendre les facteurs de coûts associés aux troubles neurocognitifs et les facteurs ayant un impact sur le parcours des

[139] D'après le site internet de la fondation MSD Avenir. https://www.msdavenir.fr/2021/10/12/projet-memora/
[140] Voir (Dauphinot et al., 2021)
[141] Etude réalisée par (Dauphinot et al., 2019)
[142] A retrouver chez (Dauphinot et al., 2022)
[143] Le descriptif des données exploitées dans la base sera détaillé dans la section suivante.
[144] Une plainte cognitive est employée pour désigner le mécontentement exprimé par un patient à l'égard d'une diminution subjective d'une ou plusieurs capacités cognitives dans la vie quotidienne.

patients, et donc à conduire des politiques de santé avec des preuves persistantes.

Notre méthodologie repose sur le regroupement des séries temporelles et les chaînes de Markov pour stratifier les groupes de patients en fonction du type de soins reçus et évaluer la probabilité de passer d'un groupe de patients à un autre au fil du temps. Dans cette étude, nous utilisons le coût des soins facturés au système national français d'assurance maladie comme indicateur de la consommation des ressources de santé et donc des soins fournis au patient.

3.2 Méthodologie

SYNOPSIS Où nous explicitons les données à disposition, les variables extraites et l'approche déployée.

3.2.1 DESCRIPTION DES DONNEES ET DE LA POPULATION A L'ETUDE

Cette section présente le contexte en vie réelle des données Memora, puis synthétise les tables principales de Memora et de la CPAM à partir desquelles les variables de notre analyse seront extraites.

Informations générales sur la base Memora. La population étudiée est constituée de patients inscrits dans la cohorte Memora suite à une consultation Mémoire du CMRR. Pour notre analyse, nous avons inclus les patients fréquentant le Centre Mémoire pour le diagnostic et le suivi de leurs troubles neurocognitifs sur une période de 6 ans, entre 2014 et 2020. Les informations cliniques portées par la cohorte ont été enrichies de données économiques agrégées au niveau du semestre à partir de la base de données nationale des demandes de remboursements de l'assurance maladie, en utilisant la première consultation comme début de la trajectoire et jusqu'à 4 ans.

Les variables obtenues à partir des deux bases de données peuvent être regroupées en plusieurs grandes catégories : caractéristiques cliniques et sociodémographiques, performances neuropsychologiques, maladies chroniques, hospitalisations, actes médicaux, consultations médicales et pharmacie (voir Figure 13 ci-dessous).

A chaque consultation Mémoire, le patient génère une ligne dans la base de données du centre. La plupart des informations sont saisies au cours de la visite initiale et n'évolueront pas ou peu au cours du parcours (e.g., les informations socio-démographiques ou l'historique des comorbidités). Une à plusieurs entrées de cette base sont mises à jour au cours des visites de suivi, par exemple les scores d'évaluation du déclin cognitif et de l'autonomie ou encore le diagnostic étiologique (cause de la maladie) s'il a pu être posé. Une à deux consultations en moyenne sont effectuées par an, au cours desquelles tous les scores ne sont pas systématiquement réévalués.

En parallèle, et en dehors de ces visites au Centre Mémoire, le patient va effectuer un certain nombre de contacts avec les services de santé : consultations en médecine de ville, achat de médicaments, examens d'imagerie ou encore hospitalisations pour en citer quelques-uns. Ces contacts sont enregistrés par la Caisse Primaire d'Assurance Maladie sous trois informations : le type de consommations, le nombre et leur coût pour l'Assurance Maladie et pour le patient. Ces entrées sont ensuite agrégées par période de six mois pour constituer la base de données des consommations.

Figure 13 : Présentation de l'organisation générale de la base de données et des variables clefs de l'étude

Présentation des tables Memora. La base de données initiale Memora est composée de sept tables : BNA (Banque Nationale Alzheimer), NPI, Contexte de vie, Traitement, MiniZarit, IADL et MMSE (cf. Figure 14 ci-dessous). Ces tables sont toutes connectées entres elles via la clef primaire « IPP » qui est l'identifiant unique du patient. Les données des traitements suivis n'étant pas exhaustives dans cette base, nous choisirons de les reconstituer à partir des données de la CPAM. La table traitement ne sera donc pas utilisée dans cette analyse.

Nous présentons dans les sections qui suivent le détail de chaque table (hors traitement) et comment celles-ci sont finalement concaténées pour obtenir la table principale BNA.

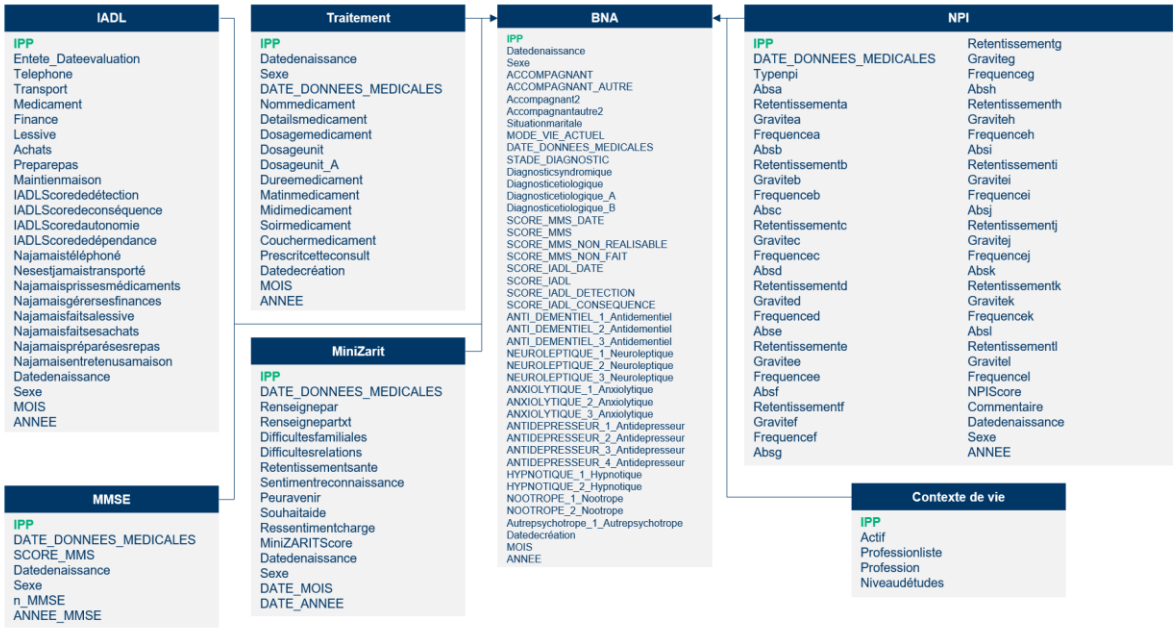


Figure 14 : Cartographie des tables et variables présentes dans la base de données Memora

La table NPI.

L'objectif de l'Inventaire NeuroPsychiatrique (NPI) est de « recueillir des informations sur la présence de troubles du comportement chez des patients souffrant de démence[145] ». Les troubles sont évalués selon douze axes de comportement[146], si la présence est avérée alors on renseigne également la fréquence d'apparition[147] et la gravité[148]. Pour chacun des douze domaines évalués, on calcule alors le score comme suit :

$$score = fréquence * gravité$$

[145] Définition et objectif posés dans le mémo « Inventaire neuropsychiatrique – version équipe soignante » de la HAS. Consultable en ligne : https://www.has-sante.fr/upload/docs/application/pdf/2013-02/08r07_memo_maladie_alzheimer_troubles_comportement_equipe_soignante_npi-es_2013-02-26_14-58-55_901.pdf

[146] Domaines évalués : Idées délirantes, hallucinations, agitation, dépression, anxiété, euphorie, apathie, désinhibition, irritabilité, comportement moteur aberrant, sommeil, modification de l'appétit et des comportements alimentaires.

[147] Echelle d'évaluation de la fréquence : 1 (quelquefois, inférieur à une fois par semaine) ; 2 (assez souvent, environ 1 fois par semaine) ; 3 fréquemment (plusieurs fois par semaine mais pas tous les jours) ; 4 (très fréquemment, tous les jours).

[148] Echelle d'évaluation de la gravité, c'est-à-dire à quel point les comportements sont invalidants pour le patient : 1 (léger, changement peu perturbant pour le patient) ; 2 (moyen, changement plus perturbant pour le patient mais sensible à l'intervention de l'aidant) ; 3 (important, changement très perturbant, insensible à l'intervention de l'aidant).

Un score supérieur à 2 est considéré comme pathologique. Par la suite, le soignant évalue également le retentissement sur l'entourage[149]. La table NPI est ainsi constituée de :

- Douze variables binaires qui marquent la présence ou l'absence du trouble du comportement sur un domaine donné,
- Pour chacun des domaines, les quatre indicateurs cités plus-haut : fréquence, gravité, score et retentissement,
- La date d'évaluation,
- L'IPP du patient concerné.

La grille complète d'évaluation telle que définie par le CMRR de Nice est publiée en **Annexe *1***.

La table MiniZarit.

La grille Mini Zarit est un outil d'évaluation de la souffrance des aidants informels dans le maintien à domicile des personnes âgées. Elle est organisée autour de sept questions, posées à l'aidant par un soignant au cours d'un entretien à domicile ou d'une consultation. Ces questions ont pour objectif d'évaluer les 7 domaines de souffrance suivants : difficulté dans la vie familiale, difficulté dans les relations et activités sociales, retentissement sur la santé de l'aidant, sentiment de ne plus reconnaître son parent, peur pour l'avenir, souhait d'être aidé et charge ressentie. Chaque axe est évalué par une note[150], puis ces dernières sont sommées pour donner un score global de 0 à 7. Un score global élevé est un marqueur de fatigue chez l'aidant. La table Mini Zarit est ainsi constituée de :

- Sept variables numériques qui correspondent à la note obtenue sur chacun des domaines évalués,
- Le score total,
- La date d'évaluation,
- L'IPP du patient concerné.

La grille complète d'évaluation du Mini-Zarit proposée par les hôpitaux universitaires de Genève est disponible en **Annexe *2***.

[149] Echelle d'évaluation du retentissement, c'est-à-dire à quel point le comportement est perturbant pour le soignant ou l'entourage au plan émotionnel : 0 (pas du tout) ; 1 (minimum) ; 2 (légèrement), 3 (modérément) ; 4 (sévèrement) ; 5 (très sévèrement, extrêmement). A noter que ce score est corrélé au MiniZazit, présenté après.
[150] Echelle d'évaluation du Mini-Zarit : 0 (jamais) ; ½ (parfois) ; 1 (souvent).

La table IADL.

Le score IADL (Instrumental Activities of Daily Living) est un outil d'évaluation qui permet de mesurer le niveau de dépendance fonctionnelle d'une personne dans ses activités de la vie quotidienne[151]. Cet indicateur n'est pas réservé uniquement aux TNC, mais est également utilisé dans le cadre de maladies chroniques, de handicaps physiques ou encore pour évaluer l'efficacité d'une stratégie thérapeutique.

Les 8 activités évaluées incluent : la capacité à utiliser un téléphone, à faire ses courses, à préparer son repas, à gérer les tâches domestiques, à faire sa lessive, les modes de transport utilisés, la capacité à prendre son traitement en autonomie, ainsi que la gestion des finances. Chacun de ses domaines est évalué entre 0 (incapacité) et 1 (capacité). Les notes individuelles sont ensuite sommées pour donner un score global compris entre 0 et 8 – un faible score indiquant un besoin d'assistance pour la vie quotidienne. La table IADL est ainsi composée de :

- Huit variables numériques qui correspondent à la note obtenue sur chacun des domaines évalués,
- Le score total,
- La date d'évaluation,
- L'IPP du patient concerné.

La grille complète d'évaluation du score IADL peut être retrouvée, en anglais, en **Annexe *3***.

La table MMSE.

Le Mini Mental State Examination (MMSE) est une échelle d'évaluation des fonctions cognitives largement utilisée dans les milieux hospitaliers et soins primaires pour estimer l'atteinte de patients souffrant de TNC. Il évalue, via trente tâches et questions, les capacités d'orientation dans le temps et l'espace, de mémoire verbale à court terme, de calcul, de langage et de praxies constructives[152].

[151] Contrairement au score de l'AVQ (Activités de la Vie Quotidienne), qui est présenté et utilisé dans le chapitre suivant, l'IADL évalue les activités qui requièrent des compétences cognitives plus avancées.

[152] Les praxies sont des capacités motrices acquises, c'est-à-dire des mouvements organisés appris dans le but d'atteindre un objectif. Il en existe plusieurs types (par exemple bucco-faciales qui regroupent les mouvements volontaires des parties du visage, comme un sourire avec les lèvres ou

Le score final est le nombre d'items exacts parmi les exercices posés, et va de 0 à 30. Un score de 23[153] ou moins est le seuil généralement accepté indiquant la présence de TNC. Il est important de noter que le score du MMSE est fortement corrélé à l'âge et au niveau d'éducation du répondant[154] et doit donc être interprété par un praticien. Une version de la grille d'évaluation du MMSE est disponible en **Annexe *4***.

La table Contexte de vie.

Cette table regroupe le niveau d'études et la profession du patient parmi une liste préétablie.

La table BNA.

BNA est la table principale qui regroupe, pour chaque patient et à chaque visite, les caractéristiques du patient et l'évolution des différents scores cités plus haut. On retrouve notamment :

- L'IPP du patient concerné,
- La date de naissance,
- Le genre du patient,
- La date de la visite,
- La présence d'un accompagnant et la nature de leur relation,
- Le stade diagnostic, une variable textuelle qui mesure le niveau de performance cognitive du patient, de « (1) plainte cognitive isolée » à « (2) trouble cognitif léger » puis « (3) trouble cognitif majeur (démence) »,
- Le diagnostic syndromique, qui définit la nature du trouble cognitif et son mode de présentation,
- Le diagnostic étiologique, qui définit la maladie à l'origine des symptômes (e. g., maladie d'Alzheimer),
- Le score MMSE,
- Le score IADL,

---

un roulement d'yeux). Les praxies constructives représentent les capacités à planifier et exécuter un mouvement dans le but d'organiser des éléments qui constitueront un dessin ou une figure finale. Définition issue du site internet https://neuronup.fr/domaines-dintervention/fonctions-cognitives/praxies.

[153] L'échelle du niveau d'atteinte en fonction du score final du MMSE : Inférieur à 10 – atteinte sévère ; 11 à 20 – atteinte modérée ; 21 à 26 – atteinte légère ; Supérieur à 27 – pas d'atteinte cognitive. Selon (*Mini-Mental State Examination (MMSE) – Strokengine*, s. d.)

[154] Comme démontré dans la littérature, il y a une relation positive entre les scores au MMSE et le niveau d'éducation. Le score médian au MMSE est de 29 pour les individus avec au moins 9 années de scolarité vs 22 pour ceux avec 0 à 4 ans. (Crum et al., 1993)

- Les médicaments prescrits (anti-démentiel, neuroleptique, anxiolytique, antidépresseur, hypnotique), bien que ces variables soient considérées comme incomplètes par les gestionnaires de la base.

A noter que certains scores sont manquants et devront être extraits depuis les tables correspondantes.

Présentation de la base CPAM. Cette base de données conséquente regroupe les différentes consommations de soins enregistrées pour chaque patient et regroupées par semestre. Les patients sont identifiés par une chaîne de caractères unique appelée RG.

Les regroupements de consommations sont effectués dans les différentes tables TAZ 100 (actes), 200 (consultations libérales), 300 (pharmacie), 501 (hospitalisations publiques), 502 (hospitalisations privées). Chaque table porte des entrées concernant :

- Le RG du patient concerné,
- L'année et semestre du regroupement,
- Le positionnement du semestre dans le parcours du patient (S1, S2, etc.),
- Le détail de la consommation (ex : hospitalisation en psychiatrie),
- La quantité consommée pendant le semestre,
- Le coût agrégé par semestre et type de consommation.

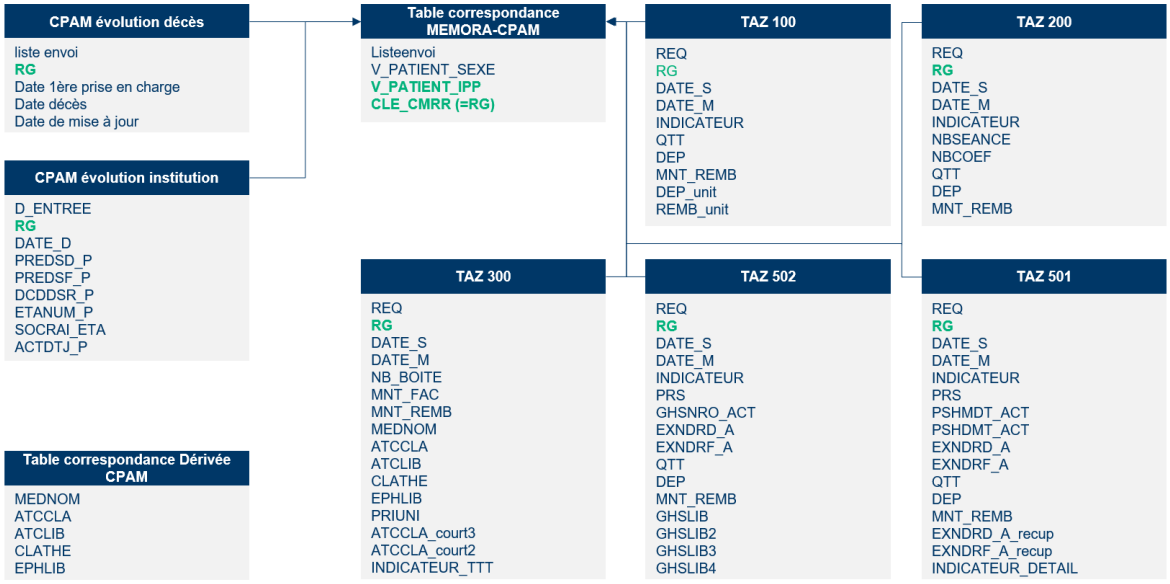


Figure 15 : Cartographie de la base de données de la Caisse Primaire d'Assurance Maladie.

La base de données regroupe par ailleurs deux tables supplémentaires qui permettent de comptabiliser les décès et l'entrée en institution de type EHPAD. La correspondance entre le RG de la CPAM et l'IPP de Memora est assuré par la table « correspondance MEMORA-CPAM ». A noter que nous disposons également des données PMSI d'hospitalisations publiques, qui portent une information plus détaillée sur les types de consommations. Nous l'utilisons pour expliciter certaines valeurs telles que 'Hospitalisations Autres'.

La jonction des bases Memora et de la CPAM ainsi que les étapes de pré-traitement décrites dans la section suivante permettront de constituer le jeu de données de l'analyse.

3.2.2 Preparation des donnees et analyse exploratoire

Pré-traitement général et mise en forme des données. L'ensemble de données a été créé en joignant les données de la cohorte MEMORA aux données des demandes de remboursements à l'aide d'un identifiant de patient unique. Afin de fusionner les tableaux, toutes les données Memora ont été agrégées au niveau du semestre et, en cas d'absence, les données cliniques renseignées lors de la dernière visite médicale ont été conservées. La jonction entre les tables BNA et les données de la CPAM est réalisée grâce à la correspondance entre l'IPP et le RG. Il est important de noter qu'un certain nombre de données sont perdues à la jonction car 45% des patients de la table Memora n'ont pas d'identifiant RG associé. Il y a deux explications à ce phénomène :

- Les données de la CPAM ont été envoyées uniquement pour les patients qui ont au moins une mesure du score IADL, ce parti pris a été établi pour une étude spécifique au suivi de cet indicateur et réalisée précédemment[155].
- La CPAM ne couvre pas tous les patients Memora (quelques personnes ne dépendent pas directement de la CPAM).

Suite à la jonction, l'ensemble de données comprend 10 695 patients, mais la population finale de l'analyse est plus petite en raison de données manquantes ou erronées.

[155] (Dauphinot et al., 2022).

Une série d'étapes de pré-traitement est réalisée sur les tables de la base Memora, synthétisée dans la Figure 16 ci-dessous.

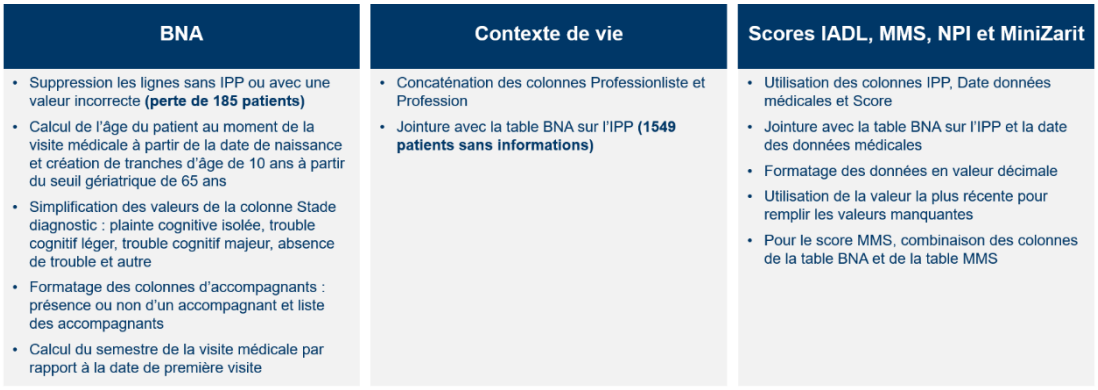


Figure 16 : Synthèse des opérations de pré-traitement réalisées sur les tables de la base Memora.

Recodage du stade diagnostic et du diagnostic étiologique.

Le stade diagnostic et le diagnostic étiologique sont des variables textuelles qui prennent un large ensemble de valeurs. Pour simplifier l'analyse et en prévision de leur transformation en variables binaires, nous les recodons selon un schéma décidé avec l'équipe Memora. Le code utilisé est consultable en **Annexe *5***.

Imputation des valeurs manquantes.

Plusieurs méthodes d'imputation des valeurs manquantes ont été utilisées en fonction du type de données. Les scores IADL, NPI, MiniZarit et MMSE comportent chacun un nombre important de valeurs manquantes à cause d'une jonction réalisée sur la date exacte (cf. valeurs en bleu foncé sur la Figure 17 ci-dessous). En effet, ces scores ne sont pas évalués systématiquement à chaque semestre, ni même à chaque visite. Pour les patients ayant enregistré plusieurs visites et dont les valeurs des scores cliniques étaient manquantes, nous avons récupéré le score connu le plus proche lorsque cela était possible (à ± 3 mois). Une jonction sur le score le plus récent, peu importe la distance temporelle enregistré, aurait permis de diminuer encore le nombre de valeurs manquantes (en gris dans la figure), mais a tendance à faire varier les scores en sous-estimant la dépendance ou l'atteinte des patients.

Pour les valeurs catégoriques manquantes restantes, nous avons créé la catégorie « Inconnu » car elles représentent une petite partie et n'auront pas d'impact significatif sur la performance du modèle. En ce qui concerne

les variables numériques manquantes, on les remplace pour l'instant par la valeur -1 pour pouvoir les repérer durant l'analyse exploratoire.

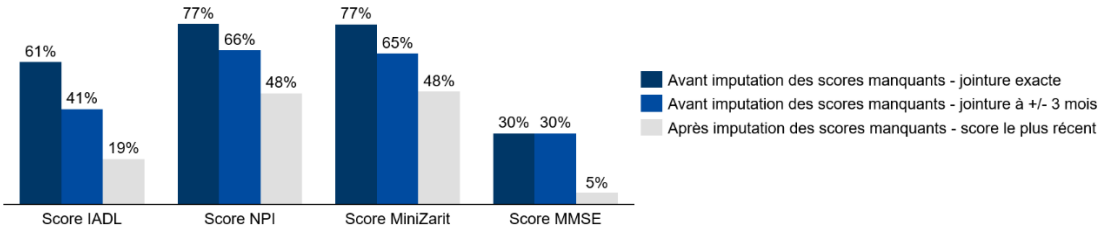


Figure 17 : Répartition des valeurs manquantes avant (bleu foncé) et après imputation des scores manquants avec une jointure à +/- 3 mois (bleu roi) vs au score le plus récent (en gris), pour les variables IADL (gauche), Score NPI (milieu gauche), MiniZarit (milieu droit), MMSE (droite).

Correction de l'IADL. Plusieurs caractéristiques ont dû être ajustées pour être exploitables. L'analyse exploratoire de l'IADL permet de mettre en lumière un biais de genre introduit par la nature du questionnaire, et reconnu dans la littérature (voir ci-après). La Figure 18 ci-dessous montre la répartition du score MMSE pour chaque valeur possible de l'IADL, en marquant par ailleurs la moyenne du MMSE pour les hommes et les femmes (barre verticale en rouge et bleu respectivement). On remarque qu'à score IADL identique, le score MMSE moyen pour les hommes est systématiquement supérieur à celui des femmes.

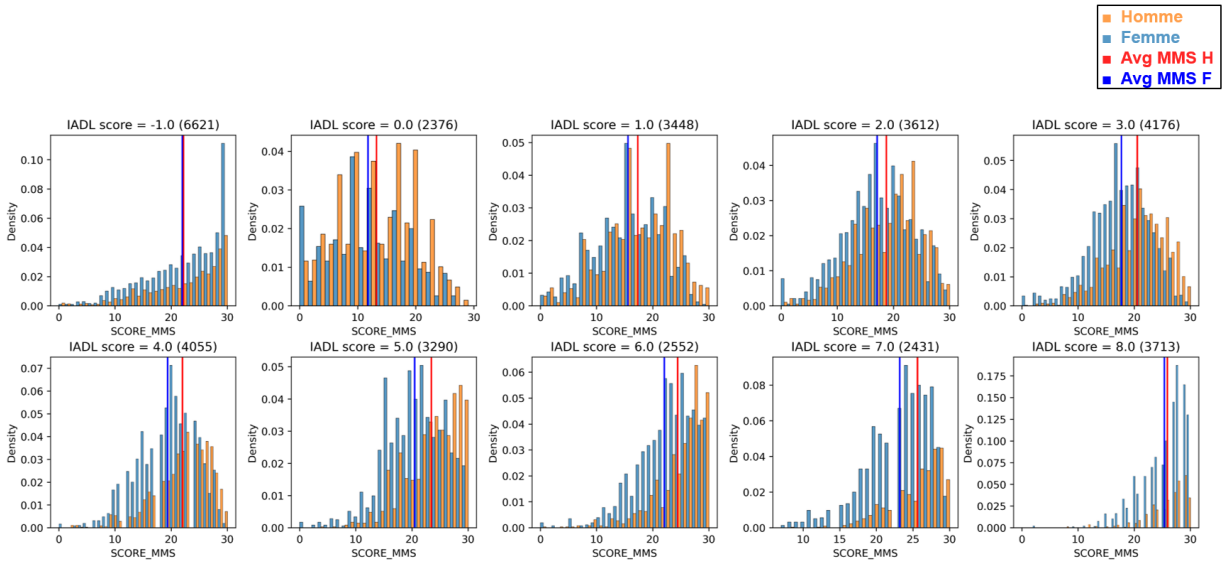


Figure 18 : Histogramme de densité du score MMSE en fonction du genre homme (en orange) et femme (en bleu clair) pour chaque IADL. Le score IADL se lit croissant du coin en haut à gauche vers la droite (de la valeur -1, valeur manquante à 8, score maximal). La moyenne du score MMSE en fonction du genre est également renseignée (en rouge pour les hommes et bleu foncé pour les femmes).

C'est-à-dire que pour un même degré d'autonomie évalué, l'atteinte cognitive est en réalité en moyenne moins sévère pour les hommes que pour les femmes, alors même que les courbes de répartition du MMS sont

de même forme de distribution. Il apparaît donc que l'IADL a tendance à surestimer la dépendance pour les hommes.

L'hypothèse principale expliquant ce biais provient de la formulation des questions évaluant certaines catégories d'activité de la vie quotidienne. A titre d'exemple, pour l'activité « Lessive », les options suivantes sont listées :

- Le patient fait sa lessive entièrement seul (1 point) ;
- Le patient lave ou rince les plus petites pièces (1 point) ;
- La lessive est entièrement réalisée par autrui (0 point).

Dans la mesure où un biais de genre peut venir influencer au sein du foyer la répartition des tâches ménagères, celui-ci est répercuté dans l'évaluation de certains patients de sexe masculin. Ce biais a par ailleurs déjà été identifié et une mesure de correction est proposée dans la littérature[156]. Nous avons corrigé l'IADL en suivant les méthodes ainsi suggérées.

Ajout des comorbidités à partir des classes ATC. Les maladies chroniques ont été reconstruites à partir de la consommation de médicaments de la base de données CPAM. Dans la table TAZ 300, les médicaments sont répertoriés sous leur nom complet et leur classification Anatomique, Thérapeutique et Chimique (ATC)[157]. Il existe dans la littérature des exemples d'implémentations de méthodologies[158] permettant de faire le lien entre les données ATC et les pathologies chroniques correspondantes. En accord avec l'équipe Memora, nous avons sélectionné les comorbidités pertinentes pour les TNC, présentées dans le code ci-dessous.

```
Index(['ACID_RELATED_DISORDERS', 'BONE_DISEASES', 'CANCER',
       'CARDIOVASCULAR_DISEASES', 'DEMENTIA', 'DIABETES_MELLITUS', 'EPILEPSY',
       'GLAUCOMA', 'GOUT_HYPERURICEMIA', 'HIV', 'HYPERLIPIDEMIA',
       'INTESTINAL_INFLAMMATORY_DISEASES', 'IRON_DEFICIENCY_ANEMIA',
       'MIGRAINES', 'PAIN', 'PARKINSON_DISEASE', 'PSYCHOLOGICAL_DISORDE RS',
       'PSYCHOSES', 'RESPIRATORY_ILLNESS', 'RHEUMATOLOGIC_CONDITIONS',
```

[156] Méthode publiée par (Dufournet et al., 2021).
[157] La classification ATC est un système de catégorisation des médicaments et thérapeutiques développé par l'OMS pour faciliter la comparaison internationale des médicaments. Ce système est basé sur la structure chimique, l'usage thérapeutique et les propriétés pharmacologiques de l'élément classé. Le code ATC est un ensemble de 7 chiffres et lettres, qui reflète la position de l'élément dans la hiérarchie des niveaux suivants : premier niveau – groupe anatomique principal (ex : système respiratoire) ; deuxième niveau – classe thérapeutique principale (ex : antibiotique) ; troisième niveau – sous-classe thérapeutique (ex : pénicillines) ; quatrième niveau – voie d'administration (ex : injection) ; cinquième niveau – principe actif. Définition issue du Centre Belge d'Information Pharmacothérapeutique.
[158] Méthode explicitée par (Huber et al., 2013)

```
       'THYROID_DISORDERS', 'TUBERCULOSIS'],
      dtype='object')
```

Par ailleurs, nous introduisons une nouvelle variable à partir de la classe ATC qui permet de simplifier l'intégration des médicaments dans le modèle d'analyse. On conserve pour chaque médicament les premier et deuxième niveaux de sa classe, ce qui permet de garder l'information du groupe anatomique principal visé et de la classe thérapeutique principale.

Recomposition du coût. A partir de la jointure des bases de données Memora et CPAM, nous sommes en mesure de recomposer le coût total de la prise en charge des patients atteints de troubles neurocognitifs (cf. Figure 19 ci-dessous). Au total, 3 950 patients suivis sur 6 572 semestres génèrent 33 M€ de consommations de soins soit environ 8 354 €/patient et 5 021 €/semestre. Les hospitalisations représentent le premier poste de coût, à hauteur de 55% du montant total de la prise en charge soit 19 M€. Les hospitalisations en gériatrie en sont le premier inducteur, suivis de près par les SMR[159] et le service de médecine générale. La Table 7 ci-dessous synthétise le coût moyen par poste d'hospitalisation et le nombre de patients concernés.

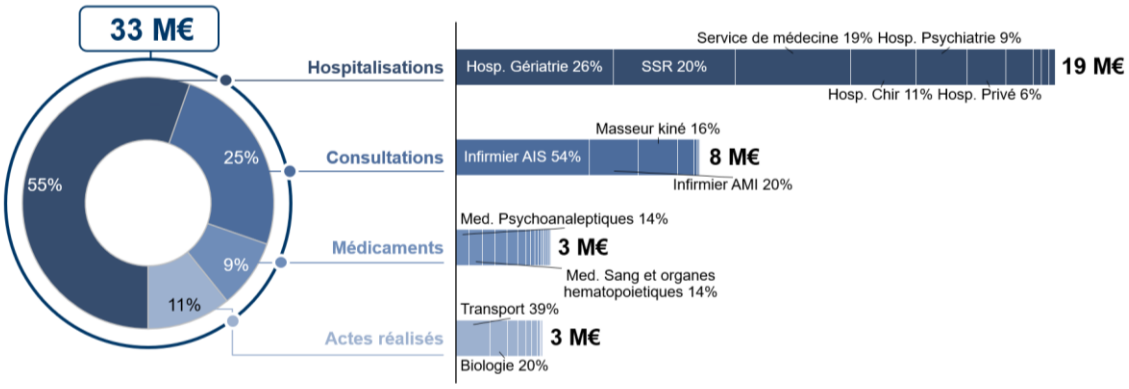


Figure 19 : Répartition des coûts de prise en charge pour la base patients Memora (3 950 patients entre 2014 et 2020).

Type d'hospitalisation	Nombre de patients	Coût total moyen par patient
Dialyse	26	25 663 €
Psychiatrie	70	22 693 €
Médecine interne	13	15 519 €
Cardiologie	17	14 794 €
Gériatrie	342	14 262 €

[159] SMR (Soins Médicaux et de Réadaptation), notés SSR (Soins de Suite et de Réadaptation) dans la figure. En janvier 2022, la dénomination SSR devient SMR suite au décret (CIRCULAIRE N°DHOS/O/2004/44..., 2004).

Neurologie	17	13 893 €
SMR	354	10 624 €
Urgences	119	7 046 €
Chirurgie	397	5 091 €
Département de médecine	1 239	2 881 €

Table 7 : Montant total moyen par patient des consommations, par type d'hospitalisation

La psychiatrie et la dialyse sont les hospitalisations les plus coûteuses, mais concernent un faible volume de patients. A contrario, le département de médecine générale est en moyenne peu coûteux mais concerne un grand volume de patients. Le deuxième poste de coût de la prise en charge est la consultation en libéral, suivi par les médicaments et les divers actes réalisés (biologie, imagerie, etc.). Les deux figures ci-dessous nous donnent un aperçu de l'évolution du coût moyen par semestre. Il apparaît que les premiers semestres de suivi sont en moyenne plus coûteux, une conclusion également retrouvée dans une étude précédemment conduite par Memora[160]. Une hypothèse avancée est que les examens, actes et consultations déployés pour poser le diagnostic lors des visites initiales se cumulent.

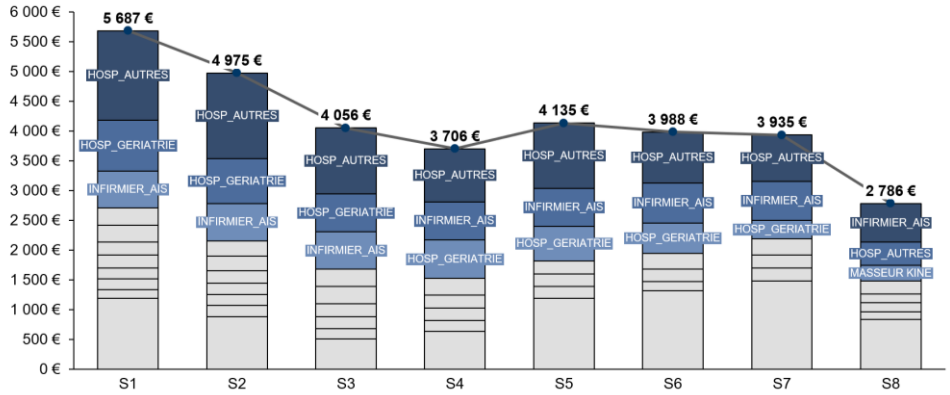


Figure 20 : Evolution du coût moyen par semestre pour les patients de la base Memora

[160] (Dauphinot et al., 2021)

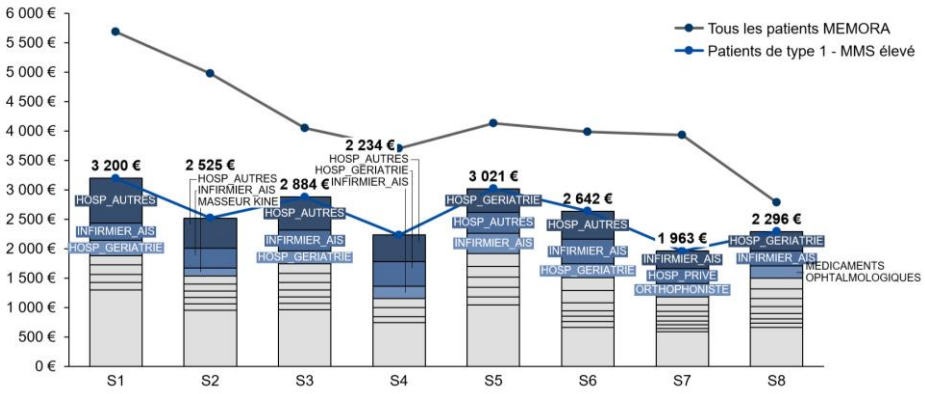


Figure 21 : Evolution du coût par semestre pour les patients ayant un MMSE élevé (supérieur à 23), vs la cohorte.

Par ailleurs, la Figure 20 nous montre également que les consommations en soins infirmiers libéraux et masseurs / kinésithérapeutes augmentent en proportion avec le temps et finissent par devenir le poste de coût prépondérant en fin de parcours (S8). Les deux figures montrent des « effets rebond » entre certains semestres, par exemple du semestre 4 au semestre 5. Il est difficile à ce stade d'en identifier la ou les causes racines. Enfin, on constate dans la Figure 21 que les patients ayant un MMSE plus élevé, c'est-à-dire une atteinte cognitive inférieure, ont des coûts moyens par semestre plus faibles que la moyenne de la cohorte.

Identification des patients avec un poids économique aberrant (*outliers*). Nous choisissons de combiner les méthodes statistiques 3σ et de l'interquartile[161] pour détecter les patients ayant un coût hors norme. 133 patients sont concernés.

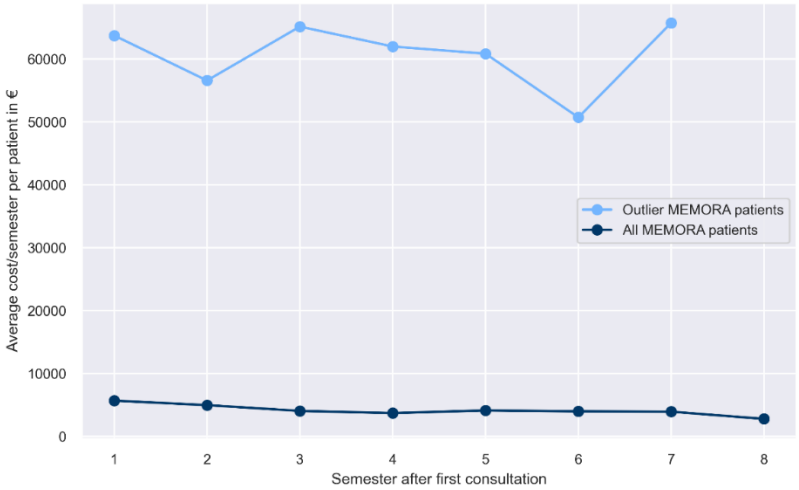


Figure 22 : Evolution moyenne du coût par semestre pour la cohorte générale (en bleu foncé) et les patients outliers (en bleu clair).

[161] La méthode de l'interquartile (IQR) est détaillée dans le chapitre suivant.

Ces patients ont effectivement un coût par semestre variant entre 50 et 60k€ contre 5k€ pour la cohorte générale (cf. Figure 22 ci-dessus). A titre d'exemple, au cours du premier semestre, le montant total des remboursements de la CPAM pour ces patients représente 28,2% du montant de la cohorte générale alors qu'ils ne représentent que 3,3% des effectifs. Nous décidons d'investiguer les actes de soins consommés par ces patients, dont les différences principales avec la cohorte générale sont résumées dans la table ci-dessous.

	Patients « outliers »	Tous les patients MEMORA sans les patients « outliers »
Durée de suivi moyenne	11 mois	20 mois
Score IADL corrigé	3,1	4,6
Retentissement	16,3	12,6
Score NPI	25,6	20,6
Score MMSE	18,1	20,6
Score MiniZarit	4	3,1
No. d'actes de biologie/semestre	33	12
No. d'actes de chirurgie/semestre	3	0,06
No. dialyse/semestre	7	0,3
No. d'acte d'infirmiers (AIS & AMI)	97 & 65	57 & 30
No. séances de masseur kiné/semestre	14	10
No. séances d'ophtalmologie & orthophonie/semestre	0,04 & 1,5	0,11 & 2,6
No. admis en urgence/semestre	2,8	0,05
No. transport/semestre	18	3,6

Table 8 : Comparatif des différences de profils patient et de consommations de soins entre les patients labellisés "outliers" et la cohorte Memora exemptée de ces mêmes patients.

Les patients « *outliers* » ont une durée de suivi moyenne plus courte que le reste de la cohorte (11 vs. 20 mois). Leur score IADL corrigé est également plus bas, ce qui marque une autonomie diminuée. Le score total NPI et son retentissement sont plus élevés, ce qui signifie des changements comportementaux plus sévères et impactant pour le patient, mais aussi son entourage. Même tendance pour le score MMSE qui est plus faible, et pour le MiniZarit qui lui est plus élevé d'une gradation par rapport à la cohorte sans « *outliers* » – ce qui traduit respectivement une atteinte cognitive plus avancée et une fatigue des aidants informels plus forte.

En termes de consommations de soins, ces patients cumulent en moyenne un nombre plus élevé d'actes, notamment pour les dialyses qui sont des procédures couramment coûteuses, mais également un plus grand nombre de consultations libérales type actes infirmiers, et donc un accompagnement à domicile. Par ailleurs, ces patients consultent 50 fois

plus aux services d'urgences et sont opérés 50 fois plus régulièrement par semestre que le reste de la cohorte. En dernier lieu, ces patients consomment plus de transport médicalisé. En lumière de ces statistiques, nous décidons de ne pas sortir ces patients de l'analyse. Il n'est pas exclu qu'ils soient classifiés dans un groupe médico-économique commun par l'algorithme.

Ensemble final de données. La corrélation entre les variables est testée au moyen du coefficient de Pearson[162]. Les scores du NPI et le MiniZarit sont les seules variables ressortant comme modérément corrélées (coefficient de Pearson égal à 0,67). Après discussion avec l'équipe médicale de Mémora, nous décidons de conserver les deux variables pour que l'interprétation des résultats reste fluide. Nous testons également la multi colinéarité via le Variance Influence Factor (VIF)[163], un outil statistique répandu qui permet de détecter si deux ou plusieurs variables indépendantes sont fortement corrélées les unes entre les autres. Le VIF d'une variable i est mesuré via l'équation suivante :

Équation i : Mesure du Variance Influence Factor d'une variable i

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

où ${R_i}^2$ représente le coefficient de détermination obtenu lors d'une régression de la variable i sur les autres variables indépendantes du modèle. Un VIF élevé traduit une forte corrélation et on considère habituellement qu'un VIF inférieur à 5 est acceptable. Les VIF varient entre 0 et 2,5 pour chacune de nos variables, ce qui signifie que nous les conserverons toutes pour l'analyse à suivre. L'ensemble de données final contient donc 3 950 patients et 6 572 semestres décrits par 182 variables, dont 70 sont des caractéristiques économiques (cf. figure ci-dessous).

[162] Le coefficient de Pearson et l'évaluation de la corrélation sont détaillés dans le chapitre suivant.
[163] Le concept de multicollinéarité et l'application du VIF pour la détecter sont présentés dans (Hair et al., 2009).

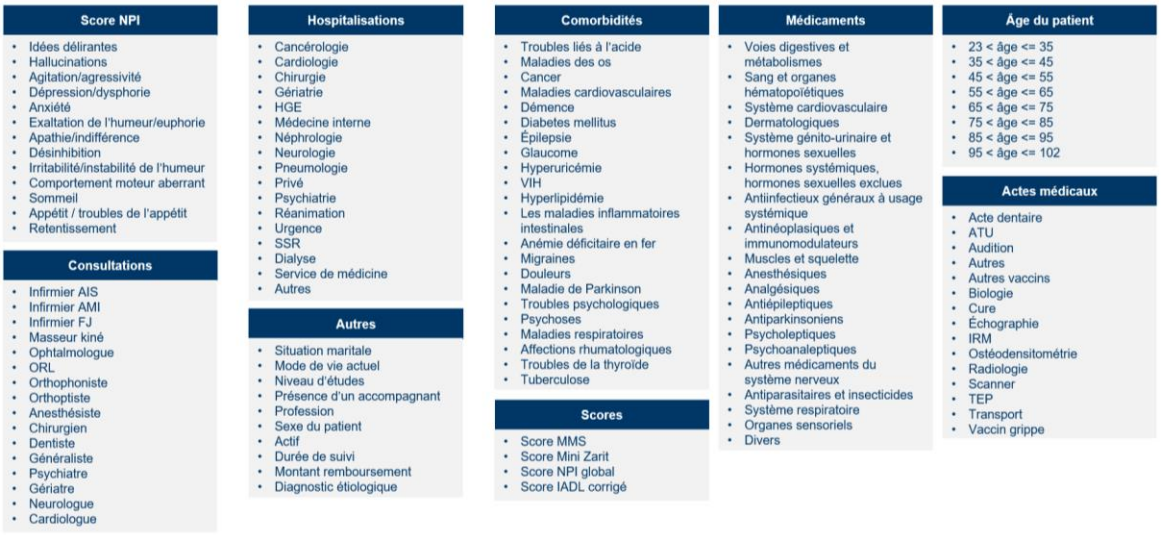


Figure 23 : Cartographie et classification des variables retenues pour l'analyse.

3.2.3 Description de l'approche deployee

Cette section présente l'approche globale déployée, où, après avoir réduit la dimension de notre jeu de données, nous avons regroupé les semestres dans le temps en fonction des coûts, puis nous avons introduit les groupes dans une chaîne de Markov afin d'estimer la probabilité de passer d'un groupe à l'autre à la période suivante. Nous avons ensuite développé une méthode d'interprétation sur mesure pour extraire les variables cliniques associées à un changement de groupe.

Réduction de la dimensionnalité. La haute dimensionnalité de notre ensemble de données est susceptible d'entraîner une forte diminution des performances pour de nombreux algorithmes, en particulier pour les tâches de classification non supervisée, qui nous intéressent ici. En effet, le volume de l'espace vectoriel de représentation augmente très rapidement avec le nombre de dimensions. Une dimensionnalité élevée conduira à un espace avec des points de données épars, à moins qu'il n'y ait une grande quantité de données disponibles. Dans ces espaces clairsemés, les points apparaissent différents les uns des autres, ce qui peut amener de mauvais résultats lorsqu'on essaie de regrouper ces données.

Pour contourner ce problème, plusieurs techniques de réduction des dimensions peuvent être envisagées, pour projeter l'espace à haute dimension dans un espace à dimension réduite. Nous avons choisi la

méthode *t-distributed stochastic neighbor embedding* (t-SNE)[164]. Il s'agit d'une technique de réduction de dimension non linéaire qui définit deux distributions de probabilité : l'une dans l'espace original et l'autre dans l'espace de faible dimension. Ces distributions sont définies de manière à ce que les points proches les uns des autres aient une forte probabilité d'être sélectionnés pour faire partie du même groupe. L'algorithme minimise ensuite la dissimilarité entre les distributions de probabilité[165]. Le résultat est un espace à deux dimensions, qui permet de représenter graphiquement l'ensemble des données.

Classification non supervisée des patients en groupes médico-économiques. La classification non supervisée, couramment appelée *clustering* dans le domaine de l'apprentissage automatique, est la tâche de grouper un ensemble de points de données où la similarité entre les points est mesurée à l'aide d'une métrique de distance dans un espace donné. Nous avons testé quatre algorithmes, qui présentent tous des avantages et des inconvénients en termes de paramètres d'entrée et d'efficacité (cf. Table 9 ci-dessous).

Méthode	Algorithme	Description	Avantages	Limites
Clustering basé sur les centroïdes	K-moyennes ou *K-means*	Les points sont assignés à leur centroïde le plus proche, défini comme le point moyen au sein de la grappe.	Simple, rapide et peu intense en calcul.	Requiert le nombre de *clusters* en entrée. Difficulté à détecter les *outliers*. Tous les points doivent être assignés à un cluster.
Clustering basé sur les centroïdes	K-médoïdes ou *K-medoids*	Les points sont assignés à leur centroïde le plus proche, défini comme le point le plus central de la grappe (médoïde).	Les centroïdes peuvent être interprétés (points réels de l'ensemble de données). Plus robuste aux outliers que K-means.	Requiert le nombre de *clusters* en entrée.
Clustering hiérarchique	BIRCH[166]	Construit un arbre de décision où chaque nœud contient de l'information sur le sous-*cluster* de ses	Un seul balayage du jeu de données est requis, ce qui le rend rapide. L'arbre augmente	Ne peut traiter que des attributs métriques. Difficulté à détecter les valeurs aberrantes. Tous les

[164] (Maaten & Hinton, 2008).
[165] (Kullback & Leibler, 1951).
[166] Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) ou, en français, Réduction Itérative Equilibrée et Classification non supervisée à l'aide de hiérarchies. Algorithme publié par (Zhang et al., 1997).

		branches dans l'arbre. Un regroupement agglomératif est ensuite appliqué à l'arbre.	la qualité des *clusters*.	points doivent être assignés à un *cluster*.
Clustering hiérarchique basé sur la densité	HDBSCAN[167]	Transforme l'espace en fonction de la densité et construit un arbre minimal qui est finalement converti en une hiérarchie de composants.	Ne nécessite pas le nombre de clusters comme paramètre d'entrée. Robuste face aux outliers, avec la capacité de les labelliser comme tels.	Performances diminuées en haute dimension. Plusieurs paramètres doivent être optimisés, sans métrique spécifique sur laquelle s'appuyer.

Table 9 : Aperçu des méthodes de classification non supervisée utilisées dans cette étude.

Optimisation des hyperparamètres et mesure de la performance de la tâche de classification non supervisée. Pour optimiser les hyperparamètres et sélectionner l'algorithme le plus efficace, nous nous sommes appuyés sur trois méthodes.

Sélection a priori du nombre de clusters cible.

Les méthodes de regroupement basées sur les centroïdes telles que K-means et K-médoïds requièrent de renseigner en entrée le nombre de clusters cible de l'algorithme. Pour optimiser cet hyperparamètre, deux méthodes sont couramment combinées[168] :

1. La méthode de l'indicateur WCSS (Within Cluster Sum of Squares). Cette méthode a pour objectif de trouver le nombre de clusters qui minimise la somme des carrées des distances entre chaque point de données et le centre de son cluster. Pour chaque valeur du nombre de clusters k, on calcule cette valeur.

2. La méthode du coude. On sélectionne ensuite la valeur de k à partir de laquelle on a un point d'inflexion, et l'ajout de clusters supplémentaires ne réduit plus de manière significative la valeur du WCSS.

Optimisation des hyperparamètres de BIRCH – le score de silhouette.

[167] Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) ou, en français, Classification non supervisée spatiale hiérarchique des applications avec bruit basée sur la densité. Algorithme publié par (Campello et al., 2013).
[168] (Umargono et al., 2019)

Le score de silhouette a été utilisé pour sélectionner le meilleur nombre de grappes pour l'algorithme BIRCH, l'objectif étant de maximiser ce score. Le score de silhouette est une mesure du degré de similitude d'un point avec son propre groupe par rapport à d'autres groupes (cohésion au sein des groupes par rapport à la séparation entre les groupes)[169].

Équation ii : Calcul du score de silhouette d'un point i donné.

$$S(i) = \frac{b_i - a_i}{max\{a_i\ , b_i\}}$$

Où a_i est la distance moyenne entre le point i et tous les autres points appartenant au même cluster (cohésion) et b_i est la distance moyenne entre le point i et tous les points dans le cluster le plus proche (séparation). La silhouette finale est la moyenne des scores de silhouette individuels.

Elle va de -1 (pire classification, les points sont dissimilaires au sein d'un même cluster et similaires entre clusters) à +1 (meilleure classification, les points sont ressemblants au sein d'un même cluster et différents entres clusters). De manière générale, un score de silhouette élevé indique une segmentation homogène et donc une meilleure qualité de la classification. A l'inverse, un score proche de zéro peut par exemple indiquer que certains clusters se chevauchent. Cette mesure a également fait partie de l'indicateur de sélection des algorithmes en termes de performances finales.

Indicateur de performance de la classification.

Nous avons choisi l'algorithme qui minimise l'évolution moyenne du coût par semestre pour la transition intra-cluster. Il reflète l'objectif global de notre étape de regroupement, c'est-à-dire qu'un patient qui reste dans le même groupe doit mettre en évidence qu'il n'y a pas de changement pertinent dans la consommation de soins pour ce patient.

Modélisation des probabilités de transition. L'étape suivante consiste à modéliser le parcours du patient comme une succession de clusters et à estimer les probabilités de transitions entre ces clusters d'un semestre à l'autre. Pour modéliser cette séquence d'événements, nous choisissons d'utiliser une chaîne de Markov.

[169] (Rousseeuw, 1987).

Une chaîne de Markov est un modèle mathématique décrivant une séquence d'événements possibles où la probabilité de chaque événement à venir ne dépend que de l'état actuel de la chaîne. Elle repose sur la propriété de Markov[170] :

$$Soit\ X = \{X_k\ ;k\ \geq 0\}\ une\ suite\ d'états\ à\ valeurs\ dans\ un\ espace\ fini\ E.$$

$$On\ dit\ que\ X\ est\ une\ chaîne\ de\ Markov\ si\ pour\ tout\ x_i \in E :$$

$$P(X_i = x_i\ |X_0 = x_0, X_1 = x_1, \ldots X_{i-1} = x_{i-1}) = P(X_i = x_i\ |X_{i-1} = x_{i-1})$$

Dans notre application, un état est représenté par l'appartenance à un cluster. Au cours du semestre suivant, un patient donné est susceptible de changer d'état, c'est-à-dire d'effectuer une transition vers un cluster qui représente un groupe médico-économique différent.

Chaque état est caractérisé par une matrice de transition, qui regroupe les probabilités de passer de cet état à un autre, et ce pour tous les états représentés, cf. équation ci-dessous.

Équation iii : Matrice de transition calculée au moment i pour toute transition de l'état x à y.

$$Q(x, y) = P(X_{i+1} = y | X_i = x)$$

Cette matrice de transition est apprise à partir d'un jeu de données d'entraînement et peut être par la suite utilisée pour prédire l'état futur d'un nouveau patient à partir de son état au semestre actuel. En d'autres termes, grâce à cette matrice, nous sommes en mesure de prédire le cluster au semestre *i* sachant que le patient est dans le cluster *k* au semestre *i-1*. La chaîne de Markov finale est alors un graphe représentant les transitions possibles entre clusters. Ce graphe évolue de semestre en semestre.

Interprétation des clusters. La dernière étape de la méthode vise à caractériser les spécificités de chaque cluster et à déterminer les facteurs de transition entre les clusters d'un semestre à l'autre. Cette étape peut être complexe et nécessite une couche algorithmique supplémentaire en raison de la réduction de la dimensionnalité réalisée en premier lieu. Les 182 variables du modèle ont été réduites à deux dimensions, on ne peut donc plus appliquer de méthode d'interprétation locale.

[170] (Lachieze-Rey, 2021)

Dans notre cas d'application, nous avons choisi de combiner quatre méthodes d'interprétation. Ces méthodes ont été sélectionnées à partir de la littérature et adaptées en fonction de notre approche :

1. Entraînement d'un arbre de décision pour prédire l'étiquette du groupe pour chaque point de données, en tenant compte des données sociodémographiques, cliniques et économiques disponibles. Il permet d'identifier les caractéristiques les plus significatives qui segmentent les patients dans un cluster spécifique.

2. Analyse de la distribution du coût par semestre et de la ventilation du coût moyen pour chaque groupe. Nous différencions les groupes sur la base de la répartition des coûts et identifions les principaux facteurs de coût pour ces patients.

3. Entrainement d'un arbre de décision par cluster pour prédire le coût par semestre[171]. L'objectif est de souligner, au sein de chaque cluster, les principaux inducteurs de coûts pour ces semestres, en particulier de mettre en évidence les différences entre les clusters.

4. Analyse de l'association de variables pour chaque cluster[172], pour identifier les caractéristiques communes partagées par les patients dans leur cluster.

Les résultats de cette méthode d'interprétation seront présentés dans la section Discussion de ce chapitre.

[171] Comme implémenté dans (Yan et al., 2019)
[172] Une méthode également appliquée dans (Hajat et al., 2021).

	Description	Objectif
1. Arbre de décisions global	Déterminer les variables les plus importantes dans la prédiction cluster de chaque consultation en utilisant le poids relatif des variables dans un arbre de décisions	Expliquer les facteurs principaux permettant de séparer les patients dans les différents clusters
2. Distribution du coût/semestre par cluster	Calculer les statistiques descriptives du coût/semestre pour chaque cluster ainsi que la répartition de ces coûts selon leur type (hospitalisation, consultation, etc.)	Différencier les clusters sur la base des coûts ainsi qu'identifier les déterminants principaux des coûts de ces patients
3. Arbre de décisions pour prédire le coût/semestre par cluster	Déterminer les variables les plus importantes dans la prédiction du coût/semestre par cluster en utilisant le poids relatif des variables dans un arbre de décisions	Identifier, au sein de chaque cluster, les déterminants principaux des coûts de ces patients, notamment pour faire ressortir les différences entre clusters
4. Associations de variables par cluster	Mesurer le pourcentage de patients qui partagent chacune des variables	Identifier les caractéristiques communes à la majorité des patients présents dans le cluster

Figure 24 : Description et objectif des étapes de la méthodologie d'interprétation déployée.

Notes sur l'implémentation. Toutes les étapes de pré-traitement et d'analyses des données ont été réalisées avec Python version 3.9.7. La préparation des données a été gérée avec les bibliothèques Numpy et Pandas, tandis que nous avons utilisé le package scikit-learn pour les étapes de réduction de dimension et de clustering. Nous avons utilisé le package spécifique hdbscan pour le clustering hiérarchique basé sur la densité et la bibliothèque pomegranate pour la chaîne de Markov. Les parcours des patients ont été représentés par des diagrammes de Sankey, générés avec sankeymatic.com.

3.3 Résultats

SYNOPSIS Où nous décrivons les performances obtenues par les quatre modèles de classification non supervisée, puis la chaîne de Markov.

3.3.1 PERFORMANCE DE LA TACHE DE CLASSIFICATION NON SUPERVISEE

L'évolution la plus faible du coût par semestre lors des transitions intra-cluster a été obtenue avec l'algorithme BIRCH (cf. Table 11 ci-dessous). Bien que les algorithmes K-Means et K-Medoids ont tous deux un score de silhouette plus performant, comme indiqué dans la Table 10, leur coût moyen de transition intra-cluster est significativement plus élevé (cf. tables ci-dessous). Nous avons poursuivi avec l'algorithme BIRCH avec un hyperparamètre de cinq clusters.

Algorithme	Méthode de sélection du nombre optimal de clusters	Nombre optimal de clusters	Silhouette Score
K-Means	Méthode WCSS et du coude	5 clusters	0.43504
K-Medoids	Méthode WCSS et du coude	5 clusters	0.43622
HDBSCAN	Le nombre de clusters n'est pas un hyperparamètre	7 clusters	-0.00698
BIRCH	Maximiser la silhouette	5 clusters	0.39037

Table 10 : Nombre optimal de clusters obtenus et score de performance (silhouette) pour chaque algorithme testé.

Transition intra-cluster	Evolution moyenne du coût/semestre	Nombre de transitions concernées
1 → 1	- 866 €	106
2 → 2	+ 24 €	750
3 → 3	- 513 €	54
4 → 4	- 504 €	225
5 → 5	- 854 €	183

Table 11 : Évolution moyenne du coût par semestre pour les transitions à l'intérieur des grappes obtenues avec l'algorithme BIRCH. La transition du cluster 1 au cluster 1 est abrégée 1 → 1.

Transition intra-cluster	Evolution moyenne du coût/semestre	Nombre de transitions concernées
1 → 1	- 902 €	5
2 → 2	- 184 €	13
3 → 3	- 8554 €	3
4 → 4	+ 688 €	7
5 → 5	- 7 €	1247
6 → 6	- 7681 €	2
7 → 7	- 60 €	233

Table 12 : Évolution moyenne du coût par semestre pour les transitions à l'intérieur des grappes obtenues avec l'algorithme HDBSCAN. La transition du cluster 1 au cluster 1 est abrégée 1 → 1.

Transition intra-cluster	Evolution moyenne du coût/semestre	Nombre de transitions concernées
1 → 1	- 293 €	257
2 → 2	+ 236 €	832
3 → 3	- 1649 €	35
4 → 4	- 982 €	110
5 → 5	- 1850 €	117

Table 13 : Évolution moyenne du coût par semestre pour les transitions à l'intérieur des grappes obtenues avec l'algorithme K-means. La transition du cluster 1 au cluster 1 est abrégée 1 → 1.

Transition intra-cluster	Evolution moyenne du coût/semestre	Nombre de transitions concernées
1 → 1	- 292 €	258
2 → 2	- 1075 €	114
3 → 3	- 1500 €	33
4 → 4	124 €	822
5 → 5	- 873 €	129

Table 14 : Évolution moyenne du coût par semestre pour les transitions à l'intérieur des grappes obtenues avec l'algorithme K-medoïds. La transition du cluster 1 au cluster 1 est abrégée 1 → 1.

Les représentations visuelles des clusters obtenus avec chaque méthode sont présentées dans la Figure 25 ci-dessous. Les points de données sont projetés en deux dimensions en utilisant les coordonnées obtenues à partir de la projection t-SNE (graphique a) et les couleurs représentent le cluster attribué à chaque point avec chaque algorithme (graphiques b à e).

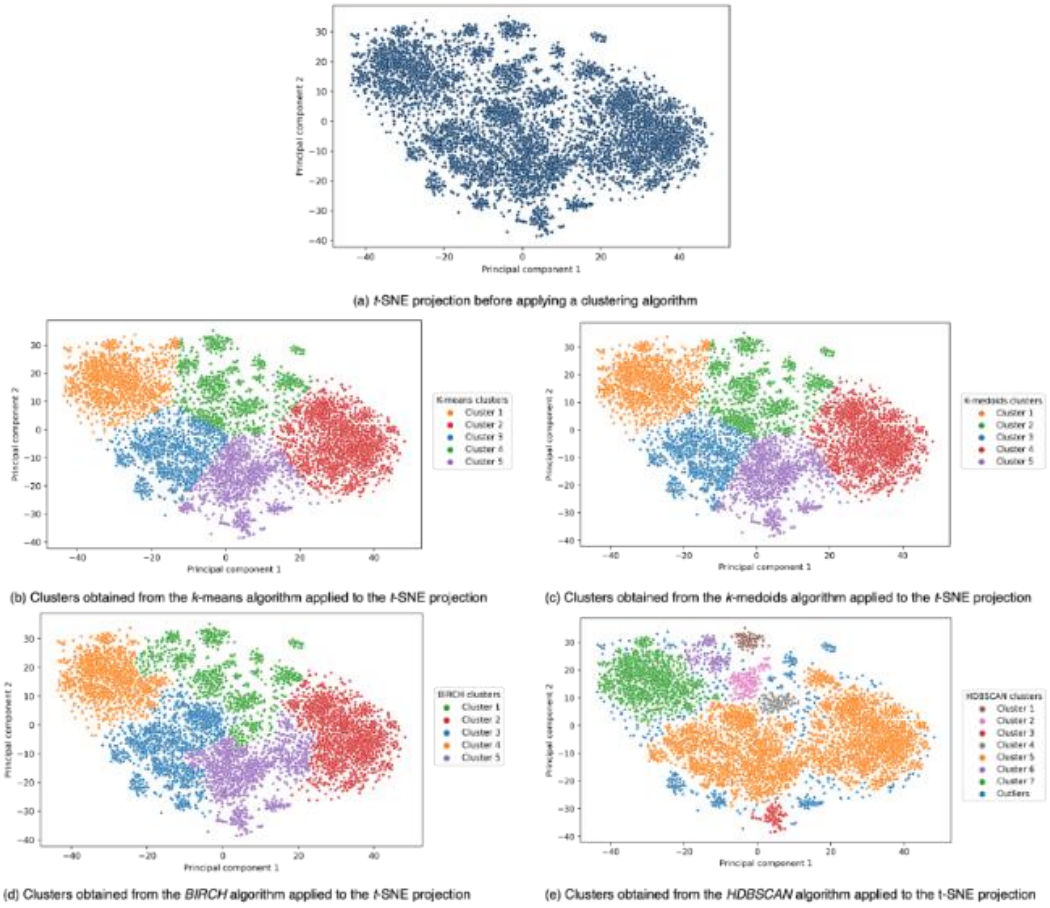


Figure 25 : (a) Projection bidimensionnelle de l'ensemble de données après entrainement de l'algorithme t-SNE. Représentation des clusters répartis par les algorithmes de classification non supervisée (b) K-means (c) K-medoids, (d) BIRCH et (e) HDBSCAN. Chaque point correspond à un semestre du parcours.

Les clusters formés par les algorithmes K-means et K-médoïds sont relativement similaires lorsque projetés en deux dimensions (cf. graphique (b) et (c) ci-dessus). La comparaison des évolutions moyennes de coût pour les transitions intra-cluster entre ces deux algorithmes, présentées dans les tables ci-dessus), montrent une différence dans la répartition des points entre les clusters : les clusters 1 et 3 semblent identiques entre les deux modèles, ce qui n'est pas le cas du cluster 2 par exemple. Ces différences peuvent être dues à la répartition des outliers entre les clusters. L'algorithme K-means, qui repose sur le calcul d'un point « moyen », est plus sensible aux valeurs aberrantes que K-médoïds.

En ce qui concerne HDBSCAN, on remarque dans la représentation en deux dimensions, que le cluster 5 capte une grande majorité des observations (cf. graphique (e) ci-dessus, en orange). Ce déséquilibre explique probablement le score négatif de silhouette, qui signe une dissimilarité entre points de données d'un même cluster.

3.3.2 Performance de la chaine de Markov

La chaîne de Markov finale est représentée ci-dessous dans la Figure 26. Le graphe dirigé est composé de 5 états pour les 5 clusters obtenus en sortie de l'algorithme BIRCH, retenu à la section précédente. Les flèches sont des arcs orientés qui illustrent les transitions entre chaque état et sont chacun associé à une probabilité.

Cette représentation nous permet de constater qu'il existe :

- Des clusters de transition, où la probabilité de départ est très élevée (à titre d'exemple dans la figure ci-dessous, le cluster 3 où la probabilité de transition[173] est de 0,81) ;
- Des clusters d'absorption, où la probabilité de départ est très faible (par exemple, le cluster 2 où la probabilité de transition est de 0,04).

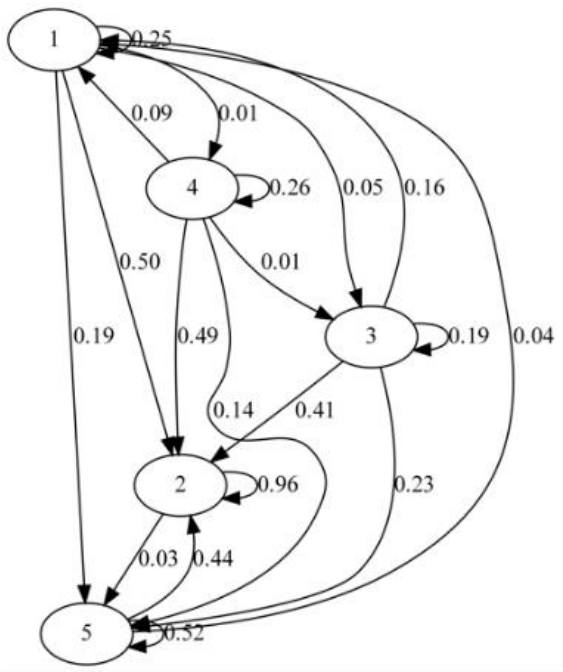


Figure 26 : Représentation graphique de la chaîne de Markov entraînée à partir des cinq clusters classifiés par BIRCH. Les transitions d'un cluster à un autre sont représentées si la probabilité de transition p ≥ 0,01.

[173] Pour un cluster donné, la probabilité de transition hors de ce cluster est la somme des probabilités associées aux arcs ayant pour origine le cluster et pour destination un cluster différent.

La chaîne de Markov a été construite en utilisant 75 % de l'ensemble de données (ensemble d'entraînement), et les 25 % restants ont été utilisés pour tester les performances de prédiction du modèle (ensemble de test). Nous précisons que la validation croisée n'est généralement pas recommandée lorsque les données sont susceptibles d'être liées temporellement entre elles, ce qui est notre cas[174]. Les mesures de performance sur les ensembles d'entraînement et de test sont représentées dans la Table 15 ci-dessous. Ce tableau montre une bonne précision du modèle pour prédire le prochain groupe d'un patient connaissant le cluster actuel, avec une exactitude (*accuracy*) et une précision sur l'ensemble de test de respectivement 85,25% et 90,95%.

Métrique[175]	Performance d'entraînement	Performance de test
Exactitude	85.64%	85.25%
F1-score	86.61%	86.41%
Précision	91.81%	90.95%
Rappel	84.36%	85.08%

Table 15 : Indicateurs de performance pour la chaîne de Markov sur le jeu d'entraînement et de test.

[174] La validation croisée repose sur l'hypothèse que toutes les observations sont indépendantes entre elles. Avec des séries temporelles, cette hypothèse n'est pas vérifiée car les observations dans un horizon temporel proche sont susceptibles d'être corrélées. (Bergmeir et al., 2018) démontrent que la validation croisée standard peut donner des résultats trompeurs lorsqu'elle est utilisée pour évaluer la performance de modèles basés sur des séries temporelles.

[175] Ces quatre métriques et leurs modes de calcul sont présentés plus en détail dans le chapitre suivant.

3.4 Discussion et perspectives

SYNOPSIS Où nous interprétons les résultats décrits dans la section précédente et notamment ceux de la classification non supervisée et des parcours patients obtenus via la chaîne de Markov.

3.4.1 INTERPRETATION DES CLUSTERS

Grâce à la méthode d'interprétation décrite dans la section précédente de présentation de l'approche déployée, nous avons pu regrouper les clusters en trois catégories (voir la Figure 27 ci-après) :

- Groupe A – patients dont les coûts ne sont pas liés à leurs troubles cognitifs (clusters 1 et 4). Le cluster 1 rassemble généralement des patients présentant plusieurs comorbidités qui nécessitent des soins (par exemple, la dialyse) et des services (par exemple, le transport) coûteux, tandis que le groupe 4 semble être le groupe "par défaut" dans lequel le reste des patients est classé. Le coût moyen des soins dans le groupe 1 est 18 % plus élevé que la moyenne de la cohorte.
- Groupe B – Patients dont les symptômes de mémoire sont généralement liés à des troubles psychiatriques (cluster 3). Ces patients ont des coûts d'hospitalisation psychiatrique très élevés (plus de 40 % des coûts psychiatriques totaux de la cohorte sont concentrés sur ce cluster). Les troubles psychiatriques diagnostiqués sont également les plus fréquents dans ce groupe, avec une forte prévalence de la dépression (30 % des patients de la cohorte totale sont dans le cluster 3).
- Groupe C – Patients dont les coûts sont liés à leurs troubles cognitifs (clusters 2 et 5) : décrits plus en détail ci-après.

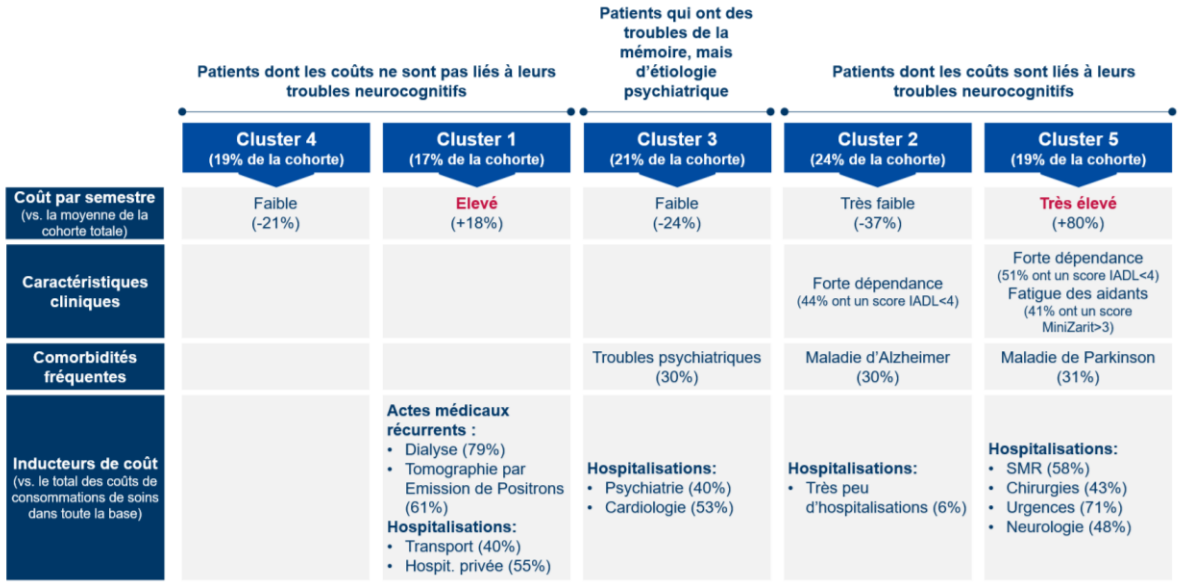


Figure 27 : Vue d'ensemble des principales caractéristiques de chaque cluster obtenu. Exemple de lecture des comorbidités les plus fréquentes : 30% des patients de la cohorte totale et diagnostiqués avec des troubles psychiatriques sont dans le cluster 3. Exemple de lecture des inducteurs de coût : 40% des coûts totaux de psychiatrie dans la cohorte sont consommés par les patients du groupe 3.

Le cluster 2 présente des patients dont les coûts sont inférieurs à ceux du reste de la cohorte (en moyenne -37%), mais dont le score IADL est plus faible (les patients ont moins d'autonomie dans les activités de la vie quotidienne). Leur profil de consommation de soins montre plus de soins infirmiers et libéraux que les autres groupes. Le diagnostic étiologique le plus représenté est la maladie d'Alzheimer, avec 30 % du total des patients diagnostiqués de la cohorte regroupés dans ce cluster.

Les patients du cluster 5 présentent les coûts les plus élevés de la cohorte (+80 % en moyenne), ainsi qu'une faible autonomie pour les tâches de la vie quotidienne. Les aidants informels de ce groupe ont tendance à afficher un score d'épuisement élevé. La maladie de Parkinson est par ailleurs plus fréquente dans le cluster 5. Le coût des soins est fortement lié aux consultations aux urgences (71 % du coût total de la cohorte des urgences), aux séjours en SMR (58% du coût total de la cohorte en SMR) aux interventions chirurgicales (43 % du coût total de la cohorte des interventions chirurgicales) et aux hospitalisations en service de neurologie (48% du coût de la cohorte). En complément, les variables « Hallucinations » du score NPI et « Vie à domicile dans sa famille » apparaissent comme prépondérantes dans la prédiction du coût au sein de ce cluster.

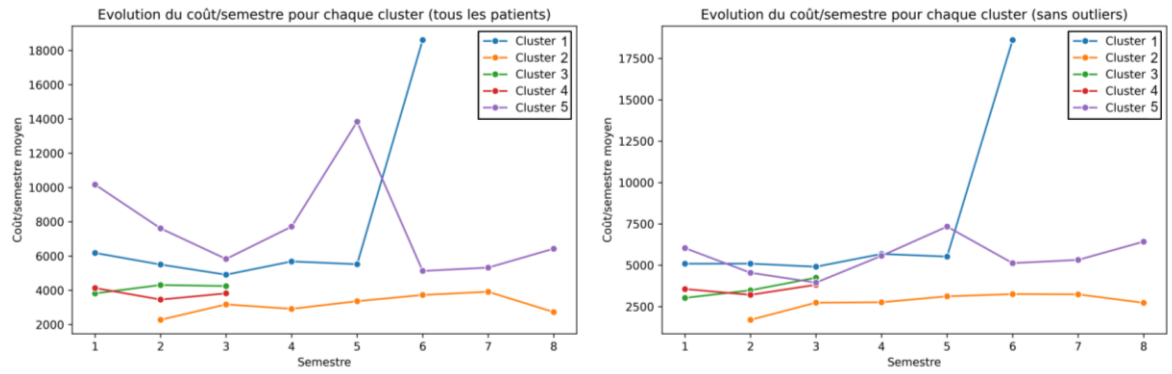


Figure 28 : Evolution du coût par semestre moyen pour chaque cluster, pour la cohorte générale (graphique de gauche) et pour la cohorte sans les patients outliers (graphique de droite).

On remarque que la plupart des patients outliers ont bien été classés dans le cluster 5 (figure ci-dessus, différence de courbe entre graphique de gauche et de droite).

Par ailleurs, les patients du cluster 3 et du cluster 4 ont des durées de suivi inférieures à 3 semestres (cf. Figure 28 ci-dessus). Cela peut être dû au fait que les patients changent de clusters, ou bien entrent en institutions, telles que l'EHPAD, changent de région et ne sont plus suivis par les CPAM d'Auvergne Rhône-Alpes, ou décèdent. A l'inverse les patients ne démarrent jamais leur suivi dans le cluster 2 (cf. Figure 29 ci-dessous), mais vont y rentrer progressivement à partir du 2ème et surtout 3ème semestres.

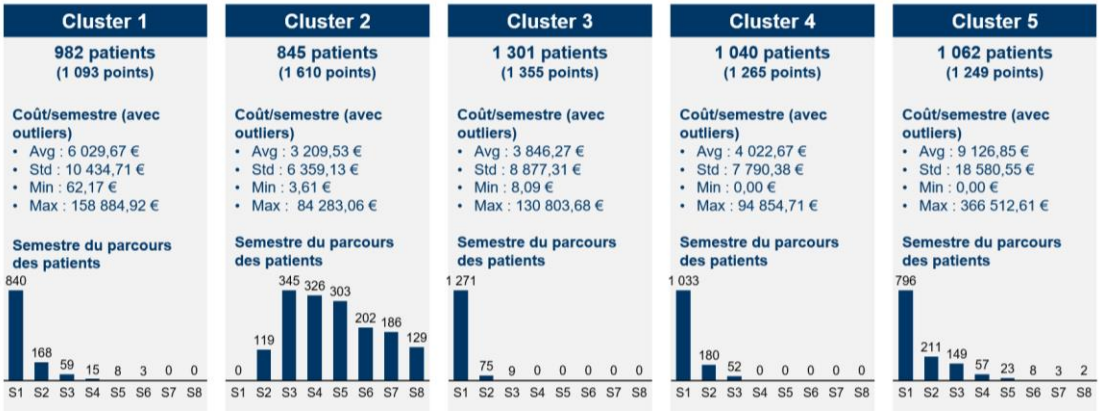


Figure 29 : Statistiques descriptives et nombre de semestres représentés en fonction du temps.

Ces groupes démontrent des tendances communes pour les patients y appartenant et particulièrement en termes de consommations de soins. Cependant, les parcours, c'est-à-dire l'enchaînement des clusters dans le temps, sont également représentatifs pour rassembler des patients et dégager des trajectoires communes.

3.4.2 INTERPRETATION DES PARCOURS PATIENT

Nous avons choisi de représenter les parcours à l'aide de diagrammes de Sankey[176] (voir la Figure 30 ci-dessous), car ils sont un excellent outil visuel pour représenter la proportion de patients qui empruntent une trajectoire donnée. Dans notre cas d'application, les trajectoires correspondent à la transposition en une dimension de notre chaîne de Markov. La taille du flux entre deux clusters est ainsi la probabilité estimée dans la matrice de transition de la chaîne pour ces mêmes états. Pour chaque transition, on renseigne les évolutions des caractéristiques cliniques notamment le score MMSE, l'IADL, le MiniZarit, le NPI et l'évolution du coût. Les événements notables, type décès ou entrée en EHPAD, sont également renseignés lorsque prépondérants.

Notre ensemble de données contient 403 trajectoires différentes. Les 18 trajectoires les plus fréquentes couvrent plus de 80 % des parcours des patients. Elles ont été examinées avec l'équipe multidisciplinaire Memora (épidémiologie, neurosciences, pharmacie, gériatrie) afin de tenter de les relier à la pratique médicale.

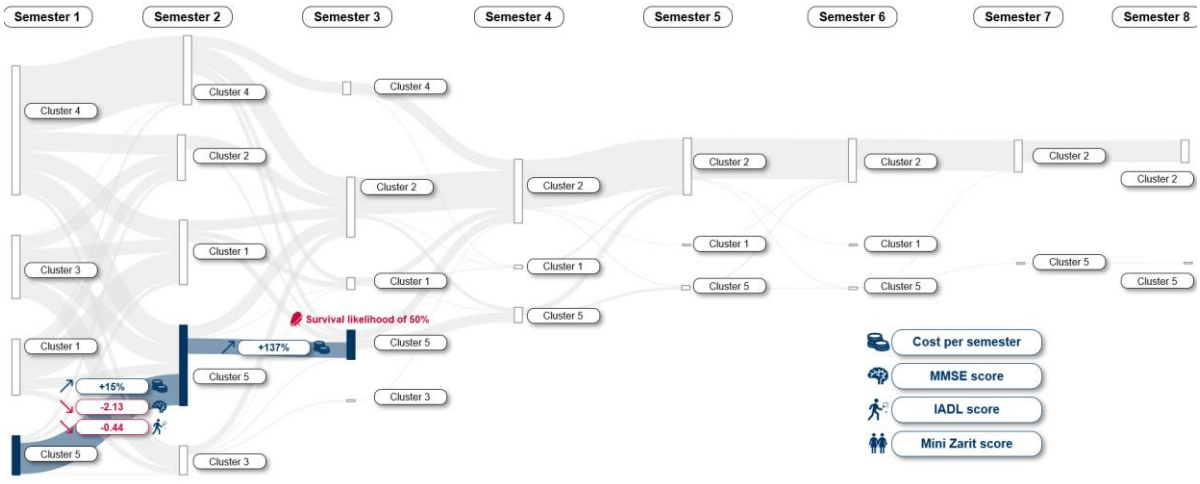


Figure 30 : Exemple de parcours cluster 5 → 5 → 5 pour des patients restant en suivi pendant trois semestres.

[176] Un diagramme de Sankey est un type de visualisation graphique de données. Il représente des flux entre des entités, et est couramment utilisé pour modéliser des transferts en économie, énergie, et gestion des processus. Ce diagramme utilise des courbes de largeurs différentes pour représenter la taille relative des flux et permet d'en suivre le parcours. Les entités sont, elles, représentées sous forme de nœuds.

La Figure 30 ci-dessus est un exemple de parcours commun pour les patients du cluster 5 suivis pendant trois semestres. Après un semestre seulement, leurs scores MMSE et IADL diminueront respectivement de 2,13 et 0,44 points (déclin cognitif plus important, perte d'autonomie). Après un an, leurs coûts de soins augmenteront de 137%, principalement en raison des hospitalisations d'urgence. Un semestre plus tard, leur probabilité de survie n'est plus que de 50 %.

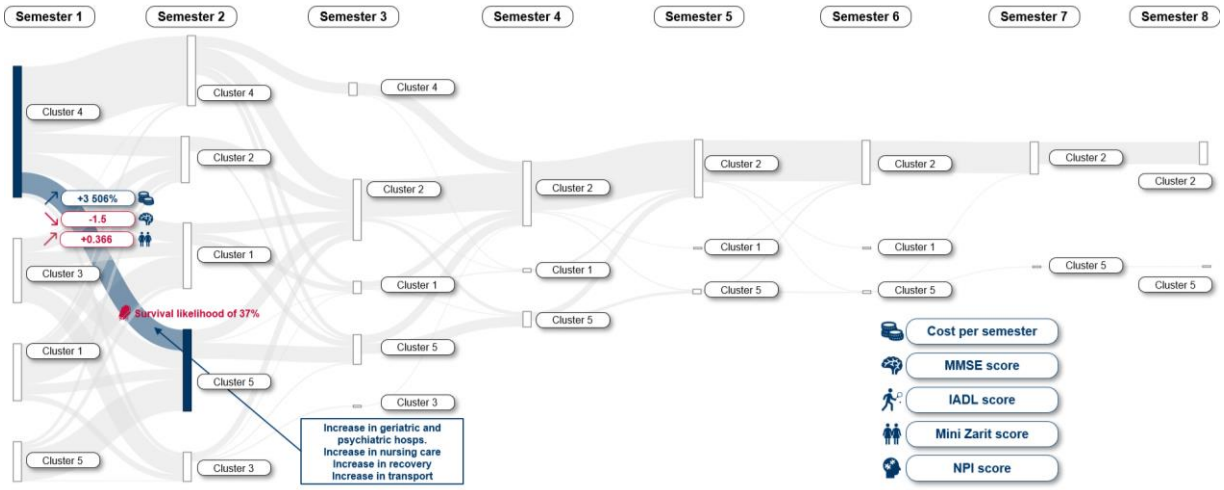


Figure 31 : Exemple de parcours cluster 4 → 5 pour des patients restant en suivi pendant deux semestres.

La Figure 31 ci-dessus est un exemple de parcours emprunté par des patients ayant un déclin cognitif notable du cluster 4 au cluster 5. En un semestre, leur score MMSE va diminuer en moyenne de 1,5 points tandis que le score MiniZarit (fatigue des aidants) augmente de 0,37. L'indicateur le plus parlant est celui du coût, qui va exploser d'en moyenne +3 506%, largement tiré par des hospitalisations gériatriques et psychiatriques à répétition, des dépenses fortes liées au transport spécialisé du patient et également des admissions en SMR. Un an après le début du suivi, la probabilité de survie est de 37%.

La Figure 32 ci-dessous est également un exemple de parcours court pour un patient qui va rester dans le cluster 4. Il n'y a effectivement pas d'évolution notable dans le coût au global, en revanche la nature des consommations de soins va évoluer, avec une augmentation des soins infirmiers à domicile et des séjours en SMR – parallèlement au score NPI (symptômes de troubles comportementaux) qui va s'élever de 0,89 point en moyenne. A bout d'un an de suivi, la probabilité d'admission en EHPAD est de 56%.

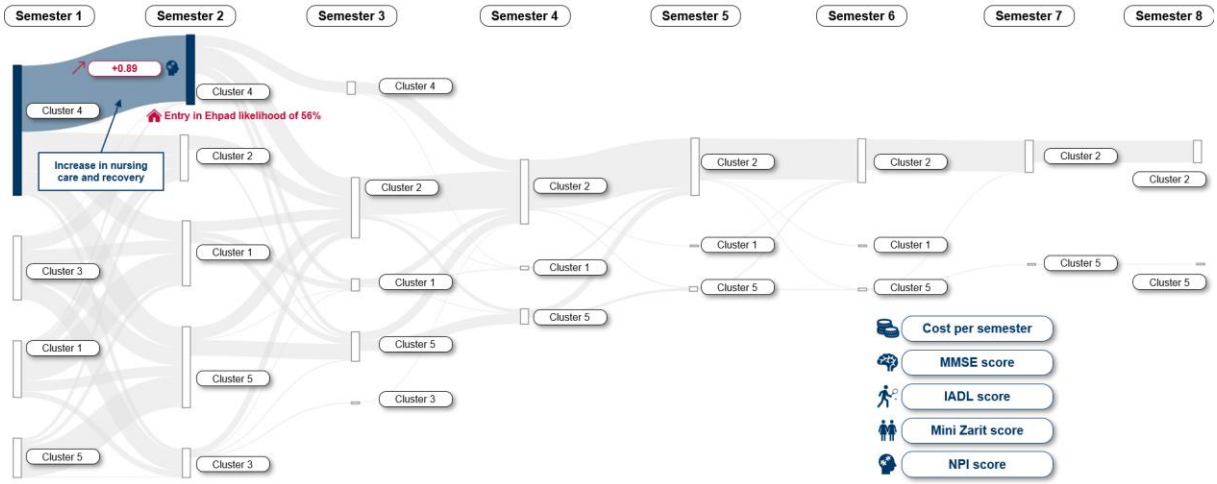


Figure 32 : Exemple de parcours cluster 4 → 4 pour des patients restant en suivi pendant deux semestres.

Nous présentons également un exemple de parcours long dans la Figure 33 ci-dessous. 48 semestres sont concernés par cette trajectoire longue – peu de patients atteignent une telle durée de suivi. Ces patients sont tous atteints de la maladie d'Alzheimer. Au bout d'un an de suivi, on note un déclin cognitif marqué, avec une perte en moyenne de 1,83 de MMSE, suivie par une diminution de 0,95 point pour l'IADL au semestre suivant. A la suite de quoi, le patient évolue vers le cluster 2, où ses consommations de soins vont se concentrer autour des accompagnements à domicile : soins infirmiers, physiothérapeutes.

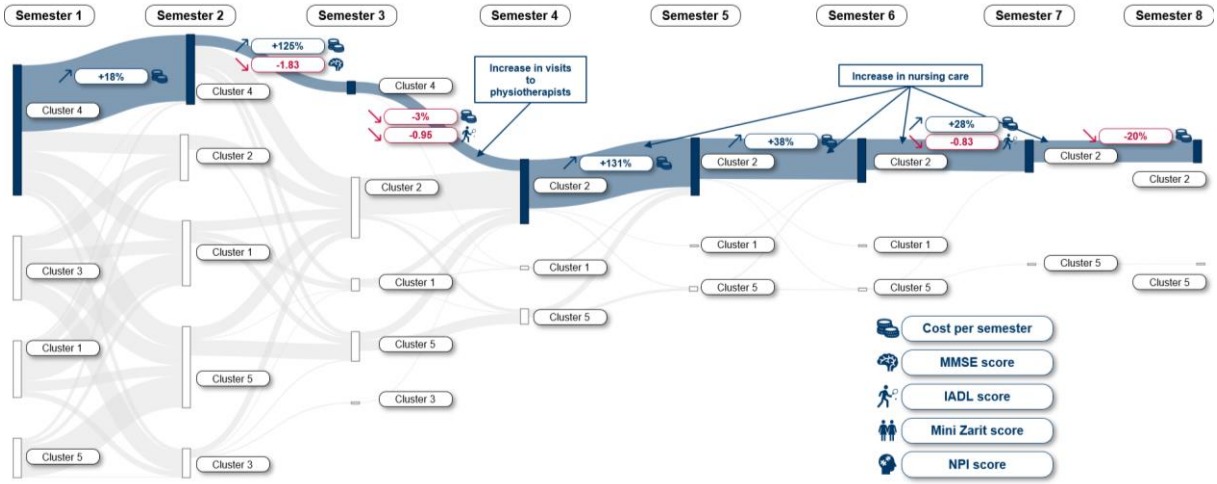


Figure 33 : Exemple de parcours cluster 4 → 4 → 4 → 2 → 2 → 2 → 2 → 2 pour des patients restant en suivi pendant huit semestres.

Parmi les quelques parcours représentés, on dégage des tendances représentatives de l'évolution de ces troubles. Il est cependant important de noter que tous les parcours ne sont pas directement interprétables, avec parfois peu ou pas d'évolution sur les indicateurs cliniques et économiques pendant plusieurs semestres.

Par ailleurs, l'interprétation de ces parcours reste un exercice manuel, à réaliser avec une équipe de spécialistes.

3.4.3 Forces et faiblesses de l'etude et perspectives futures de recherche

La cohorte MEMORA apporte un potentiel exceptionnel en matière de recherche sur les parcours patient. Nous avons pu appliquer une méthodologie permettant d'identifier les inducteurs prépondérants dans la prise en charge et la consommation de soins pour les patients atteints de troubles neurocognitifs. Dans un contexte où il n'est pas possible de dégager des groupes médico-économiques a priori, nous avons dû classifier les patients dans des clusters médico-économiques pour ensuite prédire leur parcours. Nous pensons que ce travail a des applications prometteuses pour les praticiens, les établissements de santé et les patients eux-mêmes. La prédiction du parcours du patient et de ses variations pourrait aider à fournir des soins plus efficaces et adaptés, et à anticiper les résultats. Les acteurs de la santé publique pourraient également bénéficier de ces approches pour prendre des décisions étayées par des données réelles sur les coûts et les ressources

Si cette méthodologie donne une bonne performance générale, il est cependant important de souligner les aspects limitants de notre étude.

1. Nombre d'observations : beaucoup de patients ont un parcours très court (inférieur à deux semestres), ce qui rend l'étude des parcours longs restreinte à une petite fenêtre d'observation. Il reste encore à prouver la généricité de la méthode sur une base de données à plus grande échelle, mais également présentant des séries temporelles plus longues.

2. Interprétation : la classification non supervisée est une tâche pour laquelle l'interprétation des clusters est souvent complexe, notamment à automatiser. Notre méthode d'interprétation a posteriori permet de dresser un portrait des clusters et des parcours obtenus. En revanche, elle est longue et demande l'implication active d'experts soignants. L'essor de l'apprentissage automatique (AutoML pour Automated Machine Learning), en particulier dans le domaine de l'interprétabilité, ouvre également des perspectives pour l'automatisation de cette tâche.

3. Niveau de détail : la base de données Memora offre une richesse majeure dans la recomposition des trajectoires cliniques et économiques des patients. Les données CPAM, par construction, restent cependant agrégées au semestre et ne permettent pas de connaître le détail des actes réalisés. La méthodologie doit encore être testée à un niveau temporel beaucoup plus détaillé.

C'est pour explorer la question du niveau de détail que nous présentons le chapitre suivant. L'objectif est de maintenant prédire le parcours de soins au sein d'une même structure de santé et au cours d'un même séjour.

4

Prédire le parcours de soins, exemple de l'hospitalisation à domicile

Ce quatrième chapitre décrit un cas d'application dans un contexte d'hospitalisation à domicile dans la région Auvergne Rhône-Alpes.

Contenu

4.1 Description du contexte

SYNOPSIS Où nous décrivons le positionnement de l'hospitalisation à domicile dans l'offre de soins en France, et où nous présentons le partenaire hospitalier de cette étude et ses spécificités.

4.1.1 INTRODUCTION ET BREF HISTORIQUE DE L'HOSPITALISATION A DOMICILE

Cette section pose une définition de l'HAD, détaille sa raison d'être dans l'offre de soins en France, décrit les patients cibles et leurs modes de prise en charge, et conclut en présentant les opportunités et les freins liés au développement de l'HAD.

Qu'est-ce que l'HAD ? L'hospitalisation à domicile (HAD) est une forme de prise en charge qui permet d'effectuer un ensemble de soins médicaux et paramédicaux au sein du lieu de vie d'un patient. L'HAD est porté par des structures de santé qui affichent deux objectifs principaux : permettre aux hôpitaux conventionnels de se repositionner en tant que plateau médical technique en raccourcissant ou en évitant l'hospitalisation, tout en offrant au patient une garantie de qualité et de continuité des soins dans un environnement familier. L'HAD couvre par ailleurs une gamme de soins techniques et complexes, ce qui en fait une alternative particulièrement adaptée aux populations vulnérables et de tous âges, qui souffrent de polypathologies lourdes, évolutives, chroniques et/ou instables[177]. Le personnel médical et infirmier peut être salarié ou libéral suivant l'organisation de la structure.

Depuis la circulaire DHOS/O n°2004-44 de 2004[178], l'HAD est reconnue comme une structure hospitalière à part entière et substitut à l'hospitalisation avec hébergement. Dans ce cadre législatif, ces établissements sont certifiés par la Haute Autorité de la Santé dans les mêmes conditions que les structures classiques. L'HAD existe sous de nombreuses formes et peut notamment avoir un statut public ou privé,

[177] Définition issue du site de la Fédération Nationale des Etablissements d'Hospitalisation à Domicile (FNEHAD). Qu'est-ce que l'HAD ? https//www.fnehad.fr/quest-ce-que-lhad/
[178] Voir Circulaire N°DHOS/O/2004/44, Pub. L. No. DHOS/O/2004/44 (2004). https://sante.gouv.fr/IMG/pdf/circulaire_44_040204.pdf

lucratif ou associatif, autonome ou rattaché à une structure de santé hospitalière.

L'HAD face à la restructuration de l'offre de soins en France. Ces établissements de santé se sont développés dans les années 1950 en France sous l'impulsion du Dr Siguier de l'hôpital Ténon (Paris). Le modèle est initialement inspiré de l'hôpital « Home Care » créé par le Dr Bluestone à New York pour répondre à l'explosion du nombre de ses patients[179]. Le modèle séduit depuis, et entre 2006 et 2016, ce sont 143 structures qui ont ouvert en France, tandis que le nombre de journées d'hospitalisation à domicile enregistrées par an a bondi de 3 millions dans le même laps de temp[180]. L'HAD a par la suite poursuivi sa forte croissance parallèlement à la restructuration de l'offre de soins en France.

Entre 2009 et 2019, 23 379 lits d'hospitalisation complète ont été fermés (soit une baisse de 10%, cf. section 1.2.3), tandis qu'en HAD, les capacités d'accueil progressaient de 7,1%, comptabilisant un total de 17 400 patients pouvant être traités simultanément sur le territoire[181]. En 2020, l'HAD représente 6,6 millions de journées réparties sur 153 500 patients, soit une durée moyenne de séjour de 42,9 jours[182]. A titre de comparaison sur la même période, l'hospitalisation conventionnelle a pris en charge 10,9 millions de patients pour un séjour moyen de 4,2 jours[183]. L'HAD représente donc 9% des journées d'hospitalisation mais seulement 1,4% des patients pris en charge sur le territoire.

Modes de prises en charge et patients cibles. L'HAD couvre au total 29 modes de prise en charge. Le patient type est admis avec un mode de prise en charge principal (MPP) et un mode de prise en charge associé (MPA) parmi la liste présentée Figure 35 ci-dessous. Les soins palliatifs et les pansements complexes sont les modes de prise en charge les plus répandus. Ils couvrent à eux seuls plus de 50% de l'activité 2020, une part qui est stable depuis une dizaine d'années. Par ailleurs, 1/3 des journées d'HAD concerne des pathologies cancéreuses.

[179] Selon (Raffy-Pihan, 1997).
[180] Chiffres cités par (Mauro, 2017).
[181] Selon « Les chiffres clefs de l'offre de soins 2018 ». (2019). Direction Générale de l'Offre de Soins (DGOS).
[182] Chiffres tirés du suivi de l'activité hospitalière nationale 2020, champ HAD publié par l'Agence Technique de l'Information sur l'Hospitalisation (ATIH) via le site ScanSanté. https://www.scansante.fr/applications/analyse-activite-nationale
[183] Chiffres tirés du suivi de l'activité hospitalière nationale 2020, champ MCO publié par l'ATIH via le site ScanSanté.

L'évaluation médicale du patient à l'entrée donne lieu à l'attribution d'un Indice de Karnofsky[184] (IK), correspondant à son état de santé et son degré de dépendance aux soins. En 2020, plus de 65% des journées d'HAD sont consacrées à des patients dépendants ou très dépendants aux soins, avec un IK inférieur à 40% (cf. Figure 35 ci-dessous). Ces patients sont par ailleurs âgés en moyenne de 67,8 ans et plus de la moitié sont âgés de 70 ans et plus[185]. 17,5% des patients sont pris en charge depuis l'EHPAD. Au total sur cette même année, 153 500 patients ont été hospitalisés à domicile pour un total de 6,6 millions de journées de soins. Jusqu'à présent, la prise en charge type de l'HAD est donc centrée sur des patients âgés, malades chroniques et dépendants, généralement en transfert depuis l'hospitalisation classique.

Mode de prise en charge	Nombre de journées en 2020 (en milliers)	Part du mode de prise en charge dans l'activité 2020
04-Soins palliatifs	1 787,60	27%
09-Pansements complexes et soins spécifiques (stomies compliquées)	1 640,80	25%
14-Soins de nursing lourds	536	8%
03-Traitement intraveineux	390,7	6%
06-Nutrition entérale	371	6%
08-Autres traitements	359,3	5%
07-Prise en charge de la douleur	259,6	4%
13-Surveillance postchimiothérapie anticancéreuse	248,6	4%
05-Chimiothérapie anticancéreuse	161,2	2%
10-Posttraitement chirurgical	139,2	2%
01-Assistance respiratoire	131,9	2%
02-Nutrition parentérale	131,9	2%
19-Surveillance de grossesse à risque	114,1	2%
12-Rééducation neurologique	85,1	2%
15-Education du patient et/ou entourage	83,3	1%
11-Rééducation orthopédique	64,8	1%
21-Post-partum pathologique	58,4	1%
22-Prise en charge du nouveau-né à risque	28,7	1%
24-Surveillance d'aplasie	6,5	
17-Surveillance de radiothérapie	4,6	
18-Transfusion sanguine	2,5	
29-Sortie précoce de chirurgie	2,1	
Total	6 607,90	

Figure 34 : Répartition des journées HAD réalisées en 2020 par mode de prise en charge principal. Source : ATIH ScanSanté, activité hospitalière nationale 2020, champ HAD.

[184] L'indice de Karnofsky décrit, sur une échelle synthétique de 0% (décès) à 100% (aucun signe ou symptôme de la maladie), l'état de santé global du patient, l'aide dont il a besoin pour les gestes de la vie courante (besoins personnels, habillage, etc.) et les soins médicaux qu'il requiert. Définition issue de l'ATIH, à retrouver sur atih.sante.fr

[185] Selon « Analyse de l'activité hospitalière 2020. HAD ». (2021). Agence Technique de l'Information sur l'Hospitalisation (ATIH).

	Indice de Karnofsky	Nombre de journées 2020 (en milliers)
10%	Le patient est moribond, le processus fatal progressant rapidement	70,1
20%	Le patient, très malade, nécessite un traitement de soutien actif	762,2
30%	Le patient est sévèrement handicapé	1 795,1
40%	Le patient est handicapé et nécessite une aide et des soins particuliers	1 710,7
50%	Le patient nécessite une aide suivie et des soins médicaux fréquents	1 147,5
60%	Le patient nécessite une aide occasionnelle, mais peut prendre en charge la plupart des soins personnels	683,0
70%	Le patient est capable de se prendre en charge, mais est incapable de mener une vie normale ou de travailler	305,1
80%	Le patient est capable de réaliser les activités normales de la vie quotidienne avec effort, quelques symptômes ou signes mineurs	77,1
90%	Le patient est capable de réaliser les activités normales de la vie quotidienne, symptômes ou signes mineurs de la maladie	20,2
100%	Le patient ne présente aucun signe ou symptôme de maladie	37,1
Total		6 607,9

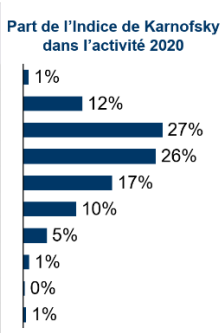


Figure 35 : Répartition des journées HAD réalisées en 2020 par Indice de Karnofsky. Source : ATIH ScanSanté, activité hospitalière nationale 2020, champ HAD

Il est cependant à noter que l'HAD a été fortement mobilisée pour répondre à la crise sanitaire : 14 500 patients ont été hospitalisés pour une prise en charge de la COVID-19 en 2020, ce qui représente 3% de l'ensemble des journées (cf. Figure 36 ci-dessous) et 8% des patients[186]. L'ATIH estime que l'impact de la crise contribue à 56,9% de l'augmentation globale du nombre de patients en HAD entre 2019 et 2020, ce qui montre que ces structures de santé sont également parfaitement positionnées pour répondre à des vagues épidémiques aiguës, tant pour la prise en charge de patients atteints que pour décongestionner la MCO et libérer des lits.

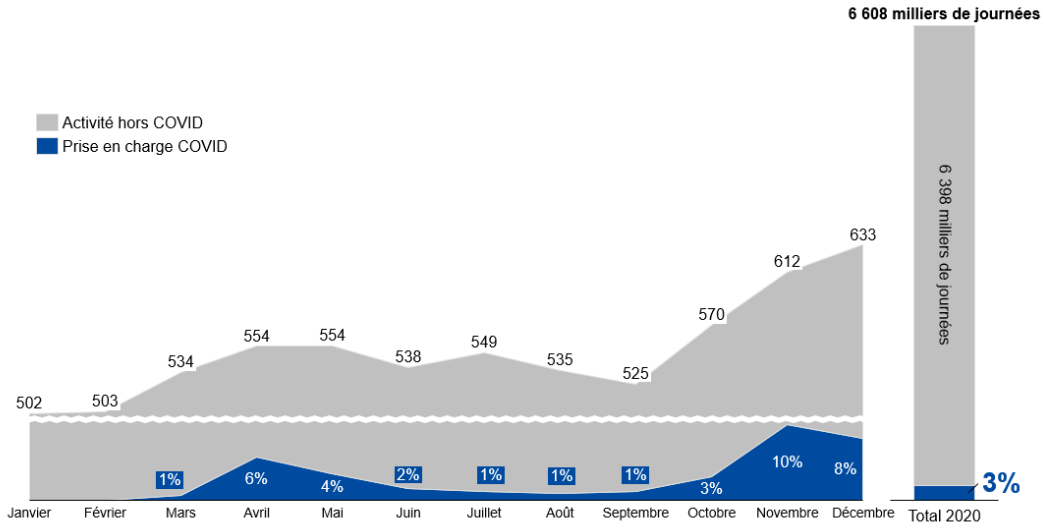


Figure 36 : Répartition des journées HAD COVID et hors COVID réalisées en 2020 par mois (graphique de gauche) et au total sur l'année (graphique de droite). Source : ATIH ScanSanté.

[186] Selon « Analyse de l'activité hospitalière 2020. HAD ». (2021). Agence Technique de l'Information sur l'Hospitalisation (ATIH).

L'HAD poursuit son développement. La feuille de route HAD 2021-2026 publiée par le ministère de la santé[187] affiche clairement la volonté de développer ce mode de prise en charge notamment pour répondre aux besoins d'une population vieillissante et accédant parfois difficilement aux plateaux techniques de soins.

Cette stratégie repose sur 7 axes :

1. Améliorer la connaissance de l'HAD et l'attractivité de cette activité : ce type de prise en charge est encore méconnu du grand public et des médecins prescripteurs, notamment sur les compétences cliniques disponibles dans ce type de structure.
   Une action concrète : « animer le réseau des ARS sur l'HAD et instituer un partage d'expériences afin que les bonnes pratiques et les organisations innovantes se développement sur tout le territoire national. »

2. Renforcer la place des HAD dans l'organisation territoriale sanitaire : l'HAD est un levier clef pour répondre aux enjeux démographiques de la France, l'objectif est de favoriser la complémentarité des activités entre les acteurs du soin et de positionner l'HAD entre le secteur de l'hospitalisation conventionnelle et l'ambulatoire.
   Une action concrète : « inscrire aux CPOM (contrats pluriannuels d'objectifs et de moyens)[188] des établissements de santé un objectif de recours à l'HAD. »

3. Développer l'articulation entre l'HAD et le secteur social et médico-social, et renforcer le rôle de l'HAD dans les parcours des personnes vulnérables : l'HAD a un rôle unique et une réelle expertise dans la coordination des acteurs médicaux et paramédicaux du soin pour des prises en charges complexes qui nécessitent un plan de soins divers. L'enjeu de cet axe est de renforcer cette coopération, en particulier pour les personnes âgées, les publics en situation de handicap et/ou de précarité.

[187] Publié par le Ministère des solidarités et de la santé. (2021). Feuille de route 2021-2026 HAD. https://sante.gouv.fr/IMG/pdf/feuille-de-route-had-2022-05-01-2.pdf
Les exemples d'actions concrètes sont issues de ce document.

[188] Les CPOM sont des contrats signés entre les Agences Régionales de Santé (ARS) et les établissements. Valables pour une durée de 5 ans, les CPOM listent notamment les autorisations dont dispose l'établissement dont les implantations géographiques, les activités spécifiques qui lui sont reconnues ainsi que les financements octroyés et les objectifs auxquels ceux-ci sont reliés. Définition issue du site des ARS. https://www.ars.sante.fr/les-contrats-pluriannuels-dobjectifs-et-de-moyens-1

Une action concrète : « Accompagner, dans le respect du parcours patient et en lien avec les équipes de coordination en gériatrie, le déploiement par les HAD en EHPAD d'offres de soins, telles que les chimiothérapies injectables ou les transfusions au bénéfice des résidents malades. »

4. Renforcer la qualité et la pertinence de la prise en charge en HAD : depuis sa création, l'HAD a développé une expertise autour de la réalisation d'actes techniques complexes, tout en assurant une continuité des soins. L'objectif est de favoriser l'implémentation et la diffusion des bonnes pratiques cliniques entre les établissements.
   Une action concrète : « Adapter la règlementation relative au stockage des médicaments à l'environnement du domicile. »

5. Faire de la e-santé et du numérique un levier de diversification des prises en charge : au même titre que les hôpitaux, les HAD informatisent leurs processus et tentent d'introduire des applications numériques dans la prise en charge de leurs patients. Cet axe souhaite stimuler cette dynamique en permettant aux acteurs les moins matures de bénéficier du savoir-faire des plus avancés.
   Une action concrète : « Promouvoir le recours en HAD aux objets connectés à domicile. »

6. Permettre au patient et à ses aidants d'être acteurs dans le parcours HAD : les soins au domicile du patient favorisent l'opportunité d'impliquer les aidants et les patients dans la prise de décision, mais également d'apporter un soutien et une écoute dans l'accompagnement de la maladie.
   Une action concrète : « Accompagner les patients par l'éducation thérapeutique pour leur permettre d'autogérer l'administration de leurs thérapeutiques notamment dans le cas de maladies chroniques et de traitements connus de longue date par les malades et leur entourage. »

7. Développer la recherche et l'innovation en HAD : de par sa spécificité d'intervention en lieux de vie des patients, ces prises en charges sont un terrain original de mise en œuvre de thérapeutiques innovantes. L'HAD est cependant encore souvent ignorée dans les processus d'innovation et les protocoles de recherche.

Une action concrète : « Mobiliser l'HAD pour initier des projets de recherche dans le champ du domicile en particulier dans le champ des soins palliatifs. »

Cette stratégie de développement sert plusieurs objectifs dont celui d'offrir une alternative à l'hospitalisation conventionnelle et notamment favoriser la décongestion des hôpitaux, celui d'assurer une continuité des soins de qualité élevée entre l'hôpital et le domicile, mais également celui de réduire les dépenses de santé.

La place de l'HAD face à l'hôpital conventionnel. Plusieurs études récentes[189],[190], explorent le positionnement historique de l'HAD face à l'hospitalisation avec hébergement, mais également face à la médecine de ville, synthétisé dans la figure 14 ci-dessous.

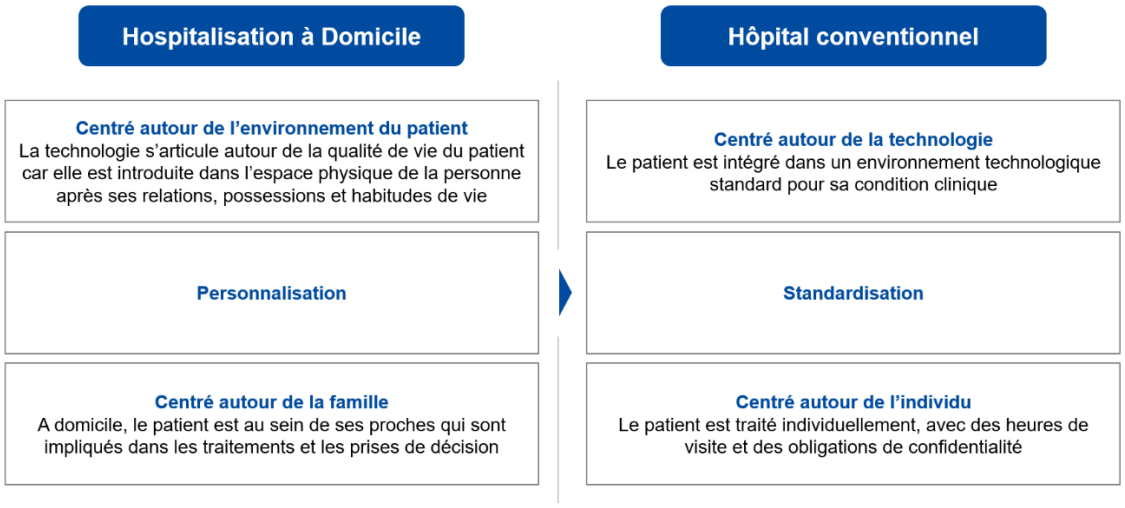


Figure 37 : Positionnement de l'HAD et de la médecine conventionnelle par rapport au patient et son environnement.

Les différences se font également ressentir dans le modèle de coûts. L'HAD fonctionne à la tarification à l'activité (T2A), mais est aussi moins chère. En 2021, le coût moyen d'une journée HAD facturé à l'assurance maladie était de 225€, un chiffre quatre fois inférieur à celui de l'hôpital conventionnel[191].

En réalité, l'interface entre l'HAD et le reste du réseau de santé est principalement axée sur le flux entrée/sortie de patients : l'admission en HAD se fait obligatoirement sur prescription, d'un médecin libéral ou

[189] Voir (Salter, 2015).
[190] Voir (Leung et al., 2016).
[191] Chiffres cités dans « Rapport d'activité 2021-2022 ». Fédération Nationale des Etablissements d'Hospitalisation à Domicile (FNEHAD). (2022)

hospitalier. Dans la pratique, le canal d'entrée principal est l'hôpital avec plus de 55% des demandes en 2021.

Or, selon la FNEHAD, près de 9 séjours sur 10 initiés depuis le domicile et prescrits par le médecin traitant permettent d'éviter l'hospitalisation conventionnelle. Pour améliorer la qualité et la continuité des soins pour ces patients, il est nécessaire d'identifier les facteurs encourageants d'une part et bloquants de l'autre, à la prescription de l'HAD par les médecins généralistes et au développement de ce mode de prise en charge des patients.

L'hospitalisation à domicile face à ses problématiques. Fin 2013, le gouvernement affiche dans son plan de développement un objectif de doublement de l'activité en HAD d'ici à 2018 : passer de 0,6% à 1,2% du nombre de séjours réalisés en hospitalisation avec hébergement, soit 30 à 35 patients pris en charge par jour pour 100 000 habitants[192]. Force est de constater que cet objectif n'est pas atteint, avec une moyenne sur le territoire de 21,3 patients par jour pour 100 000 habitants en 2018. Ce taux de couverture augmente globalement, mais il reste de très grandes disparités entre les régions, avec des statistiques allant de 13 à 96 patients/jour/100 000 habitants (respectivement en Franche-Comté et en Guadeloupe).[193]

Le développement de l'HAD fait face à des freins d'ordre stratégique, opérationnel et organisationnel. D'une part, les médecins prescripteurs de l'admission peuvent craindre d'engager leur responsabilité sur des prises en charge qui se révèlent complexes et engendrent d'importants risques cliniques[194]. De l'autre, l'HAD connaît les mêmes pressions de coût et d'efficience que l'hospitalisation classique. Les deux principaux postes de la structure de coûts sont pour l'HAD :

1. Les prestations de soins facturées par le personnel libéral et salarié intervenant au domicile du patient ;
2. Les produits pharmaceutiques, consommables et dispositifs médicaux.

[192] Selon le rapport de la Cour des Comptes. « L'hospitalisation à domicile : Évolutions récentes ». (2015). [Communication à la commission des affaires sociales et à la mission d'évaluation et de contrôle des lois de financement de la sécurité sociale de l'Assemblée Nationale]. Cour des Comptes.
[193] Chiffres cités dans « Rapport d'activité 2017-2018 ». Fédération Nationale des Etablissements d'Hospitalisation à Domicile (FNEHAD). (2019)
[194] A retrouver chez (Sentilhes-Monkam, 2005).

A l'inverse, les recettes proviennent de la T2A, sous la forme d'une allocation journalière attribuée par patient en fonction des trois descripteurs de son état clinique : les modes de prise en charge principal et associé et l'indice de Karnofsky. Une étude[195] tente d'établir un lien entre les postes de coûts, les recettes et les descripteurs du profil patient. L'absence de corrélation entre ces critères montre que les inducteurs sont autres, par exemple la pathologie ou la condition sociale du patient. De fait, le cadre de santé et/ou le médecin coordonnateur établit à l'admission du patient un plan de soins contenant notamment la nature des soins prodigués, la fréquence des visites et le personnel soignant intervenant.

En pratique, de nombreuses dérives sont observées entre le plan initial planifié et les soins effectivement réalisés. Ces contraintes de coût et d'efficience demandent notamment d'optimiser non seulement la prescription des traitements, des examens et des soins mais aussi leur délivrance et donc tous les flux logistiques et d'informations entre les différents acteurs internes et externes.

Ces interfaces entre acteurs amènent d'autres problématiques, notamment[196] :

- Collaboration entre les acteurs du réseau : à la fois lors de la transmission des informations cliniques entre le personnel soignant libéral et salarié, mais à l'interface avec les prestataires de soins tels que les pharmacies et les laboratoires ;
- Allocation des ressources et planification des tournées de soins ;
- Complexité des systèmes d'information pour la gestion numérique des ordonnances et du dossier du patient : du fait de la multiplication des utilisateurs libéraux utilisant le système, la formation au logiciel et la mise à jour des dossiers de manière uniforme restent deux points durs. Les utilisateurs libéraux sont par ailleurs nombreux du fait de la dimension du territoire couvert qui demande de faire appel à de multiples cabinets de praticiens.

[195] Décrite par (Besombes et al., 2017).
[196] Citées par (Ben Bachouch et al., 2012).

D'autres études se penchent également sur les obstacles à la prescription de l'HAD par les médecins généralistes libéraux[197], mais également à l'exercice en HAD par ces mêmes médecins[198] :

- Mauvaise connaissance du circuit d'entrée : les demandes sont chronophages et l'information n'est pas en accès direct ;
- Mauvaise connaissance du rôle du médecin traitant : des questionnements subsistent sur la coordination avec le médecin de l'HAD, la transmission de l'information clinique et la rémunération du médecin traitant s'il effectue des visites à domicile ;
- Evolution de l'exercice médical : la pénurie de médecins généralistes et une volonté d'avoir des conditions de travail plus stables sont difficilement compatibles avec l'investissement lourd des visites multiples au domicile, parfois urgentes et non-programmées, et qui viennent perturber le fonctionnement d'un cabinet.

4.1.2 Presentation du partenaire de l'etude et de sa typologie de patients

Cette section décrit la structure de santé partenaire de cette étude, ainsi que son organisation autour du processus de soins.

Soins et Santé, partenaire HAD de ce cas d'application. Soins et Santé est un établissement de santé privé d'intérêt collectif (ESPIC) créé en 1972 et localisé en banlieue lyonnaise, à Rillieux-la-Pape. La structure assure une prise en charge à domicile dans les départements du Rhône, Nord Isère et dans quelques communes de l'Ain (cf. Figure 38 ci-dessous). En 2018, Soins et Santé prend en charge 400 patients par jour en moyenne, sur un territoire d'une superficie avoisinant les 8 000 km². La structure est un acteur majeur de la santé dans la région, offrant une gamme complète de soins à domicile pour des patients chroniques mais aussi aigus. Elle emploie une centaine de salariés qui coordonnent les dossiers patients puis fait appel à des cabinets libéraux d'infirmiers, kinésithérapeutes et médecins généralistes pour réaliser les soins établis au cours du projet thérapeutique.

[197] Mentionnés par (Leung et al., 2016).
[198] Mentionnés par (Besombes et al., 2017).

Depuis quelques années, l'établissement développe également la prise en charge dans des lieux de vie alternatifs tels que les EHPAD. Afin de cibler les étapes clefs du parcours de soins, nous avons réalisé une cartographie de la trajectoire du patient de son admission à sa sortie.

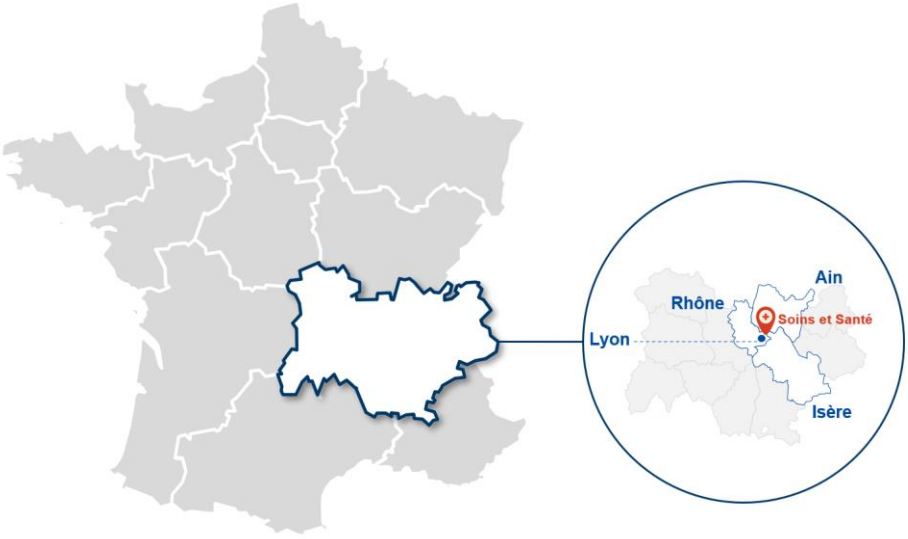


Figure 38 : Localisation et zones d'intervention de l'HAD Soins et Santé

Fonctionnement général de l'HAD Soins et Santé. La structure de santé fonctionne autour de trois piliers :

- la cellule d'éligibilité qui s'occupe de l'admission du patient,
- l'hospitalisation par secteur qui permet de coordonner les soins,
- les soignants libéraux, qui dispensent les soins au domicile du patient.

Des fonctions transverses, notamment administratives et financières, mais également paramédicales (ex : diététicien et hygiène), interagissent avec ces pôles et fournissent un support à la pratique clinique tout au long de l'hospitalisation d'un patient (cf. Figure 39, ci-dessous).

Admission : la cellule d'éligibilité.

La cellule d'éligibilité est une plateforme de préadmission pour les patients de l'HAD, qui vise à centraliser les demandes à la suite d'une prescription hospitalière ou provenant d'un médecin généraliste. Le personnel de la plateforme est composé de trois IDEC[199] de préadmission et de sept infirmières.

[199] Un IDEC est un Infirmier Diplômé d'Etat de Coordination. Il s'agit d'un infirmier cadre qui encadre l'équipe soignante (infirmiers, aides-soignants, aides médico-psychologiques), organise, priorise et contrôle les soins et leur traçabilité.

Leur rôle est de réceptionner la demande, d'évaluer la pertinence de l'hospitalisation à domicile dans le projet thérapeutique du patient puis, le cas échéant, de préparer l'admission du patient au sein de l'établissement. Le principal circuit de demande d'admission est l'outil « Trajectoire » développé par l'ARS[200] de la région Rhône-Alpes pour organiser les transferts de patients entre établissements. La demande est remplie par un service hospitalier conventionnel et comporte des informations sur le dossier du patient et le type de prise en charge demandé. En fonction de la complexité de la prise en charge, l'IDEC peut être amené à se déplacer au domicile du patient ou au sein de l'hôpital adresseur pour évaluer les besoins en soins. La cellule éligibilité se charge de dresser un plan de soins et donc de contacter les intervenants libéraux qui réaliseront les soins. Le circuit d'admission par la médecine de ville est lui initié par une prescription d'un médecin généraliste qui contacte généralement directement l'HAD.

Pour finaliser l'admission, le projet thérapeutique au sein de l'HAD et le plan de soins prévisionnel doivent être validés par un cadre de santé, un médecin et un assistant social. A la suite de cette validation, une visite d'entrée est planifiée au domicile du patient par l'IDEC de suivi pour enclencher la prise en charge. Cet IDEC de suivi fait partie d'un secteur chargé de la coordination des soins, comme décrit dans le paragraphe suivant.

Coordination des soins : les secteurs.

Le territoire d'intervention est découpé en 3 secteurs : Centre, Ouest et Est. Une équipe pluridisciplinaire de soignants est affectée à chacun de ces secteurs pour y suivre les patients pris en charge dans la zone correspondante. Chaque secteur compte en moyenne 120 patients, qui sont attribués en fonction de leur lieu de vie. Un secteur est piloté par un médecin gériatre, appelé « médecin coordonnateur », qui s'appuie sur un ensemble de personnels soignants et administratifs incluant IDEC de suivi, infirmier et secrétaire pour chaque secteur ainsi qu'un psychologue et assistant social en transverse (voir l'organigramme en Figure 39 ci-dessous).

[200] Les Agences Régionales de Santé (ARS) sont des établissements publics placés sous la tutelle du Ministère chargé des affaires sociales et de la santé. Elles assurent un pilotage unifié de la santé dans leur région et sont en charges de réguler l'offre de santé sur leur territoire. Définition issue de https://www.ars.sante.fr/quest-ce-quune-agence-regionale-de-sante

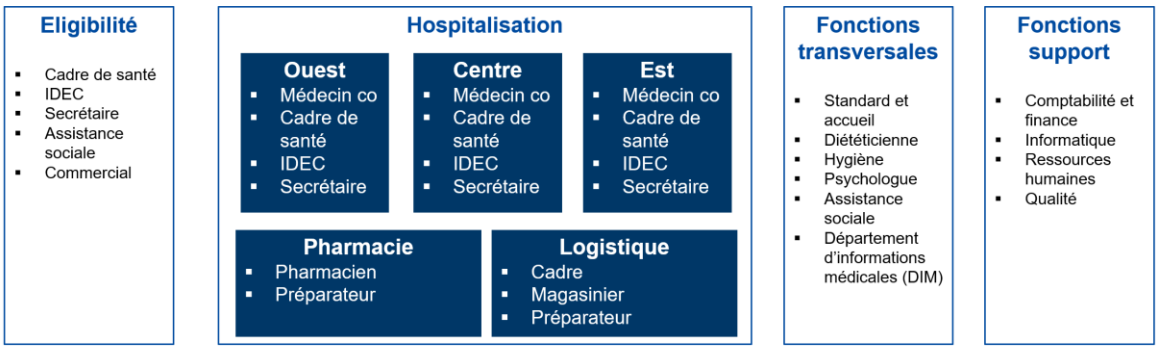


Figure 39 : Organigramme organisationnel de Soins et Santé

Cette équipe se charge d'émettre un avis à l'admission d'un patient, puis de suivre l'évolution de son plan de soins tout au long de son séjour. Elle est notamment l'interlocuteur privilégié des soignants libéraux qui effectuent les soins aux domiciles du patient, mais se charge également de la planification et du suivi des consultations en ambulatoire du patient (cf. figure 17 ci-dessous).

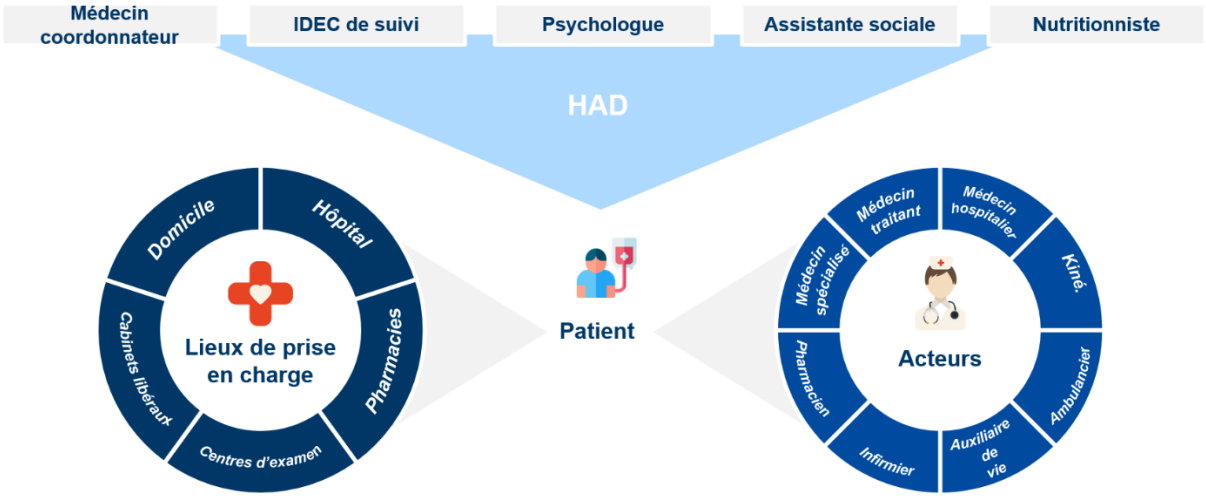


Figure 40 : Rôle et positionnement de l'HAD - coordination des soins et des lieux de prise en charge

A l'admission du patient, l'IDEC de suivi planifie une visite à domicile pour présenter l'objectif de la prise en charge et les soins à réaliser au patient et le cas échant à ses aidants, installer le matériel de santé déjà livré et rencontrer les intervenants libéraux. Au cours du suivi, des visites sont effectuées à intervalle régulier pour réaliser un bilan clinique et mettre à jour le plan de soins au besoin.

Réalisation des soins : les intervenants libéraux.

Le projet thérapeutique du patient comporte les intervenants à mobiliser, les types de soins planifiés et la fréquence nécessaire. La lourdeur du plan de soins varie selon le mode de prise en charge, l'état de santé du patient mais également son degré d'autonomie. A titre d'exemple, la grande majorité des hospitalisations à domicile nécessitent entre une à trois visites par jour. Les soignants impliqués sont des médecins généralistes, des kinésithérapeutes, des orthophonistes, des psychologues, des infirmiers, des aides-soignants ou encore des techniciens de laboratoire. A noter que Soins et Santé dispose d'une pharmacie et d'un service de logistique en interne qui se chargent respectivement de la livraison des médicaments prescrits au patient et du matériel de santé nécessaire.

Le détail du plan de soins, y compris la fréquence des visites et les soins réalisés, est susceptible d'évoluer fortement au cours du séjour. L'origine de ces variations est multifactorielle : elle peut provenir d'un changement de l'état de santé du patient comme de facteurs extérieurs plus difficiles à quantifier.

4.1.3 Problematique a l'etude

De par le fonctionnement inhérent à l'hospitalisation à domicile, les équipes de suivi prennent en charge un grand nombre de patients simultanément. Les variations dans les plans de soins sont fréquentes, de part le profil complexe et multi-pathologique des patients accueillis en HAD, et peuvent fortement perturber les activités quotidiennes de l'équipe. Pour alléger cette charge, il est crucial de pouvoir anticiper ces évolutions mais également d'identifier les facteurs qui y sont associés.

Dans ce cas d'application, nous cherchons à prédire la trajectoire de soins d'un patient au sein de l'établissement et les inducteurs de coût de son séjour. Les données médico-administratives et de facturation sont un excellent proxy permettant de recomposer le parcours de soins. L'objectif du modèle est de prédire le nombre de visites en fonction de l'intervenant au domicile et d'identifier les variables issues du dossier patient influençant le coût du séjour.

Nous projetons par ailleurs que ce type d'application permette d'ouvrir la voie à la résolution de certains problèmes que connaît le développement

de l'hospitalisation à domicile en France, notamment en termes d'allocation des ressources, de logistique et de collaboration entre les différents acteurs du réseau. Comme identifié dans le plan HAD 2021-2026, le numérique est un excellent levier d'optimisation des parcours. Développé dans les structures à domicile, il devrait permettre de positionner ces dernières au cœur de la prise en charge et de l'accompagnement des patients vulnérables.

4.2 Méthodologie

SYNOPSIS Où nous explicitons les données à disposition, les variables extraites et l'approche déployée.

4.2.1 DESCRIPTION DES DONNEES ET DE LA POPULATION A L'ETUDE

Cette section décrit les données renseignées dans le logiciel de suivi des patients à partir duquel nous extrayons les variables, l'identification des variables liées au parcours de soins sous l'angle des coûts et la sélection finale des variables pour la suite de l'analyse.

Présentation du logiciel AtHome et des variables extractibles. AtHome est un logiciel dédié à la gestion et au suivi des patients en hospitalisation à domicile, édité par la société Arcan Systems. Cet ERP[201] hospitalier permet de structurer le dossier du patient, d'organiser les soins à domicile et garantit une traçabilité des actes qui sont réalisés.

Onglet Administratif.

L'onglet administratif synthétise les informations sur la provenance du patient en termes d'adresseur, son état civil et ses informations de contact. Les aidants et leur lien avec le patient sont également listés.

Onglets Médical et Soins.

Le médecin coordonnateur et les IDEC remplissent lors de l'entrée l'anamnèse clinique[202] du patient, dont les pathologies et les motifs d'admission. Les comptes-rendus médicaux et de soins (infirmiers, psychologie) y sont également stockés. Le logiciel permet de remplir un certain nombre d'évaluations cliniques telles que les signes vitaux, mais

[201] Un système ERP (Enterprise Resource Planning, ou progiciel de gestion intégré en français) est un type de logiciel que les entreprises utilisent pour gérer leurs activités quotidiennes telles que la comptabilité, les achats, la gestion de projet, la gestion des risques et la conformité, ainsi que les opérations de *supply chain*. Dans le secteur de la santé, ces logiciels permettent d'intégrer tous les aspects d'admission, de suivi et de facturation du patient. Définition issue du site Oracle. https://www.oracle.com/fr/erp/what-is-erp/

[202] L'anamnèse est un la synthèse d'un interrogatoire réalisé pour un patient donné et donc l'objectif est de dresser l'historique de la maladie et des antécédents médicaux et chirurgicaux. Définition issue du site internet Dictionnaire Médical. https://www.dictionnaire-medical.fr/definitions/594-anamnese/

ces derniers ne sont pas utilisés par les soignants, qui préfèrent noter ces informations directement dans les comptes-rendus.

Codage de la séquence de soins.

En hospitalisation à domicile, le séjour d'un patient est découpé en une à plusieurs séquences de soins qui vont conditionner le niveau de remboursement versé par l'Assurance Maladie à la structure de santé. Une séquence de soins est définie par un groupe homogène de prise en charge[203], c'est-à-dire par une combinaison de trois variables : un mode de prise en charge principal (MPP), un mode de prise en charge associé (MPA) et l'indice de Karnofsky (IK)[204]. La séquence de soins est renseignée dans l'ERP par le médecin du Département d'Informations Médicales (DIM) de l'établissement et déclenche le financement du séjour, proportionnellement à sa durée et jusqu'à sa fin (cf. Table 16 ci-dessous).

Patient	Date de début	Date de fin	MPP	MPA	IK	Montant	Motif de fin de séquence
A	01/01	02/01	Soins palliatifs	Prise en charge de la douleur	20	813,26 €	Décès
B	16/08	29/08	Post traitement chirurgical	Pas de MPA	50	2 629,40 €	IK dégradé à 40
B	30/08	16/10	Post traitement chirurgical	Pas de MPA	40	7 822,78 €	Fin d'HAD

Table 16 : Exemples de séquences de soins pour deux patients de l'HAD Soins et Santé

Recomposition du coût du parcours patient. Des données récupérées du service comptable de la structure sur l'année 2018 nous donnent un aperçu des dépenses (cf. figure 18 ci-dessous).

[203] Un groupe homogène de prise en charge (GHPC) est une classification médico-économique qui permet de relier le mode de prise en charge d'un patient à une valorisation économique de son séjour. Chaque séquence est reliée à un GPHC. La séquence reste la même tant que la valeur des trois variables (MPP, MPA, IK) ne change pas ; la séquence change aussitôt que l'une au moins de ces variables est modifiée. Le patient entre alors dans une nouvelle séquence de soins qui est valorisée différemment. Selon le « Guide méthodologique de production des recueils d'informations standardisés de l'hospitalisation à domicile » produit par l'ATIH. Consultable en ligne : https://www.atih.sante.fr/node/4443/edit

[204] La liste des 29 modes de prises en charge ainsi que la signification et les valeurs de l'indice de Karnofsky sont détaillées plus haut dans la section 4.1.1 Introduction et bref historique de l'hospitalisation à domicile.

Sur un total de 25,6 M€ en 2018, nous estimons que 18,1 M€ soit 71% des dépenses totales, sont directement liées au parcours des patients. Le premier poste de coût, à hauteur de 9,7M€ (38%), sont les prestations à domicile du patient, notamment infirmières, kiné et soins de nursing. Les produits pharmaceutiques et consommables suivent derrière, notamment les médicaments eux-mêmes mais également les traceurs[205].

Les trois postes de coûts suivants pèsent dans le parcours d'une moindre manière et concernent les salaires médicaux fixes de la structure, la location du matériel de santé et des charges diverses incluant par exemple le transport des patients.

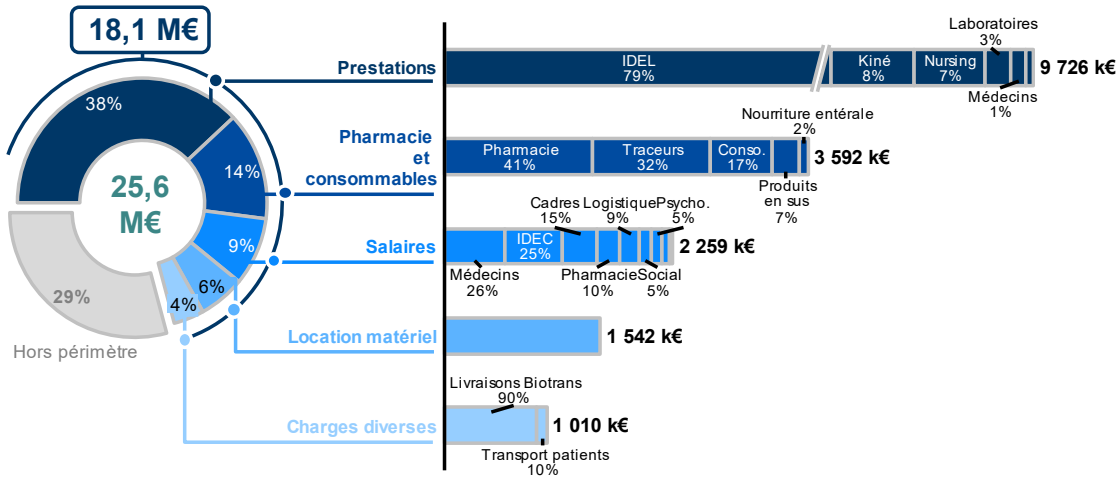


Figure 41 : Périmètre des coûts de la structure de soins liés au parcours patient (en dégradé de bleus) et non liés au parcours de soins (en gris). Zoom sur les coûts liés au parcours. Source : Soins et Santé, analyse comptable recomposée sur l'année 2018.

Parmi les coûts classés hors périmètre (7,5M€ soient 29% des dépenses totales en 2018), on retrouve principalement des dépenses liées à des salaires fixes de la structure, par exemple pour des fonctions administratives et des charges diverses (cf. Figure 42 ci-dessus).

[205] Un traceur est un produit radioactif qui, une fois injecté dans le sang, peut être visualisé dans l'organisme du patient lors de différents examens d'imagerie médicale (TEP, scintigraphie). Le traceur, en se fixant sur différents organes, permet d'en analyser le fonctionnement. Ce produit est notamment utilisé lors de prises en charges en oncologie. Définition issue de l'institut national du cancer, consultable en ligne e-cancer.fr/dictionnaire/

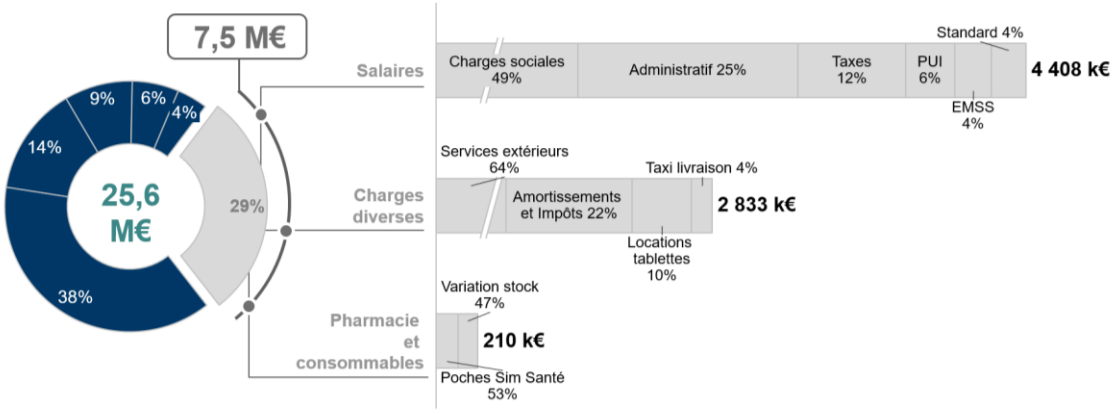


Figure 42 : Périmètre des coûts de la structure de soins liés au parcours patient (en bleu) et non liés au parcours de soins (en gris). Zoom sur les coûts non liés au parcours. Source : Soins et Santé, analyse comptable recomposée sur l'année 2018.

Le poids respectif des coûts directement liés et non directement liés au parcours patient peut varier selon les années (illustration de l'évolution 2017 à 2018 dans la Figure 43 ci-dessous). En revanche la structure des coûts du parcours de soins semble rester identique, notamment en termes d'importance relative des inducteurs économiques.

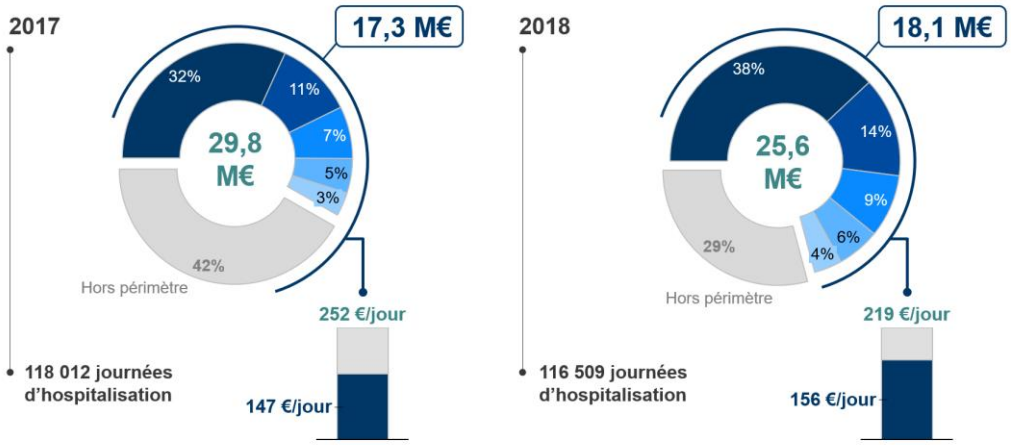


Figure 43 : Recomposition des coûts journaliers directement liés (en bleu) et non directement liés (en gris) au parcours patient sur les années 2017 (graphique de gauche) et graphique de droite). Source : Soins et Santé, analyse comptable.

En recomposant les postes de coûts directement imputables au parcours patient, nous avons délimité le périmètre de l'analyse. Il s'agit à présent d'identifier quel périmètre les données présentes dans le système AtHome permettront de couvrir.

Les variables exploitables dans AtHome et leur sélection. Au cours d'une série d'entretiens avec le responsable du service informatique de Soins et Santé, ayant eu lieu au cours du dernier trimestre 2019, nous

avons cartographié le modèle de relations des données traitées dans le progiciel AtHome (cf. Figure 44 ci-dessous). Il existe au total 10 tables et 86 variables exploitables dans la base de données. Les variables sélectionnées pour la suite de l'analyse ont été identifiées par deux sources principales. Les grandes catégories de variables à cibler ont premièrement été déterminées grâce à la recomposition des postes de coûts les plus significatifs.

Par la suite, des entretiens avec le médecin coordonnateur principal de l'HAD ainsi que plusieurs IDEC de suivi ont permis d'identifier les variables d'intérêt les plus pertinentes (en bleu dans la cartographie ci-dessous), et notamment d'exclure celles ne présentant pas d'intérêt pour l'analyse ou pouvant être peu mises à jour dans le système d'information au cours du séjour. A noter que la variable d'association entre toutes les tables est l'identifiant unique du séjour, il s'agit donc de la clef primaire à partir de laquelle seront faites les jonctions.

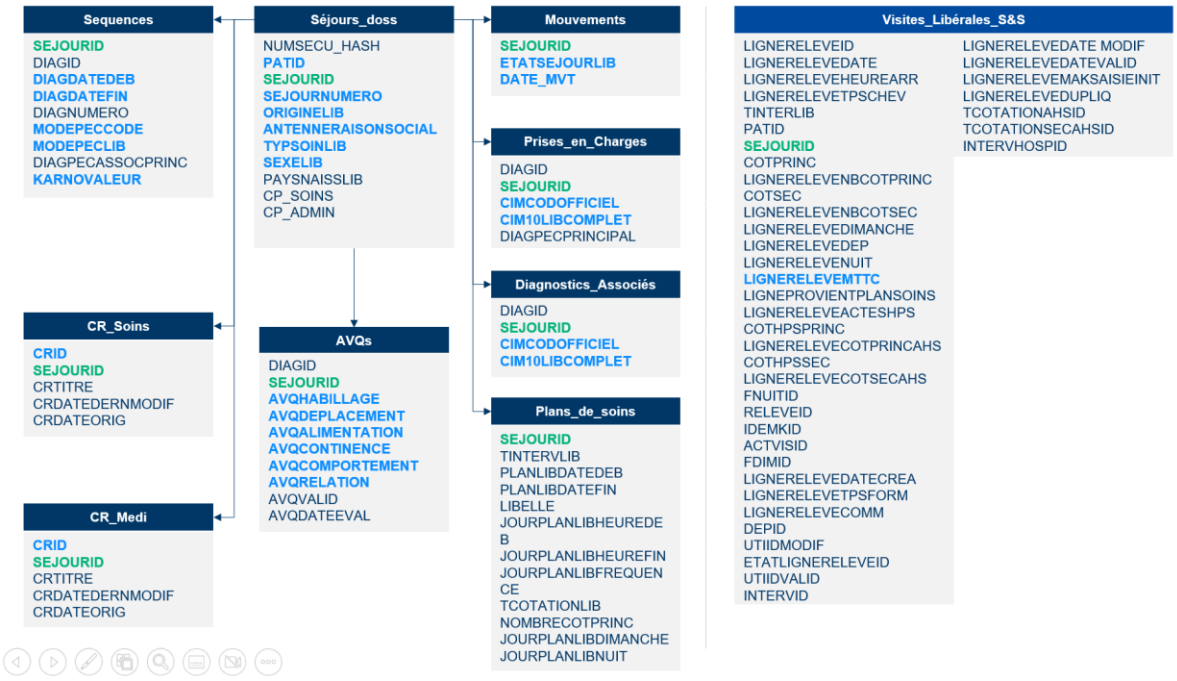


Figure 44 : Cartographie des tables et des données AtHome. Légende : en vert la clef principale reliant les tables ; en bleu les variables sélectionnées pour l'analyse.

« Sejours_doss » : la table principale.

Sejours_doss est la table principale de la base de données. Elle contient toutes les données administratives liées au patient et à son séjour incluant le numéro de sécurité sociale (crypté lors de la transmission pour garantir l'anonymat), l'identifiant du séjour, le canal d'entrée en HAD, la situation familiale, le genre du patient et le type de soins pour lesquels le patient est admis. Un patient peut avoir plusieurs séjours enregistrés à son actif.

Nom de la variable	Type de variable	Définition	Intervalle de valeurs
SEJOURID	Float	Identifiant unique du séjour (clé primaire)	[1, 50968]
PATIENTID	Integer	Identifiant unique du patient (clé secondaire)	[203,33231]
SEJOURNUM	Float	Classement du séjour (numéro)	[1, 157]
ORIGINE	String	Canal d'entrée du patient (service adresseur ou médecin prescripteur)	[Médecin hospitalier, Médecin traitant, NA]
TYPESOINS	String	Type de soins reçus	[Continu, Ponctuel, Réadaptation, NA]
SEXE	String	Genre du patient	[Masculin, Féminin, NA]
DATE_NAISSANCE	DateTime	Date de naissance	
SITUATIONFAM	String	Situation familiale du patient	[Marié(e), Veuf(ve), Célibataire, Divorcé(e), Concubin, Séparé(e), NA]

Table 17 : Dictionnaire des variables de la table Sejours_doss

La table brute contient 50 709 entrées sur une période de 1997 à 2020, chaque ligne correspondant à un séjour.

« Mouvements » : la table des événements administratifs.

Mouvements est une table qui synthétise les événements administratifs notables du séjour, c'est-à-dire la série d'états dans lesquels le dossier va se trouver au long du séjour et la date d'enregistrement de ces états : demande en cours (DDE), patient en cours de suivi (SUIVI), fin du séjour (FIN). On y enregistre également les décès (DCD). Ces états successifs sont appelés « mouvements du dossier ».

Nom de la variable	Type de variable	Définition	Intervalle de valeurs
SEJOURID	Float	Identifiant unique du séjour (clé secondaire)	[1, 26118]
ETATSEJOUR	String	Libellé administratif du mouvement	[DDE, SUIVI, DCD, FIN]
DATE_MVT	DateTime	Date du mouvement	De 1997 à 2020

Table 18 : Dictionnaire des variables de la table Mouvements

La table brute contient 65 535 entrées, chaque ligne correspondant à un mouvement. Chaque séjour enregistré possède au moins un mouvement, le plus grand nombre de mouvements enregistrés pour un seul séjour est de 20. Le plus vieux mouvement enregistré date de 1997, et le plus récent de 2020. A noter que dans cette table, beaucoup de séjours ne sont pas représentés (max(SEJOURID) = 26118), ce qui peut être dû à des erreurs administratives dans le relevé des états.

« Prises_en_charges » : la table des diagnostics principaux.

Prises_en_charges est une table qui enregistre les diagnostics à l'admission du patient. Chaque diagnostic est identifié par un code (3 à 5 caractères) et un libellé – par exemple 'C34', 'Tumeur maligne des bronches et du poumon'. Pour un même séjour, on distingue le diagnostic principal de ses diagnostics associés via la variable DIAGPECPRINCIPAL.

Nom de la variable	Type de variable	Définition	Intervalle de valeurs
DIAGID	Integer	Identifiant unique du diagnostic (clé primaire)	[2, 59341]
SEJOURID	Integer	Identifiant unique du séjour (clé secondaire)	[1, 50630]
CIMCODE	String	Code CIM-10	1633 codes diagnostics et les libellés correspondants de la classification CIM-10 sont représentés[206]
CIM10LIBELLE	String	Libellé du diagnostic correspondant au code CIM-10	

[206] La Classification Internationale des Maladies (CIM) est la classification médicale permettant le codage en morbi-mortalité proposée et recommandée par l'OMS. Elle associe un code de 3 à 5 chiffres et lettres à une pathologie. Elle est soumise à des révisions périodiques, la 10e révision (CIM-10) a été diffusée en 1992. La classification et son fonctionnement seront détaillés dans la section

DIAGPECPRINCIPAL	String	Différencie le diagnostic principal (O pour oui) des diagnostics associés (N pour non)	['O', 'N']

Table 19 : Dictionnaire des variables de la table Diagnostics

La table brute contient 75 345 entrées, chaque ligne correspondant à un diagnostic.

« Diagnostics_assoc » : la table des diagnostics associés.

Diagnostics_assoc regroupe les diagnostics supplémentaires pouvant être renseignés au cours du séjour. Ils sont identifiés de la même manière que dans la table Diagnostics.

Nom de la variable	Type de variable	Définition	Intervalle de valeurs
DIAGID	Integer	Identifiant unique du diagnostic (clé primaire)	[56, 59340]
SEJOURID	Integer	Identifiant unique du séjour (clé secondaire)	[4, 50609]
CIMCODE	String	Code CIM-10	255 codes diagnostics et les libellés correspondants de la classification CIM-10 sont représentés
CIM10LIBELLE	String	Libellé du diagnostic correspondant au code CIM-10	

Table 20 : Dictionnaire des variables de la table Diagnostics_assoc

La table brute contient 62 427 entrées, chaque ligne correspondant à un diagnostic.

« Sequences » : la table des séquences de soins.

Sequences est une table qui enregistre la succession des séquences d'un même séjour. Le fonctionnement et la codification des séquences ont été présentés dans la section précédente. Pour une même séquence, on différencie le MPP du MPA via la variable PECPRINCIPAL.

suivante. Définition issue de la page INSERM CépiDc. Consultable en ligne : https://www.cepidc.inserm.fr/causes-medicales-de-deces/classification-internationale-des-maladies-cim

Nom de la variable	Type de variable	Définition	Intervalle de valeurs
SEJOURID	Integer	Identifiant unique du séjour (clé secondaire)	[1, 24229]
SEQID	Integer	Identifiant unique de la séquence (clé primaire)	[2, 31240]
SEQDATEDEB	DateTime	Date de début de la séquence	-
SEQDATEFIN	DateTime	Date de fin de la séquence	-
SEQNUMERO	Integer	Classement de la séquence (numéro)	[0, 87]
MODEPECCODE	Integer	Code du MPP/MPA	Valeurs possibles décrites dans la section précédente
MODEPECLIB	String	Libellé du MPP/MPA	
PECPRINCIPAL	String	Différencie le MPP (O pour oui) du MPA (N pour non)	['O', 'N']
KARNOVALEUR	Integer	Indice de Karnofsky	Echelle de valeur décrite dans la section précédente

Table 21 : Dictionnaire des variables de la table Sequences

La table brute contient 82 923 entrées, chaque ligne correspondant à un MPP ou MPA.

« Visites_Liberales_S&S » : la table des visites libérales.

La table soins regroupe les visites libérales effectuées au domicile ou lieu de prise en charge du patient. On relève la date, l'heure et le temps passé au chevet du patient – ainsi que l'intervenant ayant réalisé la visite. Ces relevés sont enregistrés par les libéraux eux-mêmes qui facturent le soin à la structure de santé. Ce montant est ensuite associé dans le système au séjour et patient correspondant.

Nom de la variable	Type de variable	Définition	Intervalle de valeurs
LIGNERELEVEID	Integer	Identifiant unique de la visite (clé primaire)	[1, 3759753]
LIGNERELEVEDATE	DateTime	Date de la visite	-
LIGNERELEVEHEURE	DateTime	Heure d'arrivée	-

LIGNERELEVETPSCHEV	DateTime	Temps au chevet du patient	[1 min, 900 min]
SEJOURID	Integer	Identifiant unique du séjour (clé secondaire)	[2, 50927]
INTERVENANT	String	Type d'intervenant à domicile	[IDE Libéral(e), Kinésithérapeute, Médecin traitant, Orthophoniste, NA]
LIGNERELEVEMTTC	Float	Montant facturé lors de la visite (en €)	[0, 7540]

Table 22 : Dictionnaire des variables de la table Soins

La table brute contient 3 759 753 entrées, chaque ligne correspondant à une visite au lieu de vie du patient.

4.2.2 Preparation des donnees et analyse exploratoire

Cette section décrit les étapes préliminaires de préparation des données et présente la table finale utilisée pour l'analyse ainsi que la distribution statistique des principales variables d'intérêt.

Pré-traitement général et mise en forme des données. La clef d'identification conjointe aux tables étant l'identifiant du séjour, nous avons pris le parti de construire un jeu de données dont chaque ligne représentera un séjour. Les valeurs manquantes sont traitées en fonction de leur type :

- Pour les variables numériques continues (âge et durée moyenne au chevet du patient) dont la proportion de valeurs manquantes est faible, on remplace celles-ci par la valeur moyenne de la colonne. L'Indice de Karnofsky, qui est une variable pressentie comme fortement corrélée aux soins et pour laquelle la proportion de manquants est élevée, sera traité plus tard dans cette section.
- Pour toutes les données qualitatives, on crée une nouvelle catégorie « non renseigné ».

Données de diagnostics principal et associés.

Les données de diagnostic sont identifiées par le code CIM-10 et son libellé correspondant, qui sont donc des données qualitatives difficiles à exploiter.

Le code est composé d'une lettre de A à Z, qui correspond à la catégorie majeure du diagnostic, suivie de deux à quatre chiffres précisant le diagnostic et sa sous-catégorie (cf. illustration dans la Figure 45 ci-dessous).

Figure 45 : Illustration de la décomposition du code CIM-10 de la tumeur maligne de la tête du pancréas.

La classification CIM-10 complète regroupe plus de 12 000 diagnostics différents. Dans les données de l'étude en sont représentées environ 2 000. Or, il n'y a pas de hiérarchie de la sévérité entre ces codes, c'est-à-dire à titre d'exemple qu'on ne peut pas comparer numériquement C250 à D46. Deux stratégies sont communément appliquées pour traiter une variable qualitative :

1. La transposition en variables binaires (*one-hot encoding*) : il y aura autant de variables que de valeurs de la variable qualitative originelle. Ce qui signifierait d'introduire dans notre cas environ 2 000 nouvelles variables.

2. La transformation en variable catégorique, par exemple une échelle numérique hiérarchique (non applicable ici, comme décrit plus haut).

Nous choisissons d'introduire une nouvelle variable CMD qui est la catégorie majeure de diagnostic (première lettre du code). Cette variable ne porte pas le détail du diagnostic ni sa gravité, mais cette dernière est par ailleurs représentée par l'IK. A ce stade et en vue de l'analyse exploratoire, nous conservons la série de diagnostics détaillés en agrégant les codes CIM-10 dans un tuple, le premier élément du tuple étant celui identifié comme le diagnostic principal et est l'information clef à conserver. La nouvelle variable CMD est également un tuple formé à partir de la catégorie majeure des éléments du tuple Diagnostics. Les doublons de catégories majeures au sein d'un même tuple sont supprimés.

Nous instancions également une nouvelle variable Nb_diag qui comptabilise le nombre de diagnostics enregistrés pour un seul séjour, incrémenté à partir de la longueur du tuple Diagnostics.

SEJOURID	DIAGNOSTIC	CODE_DIAG	CMD	NB_DIAG
1	(Tumeur maligne secondaire des os et de la moe...	(C795,)	(C,)	1
2	(Sclérose en plaques, Ulcère de décubitus et z...	(G35, L89)	(G, L)	2
3	(Tumeur maligne de la vésicule biliaire,)	(C23,)	(C,)	1
4	(Sclérose en plaques, Surveillance de colostom...	(G821, L89, Z433, I10, G35, E119)	(E, G, L, I, Z)	6
5	(Séquelles d'accident vasculaire cérébral, non...	(G819, I694)	(G, I)	2

Figure 46 : Table Diagnostics agrégée et retraitée.

Données de séquence.

Nous enlevons les dates de début et de fin pour ne garder que la durée de la séquence en jours. Nous stockons la séquence sous la forme d'un tuple (MPP, MPA). Lorsque plusieurs séquences existent, nous les concaténons pour obtenir un tuple de tuples. L'idée première était de raisonner par séquence, mais pour deux raisons nous décidons de ne pas le faire :

- Il est complexe de relier les soins à une séquence car la date du relevé de la visite n'est pas forcément la date à laquelle la visite a été réalisée, mais la date à laquelle elle a été enregistrée. Pour certains libéraux, il peut y avoir un décalage de quelques jours. L'imputation d'une visite à une séquence pourrait ainsi être faussée.
- Le mode de prise en charge n'est pas complètement représentatif des soins reçus par le patient : il s'agit d'une séquence économique, codée a posteriori par le médecin DIM pour déclencher un remboursement.

L'indice de Karnofsky et ses variations sont également stockés dans un tuple indépendant, pour être traité comme une variable numérique.

Nous instancions une nouvelle variable DEG qui représente le degré maximal de dégradation de l'IK au cours du séjour.

Données de soins.

Les données de soins, qui sont les relevés des visites libérales au lieu de vie du patient, sont des données terrain remplies par un large panel d'utilisateurs (plusieurs centaines de libéraux différents interviennent chaque année). Le nettoyage et l'harmonisation des données a donc nécessité un effort important[207].

Les visites sont regroupées par séjour, par intervenant et par position dans le séjour (J0, J1, etc. ; cf. Figure 47). On introduit la variable « durée des soins », recomposée à partir de la différence entre le premier et le dernier jour de soins. Cette valeur sera utilisée comme proxy de la durée de séjours, bien que cette dernière soit identifiable dans la table Mouvements, et ce, pour deux raisons. Premièrement, une forte proportion de séjours et visites ne sont pas représentés dans cette table, ce qui génère des valeurs manquantes. Deuxièmement, certains patients, notamment en chimiothérapie à domicile, ont des séjours courts et répétés sous la forme de cycles. Un cycle peut durer 5 à 7 jours et être interrompu pendant plusieurs jours sans qu'il y ait fin administrative du séjour. Pour ces patients, la distinction entre les durées de séjour et de soins effectifs a son importance pour recomposer un coût moyen journalier représentatif.

	TINTERVLIB	IDE Libéral(e)	Kinésithérapeute	Médecin traitant	Orthophoniste
SEJOURID	JOUR				
2	0	3	0	0	0
	1	3	0	0	0
	2	3	0	0	0
	3	3	0	0	0
	4	3	1	0	0

Figure 47 : Illustration de la table "Visites" qui regroupe le nombre de visites au domicile par séjour, par jour et par intervenant.

Présentation de la table finale et distributions statistiques des principales variables. La Table 23 ci-dessous synthétise les 23 variables

[207] Cette difficulté fait écho aux problématiques de développement de l'hospitalisation à domicile remontées par (Ben Bachouch et al., 2012).

retenues pour l'analyse. La Table 24 ci-après décrit les statistiques communes des variables numériques.

Nom de la variable	Type de variable	Définition	Intervalle de valeurs
SEJOURID	Float	Identifiant unique de la visite (clé primaire)	[2 ; 59927]
PATIENTID	Float	Identifiant unique du patient (clé secondaire)	[203 ; 33190]
NUM_SEJOUR	Float	Classement du séjour (numéro)	[0 ; 157]
ORIGINE	String	Canal d'entrée du patient (service adresseur ou médecin prescripteur)	[Médecin hospitalier, Médecin traitant, Non renseigné]
SOINS	String	Type de soins reçus	[Continu, Ponctuel, Réadaptation, Non renseigné]
SEXE	String	Genre du patient	[Féminin, Masculin, Non renseigné]
AGE	Float	Age au moment de l'admission	[0 ; 110]
FAMILLE	String	Situation familiale du patient	[Marié(e), Veuf(ve), Célibataire, Divorcé(e), Concubin, Séparé(e), Non renseigné]
CODE_DIAG	Tuple de strings	Codes CIM-10 des diagnostics du patient	(Code principal, Code associé 1, …)
DIAGNOSTIC	Tuple de strings	Libellé CIM-10 correspondant aux codes	(Libellé principal, Libellé associé 1, …)
CMD	Tuple de string	Catégories majeures de chaque diagnostic	(Catégorie principale, Catégorie associée 1, …)
NB_DIAG	Float	Nombre de diagnostics principal et associés du patient	[0 ; 9]
DUREE_SEQ	Tuple de DateTIme	Durées successives des séquences	(Durée 1, Durée 2, …)
CODE_PEC	Tuple de floats	Codes des MPP et MPA successifs	(Code MPP1, code MPA 1, Code MPP2, Code MPA2, …)

MODE_PEC	Tuple de strings	Libellés des MPP et MPA successifs	(Libellé MPP1, Libellé MPA1, …)
DUREE_SOINS	Float	Durée de séjours recomposée à partir de la différence [dernier – premier jour de soins]	|0 ; 4451]
NB_JOURS_SOINS	Float	Nombre de jours où au moins une visite a été enregistrée	[1 ; 4059]
COUT_TOTAL	Float	Somme du coût des visites facturées	[2 ; 401174]
TEMPS_CHEVET	Float	Somme du temps passé au chevet du patient	[1 ; 497150]
COUT_JOURNEE	Float	Coût journalier moyen	[2 ; 1240]
IK	Float	Indice de Karnofsky à l'admission	[0 ; 100]
DEG	Float	Evolution maximale de l'indice au cours du séjour	[-60 ; 70]

Table 23 : Dictionnaire des variables de la table finale

Nom de la variable (unité)	Moyenne	σ	Min	Max	IQ1	Médiane	IQ3
NUM_SEJOUR (séjours)	5,4	13,8	0	157	1	2	4
AGE (années)	67,8	16,5	0	110	59	71	80
NB_DIAG (diagnostics)	2,3	1,3	0	9	1	2	3
DUREE_SOINS (jours)	47,2	104,1	0	4 451	6	18	48
NB_JOURS_SOINS (jours)	44,9	94,9	1	4 059	6	18	46
COUT_TOTAL (€)	3 322	8 263,9	2	401 174	322,3	1 138,8	3 145,8
TEMPS_CHEVET (minutes)	4 161,4	10 399,6	1	497 150	450	1 460	3 924
COUT_JOURNEE (€/jour)	70,1	40,3	2	1 240,6	40,2	62,6	91,1

IK (%)	46,6	18,6	10	100	30	50	60
DEG (%)	1,4	8,33	- 60	70	0	0	0

Table 24 : Distribution statistique des variables numériques de la table finale. Dans l'ordre : moyenne, écart-type, valeur minimale, valeur maximale, première valeur interquartile, médiane, troisième valeur interquartile.

L'objectif est d'identifier les valeurs pouvant apparaître comme aberrantes et susceptibles de pénaliser le modèle. En effet, à la vue des indicateurs statistiques, de nombreuses distributions sont asymétriques.

Nombre de séjours par patient.

Si plus de 75% des séjours observés sont parmi les 4 premiers séjours du patient, certains peuvent en revanche comptabiliser plusieurs dizaines de séjours. Il s'agit habituellement de patients pris en charge pour un à plusieurs cycles de chimiothérapie à domicile. Ces cycles sont organisés en série de 5 à 7 jours à la suite desquels les patients « sortent » puis sont réadmis au cycle suivant[208].

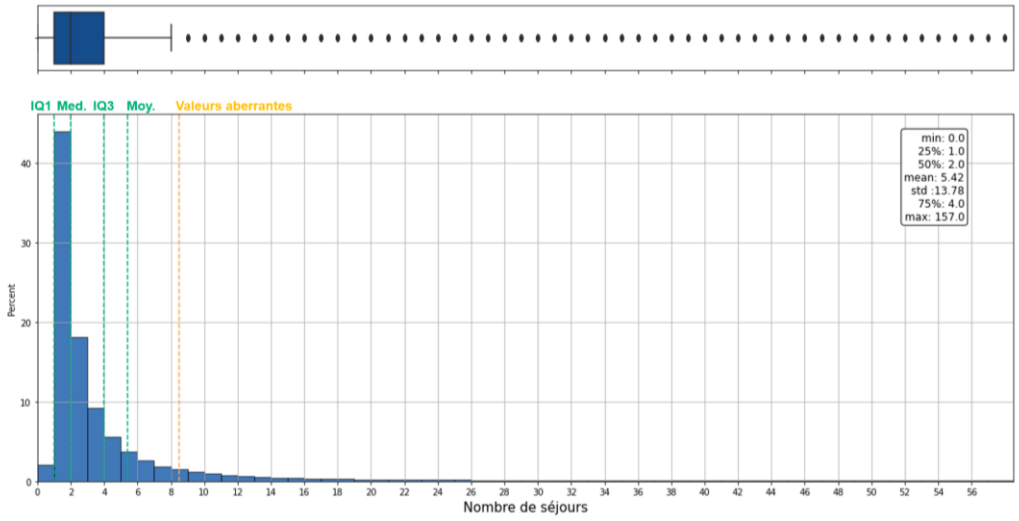


Figure 48 : Distribution et seuil de valeurs aberrantes du nombre de séjours par patient. Boîte à moustaches (graphique du haut) et histogramme (graphique du bas).

[208] Protocole de chimiothérapie sous-cutanée par Azacitidine (Vidaza) en hospitalisation à domicile, décrit par le réseau régional de cancérologie des Hauts-de-France. Consultable en ligne : https://www.onco-hdf.fr/app/uploads/2022/02/Annexe-VI-Modele-type-AZACITIDINE-VIDAZA-EN-HAD-VF.pdf

La méthode de l'écart interquartile[209] fait ressortir 3 828 séjours[210] comme des valeurs aberrantes (au-delà du 8ème séjour). Ces « séjours aberrants » correspondent à 445 patients dont plus de la moitié ont un MPP ou MPA 'chimiothérapie anticancéreuse '. On décide donc de conserver ces valeurs.

Distribution de l'âge.

La moyenne de l'âge à l'admission est de 67,8 ans et est proche de la médiane de 70 ans. Par ailleurs, avec un écart-type de 16,4, on peut voir que la distribution est plutôt concentrée en tranches d'âges similaires. 50% des patients ont un âge à l'admission compris dans une fourchette de 21 ans, entre 59 et 80 ans.

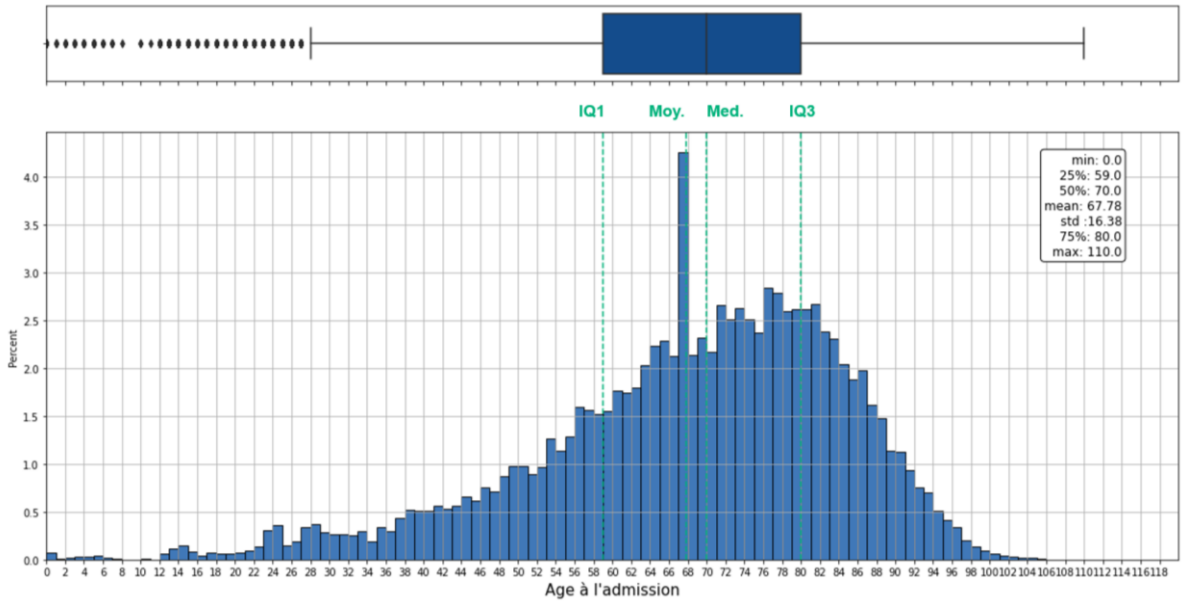


Figure 49 : Distribution et seuil de valeurs aberrantes de l'âge à l'admission par séjour. Boîte à moustaches (graphique du haut) et histogramme (graphique du bas).

Nombre de jours de soins.

Le nombre effectif de jours de soins est en moyenne légèrement inférieur à la durée de séjours (2,3 jours d'écart en moyenne). La différence se

[209] Cette méthode statistique répandue, permet de détecter les valeurs aberrantes à partir de la mesure de l'étendue interquartile (différence entre le troisième et le premier quartile). Méthode décrite dans (Kremp, 1995).
[210] Soit environ 6% du total de séjours (59 926 séjours après jonction de la table finale).

situe principalement sur les séjours longs : le premier quartile et la médiane sont identiques pour les deux variables.

Par ailleurs, la moyenne du nombre de jours de soins est très proche du troisième quartile (respectivement 44,9 jours et 46 jours). On note également un écart-type très élevé (sigma = 94,9). La distribution graphique ci-dessous confirme que la durée totale des soins pendant un séjour est extrêmement variable selon les patients, probablement dû à la variété des profils patients accueillis (e.g., une antibiothérapie qui se compte en jours vs. des soins palliatifs qui durent plusieurs mois).

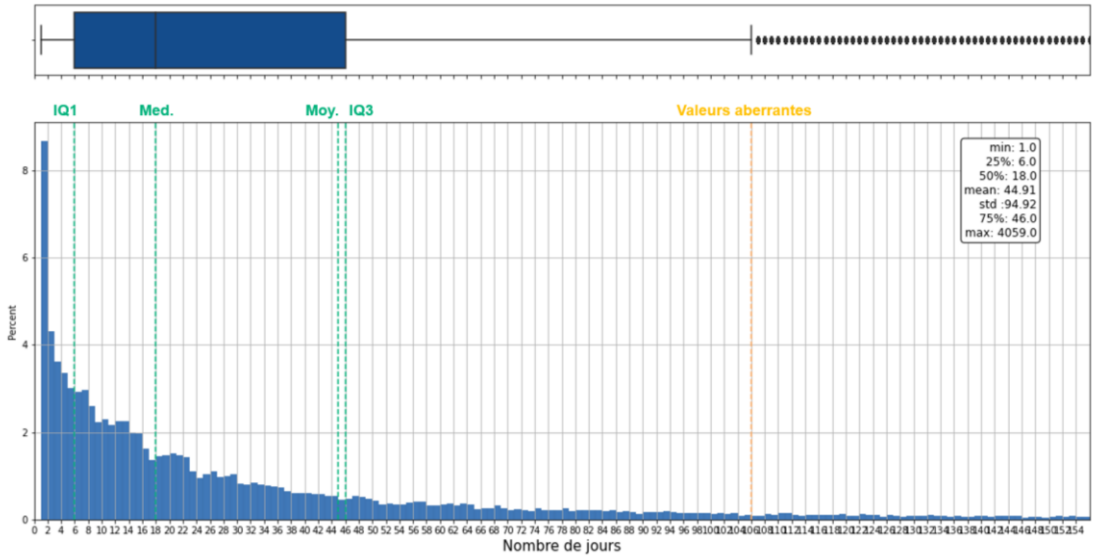


Figure 50 : Distribution et seuil de valeurs aberrantes de la durée des soins par séjour. Boîte à moustaches (graphique du haut) et histogramme (graphique du bas).

Coût total et coût par jour moyen.

Le coût total est très dispersé, car il est proportionnel au nombre de jours de soins (cf. Figure 51). On note par ailleurs les mêmes phénomènes statistiques : écart-type élevé dans les mêmes proportions, et moyenne très proche du troisième quartile. En résultat de quoi, 3 472 valeurs sont considérées comme aberrantes. On se penche donc plutôt sur le coût journalier.

La distribution du coût journalier moyen tend vers une gaussienne asymétrique à droite. Seulement 886 (2,6%) valeurs sont considérées comme aberrantes. On raisonnera par la suite plutôt au travers du coût journalier moyen, qui par ailleurs est une valeur qui peut être estimée en

début de séjour via le plan de soins initial établi par la cellule d'éligibilité. On note également que l'écart-type est faible, et on retrouve une médiane et une moyenne proches (respectivement 62,6 et 70,1 €/jour).

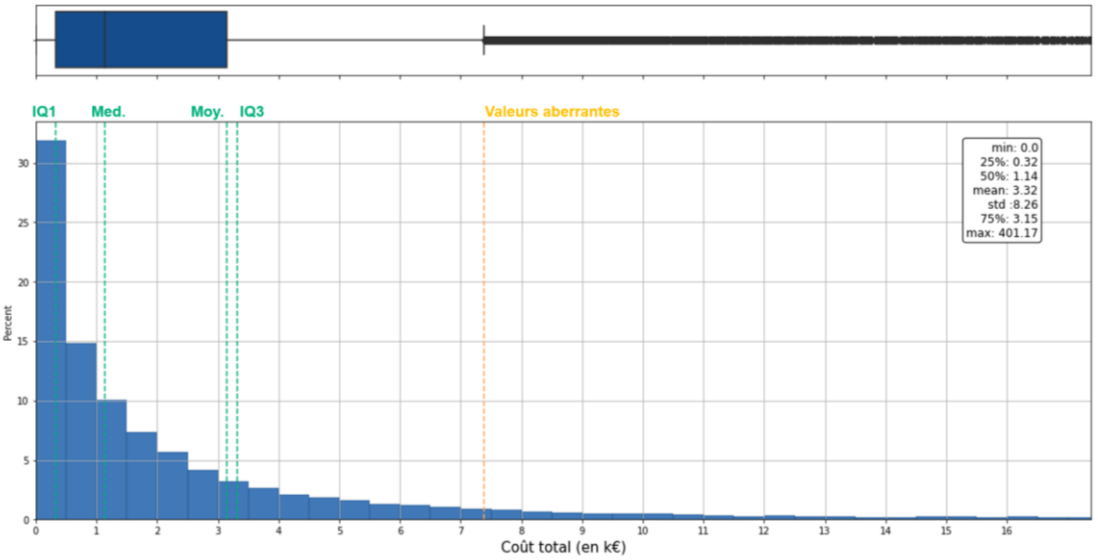


Figure 51 : Distribution et seuil de valeurs aberrantes du coût total par séjour. Boîte à moustaches (graphique du haut) et histogramme (graphique du bas).

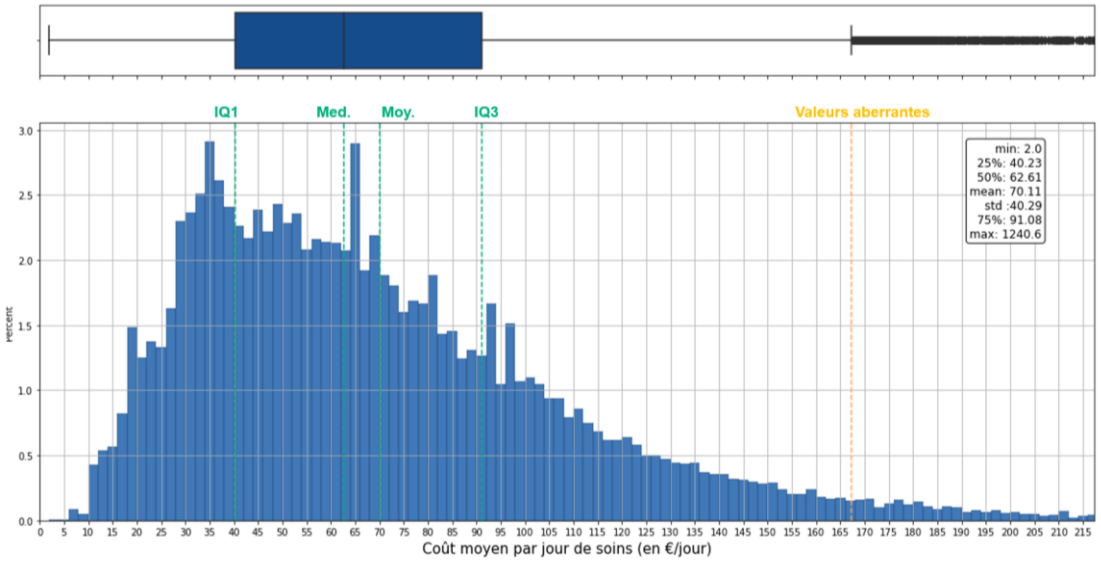


Figure 52 : Distribution et seuil de valeurs aberrantes du coût moyen journalier par séjour. Boîte à moustaches (graphique du haut) et histogramme (graphique du bas).

Indice de Karnofsky à l'admission et amplitude de dégradation en cours de séjour.

Seulement 16 179 séjours ont au moins un indice de Karnofsky enregistré, soient 48% du nombre total de séjours. Une méthode d'estimation des IK

restants sera détaillée dans la section suivante. Sur les quelque 16 000 séjours, la distribution est équilibrée, avec la moitié des valeurs comprises entre 30 et 60%.

La moyenne est légèrement plus basse que la médiane, ce qui montre une plus grande occurrence de patients ayant un IK inférieur à 50 lors de l'admission à l'établissement, en ligne avec le profil de patients cibles de l'hospitalisation à domicile (cf. Figure 53).

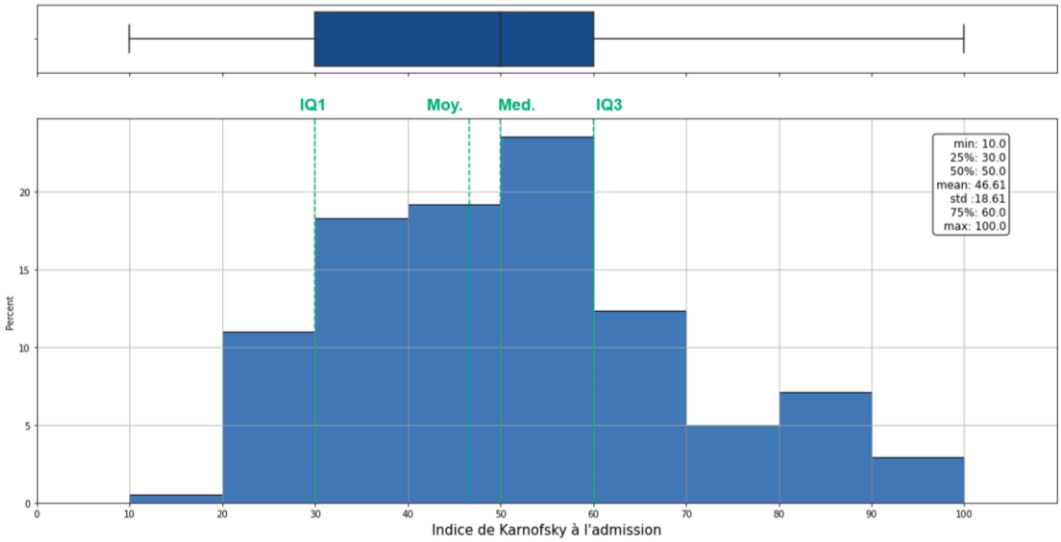


Figure 53 : Distribution et valeurs interquartiles de l'indice de Karnofsky à l'admission. Boîte à moustaches (graphique du haut) et histogramme (graphique du bas).

Pour ces séjours, l'indice n'est revisité que dans 24% des cas. Les dégradations sont presque deux fois plus fréquentes que les améliorations. Si une évolution est enregistrée, il s'agit généralement d'une variation d'un palier (ex : de 50 à 60% ou de 30 à 20%, cf. Figure 54). A noter que le nombre de paliers de la variation dépend également de l'indice initial : un patient avec un IK faible à l'entrée ne pourra pas avoir un écart élevé même si son état se dégrade.

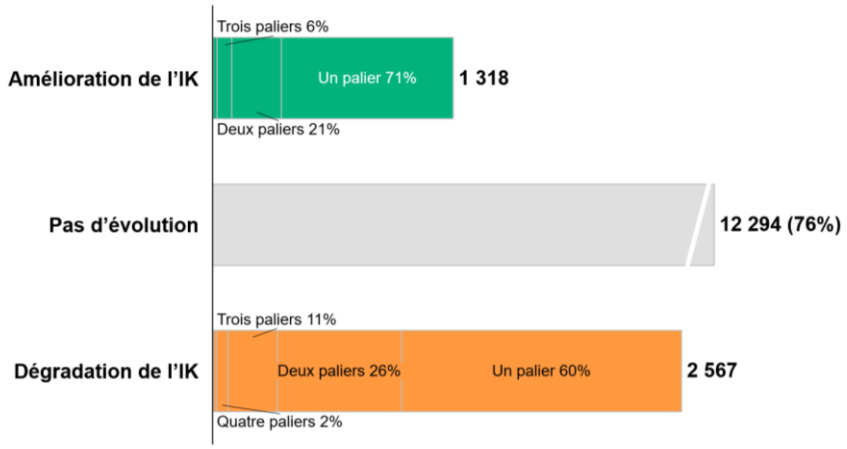


Figure 54 : Classification des degrés d'évolution de l'indice de Karnofsky au cours du séjour (dégradation / amélioration) et nombre de paliers comptabilisés.

Modes de PEC les plus courants.

Les dix prises en charges les plus courantes couvrent 70% des séquences totales renseignées. Les trois prises en charge les plus répandues sont les soins palliatifs, la prise en charge de la douleur et la surveillance post-chimiothérapie anti cancéreuse (34% des séjours). A noter que 18% des séjours n'ont ni MPP, ni MPA.

Cette distribution confirme un panel varié de prises en charges au sein de l'HAD et donc des patients aux profils divers.

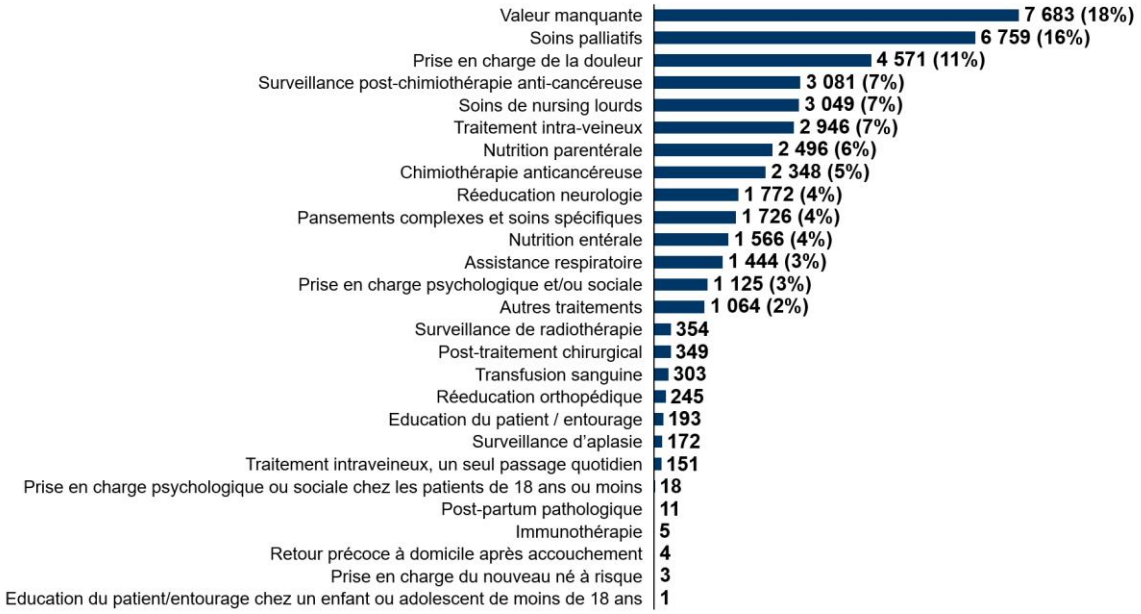


Figure 55 : Occurrence (valeur absolue et % du total) des modes de prises en charges.

Diagnostics les plus courants, catégories majeures de diagnostics et nombre de diagnostics associés.

On s'intéresse premièrement aux diagnostics principaux seulement, soit 48 804 diagnostics. Les dix diagnostics les plus fréquents ne couvrent que 30% des diagnostics principaux, ce qui montre un panel de diagnostics très large (cf. Figure 56). Les diagnostics associés sont eux moins étalés : les dix plus fréquents représentent environ 56% des diagnostics associés totaux, cependant la diversité de catégories majeures de diagnostics y est plus représentée (cf. Figure 57).

La CMD la plus fréquemment enregistrée est celle des tumeurs malignes (CMD C). 40% des diagnostics y sont associés, ce qui confirme bien le positionnement de l'HAD (cf. Figure 58).

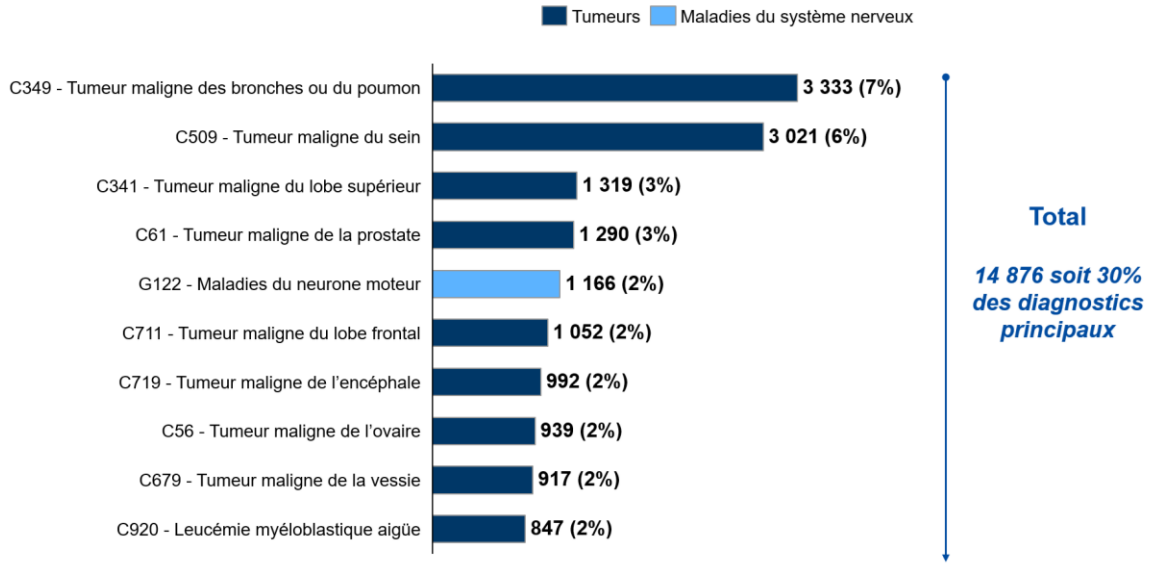


Figure 56 : Occurrence (valeur absolue et % du total) des dix diagnostics principaux les plus fréquents. Code diagnostic et libellé CIM-10, classés par catégories majeures de diagnostics (dégradés de bleu).

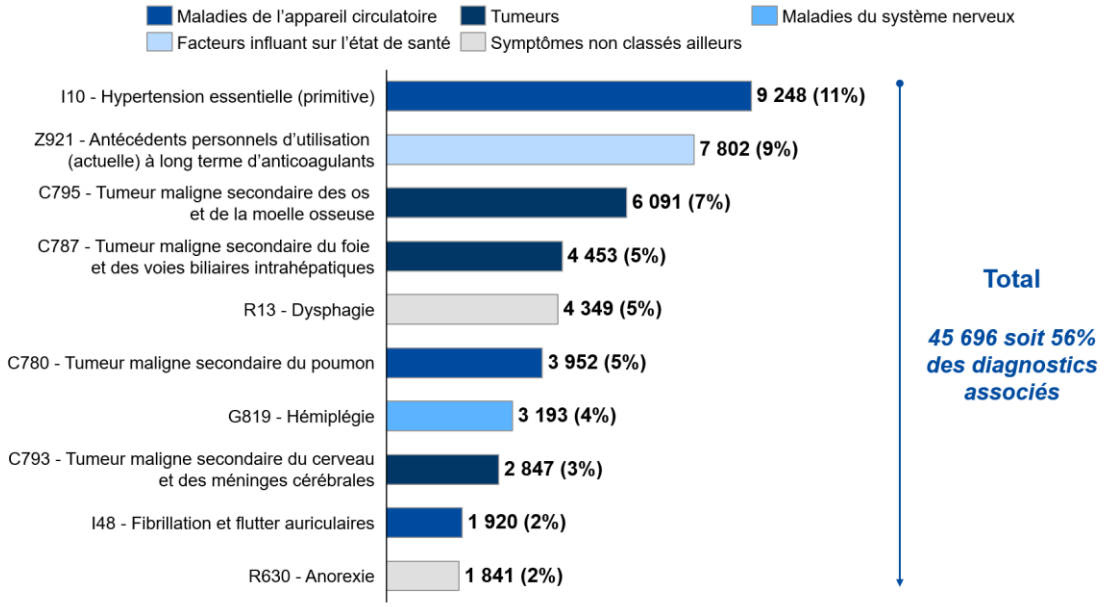


Figure 57 : Occurrence (valeur absolue et % du total) des dix diagnostics associés les plus fréquents. Code diagnostic et libellé CIM-10, classés par catégories majeures de diagnostics (dégradés de bleu).

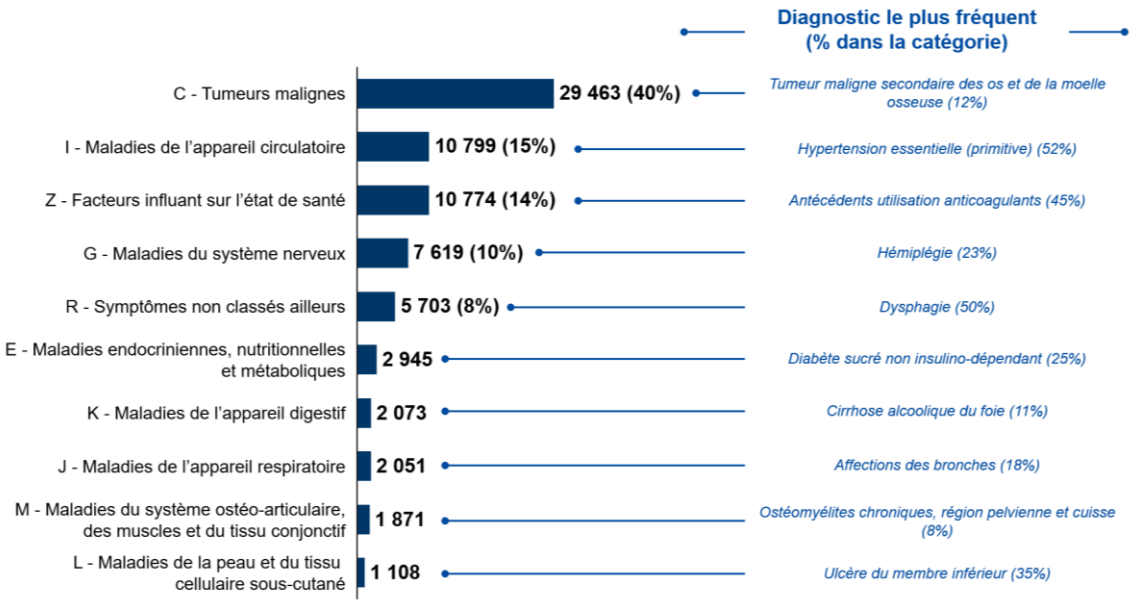


Figure 58 : Occurrence (valeur absolue et %) des 10 catégories majeures de diagnostics les plus fréquentes (gauche). Diagnostic le plus fréquent et % d'apparition dans chaque CMD (droite).

Pour conclure l'analyse exploratoire, les variables qualitatives caractérisant le mode de prise en charge et la catégorie majeure du diagnostic sont transposées en variables binaires (*one-hot encoding*).

Imputation des variables manquantes. Les colonnes correspondant à la séquence de soins du patient, c'est-à-dire les modes de prise en charge et l'indice de Karnofsky, ont une proportion de valeurs manquantes proche de 50%. Or, ce sont des variables d'intérêt pour le modèle, nous choisissons donc d'estimer les valeurs de ces variables au moyen d'un modèle d'apprentissage automatique.

MissForest est une méthodologie qui permet d'implémenter un algorithme d'imputation de valeurs manquantes en utilisant des forêts aléatoires, communément appelées *Random Forests*[211]. Cet algorithme est ensembliste et repose sur la combinaison de plusieurs arbres de décisions (cf. pseudo-code en *Annexe 6*) :

1. Sélection aléatoire d'un sous-ensemble d'échantillons et de variables à partir du jeu de données originel.
2. Construction d'un arbre de décision : pour la variable présentant des valeurs manquantes, l'algorithme découpe l'échantillon en deux groupes de manière à optimiser le gain d'information tiré à chaque embranchement[212]. Le processus de division se réitère jusqu'à ce qu'un critère d'arrêt soit rencontré (généralement la profondeur de l'arbre à atteindre, renseignée en hyperparamètre d'entrée).
3. Répétition des étapes 1 et 2 qui permettent d'aboutir à un ensemble d'arbres de décisions.
4. Prédiction de la valeur cible à partir d'un « vote majoritaire », c'est-à-dire la médiane des prédictions des arbres.

Cette méthodologie présente par ailleurs de meilleures performances par rapport à des méthodes classiques d'imputation et notamment dans des contextes où les relations entre variables sont complexes et non linéaires, et où la dimension des données est élevée[213].

Mesure de corrélation des variables numériques. La régression linéaire est une méthode statistique sensible à la corrélation de ses

[211] Présenté par (Stekhoven & Bühlmann, 2012).
[212] Selon (Dixneuf, 2019).
[213] Selon (Stekhoven & Bühlmann, 2012).

variables mesurées (ou encore appelées *variables indépendantes*[214]). Dans la mesure où l'introduction de variables corrélées est susceptible d'entraîner une instabilité des coefficients de régression et une imprécision dans les estimations des effets de chaque variable indépendante sur la variable dépendante, il est important de quantifier cette corrélation[215].

Le coefficient de corrélation de Pearson est une mesure répandue pour évaluer le degré de colinéarité entre deux variables numériques continues[216]. Ce coefficient est le rapport entre la variation conjointe des deux variables par rapport à leur moyenne respective (covariance) et le produit des écarts-types des variables, qui mesure leur dispersion autour de leur moyenne respective (cf. Équation iv ci-dessous). Il est utilisé pour déterminer si ces deux variables sont associées, et si oui, dans quelle mesure. Sa valeur est comprise entre -1 (corrélation négative parfaite) et 1 (corrélation positive parfaite). Un coefficient de 0 indique l'absence de corrélation.

Équation iv : Coefficient de corrélation de Pearson entre deux variables X et Y, de moyennes respectives $\bar{X}$ et $\bar{Y}$ sur un échantillon de taille n.

$$r = \frac{COV(X,Y)}{\sigma(X) * \sigma(Y)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} * \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

La librairie Seaborn de Python permet d'implémenter une matrice visuelle qui synthétise la valeur de ce coefficient pour chaque paire de variables du jeu de données.

[214] Les variables dépendante et indépendante sont des termes utilisés en statistiques et en sciences sociales pour désigner les types de variables d'une analyse. Les variables dites *indépendantes* sont celles dont on attend qu'elles influencent les variables dites *dépendantes*. A titre d'exemple : l'on souhaite déterminer si de fortes concentrations en gaz d'échappement ont un impact sur l'incidence de l'asthme chez les enfants. La concentration en gaz est la variable indépendante et l'asthme la variable dépendante. Définition extraite de la National Library of Medicine de la NIH, section « Statistiques de la santé ». Consultable en ligne nlm.nih.gov

[215] Selon (Thery et al., 1982).

[216] Définition et équations décrites dans (Schober et al., 2018).

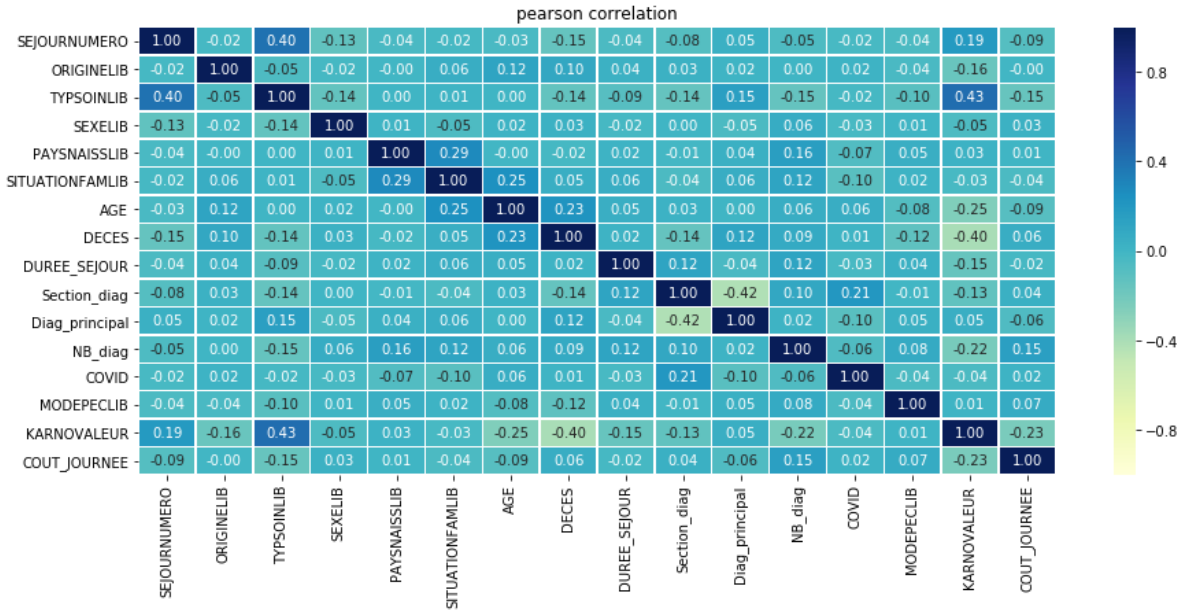


Figure 59 : Matrice de corrélation de Pearson pour chaque paire de variables numériques du jeu de données.

La Figure 59 ci-dessus ne montre pas d'évidence de corrélation linéaire entre nos variables numériques : la valeur absolue du coefficient de Pearson est inférieure à 0,43 pour toutes les associations en dehors de la diagonale. Nous décidons donc de toutes les conserver.

Normalisation du jeu de données. La normalisation est une méthode de pré-traitement des données qui est un préalable à certains algorithmes. Elle consiste à réajuster l'échelle des variables en les transformant de sorte qu'elles aient une échelle de comparaison commune.

Il est recommandé de normaliser les variables avant de réaliser une régression pour plusieurs raisons[217], notamment :

1. Améliorer la convergence de l'algorithme qui permet d'estimer les coefficients de régression,
2. Permettre la comparaison de l'effet relatif des variables indépendantes sur la variable dépendante via les coefficients de régression,
3. Réduire les effets des valeurs aberrantes qui sont observées sur plusieurs variables dans notre cas (cf. section précédente).

[217] A noter que cette méthode est à proscrire pour des algorithmes de réduction de la dimension (ex : Principal Component Analysis) et n'a pas d'effets pour des modèles basés sur des seuils tels que les arbres de décisions. Approches et impacts investigués dans (Singh & Singh, 2020).

Plusieurs méthodes de normalisation existent, on fait le choix d'utiliser la normalisation par l'amplitude interquartile (IQR)[218]. Cette méthode transforme les données de sorte que la différence entre le troisième et le premier quartile soit égale à un (cf. Équation v ci-dessous). En se basant sur l'écart interquartile et la médiane, elle est plus robuste aux valeurs aberrantes que des méthodes utilisant la moyenne et l'écart-type. Elle est implémentable en Python via la librairie RobustScaler de Scikit Learn.

Équation v : Normalisation par l'amplitude interquartile X_{norm} de la variable X, de premier quartile Q1(X) et de troisième quartile Q3(X)

$$X_{norm} = \frac{X - Q1(X)}{Q3(X) - Q1(X)}$$

Répartition des classes représentées dans les données d'entraînement et de test. Chaque classe à prédire représente un nombre de visites effectuées au domicile du patient pour un jour donné. Ce nombre est habituellement compris entre 0 (« pas de visites ») et 4 (« 4 visites par jour »), mais peut aller jusque 15 (cf. Table 25 ci-dessous).

Classe	Occurrence d'observations en J0
0	886 (3%)
1	18 744 (59%)
2	9 593 (30%)
3	2 140 (7%)
4	468 (1%)
5	93 (0%)
6	36 (0%)
7	15 (0%)
9	4 (0%)

[218] Méthode décrite dans (Gallón et al., 2013).

8	2 (0%)
15	1 (0%)

Table 25 : Occurrence et part en % du total d'observations par classe - exemple de la première journée de soins J0

On considère – en accord avec la cadre de santé responsable de l'élaboration du plan de soins – que les patients qui reçoivent 5 et plus visites de l'infirmière libérale par jour sont des « anomalies ». Celles-ci peuvent être dues au fait que plusieurs visites aient été dupliquées par l'intervenant car plusieurs actes de soins ont été réalisés au même moment. Le total de ces observations représentent moins de 1% des visites, on choisit donc de les supprimer. Par la suite, il faudra systématiquement s'assurer que le même nombre de classes est bien représenté à la fois dans les données d'entrainement et de test.

4.2.3 Description de l'approche deployee

Cette section présente les méthodes implémentées pour prédire le coût de journée et les visites au domicile du patient.

Poids relatif des variables indépendantes dans la prédiction du coût de journée. Le coût de journée et le nombre de visites quotidiennes d'un intervenant sont étroitement corrélés.

Les variables inductrices du coût journalier et du nombre de visites[219] sont donc susceptibles d'être les mêmes. Pour dégager les caractéristiques du profil patient susceptibles d'influencer elles aussi le coût journalier, nous appliquons un modèle d'apprentissage ensembliste, le *gradient boosting*[220]. Cet algorithme combine plusieurs modèles de prédiction simples, appelés également *weak learners*, pour améliorer la performance globale de la prédiction. Son fonctionnement repose sur une série de taille *N* prédéfinie d'initialisation, d'entraînement et de comparaison :

1. Initialisation d'un modèle de prédiction simple, en général un arbre de décision peu profond pour estimer la variable cible en fonction des variables d'entrée.

2. Calcul des résidus, c'est-à-dire de la différence entre chaque valeur réelle et prédite de la variable dépendante.

[219] N.B. Le nombre de visites par patient et par jour sera la variable finale à prédire.
[220] Approche décrite par (Mason et al., 1999).

3. Entraînement d'un nouveau modèle simple sur les résidus obtenus en étape 2. Cette étape permet d'obtenir une prédiction de l'erreur par variable. Ce second modèle est ajouté à l'existant pour donner une prédiction pondérée (« valeur prédite + erreur prédite »)

4. Répétition des étapes 2 et 3 jusqu'à ce qu'un critère d'arrêt soit atteint (ex : seuil pour l'erreur de validation ou nombre de modèles à entraîner prédéfini en hyperparamètre).

5. Prédiction finale à partir de l'agrégation des modèles simples.

Cet enchaînement est décrit dans le pseudo-code en *Annexe 7*.

La méthode feature_importances_ de la librairie ScikitLearn peut être appelée sur un modèle et renvoie un tableau de valeurs estimant l'importance relative de chaque variable indépendante dans la prédiction de la variable dépendante. La manière dont l'importance est évaluée varie selon le modèle d'apprentissage automatique utilisé. Dans l'approche que nous avons appliquée, l'importance est calculée en mesurant l'importance relative de chaque variable à la réduction de la somme des résidus. Il s'agit, pour chaque *weak learner*, de mesurer l'erreur avant et après la division d'un nœud de l'arbre en deux sous-groupes – puis de normaliser les contributions relatives pour que leur somme cumulée soit égale à 1. L'importance estimée de chaque variable est également un facteur de décision pour leur inclusion dans le modèle de prédiction des soins.

Prédiction des semaines de soins. Les jours de visites sont regroupés par semaine (de J0 à J6, puis J7 à J13, etc.). L'objectif est de prédire les visites quotidiennes par type d'intervenants, semaine par semaine. A chaque semaine terminée, on garde trace des visites réalisées lors des 7 derniers jours de visites et ceux-ci sont intégrés dans le jeu de données en tant que variables numériques. L'objectif est d'évaluer la pertinence des modèles de prédiction pour les scénario suivants :

- Scénario 1 – sans historique de visite : pour la première semaine de soins sans autres informations a priori que le profil patient,

- Scénario 2 – avec l'historique d'une semaine : pour la seconde semaine de soins sachant les visites effectuées la semaine précédente,

- Scénario 3 – avec l'historique d'une semaine, à un horizon temporal plus éloigné : pour la troisième semaine et plus sachant les visites effectuées lors de la première semaine. L'objectif est d'évaluer la stabilité de la performance à mesure que l'horizon temporel s'élargit.

Les variables à prédire sont des nombres entiers. On choisit de traiter le problème comme une tâche de classification multi-classe en quatre sous-groupes, représentant le nombre exact de visites possibles par jour par intervenant. Trois modèles sont sélectionnés parmi la littérature : *Support Vector Machine Classifier*, *Random Forest Classifier* et *Gradient Boosting Classifier*. Ces algorithmes seront entraînés, testés puis comparés dans la section suivante.

Support Vector Machine Classifier.

Cet algorithme, communément abrégé SVM, est un modèle de classification binaire répandu qui partitionne un échantillon de valeurs en deux sous-groupes[221]. La partition est effectuée par un hyperplan de dimension (M-1), M étant le nombre de variables d'entraînement. L'objectif de l'algorithme est de maximiser la marge, définie comme la distance entre l'hyperplan et les échantillons les plus proches de chaque classe. Ces derniers sont appelés les vecteurs de support et sont critiques pour la définition de la frontière de classification. L'hyperplan a pour équation $w.X + b = 0$, où w et b sont respectivement les vecteurs de poids et de biais et sont appris par l'algorithme (cf. pseudo-code en *Annexe 8*).

Plusieurs stratégies existent pour adapter ce modèle en classification multi-classe. On retiendra la méthode « One vs. One » [222] qui consiste à entrainer un classificateur pour chaque paire de classes distinctes soit 10 algorithmes[223] dans notre cas. La classification finale d'une instance est réalisée en sélectionnant la classe majoritaire parmi les prédictions de tous les classificateurs. Il est important de noter qu'il n'y a pas de garantie absolue qu'une classe majoritaire émerge, dans certains cas et particulièrement si les données présentent un chevauchement dans les classes, il est possible que plusieurs classes obtiennent le même nombre de prédictions. Le traitement de ce cas particulier peut être réalisé de plusieurs manières : par une hiérarchisation de l'importance des classes,

[221] Approche décrite par (Evgeniou & Pontil, 2001).
[222] Méthode décrite dans (Duan et al., 2007).
[223] 5 classes soit $\frac{5*(5-1)}{2} = 10$ paires de classes distinctes.

une sélection aléatoire de la classe gagnante parmi les majoritaires ou encore l'introduction d'une métrique de sélection supplémentaire. Nous choisissons cette dernière stratégie en raison d'un déséquilibre déjà présent dans la répartition des classes et pour éviter d'introduire un biais supplémentaire. En cas d'égalité, la prédiction présentant une confiance élevée, c'est-à-dire une distance à la frontière de décision la plus faible, sera sélectionnée.

Hyperparamètre	Définition	Impact	Intervalle usuel
Paramètre de régularisation C	Paramètre de tolérance pour les erreurs de classification	Plus C est élevé, plus la frontière de classification est complexe et le risque de surajustement est grand.	[0,1 ; 1 000]
Kernel	Fonction noyau utilisée pour transformer les données d'entrée en un espace de dimension supérieure	-	Radial Basis Function et Linear
Gamma (si Kernel = Radial Basis Function)	Plage d'influence des données d'entrainement sur les données de test	Une valeur élevée signifie que les données d'entrainement auront une forte influence, et que la frontière de décision sera plus complexe.	[0,1 ; 10]

Table 26 : Explicitation, impact et intervalle usuel des principaux hyperparamètres du Support Vector Machine Classifier.

Random Forest Classifier.

Cet algorithme de classification repose sur plusieurs arbres de décisions, chacun construit à partir d'un échantillon aléatoire des données d'entraînement couplé à un sous-groupe des variables d'entrée. Les prédictions individuelles des arbres sont ensuite agrégées et la classe majoritaire d'une instance donnée devient sa prédiction finale (cf. pseudo-code en *Annexe 9*).

Cet algorithme, introduit en 2001, présente de nombreux avantages par rapport à un arbre de décision simple[224]. En raison de sa nature agrégative, ce modèle réduit le risque de surentraînement et améliore donc la précision. Il est également plus robuste aux valeurs aberrantes et est en mesure de gérer des ensembles de données de dimensions élevées. Il perd en revanche en interprétabilité, les seuils de séparation variables des arbres individuels peuvent difficilement être concaténés. Les

[224] Présenté par (Breiman, 2001).

principaux hyperparamètres à optimiser sont explicités dans la Table 27 ci-dessous.

Hyperparamètre	Définition	Impact	Intervalle usuel
Nombre d'estimateurs	Nombre total d'arbres à utiliser pour construire le modèle	Un nombre élevé d'arbres augmente la précision mais aussi le temps d'entraînement.	[50 ; 5 000]
Profondeur de l'arbre	Profondeur maximale que chaque arbre est autorisé à atteindre	Une grande profondeur permet de capturer des relations plus complexes mais peut entraîner un surajustement.	[3 ; 10]
Sous-échantillonnage	Proportion de l'échantillon à utiliser pour chaque arbre	Des échantillons de grandes tailles réduisent le risque de surajustement mais aussi la précision	[0,5 ; 0,8]
Minimum de partition (nœud et feuille)	Nombre minimal d'observations nécessaires pour créer un embranchement (nœud et feuille)	La valeur du seuil minimal est un compromis entre la diminution du surajustement et la capacité à capturer des relations complexes	[1 ; 20]
Criterion	Fonction d'impureté utilisée pour évaluer l'homogénéité des classes dans chaque nœud	-	Entropie (cf. équation ci-dessous)

Table 27 : Explicitation, impact et intervalle usuel des principaux hyperparamètres du Random Forest Classifier.

L'objectif de l'algorithme est de sélectionner les variables et les seuils de divisions qui minimisent la fonction d'impureté globale sur l'ensemble des partitions. Plusieurs fonctions existent, mais l'entropie est généralement la mesure de choix lorsque les classes sont déséquilibrées.

Équation vi : Formule de l'entropie d'un nœud S mesurant l'impureté entre K classes.

$$H(S) = -\sum_{i=1}^{K} p_i log_2(p_i)$$

Où

- $S\ est\ un\ noeud\ donné\ de\ l'arbre\ et\ K\ le\ nombre\ total\ de\ classes$
- $p_i\ \ est\ la\ proportion\ d'observations\ de\ la\ classe\ i\ dans\ le\ noeud\ S$

L'entropie est comprise entre 0 : toutes les observations dans le nœud appartiennent à la même classe ; et 1 : toutes les classes sont représentées dans le nœud S.

Gradient Boosting Classifier.

Cet algorithme ensembliste fonctionne de manière similaire au *Gradient Boosting Regressor,* décrit dans le paragraphe précédent. Les principaux hyperparamètres à optimiser sont explicités dans la Table 28 ci-dessous.

Hyperparamètre	Définition	Impact	Intervalle usuel
Nombre d'estimateurs	Nombre total d'arbres à utiliser pour construire le modèle	Un nombre élevé d'arbres augmente la précision mais aussi le temps d'entraînement.	[50 ; 5 000]
Profondeur de l'arbre	Profondeur maximale que chaque arbre est autorisé à atteindre	Une grande profondeur permet de capturer des relations plus complexes mais peut entraîner un surajustement.	[3 ; 10]
Taux d'apprentissage	Contrôle la vitesse à laquelle l'algorithme apprend	Un apprentissage rapide diminue le temps d'entraînement mais rend l'algorithme plus sensible au bruit dans les données.	[0,01 ; 0,1]
Sous-échantillonnage	Proportion de l'échantillon à utiliser pour chaque arbre	Des échantillons de grandes tailles réduisent le risque de surajustement mais aussi la précision	[0,5 ; 0,8]
Minimum de partition (nœud et feuille)	Nombre minimal d'observations nécessaires pour créer un embranchement (nœud et feuille)	La valeur du seuil minimal est un compromis entre la diminution du surajustement et la capacité à capturer des relations complexes	[1 ; 20]
Fonction de perte	Type de fonction utilisée pour évaluer la qualité de la partition créée.	-	Déviance

Table 28 : Explicitation, impact et intervalle usuel des principaux hyperparamètres du Gradient Boosting Classifier.

Pour une tâche de classification multi-classes, plusieurs stratégies de fonction de perte sont possibles[225] :

1. « *One vs. Rest* » entraîne un classificateur binaire pour chaque classe et renvoie la classe associée à une prédiction de plus haute probabilité.

[225] L'algorithme de Gradient Boosting a été introduit par Friedman, qui présente, dans un article de 2002, les fonction de perte à utiliser en fonction du contexte d'implémentation (perte quadratique pour la régression, perte logarithmique pour la classification binaire et déviance pour la classification multi-classe). (Friedman, 2002)

Avantages : Facile à implémenter et performe correctement en présence de classes déséquilibrées en termes d'observations.
Inconvénients : Par construction, ne peut pas intégrer les relations entre classes. Si les données se chevauchent en termes de classe, c'est-à-dire qu'une observation peut appartenir à plusieurs classes ou que les classes sont corrélées, les résultats sont incohérents.

2. « *Multinomial* » entraîne un seul classificateur pour prédire les probabilités de toutes les classes simultanément. Les probabilités de chaque classe sont ensuite normalisées pour être interprétées comme une distribution de probabilité.
   Avantages : Plus précise que la stratégie précédente car intègre les effets entre classes.
   Inconvénients : Plus complexe en termes d'implémentation et nécessite plus de données pour être efficace.

Ces deux stratégies reposent sur la même fonction de perte, appelée déviance ou encore entropie croisée (cf. Équation vii ci-dessous). L'objectif de cette fonction est de minimiser la log-perte logarithmique entre les probabilités prédites et les étiquettes de classes réelles, ce qui donne une mesure de la qualité de la prédiction du modèle.

Équation vii : Formule de la déviance pour une instance de classe réelle y et de classe prédite avec une probabilité p.

$$D(y,p) = -2 * [\, y \log(p) + (1-y) \log(1-p)]$$

Pour une classification multi-classe, la notion de déviance est étendue et représente la somme de la déviance individuelle de chaque classe.

Équation viii : Formule de la déviance pour les matrices d'instances de classes réelles y et de classes prédites avec une probabilité p.

$$D(y,p) = -2 * \sum_{i=0}^{N} \sum_{j=0}^{K} y_{i,j} * \log(p_{i,j})$$

Où :

- $y_{i,j} = 1\ si\ i\ appartient$ à $la\ classe\ j, et\ 0\ sinon$
- $p_{i,j}\ \ est\ la\ probabilité\ prédite\ de\ l'appartenance\ de\ l'instance\ i$ à j

On choisira ici la stratégie multinomiale pour éviter la perte d'informations des relations interclasses.

Approche qualitative complémentaire. Pour donner de la profondeur à l'interprétation des résultats, nous décidons de mener une étude qualitative de suivi sur vingt patients de la structure de santé. Cette analyse complémentaire a pour objectif d'aider à la formulation d'hypothèses d'interprétation, mais aussi de comprendre comment les soignants interagissent avec et alimentent le système d'information AtHome.

Les 20 patients sont répartis entre 4 prises en charge. Ces prises en charge sont sélectionnées en fonction du nombre de séjours concernés (valeur seuil de 2 000 observations). La Figure 60 ci-dessous montre sept PEC éligibles. Nous décidons d'inclure la chimiothérapie (à domicile ou en surveillance post-chimio), les pansements complexes, les soins palliatifs et les traitements intraveineux ; ces PEC bénéficient à la fois d'un nombre suffisant d'observations tout en étant distinctes en termes de coût journalier moyen.

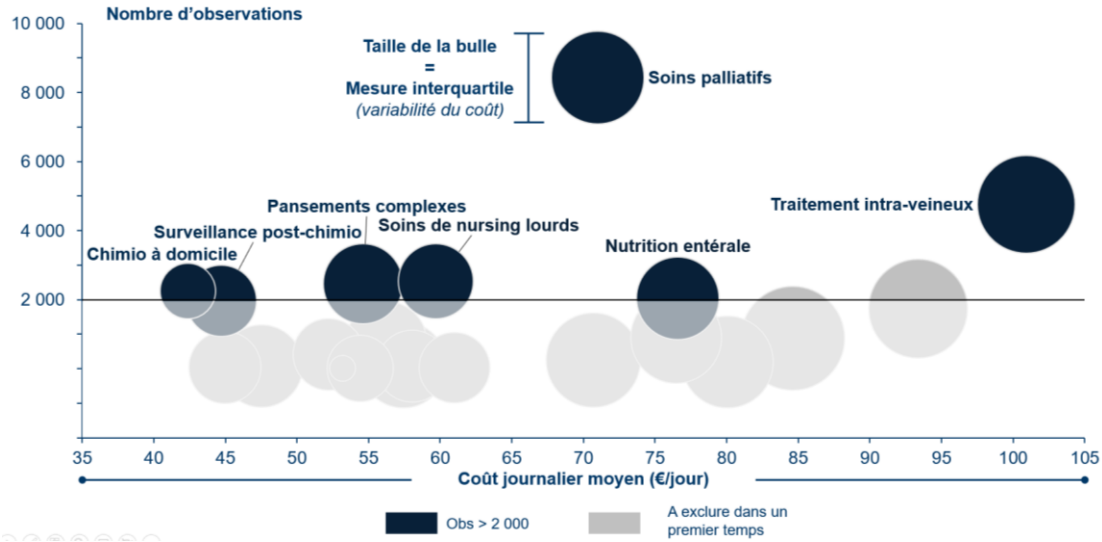


Figure 60 : Répartition des prises en charge parmi les séjours observés, en fonction du nombre de séjours et du coût journalier moyen constaté.

Les patients à inclure sont sélectionnés par la cadre de santé qui dirige la cellule d'éligibilité, au fur et à mesure de leur arrivée. Jusqu'à la fin de leur séjour ou pour une durée maximale de trois mois si celui-ci est long, on relève les variables cliniques identifiées dans la Figure 61 ci-dessous.

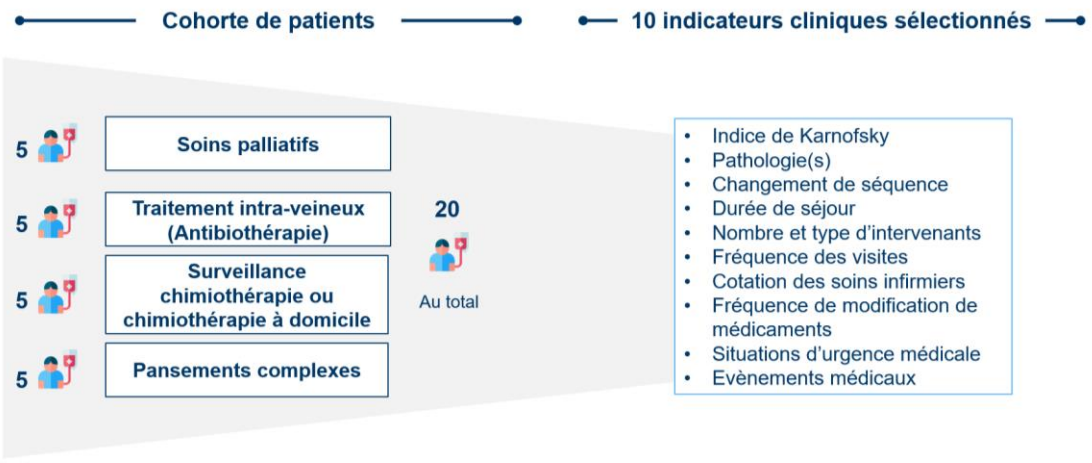


Figure 61 : Répartition initiale cible de la cohorte par PEC et indicateurs cliniques relevés.

Les données cliniques seront anonymisées et recueillies directement depuis le logiciel AtHome puis discutées avec les IDEC de suivi au cours d'entretiens hebdomadaires.

4.3 Résultats

SYNOPSIS Où nous décrivons les performances obtenues par les différents modèles testés dans les trois scénarios de données présentés dans la section précédente, ainsi que les apprentissages clefs retenus de l'étude qualitative des parcours patients.

4.3.1 PERFORMANCE DES ALGORITHMES

Cette section introduit les concepts de scores utilisés pour évaluer la performance des algorithmes. Nous explicitons ensuite ces scores pour les trois scénarios de prédiction présentés dans la section précédente : la première semaine de soins sans information à priori, puis la deuxième et la troisième semaines de soins sachant les visites effectuées au cours de la première semaine. Pour rappel, l'objectif est d'évaluer la robustesse du modèle dans trois scénarios de prédiction :

- Scénario 1 – sans historique de visite : pour la première semaine de soins sans autres informations a priori que le profil patient,
- Scénario 2 – avec l'historique d'une semaine : pour la seconde semaine de soins sachant les visites effectuées la semaine précédente,
- Scénario 3 – avec l'historique d'une semaine, à un horizon temporal plus éloigné : pour la troisième semaine et plus sachant les visites effectuées lors de la première semaine. L'objectif est d'évaluer la stabilité de la performance à mesure que l'horizon temporel s'élargit.

Evaluation de la performance des modèles de classification. Plusieurs scores peuvent être étudiés pour évaluer la performance d'un algorithme de classification multi-classe. On choisira de se concentrer sur les scores décrits ci-après, soit de manière globale pour toutes les classes, soit individuellement pour chaque classe.

Scores globaux.

L'exactitude, encore appelée précision globale ou *accuracy* en anglais, est une mesure qui évalue le pourcentage de prédictions correctes parmi le total d'observations[226]. Cette métrique est probablement l'une des plus répandues pour évaluer la performance d'un modèle. En revanche, lorsque les classes sont déséquilibrées, les prédictions peuvent être concentrées sur la classe majoritaire au détriment des autres[227]. Il est crucial de la combiner avec des scores individuels pour détecter ce phénomène.

Équation ix : Calcul de l'exactitude pour un algorithme

$$Exactitude = \frac{Vrais\ Positifs + Vrais\ Négatifs}{Nombre\ total\ d'observations}$$

Pour chaque prédiction, on mesure également la différence entre la classe prédite et la classe réelle[228]. Cette différence est sommée en valeur absolue et moyennée pour évaluer l'erreur absolue moyenne sur l'ensemble des observations.

Équation x : Calcul de l'erreur moyenne absolue sur l'ensemble des observations i de classe réelle y_i et de classe prédite $\hat{y}_i$

$$Erreur\ moyenne = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N}$$

Scores individuels.

La précision mesure la proportion d'observations réellement positives et qui ont donc été correctement labelisées parmi toutes les observations prédites comme positives (voir Équation xi ci-dessous). Une précision élevée et proche de 1 pour une classe donnée signifie qu'il y a peu de faux positifs. Cette mesure ne permet en revanche pas de savoir si la majorité des vrais positifs a bien été identifiée. Elle doit donc être combinée avec l'indicateur de rappel.

[226] A noter que dans les formules à suivre, on parlera d'observation positive pour une classe donnée si l'algorithme prédit que cette observation (ce séjour) appartient à cette même classe. Elle est négative si au contraire, on lui étiquette une classe différente. Par exemple, si l'algorithme prédit une visite par jour, alors l'observation est positive pour la classe « 1 visite/jour ». Si le patient a effectivement reçu une visite par jour, alors l'observation est un vrai positif ; dans le cas contraire, il s'agit d'un faux positif.

[227] Discussion abordée dans (Japkowicz & Shah, 2011).

[228] Dans notre cas, chaque classe est associée à une valeur numérique (le nombre de visites à réaliser). Nous définissons donc simplement la différence entre deux classes comme la différence entre le nombre de visites de chaque classe.

Équation xi : Calcul de la précision d'un algorithme pour une classe donnée i

$$Précision_i = \frac{Vrais\ Positifs_i}{Vrais\ Positifs_i + Faux\ Positifs_i}$$

Le rappel, communément appelé *recall* d'après l'anglais, évalue la proportion d'observations correctement labellisées positives parmi toutes les observations réellement positives (voir Équation xii ci-dessous). Une valeur de rappel élevée signifie qu'il y a peu de faux négatifs. Il s'agit donc d'équilibrer la précision et le rappel pour chaque classe.

Équation xii : Calcul du rappel d'un algorithme pour une classe donnée i

$$Recall_i = \frac{Vrais\ Positifs_i}{Vrais\ Positifs_i + Faux\ Négatifs_i}$$

Le F1-score d'une classe se traduit par la moyenne harmonique du rappel et de la précision de cette classe. Dans le cas où on cherche à maximiser à la fois la précision et le rappel et que les classes sont déséquilibrées, le F1-score est une excellente mesure de performance. Comme la précision et le rappel, ce score est compris entre 0 et 1 (valeur optimale). Un F1-score proche de zéro signifie que soit la précision est faible (le modèle prédit beaucoup de faux positifs), soit le rappel est faible (le modèle prédit beaucoup de faux négatifs), ou les deux.

Équation xiii : Calcul du F1-score d'un algorithme pour une classe donnée i

$$F1 - score_i = 2 * \frac{Recall_i * Précision_i}{Recall_i + Précision_i}$$

Tous ces scores seront utilisés en plusieurs occasions, sur les données d'entraînement pour optimiser et valider les hyperparamètres, puis sur les données de tests pour évaluer la robustesse du modèle selon les scénarios. Les données sont séparées aléatoirement en un jeu d'entraînement et un jeu de test (respectivement 80% et 20% des données originales[229]).

[229] Le ratio 80/20% est une règle empirique couramment utilisée, mais il n'existe pas de fondement mathématique qui justifie ce choix. L'objectif est de réaliser un compromis raisonnable entre la nécessité d'entraîner le modèle sur un ensemble de données représentatif tout en évitant le surapprentissage et une mauvaise généralisation à de nouvelles données. La nécessité de diviser le jeu de données et l'importance du compromis entre biais et variance est discuté dans (Hastie et al., 2009).

Les hyperparamètres de chaque algorithme sont optimisés au moyen d'une *Grid Search*[230], puis les modèles sont entraînés et évalués sur les métriques introduites ci-dessus par une validation croisée de 5 plis[231]. Les performances présentées par la suite sont les performances finales évaluées sur le jeu de test. Ces manipulations sont répétées pour chaque scénario de prédiction.

Prédiction de la première semaine de soins pour les IDE libérales. Les prédictions de ce premier scénario reposent sur les seules caractéristiques du patient connues à l'admission. On prédit ainsi successivement les visites requises de l'IDE au cours de la première semaine de soins (J0 à J6). Les tables complètes des scores pour chaque classe et algorithme sont disponibles en annexe.

La précision globale des algorithmes est relativement stable pour les sept premiers jours de soins (cf. Figure 62 ci-dessous). Les algorithmes Random Forest et Gradient Boosting affichent une performance similaire et supérieure au SVM (à titre d'exemple, respectivement de 0,61 et 0,62 vs 0,59 pour le premier jour de soins J0). La baisse de performance, pour les trois algorithmes, observée au J1 est probablement due à une répartition équilibrée des observations entre les classes « pas de visites » à « 3 visites par jour », contrairement à ce qui est observé dans les données d'entraînement et dans les données de tests pour les autres jours.

[230] L'algorithme de Grid Search est une technique d'optimisation courante qui permet de tester toutes les combinaisons possibles parmi une liste d'hyperparamètres spécifiée. Décrit dans (Bishop, 2006).
[231] La validation croisée k-plis est une technique d'évaluation répandue en apprentissage automatique. Elle consiste à diviser l'ensemble d'entraînement en k sous-ensembles (ou plis). Le modèle est ensuite entraîné k fois et la performance du modèle retenue est la moyenne des performances obtenues sur chaque pli. L'objectif est d'évaluer les capacités de généralisation du modèle à un semble de données différent. La performance moyenne obtenue sur les données d'entraînement peut ensuite être comparée à la performance finale sur les données de test. (Kohavi, 1995).

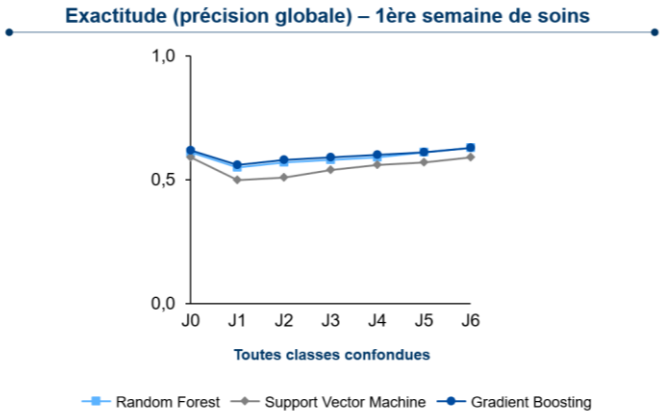


Figure 62 : Evolution du score de précision globale en fonction des jours pour les trois modèles testés

L'étude du F1-score confirme que les performances du Gradient Boosting sont meilleures pour tous les types de classes (cf. Figure 63 ci-dessous). On remarque que pour les classes majoritaires, c'est-à-dire comportant le plus d'observations, le F1-score oscille entre 0,76 et 0,9 à la fois pour le Gradient Boosting et le Random Forest. En revanche, pour les classes comportant moins d'observations, les performances se dégradent sensiblement (par exemple pour 4 visites par jour, le F1-score est de 0 pour le SVM et ne dépasse pas 0,07 pour le Gradient Boosting).

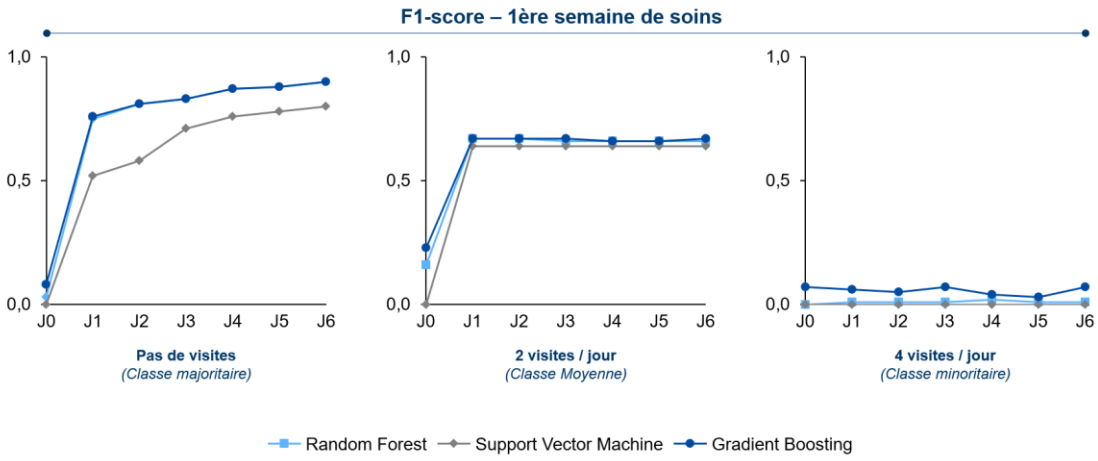


Figure 63 : Evolution du F1-score en fonction des jours pour les trois modèles testés, exemple de 3 classes "pas de visites" (à gauche), "2 visites par jour" (au milieu), "4 visites par jour" (à droite).

Les performances dégradées des classes minoritaires proviennent notamment d'un score de rappel proche de zéro pour tous les algorithmes (voir Figure 64 ci-dessous) alors même que la précision varie entre 0,62 et 0,82 pour le Gradient Boosting à titre d'exemple (voir Figure 65 ci-dessous). Une précision élevée avec un rappel approximativement nul indique un très petit nombre de faux positifs, comparés aux vrais positifs.

Par ailleurs ici, ces classes comptent très peu d'observations (e.g. la classe « 4 visites par jour »). Lorsque l'algorithme rencontre une observation courante qui n'appartient pas à cette classe, il prédit correctement la « non appartenance ». Ce phénomène a tendance à rester vrai, même pour les observations plus rares qui appartiennent effectivement à la classe minoritaire.

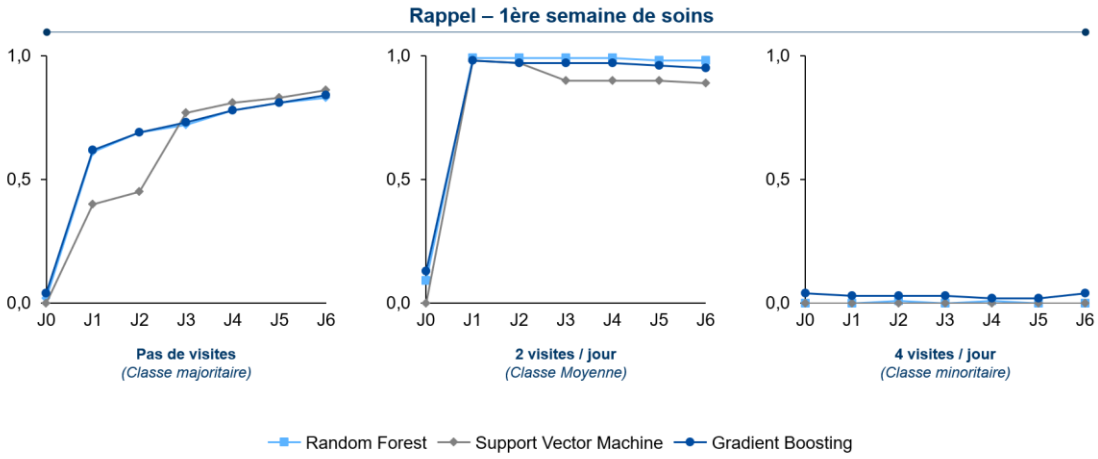


Figure 64 : Evolution du rappel en fonction des jours pour les trois modèles testés, exemples de 3 classes "pas de visites" (à gauche), "2 visites par jour" (au milieu), "4 visites par jour" (à droite).

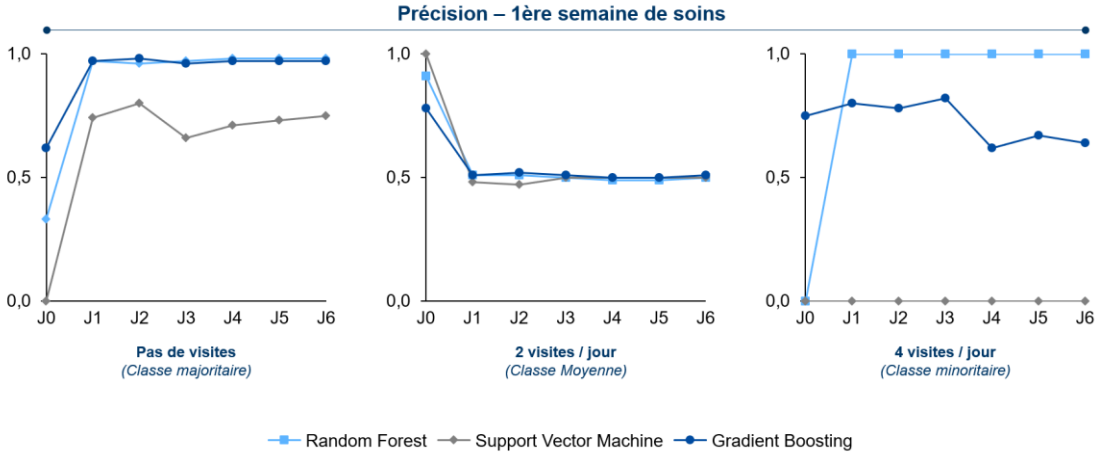


Figure 65 : Evolution de la précision en fonction des jours pour les trois modèles testés, exemples de 3 classes "pas de visites" (à gauche), "2 visites par jour" (au milieu), "4 visites par jour" (à droite).

Ces phénomènes se retrouvent également dans les matrices de confusion. Ces tables permettent de visualiser les valeurs réelles (en ligne) vs. les valeurs prédites (en colonne) des classifications. On note dans la Figure 66 ci-dessous que l'algorithme Gradient Boosting est le plus à même d'identifier les quatre classes, avec toutefois une tendance à labelliser les observations dans la classe « 1 visite par jour ».

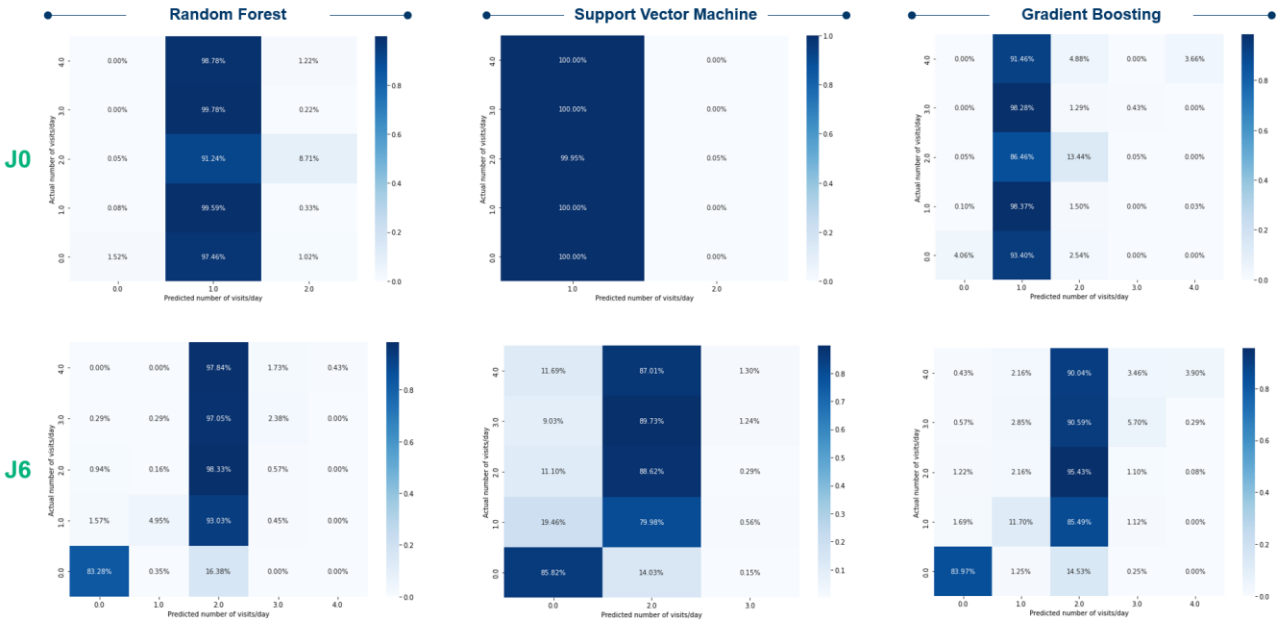


Figure 66 : Matrice de confusion (valeur réelle vs valeur prédite) pour chaque algorithme. Random Forest (à gauche), SVM (au milieu), Gradient Boosting (à droite). Deux exemples de prédictions sont montrés : J0 (en haut) et J6 (en bas).

La matrice de confusion permet de confirmer que la qualité globale de prédiction est plutôt moyenne pour les trois algorithmes, malgré une performance légèrement meilleure pour le Gradient Boosting. Pour les scénarios suivants, ce dernier sera le modèle préférentiel pour les prédictions et la présentation des résultats.

Prédiction de la deuxième semaine de soins pour les IDE libérales.
Les prédictions de ce second scénario reposent à la fois sur les caractéristiques du profil patient connues à l'admission, mais également sur les visites effectuées lors de la première semaine de soins. On ajoute donc 7 variables au jeu de données correspondant au nombre de visites reçues de J0 à J6.

En termes de performances globales, les scores sont nettement améliorés par l'ajout de l'historique des visites effectuées (cf. Figure 67 ci-dessous). Cette amélioration bénéficie à toutes les classes (comme on peut le noter dans la Figure 68 ci-dessous). A titre d'exemple, le gain de précision globale est de 23% entre J0 et J7 (respectivement 0,62 vs 0,85). L'erreur moyenne absolue elle aussi diminue de 0,49 à 0,19, soit de plus de la moitié.

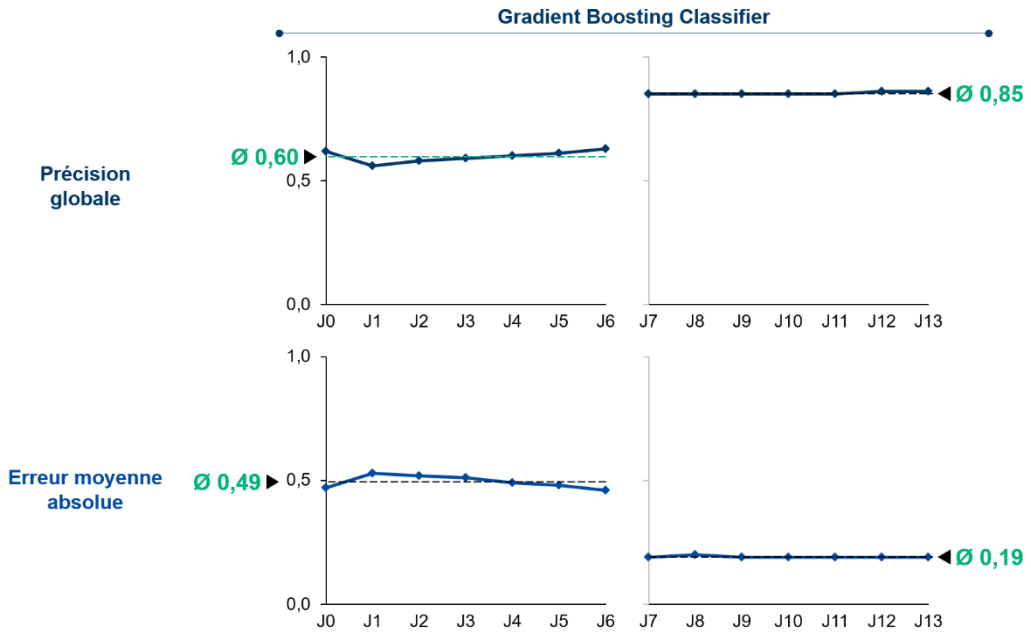


Figure 67 : Evolution des scores de performance du Gradient Boosting Classifier (toutes classes confondues) et en fonction des jours entre la 1ère semaine de soins (graphiques de gauche) et la 2ème semaine de soins (graphiques de droite). Deux scores de performance sont présentés : précision globale (graphiques du haut) et erreur moyenne absolue (graphiques du bas).

Les scores de précision par classe sont similaires et dans des fourchettes de valeurs moins variables que pour la première semaine de soins (cf. graphique du milieu dans la figure ci-dessous). On retrouve cependant l'impact de la classe minoritaire avec un score de recall pénalisé par un faible nombre d'observations.

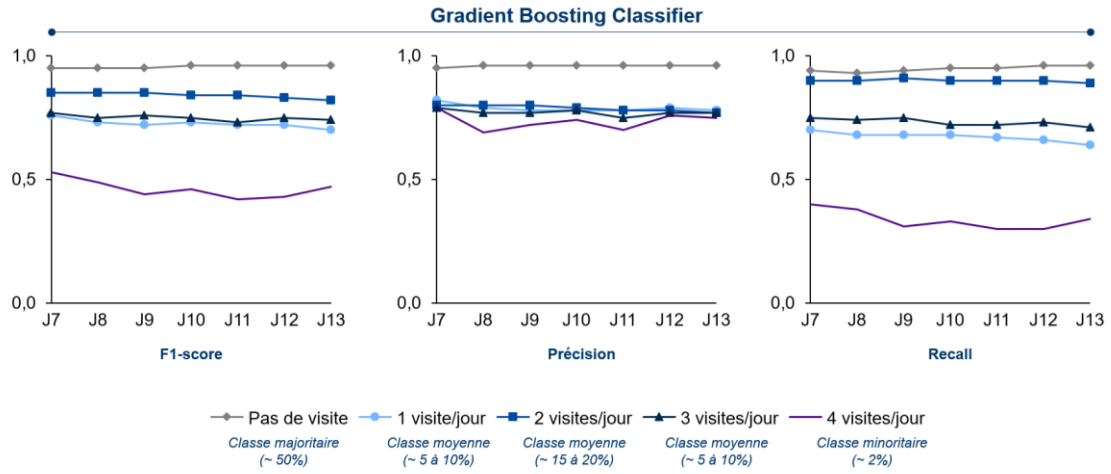


Figure 68 : Evolution des scores de performance du Gradient Boosting Classifier par classe et en fonction des jours (2ème semaine de soins). F1-score (à gauche), Précision (au milieu), Recall (à droite).

Les matrices de confusion sont par ailleurs plus équilibrées, comme on peut le noter sur les diagonales valeurs réelles vs. valeurs prédites de la figure ci-dessous.

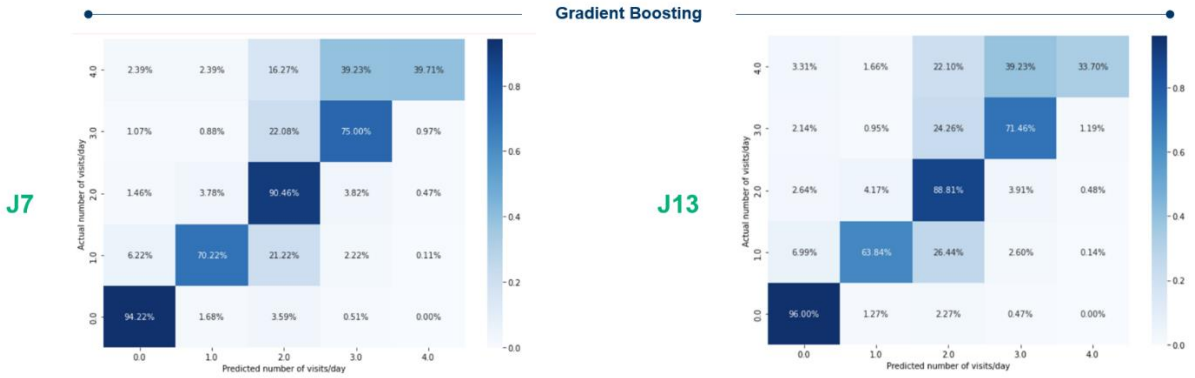


Figure 69 : Matrice de confusion (valeur réelle vs valeur prédite) pour le Gradient Boosting Classifier. Deux exemples de prédictions sont montrés : J7 (en haut) et J13 (en bas).

Prédiction de la troisième semaine de soins et plus pour les IDE libérales. Les prédictions de ce second scénario reposent à la fois sur les caractéristiques du profil patient connues à l'admission, mais également sur les visites effectuées lors de la première semaine de soins. Comme pour le scénario précédent, on ajoute 7 variables au jeu de données correspondant au nombre de visites reçues de J0 à J6. L'objectif de ce scénario est d'évaluer si les performances de prédictions restent stables à un horizon temporel supérieur : de J14 à J20, puis J27 et enfin J34.

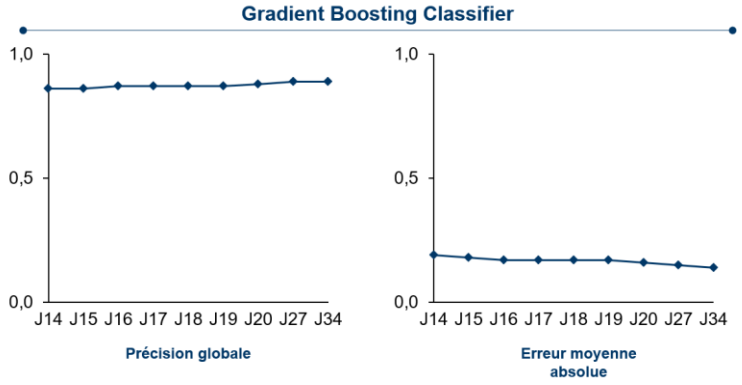


Figure 70 : Evolution des scores de performance du Gradient Boosting Classifier (toutes classes confondues) et en fonction des jours (3ème semaine de soins, fin de la 4ème semaine et fin de la 5ème semaine). Précision globale (à gauche) et Erreur Moyenne Absolue (à droite).

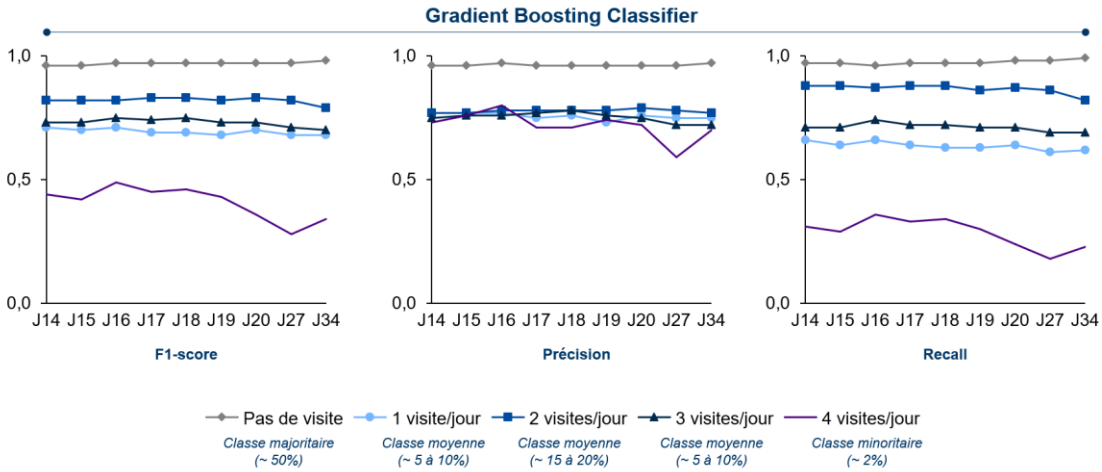


Figure 71 : Evolution des scores de performance du Gradient Boosting Classifier par classe et en fonction des jours (3ème semaine de soins, fin de la 4ème semaine et fin de la 5ème semaine). F1-score (à gauche), Précision (au milieu), Recall (à droite).

On note que les scores à la fois globaux et spécifiques à chaque classe ne subissent pas d'évolution significative, à l'exception du rappel et donc du F1-score pour la classe minoritaire « 4 visites par jour » qui diminuent de 10% en moyenne entre J14 et J34.

4.3.2 Resultat de l'etude qualitative

Cette section présente une analyse qualitative complémentaire qui a pour objectif d'aider à la formulation d'hypothèses d'interprétation, mais aussi de comprendre comment les soignants interagissent avec et alimentent le système d'information AtHome.

Synthèse des parcours suivis. Vingt patients ont été initialement sélectionnés par la cadre de santé de la cellule d'éligibilité pour inclusion dans l'étude qualitative. Un patient s'est finalement vu refuser l'admission en HAD par l'équipe médicale et a été redirigé vers le Service de Soins Infirmiers à Domicile (SSIAD), qui intervient pour des soins infirmiers moins complexes. Ce patient n'a pas été remplacé dans l'étude, ce qui ramène à dix-neuf le nombre total de parcours de soins observés (cf. Figure 72 ci-dessous). Le critère d'inclusion principal dans l'étude est le mode de prise en charge pour obtenir une répartition équilibrée entre les traitements intra-veineux, les pansements complexes, la chimiothérapie et les soins palliatifs.

L'étude comprend huit femmes et onze hommes, âgés de 23 à 94 ans avec une moyenne d'âge de 68 ans (vs. 67,7 pour la cohorte totale) et un indice de Karnofsky à l'admission compris entre 70 et 20 (moyenne de 50 vs. 46,6 pour la cohorte totale). Onze patients ont été admis avec pour diagnostic principal une tumeur maligne (CMD C), cinq patients pour une infection (CMD M), et trois pour un traumatisme (CMD S ou T).

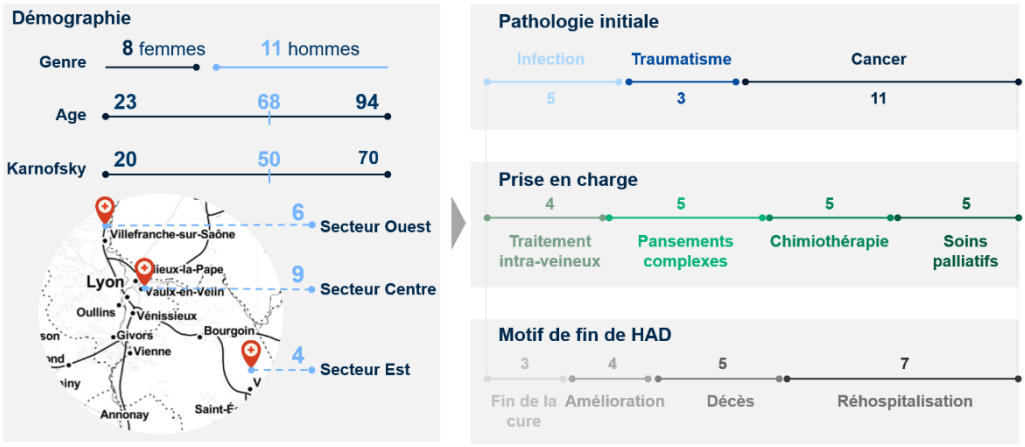


Figure 72: Synthèse de l'étude qualitative des parcours de soins en HAD. Présentation des caractéristiques démographiques, diagnostic principal à l'admission, motifs de prise en charge principal et motifs de fin d'HAD.

Les motifs de fin de séjour d'HAD sont en premier lieu une réhospitalisation en urgence (sept patients), le décès du patient (pour cinq d'entre eux), une amélioration de l'état clinique (quatre patients) ou encore l'arrêt planifié du traitement (trois patients).

Le détail des parcours est listé dans la Table 29 ci-dessous. Les noms des patients ont été modifiés pour garantir l'anonymat.

Patient	Sexe	Age	MPP et MPA	IK	Num séjour	Diagnostic principal	Motif de fin	Durée
Emilie	F	53	Antibiothérapie	50	2	Sepsis chronique de la hanche gauche	Amélioration état	7 jours
Adèle	F	61	Antibiothérapie	50	1	Osthéoartrite	Amélioration état	9 jours
Henri	M	65	Pansements complexes	60	1	Infection post-pontage	Amélioration état	13 jours
Alexandre	M	59	Antibiothérapie	50	1	Choc septique sur bactériémie	Amélioration état	38 jours
Renée	F	94	Pansements complexes et soins palliatifs	50	1	Tumeur frontale, sourcil et nez (plaie)	Décès	20 jours

Agathe	F	75	Soins palliatifs	40	1	Glioblastome pariéto-occipital gauche	Décès	44 jours
Léa	F	82	Soins palliatifs	30	3	Glioblastome du carrefour ventriculaire gauche	Décès	122 jours
Louise	F	94	Soins palliatifs	40	1	Tumeur maligne du pancréas	Décès	130 jours
André	M	81	Surveillance chimiothérapie	40	1	Glioblastome fronto-pariétal droit	Décès	55 jours
George	M	69	Chimiothérapie à domicile	70	1	LMMC I (leucémie)	Fin cure	91 jours
Edgar	M	55	Chimiothérapie et soins palliatifs	60	9	Tumeur rectale avec métastases pulmonaires	Fin cure	4 jours
Léon	M	63	Chimiothérapie à domicile	70	1	Leucémie	Fin cure	3 jours
Jacques	M	60	Antibiothérapie	20	1	Infection pulmonaire et urinaire	Urgences	7 jours
Gabriel	M	23	Pansements complexes et antibiothérapie	50	1	Fracture complexe du calcanéum	Urgences	13 jours
Luc	M	67	Soins palliatifs	30	1	Carcinome pulmonaire	Urgences	29 jours
Maurice	M	78	Pansements complexes et antibiothérapie	50	1	Escarre trochantérienne gauche	Urgences	16 jours
Jeanne	F	50	Surveillance chimiothérapie et soins palliatifs	40	3	Tumeur rectale	Urgences	10 jours
Charles	M	80	Pansements complexes	50	1	Amputation, cicatrice avec nécrose	Urgences	57 jours
Victoria	F	90	Soins palliatifs	40	1	Tumeur des bronches	Urgences	10 jours

Table 29 : Synthèse des parcours de soins observés, caractéristiques démographiques, séquence, diagnostic principal, motif de fin de séjour et durée des soins.

Les entretiens hebdomadaires avec les IDEC de suivi et les médecins coordonnateurs ont permis de mettre en lumière des problématiques dans la remontée des données dans le système et des enseignements qui peuvent expliquer certaines prédictions erronées de l'algorithme. Ces problématiques sont structurées en trois piliers :

1. Economique : les prises en charge les plus coûteuses sont celles liées à une chimiothérapie qui implique des molécules onéreuses (ex : Vidaza) et certains soins complexes, comme la thérapie VAC. Les données de facturation des soins ne permettent pas entièrement d'identifier les prises en charge les plus coûteuses.

2. Organisationnel :

   a. Certains patients devraient bénéficier de soins supplémentaires mais ne peuvent pas en raison de l'offre disponible localement.

   b. Les PEC jugées les plus complexes par les soignants sont celles associées à une difficulté d'interface avec l'hôpital et les intervenants à domicile, ou encore la mauvaise acceptation des soins par le patient ou les aidants. Cette complexité peut avoir un impact sur le nombre de visites requises mais ne se reflète pas a priori dans le profil patient.

3. Limites des données

   a. L'amélioration de l'état du patient qui souvent met fin au séjour est peu caractérisée et renseignée concrètement dans le système d'information et donc ne se reflète pas dans les données.

   b. Les modes de prises en charges principaux et associés ne représentent pas toujours la réalité des soins. Ces informations, qui caractérisent la séquence de soins, sont renseignées a posteriori par le médecin DIM. Elles reflètent une réalité économique qui est corrélée avec les soins réalisés mais aussi d'autres paramètres, e.g. la durée de séjour. A l'inverse, il peut y a avoir un changement dans les soins réalisés à domicile (par exemple, le patient contracte une infection qui requière une adaptation de son plan de soins) sans que cela entraîne une modification systématique de la séquence.

c. Les changements de fréquence de visites à domicile ne sont pas systématiquement corrélés à un changement de séquence ou à une évolution de l’indice de Karnofsky, mais plutôt à un changement dans le traitement. Pour aller plus loin, il serait nécessaire de compléter les données en incluant les traitements reçus par le patient.

4.4 Discussion et perspectives

SYNOPSIS Où nous explicitons les variables qui influencent les prédictions de l'algorithme, discutons des forces et faiblesses de l'analyse et ouvrons l'horizon des perspectives de recherche.

4.4.1 IMPACT RELATIF DES VARIABLES INDEPENDANTES DANS LES PREDICTIONS DU COUT DE JOURNEE ET DU NOMBRE DE VISITES

Seize variables ont une importance relative supérieure à 1% (cf. Figure 73 ci-dessous). La variable prépondérante dans l'explication du coût journalier est l'indice de Karnofsky, à hauteur de 15,5%. On retrouve ensuite un certain nombre de modes de prise en charge : traitement intra-veineux, nutrition parentérale, surveillance post-chimiothérapie anticancéreuse et soins palliatifs parmi les plus importants. Parmi les variables numériques : le nombre de diagnostics associés, le nombre de séjours, l'âge et l'évolution de l'indice de Karnofsky viennent influencer le coût. Enfin, certaines catégories majeures de diagnostics apparaissent comme des inducteurs économiques dans une moindre mesure (entre 5 et 1%) : maladies de l'appareil digestif (K), du système nerveux (G), endocriniennes (E), du système ostéoarticulaire (M) et les tumeurs malignes (C).

Ce même modèle appliqué au nombre de visites quotidiennes renvoie certaines similarités dans l'importance des variables, mais à des niveaux d'impact différents (cf. Figure 74 ci-dessous). A titre d'exemple l'indice de Karnofsky est classé 5ème avec un impact évalué à 6,1% (vs rang 1 et 15,5% pour le coût). Pour la prédiction des visites, l'âge est la variable prépondérante à 19,7%, ainsi que le nombre de jours de soins. Il apparaît que plus le séjour est long, moins le nombre de visites est important. On retrouve le nombre de diagnostics associés et l'évolution de l'indice de Karnofsky en cours de séjour avec des niveaux d'impact similaires. Les catégories majeures de diagnostics impactantes sont les mêmes, à l'exception de la catégorie I qui fait son apparition.

On note par ailleurs que la situation familiale influe sur le nombre de visites, contrairement au coût. La présence d'un aidant à domicile, qui est fortement corrélée à la variable « Marié(e) », est en effet susceptible de faire diminuer le nombre de visites requis au domicile. On note également une légère influence du sexe du patient. Deux hypothèses peuvent expliquer cet impact :

- Les femmes ont, en moyenne, une espérance de vie plus longue, elles ont une probabilité plus élevée de survivre à leur conjoint et donc d'être isolée à leur domicile sans aidant. Cette hypothèse est renforcée par le fait que les femmes sont sur-représentées dans la catégorie « Veuf(ve) ».
- Le genre du patient peut être corrélé au diagnostic principal pour lequel le patient est admis.

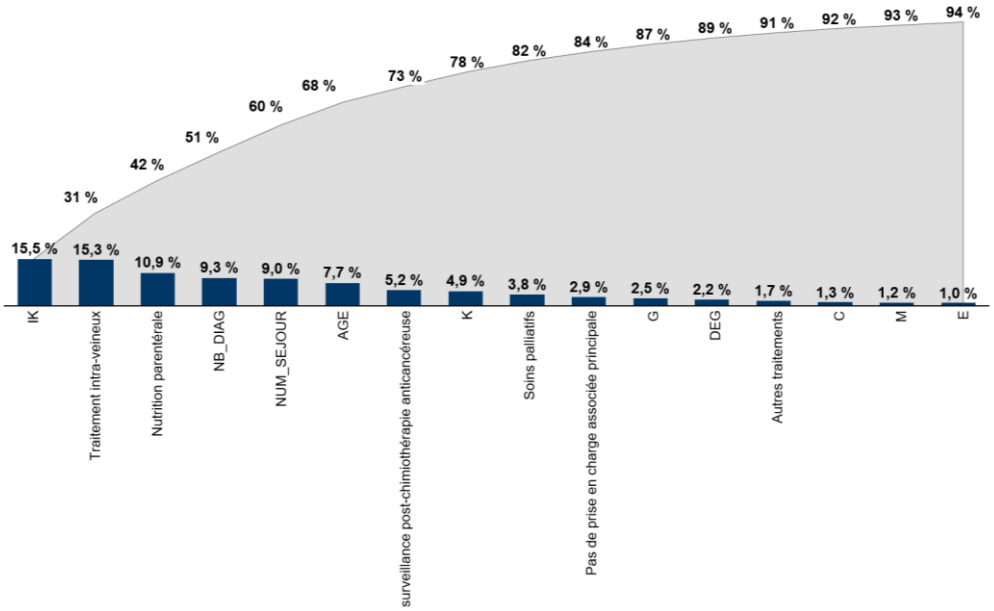


Figure 73 : Importance relative (en bleu) et cumulée (courbe grise d'arrière-plan) des variables indépendantes dans la prédiction du coût journalier. Seules les variables dont l'importance relative est supérieure à 1% sont représentées.

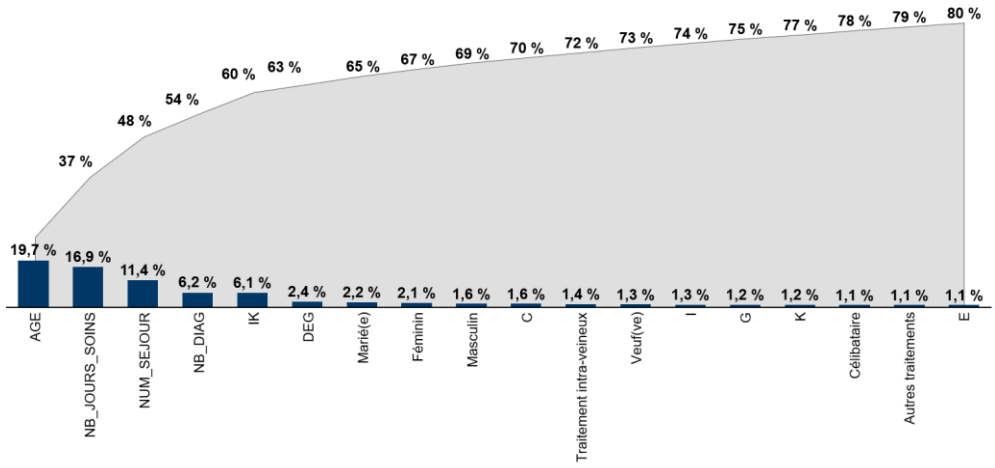


Figure 74 : Importance relative (en bleu) et cumulée (courbe grise d'arrière-plan) des variables indépendantes dans la prédiction du nombre de visites pour J0. Seules les variables dont l'importance cumulée permet de couvrir 80% de l'impact sont représentées.

On note par ailleurs que les variables et leur importance relative restent cohérentes au cours du temps de J0 à un horizon temporel plus élevé.

4.4.2 Forces et faiblesses de l'etude et perspectives futures de recherche

L'interprétation des résultats des prédictions de la première semaine de soins nous permet de conclure que les caractéristiques du patient connues à l'admission et renseignées dans le système de soins ne permettent pas d'expliciter entièrement le nombre de visites requises à domicile. Plusieurs facteurs importants sont susceptibles d'influencer les soins et ne sont pas représentés dans les données, par exemple les traitements prescrits au patient, l'acceptabilité des soins ou encore un changement de situation au domicile. Certains diagnostics peuvent apparaître en cours de séjour, par exemple une infection qui nécessiterait une hospitalisation en urgence. Ceux-ci sont généralement mentionnés dans les notes cliniques mais ne seront pas ajoutés à la liste des diagnostics associés.

L'ajout de la première semaine de soins permet cependant de renforcer la robustesse des prédictions, et ce même à un horizon temporel de plusieurs semaines. On ne remarque en effet pas de baisse de performance significative entre les prédictions de la deuxième semaine de soins et celles de la troisième. Le modèle Gradient Boosting est l'algorithme qui permet à la fois de générer des prédictions plus robustes mais également de représenter la variété de classes. Les classes minoritaires restent pour autant fortement défavorisées et ce dans tous les scénarios de prédiction. Cette pénalisation reste cependant acceptable car elle ne concerne qu'approximativement 2% des séjours.

En implémentation en vie réelle, l'algorithme nécessiterait une prise en main par un à plusieurs membres de l'équipe soignante[232]. Cet aspect est crucial pour garantir l'acceptation et l'efficacité du modèle, et doit passer par :

- La sensibilisation : la vulgarisation du fonctionnement de l'algorithme, les forces et faiblesses du modèle et la finalité de l'exercice ainsi que sa valeur ajoutée.
- La formation : bien que l'équipe soignante n'ait pas besoin de comprendre l'ensemble des détails techniques de l'algorithme, il est important de transmettre une compréhension générale des

[232] De nombreux articles abordent la question de l'implémentation d'un modèle en vie réelle dans la pratique clinique. La prise en main et la formation des utilisateurs soignants d'une clinique neurologique sont notamment discutées dans (Bruun et al., 2019).

prédictions. Un levier puissant peut être la formation pratique avec des sessions interactives qui leur permette de tester l'outil.

- Le support documentaire : des ressources bien documentées, de forme écrite (une notice d'utilisation) ou vidéo (un enregistrement d'une formation) peuvent offrir un support d'utilisation continu.
- Le feedback : l'algorithme doit continuer à vivre avec de nouvelles données et un retour structuré de l'équipe soignante sur la qualité des prédictions fournies.

Pour renforcer l'interprétabilité du modèle, des vérifications manuelles, effectuées prédiction par prédiction, pourront être nécessaires. Il est notamment possible d'intégrer à la méthode des techniques d'interprétation locale, telles que les valeurs de Shapley[233]. L'interprétation de ces valeurs par l'équipe soignante n'est possible que si la prise en main a bien été effectuée. Une étude future pourra suivre l'implémentation du modèle en vie réelle et évaluer la robustesse de l'algorithme face à de nouvelles données, ainsi que la confiance portée par les soignants dans les prédictions.

[233] SHAP (SHapley Additive ExPlanations) est une approche utilisée en apprentissage automatique nommée d'après Lloyd Shapley, lauréat du prix Nobel d'économie. La valeur de Shapley est une mesure de l'importance de chaque variable dans la prédiction d'une instance donnée. L'importance de chaque caractéristique du modèle est déterminée d'après les étapes suivantes :

1. Pour une prédiction donnée, les variables et leurs valeurs sont divisées en sous-ensembles appelés « ensembles de coopération ». Il en existe autant que de sous-ensembles possibles de variables.
2. Pour chaque ensemble de coopération, on évalue la contribution marginale de chaque variable dans la prédiction. Il s'agit de la différence entre la prédiction du modèle avant et sans la caractéristique d'intérêt (une prédiction inchangée signifie que la variable a une probabilité faible d'être corrélée au résultat final).
3. La valeur de Shapley d'une variable donnée est la moyenne de ses contributions marginales sur tous les ensembles de coopération possibles.

Pour une prédiction donnée, on est donc en mesure d'identifier les variables qui ont influencé la prise de décision de l'algorithme. Il est cependant à noter que le calcul des valeurs de Shapley est coûteux en termes d'intensité de computation. Par ailleurs, il reflète le poids relatif des variables dans une prédiction du modèle, qui peut être erronée.

5

Conclusion et perspectives futures de recherche

Ce cinquième chapitre met en perspective les travaux de thèse par rapport à la littérature scientifique existante, discute des limitations de l'étude, propose des perspectives de recherche futures et dresse un bilan au regard des questions soulevées dans les chapitres 1 et 2.

Contenu

5.1 Conclusion et mise en perspective des travaux

Ces travaux de recherche ont été animés par l'investigation d'une problématique principale, introduite dans le premier chapitre :

« COMMENT CARACTERISER ET ANTICIPER LES PARCOURS PATIENTS ET DE SOINS SOUS L'ANGLE DES COUTS ? »

L'enjeu essentiel est d'apporter des pistes de résolution à des difficultés rencontrées par les acteurs de santé dans l'organisation des soins, en particulier pour les patients atteints de maladies chroniques. L'optimisation des parcours patients est un levier d'efficience puissant pour comprendre et anticiper les facteurs et événements impactant le recours et la consommation des soins. L'objectif est de soulager les pressions opérationnelles des ressources hospitalières tout en améliorant la qualité et le confort de la prise en charge des patients.

Nous avons proposé une approche basée sur l'intelligence artificielle en plusieurs étapes : estimation du coût du parcours, classification des patients en sous-groupes médico-économiques cohérents entre eux puis modélisation et prédiction du parcours. Cette approche a été adaptée et testée sur deux cas d'application : la prédiction du parcours patient pour des prises en charge de troubles neurocognitifs, et la prédiction du parcours de soins pour des séjours en hospitalisation à domicile. Elle a été déclinée de la manière suivante sur ces deux cas d'application :

1. Prédiction du parcours patient, exemple des troubles neurocognitifs : l'ensemble des parcours a été projeté en deux dimensions avec l'algorithme *t-SNE* ; les parcours ont été segmentés par coût et type de consommations grâce à une classification non supervisée, pour laquelle quatre algorithmes ont été évalués (K-means, K-médoïds, HDBSCAN et BIRCH). BIRCH a permis d'obtenir à la fois un score de silhouette et un coût moyen de transition intra-cluster plus performants ; puis la probabilité de transition entre les clusters a été modélisée au moyen d'une chaîne de Markov.

2. Prédiction du parcours de soins, exemple de l'hospitalisation à domicile : les parcours de soins sont déjà regroupés en groupes médico-économiques communs au moyen de la séquence de soins (MPP, MPA et IK) ; les parcours de soins et leur coût ont été recomposés à partir des visites quotidiennes au domicile du patient ; trois modèles d'apprentissage automatique classiques (SVM, Random Forests et Gradient Boosting) ont été évalués pour prédire le nombre de visites requis sous la forme d'une classification multi-classes ; Gradient Boosting est l'algorithme permettant d'obtenir les meilleures performances, sa performance a été évaluée sur plusieurs scénarios de prédiction (première semaine de soins sans information a priori sur le plan de soins, deuxième de soins avec l'historique de visites de la première semaine et troisième semaine de soins avec un historique de visites éloigné dans le temps).

L'analyse de la littérature scientifique existante avait permis d'identifier des apports à couvrir sur trois axes principaux :

1. La sélection des variables d'intérêt : il s'agit intégrer plusieurs indicateurs cliniques qui permettent de caractériser la progression de l'état du patient et d'interpréter des variations dans les coûts de prise en charge ou les trajectoires ; il convient également d'inclure les interventions et soins réalisés dans l'analyse des parcours.

2. Les choix méthodologiques : il semble en particulier important de veiller à choisir une méthode de classification qui prenne à la fois en compte une répartition déséquilibrée des observations entre les classes et qui puisse labelliser les observations « bruit » comme telles ; par ailleurs, il convient de combiner des approches d'estimation des coûts macro et micro pour aboutir à une évaluation exhaustive.

3. Une interprétation et une évaluation des résultats robuste : implémentation sur des données en vie réelle ; l'emploi d'une méthode robuste d'interprétation des clusters et possibilité de faire évaluer la cohérence des trajectoires par une équipe de soignants et d'experts.

Ces apports, en ligne avec les verrous scientifiques précédemment identifiés, ont pu être traités pour chaque cas d'application.

Une attention sur la sélection des variables d'intérêt.

L'estimation macro des coûts a été recomposée pour les chapitres 3 et 4 à partir respectivement de l'agrégation des trajectoires individuelles au sein de la cohorte et de l'analyse comptable annuelle de l'établissement de soins à domicile. Cette étude préliminaire a permis d'identifier les principaux inducteurs de coût du parcours patient ou de soins et donc d'orienter la sélection des variables au cours des phases de pré-traitement des jeux de données. Par ailleurs, nous avons porté une attention particulière à l'intégration d'indicateurs cliniques spécifiques aux profils pathologiques observés dans les données (MMSE, IADL, Indice de Karnofsky, etc.).

Une méthodologie sur-mesure.

Plusieurs méthodes de classification supervisée et non supervisée ont été testées au cours de ces travaux de recherche. Le cas d'application traité dans le chapitre 3 présentait un grand nombre d'observations « bruit », ce qui a eu une influence sur les performances des méthodes de clustering classiques basées sur les centroïdes (K-moyennes et K-médoïdes). Des méthodes basées sur la densité, et notamment l'algorithme BIRCH, ont pu conduire à des résultats améliorés. Les clusters obtenus sont par ailleurs équilibrés en termes de répartition des observations et n'ont pas impacté la construction de la matrice de transition de la chaîne de Markov.

Le déséquilibre des classes a lui été un obstacle à la performance des algorithmes du cas d'application traité dans le chapitre 4. Des méthodes d'*oversampling* n'ont pas permis d'améliorer les scores d'évaluation, et cet aspect mériterait d'être investigué dans des recherches futures.

En ce qui concerne l'analyse des coûts, nous avons recomposé, dans chacun des cas d'applications, le coût des trajectoires individuelles ainsi que celui du service à partir des analyses comptables. Cette double approche, *top down* et *bottom up*, permet de mettre en regard les inducteurs de coûts impactant au niveau global et au niveau individuel à l'échelle du patient.

Une démarche d'interprétation robuste.

L'interprétabilité des résultats issus des algorithmes d'Intelligence Artificielle peut être un obstacle majeur à l'implémentation de ce type de solutions dans des systèmes socio-organisationnels réels. C'est d'autant

plus vrai pour les techniques d'apprentissage non supervisées, pour lesquelles il est essentiel d'ajouter une couche d'interprétation afin de permettre une confrontation de la cohérence des résultats avec l'expertise terrain des soignants. Dans les deux cas d'application, nous nous sommes employés à examiner et évaluer les parcours obtenus avec une équipe médicale pluridisciplinaire.

En synthèse, notre travail a permis de couvrir les grandes problématiques soulevées dans la littérature scientifique au sujet de la prédiction des parcours patients sous l'angle des coûts – tout en testant une approche combinée sur deux contextes de parcours différents et sur des données de santé en vie réelle dans les deux cas. Notre approche permet un apport significatif en tant que support de décision clinique. Implémentée dans des services hospitaliers, elle pourrait notamment permettre l'anticipation des variations dans les parcours et la prise de décision éclairée sur les stratégies de soins à adopter ou encore les ressources à adapter. Nos travaux pourront également apporter un outil significatif aux décisionnaires de santé publique pour identifier la variété des parcours patients souffrant d'une même pathologie, et mesurer la valeur ajoutée d'une thérapeutique ou d'une prise en charge dans la trajectoire.

5.2 Limitations et perspectives futures

L'approche déployée et testée présente un certain nombre d'aspects limitants, dont quelques-uns ont déjà été évoqués dans les chapitres 3 et 4. Ces limitations se déclinent notamment sur les axes suivants :

1. Nombre d'observations : dans chacun des cas d'application, le nombre de parcours observés varie entre des ordres de grandeur de 10 000 à 30 000. Ce volume de données, relativement petit, est limitant pour l'application d'un certain nombre de techniques d'intelligence artificielle, dont l'apprentissage profond et les réseaux de neurones par exemple.

2. Interprétation : les résultats des algorithmes restent complexes à interpréter, malgré l'application d'une méthode a posteriori. Celle-ci demande un niveau de connaissances encore élevé du fonctionnement des algorithmes et devrait être adaptée si l'on envisage l'implémentation de ce type de solution dans la pratique clinique, en simplifiant l'interface utilisateur ou encore en assurant la formation technique des équipes soignantes. Cette question reste un frein majeur pour l'adoption de ces technologies.

3. Biais du prescripteur : si certains biais, et notamment de genre, ont été traités au cours des travaux, l'étude est restée focalisée sur le contexte de la région Auvergne Rhône-Alpes, où sont implantées les structures de santé concernées par ces analyses. Or, les parcours peuvent être largement influencés par des protocoles mis en place localement, ou encore des habitudes de prescription. Ce phénomène est par ailleurs connu dans la littérature sous le nom de biais du prescripteur[234]. Nous n'avons pas pu évaluer la présence ou non de ce biais dans nos données, et l'étude pourrait être approfondie avec une comparaison des parcours entre les régions.

4. Données disponibles : les analyses sont toujours limitées par les données disponibles. Les parcours sont susceptibles d'être

[234] (Danaei et al., 2013) discutent du biais du prescripteur dans le contexte de la recherche sur l'efficacité comparative entre deux traitements, où les thérapeutiques sont souvent non aléatoires et peuvent être influencés par les caractéristiques du patient.

influencés par des facteurs externes à la prise en charge et difficiles à capter, tels que l'acceptabilité et l'adhérence aux soins par exemple.

5. Utilisation des consommations facturées comme proxy : nous avons recomposé les parcours patient et de soins à partir des consommations de soins réalisées. Comme discuté en introduction dans le chapitre 1 et observé dans les limitations du chapitre 4, ce choix peut amener un biais d'observation. A titre d'exemple, un patient peut nécessiter une consultation en urgence mais renoncer pour une raison financière. La notion de besoin en soins n'apparaît pas systématiquement dans les données de consommations.

6. Généralisation à des parcours nouveaux : de futures recherches pourront investiguer à l'application de ces travaux dans un contexte de santé nouveau et notamment avec des profils de patients différents.

Ces limitations pourront notamment être traitées dans des recherches futures pour améliorer la performance de l'approche, mais aussi tirer des conclusions sur la généricité de la méthodologie.

Bibliographie

Aalst, W. M. P. van der. (2016). *Process Mining : Data Science in Action*. Springer Berlin Heidelberg.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2015). *TensorFlow : A system for large-scale machine learning*. Software available from tensorflow.org.

Abu Zwaida, T., Pham, C., & Beauregard, Y. (2021). Optimization of Inventory Management to Prevent Drug Shortages in the Hospital Supply Chain. *Applied Sciences*, *11*(6), Article 6. https://doi.org/10.3390/app11062726

Abul-Husn, N. S., & Kenny, E. E. (2019). Personalized Medicine and the Power of Electronic Health Records. *Cell*, *177*(1), 58-69. https://doi.org/10.1016/j.cell.2019.02.039

Alami, H., Lehoux, P., Auclair, Y., Guise, M. de, Gagnon, M.-P., Shaw, J., Roy, D., Fleet, R., Ahmed, M. A. A., & Fortin, J.-P. (2020). Artificial Intelligence and Health Technology Assessment : Anticipating a New Level of Complexity. *Journal of Medical Internet Research*, *22*(7), e17707. https://doi.org/10.2196/17707

Alami, H., Lehoux, P., Denis, J.-L., Motulsky, A., Petitgand, C., Savoldelli, M., Rouquet, R., Gagnon, M.-P., Roy, D., & Fortin, J.-P. (2020). Organizational readiness for artificial intelligence in health care : Insights for decision-making and practice. *Journal of Health Organization and Management*, *35*(1), 106-114. https://doi.org/10.1108/JHOM-03-2020-0074

Amisha, Malik, P., Pathania, M., & Rathaur, V. K. (2019). Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care*, *8*(7), 2328-2331. https://doi.org/10.4103/jfmpc.jfmpc_440_19

Anguis, M., Bergeat, M., Pisarik, J., Vergier, N., Chaput, H., Laffeter, Q., Legendre, B., Dixte, C., & Barlet, M. (2021). *Quelle démographie récente et à venir pour les professions médicales et pharmaceutique ?* (Nº 76; Les dossiers de la DREES).

Anh Luong, D. T., & Chandola, V. (2017). A K-Means Approach to Clustering Disease Progressions. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 268-274. https://doi.org/10.1109/ICHI.2017.18

Ankri, J., Van Broeckhoven, C., Renucci, A., Hesse, C., & Martineau, A. (2013). *Evaluation du plan Alzheimer 2008-2012*. Ministère des Affaires sociales et de la Santé. http://www.vie-publique.fr/rapport/33267-evaluation-du-plan-alzheimer-2008-2012

Appleton, D. R., French, J. M., & Vanderpump, M. P. J. (1996). Ignoring a Covariate : An Example of Simpson's Paradox. *The American Statistician*, *50*(4), 340-341. https://doi.org/10.1080/00031305.1996.10473563

Augusto, V., Xie, X., Prodel, M., Jouaneton, B., & Lamarsalle, L. (2016). Evaluation of discovered clinical pathways using process mining and joint agent-based discrete-event simulation. *2016 Winter Simulation Conference (WSC)*, 2135-2146. https://doi.org/10.1109/WSC.2016.7822256

Barlet, M., Coldefy, M., Collin, C., & Lucas-Gabrielli, V. (2012). L'accessibilité potentielle localisée (APL) : Une nouvelle mesure de l'accessibilité aux médecins généralistes libéraux. *Etudes et Résultats*, *795*.

Barnett, G. O., Cimino, J. J., Hupp, J. A., & Hoffer, E. P. (1987). DXplain : An Evolving Diagnostic Decision-Support

System. *JAMA*, *258*(1), 67-74. https://doi.org/10.1001/jama.1987.03400010071030

Bellini, V., Guzzon, M., Bigliardi, B., Mordonini, M., Filippelli, S., & Bignami, E. (2019). Artificial Intelligence : A New Tool in Operating Room Management. Role of Machine Learning Models in Operating Room Optimization. *Journal of Medical Systems*, *44*(1), 20. https://doi.org/10.1007/s10916-019-1512-1

Ben Bachouch, R., Guinet, A., & Ruiz, A. (2012). Problématiques de gestion des structures de soins à domicile : Un état des investigations. *Conférence GISEH 2012*. 6ème conférence en Gestion et Ingéniérie des Systèmes Hospitaliers, Québec, Canada.

Bérard, A., Gervès, C., & Aquino, J.-P. (2015). *Combien coûte la maladie d'Alzheimer ?*

Bergmeir, C., Hyndman, R., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*(C), 70-83.

Besombes, B., Dubost, E., Guinet, A., & Marcon, E. (2017, janvier 1). Internalisation/externalisation de la fonction médicale. *Gestions hospitalières*. https://gestions-hospitalieres.fr/internalisationexternalisation-de-fonction-medicale/

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (Springer). https://link.springer.com/book/9780387310732

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bruun, M., Frederiksen, K. S., Rhodius-Meester, H. F. M., Baroni, M., Gjerum, L., Koikkalainen, J., Urhemaa, T., Tolonen, A., van Gils, M., Rueckert, D., Dyremose, N., Andersen, B. B., Lemstra, A. W., Hallikainen, M., Kurl, S.,

Herukka, S.-K., Remes, A. M., Waldemar, G., Soininen, H., … Hasselbalch, S. G. (2019). Impact of a clinical decision support tool on prediction of progression in early-stage dementia : A prospective validation study. *Alzheimer's Research & Therapy*, *11*(1), 25. https://doi.org/10.1186/s13195-019-0482-3

Büla, C., David, S., Stiefel, F., & Berney, A. (2007). Le diagnostic différential des troubles cognitifs en médecine de premier recours. *Rev Med Suisse*, *098*, 389-395.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Éds.), *Advances in Knowledge Discovery and Data Mining* (p. 160-172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Chambaud, L. (2018). Accès aux soins : Éléments de cadrage. *Regards*, *53*(1), 19-28. https://doi.org/10.3917/regar.053.0019

Chang, S., Long, S. R., Kutikova, L., Bowman, L., Finley, D., Crown, W. H., & Bennett, C. L. (2004). Estimating the cost of cancer : Results on the basis of claims data analyses for cancer patients diagnosed with seven types of cancer during 1999 to 2000. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *22*(17), 3524-3530. https://doi.org/10.1200/JCO.2004.10.170

Charavel, C., Mauro, L., & Seimandi, T. (2018). *Les soins de suite et de réadaptation entre 2008 et 2016—Forte progression de l'activité, en réponse au vieillissement de la population | Direction de la recherche, des études, de l'évaluation et des statistiques* (N° 30; Les dossiers de la DREES). https://drees.solidarites-

sante.gouv.fr/publications/les-dossiers-de-la-drees/les-soins-de-suite-et-de-readaptation-entre-2008-et-2016
Chen, C. (2004). *Using Random Forest to Learn Imbalanced Data*. https://www.semanticscholar.org/paper/Using-Random-Forest-to-Learn-Imbalanced-Data-Chen/2138b37bfced70599d26dfccbf93a8e7a4b7ad85
Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA Journal of Ethics*, *21*(2), 167-179. https://doi.org/10.1001/amajethics.2019.167
Choné, P., Coudin, E., & Pla, A. (2019). Does the Provision of Physician Services Respond to Competition? *Working Papers*, Article 2019-20. https://ideas.repec.org//p/crs/wpaper/2019-20.html
Cresswell, K., & Sheikh, A. (2009). The NHS Care Record Service (NHS CRS) : Recommendations from the literature on successful implementation and adoption. *Informatics in Primary Care*, *17*(3), 153-160. https://doi.org/10.14236/jhi.v17i3.730
Crocq, M.-A., & Guelfi, J.-D. (2015). *DSM-5 : Manuel diagnostique et statistique des troubles mentaux* (5e éd). Elsevier Masson.
Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *JAMA*, *269*(18), 2386-2391.
Dabek, F., Chen, J., Garbarino, A., & Caban, J. J. (2015). Visualization of longitudinal clinical trajectories using a graph-based approach. *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare*, 1-7. https://doi.org/10.1145/2836034.2836039
Danaei, G., Rodríguez, L. A. G., Cantero, O. F., Logan, R., & Hernán, M. A. (2013). Observational data for comparative

effectiveness research : An emulation of randomised trials of statins and primary prevention of coronary heart disease. *Statistical Methods in Medical Research*, *22*(1), 70-96. https://doi.org/10.1177/0962280211403603

Dauphinot, V., Garnier-Crussard, A., Moutet, C., Delphin-Combe, F., Späth, H.-M., & Krolak-Salmon, P. (2021). Determinants of Medical Direct Costs of Care among Patients of a Memory Center. *The Journal of Prevention of Alzheimer's Disease*, *8*(3), 351-361. https://doi.org/10.14283/jpad.2021.16

Dauphinot, V., Moutet, C., Rouch, I., Verdurand, M., Mouchoux, C., Delphin-Combe, F., Gaujard, S., & Krolak-Salmon, P. (2019). A multicenter cohort study to investigate the factors associated with functional autonomy change in patients with cognitive complaint or neurocognitive disorders : The MEMORA study protocol. *BMC Geriatrics*, *19*, 191. https://doi.org/10.1186/s12877-019-1204-1

Dauphinot, V., Potashman, M., Levitchi-Benea, M., Su, R., Rubino, I., & Krolak-Salmon, P. (2022). Economic and caregiver impact of Alzheimer's disease across the disease spectrum : A cohort study. *Alzheimer's Research & Therapy*, *14*(1), 34. https://doi.org/10.1186/s13195-022-00969-x

Dean, J. (2022). A Golden Decade of Deep Learning : Computing Systems & Applications. *Daedalus*, *151*(2), 58-74. https://doi.org/10.1162/daed_a_01900

Decoupigny, F., Pérez, S., & Yordanova, D. (2007). Modélisation de l'accessibilité aux soins. *Territory in movement Journal of geography and planning*, *4*, Article 4. https://doi.org/10.4000/tem.886

Delattre, E., & Dormont, B. (1999). Induction de la demande de soins par les médecins libéraux français. Etude microéconométrique sur données de panel. *THEMA*

Working Papers, Article 99-21. https://ideas.repec.org//p/ema/worpap/99-21.html

Derevitskii, I. V., & Kovalchuk, S. V. (2019). Analysis course of the disease of type 2 diabetes patients using Markov chains and clustering methods. *Procedia Computer Science*, *156*, 114-122. https://doi.org/10.1016/j.procs.2019.08.186

Dicuonzo, G., Donofrio, F., Fusco, A., & Shini, M. (2023). Healthcare system : Moving forward with artificial intelligence. *Technovation*, *120*, 102510. https://doi.org/10.1016/j.technovation.2022.102510

Dilsizian, S. E., & Siegel, E. L. (2013). Artificial Intelligence in Medicine and Cardiac Imaging : Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment. *Current Cardiology Reports*, *16*(1), 441. https://doi.org/10.1007/s11886-013-0441-8

Dixneuf, P. (2019). *Analyse de la performance de la méthode d'imputation de données manquantes missForest et application à des données environnementales*. ÉCOLE DE TECHNOLOGIE SUPÉRIEURE UNIVERSITÉ DU QUÉBEC.

Duan, K.-B., Rajapakse, J. C., & Nguyen, M. N. (2007). One-Versus-One and One-Versus-All Multiclass SVM-RFE for Gene Selection in Cancer Classification. In E. Marchiori, J. H. Moore, & J. C. Rajapakse (Éds.), *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics* (p. 47-56). Springer. https://doi.org/10.1007/978-3-540-71783-6_5

Dufournet, M., Moutet, C., Achi, S., Delphin-Combe, F., Krolak-Salmon, P., Dauphinot, V., Krolak-Salmon, P., Dauphinot, V., Delphin-Combe, F., Makaroff, Z., Federico, D., Coste, M.-H., Rouch, I., Dorey, J.-M., Lepetit, A., Danaila, K., Vernaudon, J., Bathsavanis, A., Sarciron, A., … the

MEMORA group. (2021). Proposition of a corrected measure of the Lawton instrumental activities of daily living score. *BMC Geriatrics*, *21*(1), 39. https://doi.org/10.1186/s12877-020-01995-w

Elkin, P. L., Liebow, M., Bauer, B. A., Chaliki, S., Wahner-Roedler, D., Bundrick, J., Lee, M., Brown, S. H., Froehling, D., Bailey, K., Famiglietti, K., Kim, R., Hoffer, E., Feldman, M., & Barnett, G. O. (2010). The introduction of a diagnostic decision support system (DXplain^{TM}) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging Diagnostic Related Groups (DRGs). *International Journal of Medical Informatics*, *79*(11), 772-777. https://doi.org/10.1016/j.ijmedinf.2010.09.004

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.

Evgeniou, T., & Pontil, M. (2001). *Support Vector Machines : Theory and Applications* (Vol. 2049, p. 257). https://doi.org/10.1007/3-540-44673-7_12

Fosch-Villaronga, E., Drukarch, H., Khanna, P., Verhoef, T., & Custers, B. (2022). Accounting for diversity in AI for medicine. *Computer Law & Security Review*, *47*, 105735. https://doi.org/10.1016/j.clsr.2022.105735

Frédéric Pierru, B. D. (2020). La tarification à l'activité (T2A) à la française. *Revue française d'administration publique*, *174*(2), 487-497. https://doi.org/10.3917/rfap.174.0191

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gallón, S., Loubes, J.-M., & Maza, E. (2013). Statistical properties of the quantile normalization method for density

curve alignment. *Mathematical Biosciences*, *242*(2), 129-142. https://doi.org/10.1016/j.mbs.2012.12.007

Geitona, M., Androutsou, L., & Theodoratou, D. (2010). Cost estimation of patients admitted to the intensive care unit : A case study of the Teaching University Hospital of Thessaly. *Journal of Medical Economics*, *13*(2), 179-184. https://doi.org/10.3111/13696991003684092

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits on Translational Science Proceedings*, *2020*, 191.

Gönel, A. (2020). Clinical biochemistry test eliminator providing cost-effectiveness with five algorithms. *Acta Clinica Belgica*, *75*(2), 123-127. https://doi.org/10.1080/17843286.2018.1563324

Guide parcours de soins maladie de Parkinson. (s. d.). Haute Autorité de Santé. Consulté 2 mai 2023, à l'adresse https://www.has-sante.fr/jcms/c_1242645/fr/guide-parcours-de-soins-maladie-de-parkinson

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate Data Analysis* (7th edition). Pearson.

Hajat, C., Siegal, Y., & Adler-Waxman, A. (2021). Clustering and Healthcare Costs With Multiple Chronic Conditions in a US Study. *Frontiers in Public Health*, *8*. https://www.frontiersin.org/articles/10.3389/fpubh.2020.607528

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Huang, C.-W., Lu, R., Iqbal, U., Lin, S.-H., Nguyen, P. A., Yang, H.-C., Wang, C.-F., Li, J., Ma, K.-L., Li, Y.-C., & Jian, W.-S. (2015). A richly interactive exploratory data analysis and

visualization tool using electronic medical records. *BMC Medical Informatics and Decision Making*, *15*(1), 92. https://doi.org/10.1186/s12911-015-0218-7

Huber, C. A., Szucs, T. D., Rapold, R., & Reich, O. (2013). Identifying patients with chronic conditions using pharmacy data in Switzerland : An updated mapping approach to the classification of medications. *BMC Public Health*, *13*, 1030. https://doi.org/10.1186/1471-2458-13-1030

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., & Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(2), 193-209. https://doi.org/10.1111/1467-9884.00351

Japkowicz, N., & Shah, M. (2011). *Evaluating Learning Algorithms : A Classification Perspective*. Cambridge University Press. https://doi.org/10.1017/CBO9780511921803

Kapadia, A. S., Vineberg, S. E., & Rossi, C. D. (1985). Predicting course of treatment in a rehabilitation hospital : A Markovian model. *Computers & Operations Research*, *12*(5), 459-469. https://doi.org/10.1016/0305-0548(85)90018-8

Karabatsou, D., Tsironi, M., Tsigou, E., Boutzouka, E., Katsoulas, T., & Baltopoulos, G. (2016). Variable cost of ICU care, a micro-costing analysis. *Intensive and Critical Care Nursing*, *35*, 66-73. https://doi.org/10.1016/j.iccn.2016.01.001

Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, *92*(4), 807-812. https://doi.org/10.1016/j.gie.2020.06.040

Kausch, S. L., Lobo, J. M., Spaeder, M. C., Sullivan, B., & Keim-Malpass, J. (2021). Dynamic Transitions of Pediatric Sepsis : A Markov Chain Analysis. *Frontiers in Pediatrics*,

9. https://www.frontiersin.org/articles/10.3389/fped.2021.743544

Keel, G., Savage, C., Rafiq, M., & Mazzocato, P. (2017). Time-driven activity-based costing in health care : A systematic review of the literature. *Health Policy (Amsterdam, Netherlands)*, *121*(7), 755-763. https://doi.org/10.1016/j.healthpol.2017.04.013

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, 1137-1143.

Kremp, E. (1995). Nettoyage de fichiers dans le cas de données individuelles : Recherche de la cohérence transversale. *Économie & prévision*, *119*(3), 171-193. https://doi.org/10.3406/ecop.1995.5738

Krenn, M., Buffoni, L., Coutinho, B., Eppel, S., Foster, J. G., Gritsevskiy, A., Lee, H., Lu, Y., Moutinho, J. P., Sanjabi, N., Sonthalia, R., Tran, N. M., Valente, F., Xie, Y., Yu, R., & Kopp, M. (2022). *Predicting the Future of AI with AI : High-quality link prediction in an exponentially growing knowledge network* (arXiv:2210.00881). arXiv. http://arxiv.org/abs/2210.00881

Kreuzthaler, M., Martínez-Costa, C., Kaiser, P., & Schulz, S. (2017). Semantic Technologies for Re-Use of Clinical Routine Data. *Studies in Health Technology and Informatics*, *236*, 24-31.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86. https://doi.org/10.1214/aoms/1177729694

Lachieze-Rey, R. (2021). *Chaînes de Markov (et applications)*.

Lapinte, A., Legendre, B., & statistiques -DREES, D. de la recherches des études de l'évaluation et des. (2021).

Renoncement aux soins : La faible densité médicale est un facteur aggravant pour les personnes pauvres. *Etudes et résultats - DREES*, *n° 1200*, 98p.

Lebcir, R., Hill, T., Atun, R., & Cubric, M. (2021). Stakeholders' views on the organisational factors affecting application of artificial intelligence in healthcare : A scoping review protocol. *BMJ Open*, *11*(3), e044074. https://doi.org/10.1136/bmjopen-2020-044074

Leung, I., Casalino, E., Pateron, D., Grateau, G., Garandeau, E., & de Stampa, M. (2016). Participation des médecins généralistes dans les prises en charge de leurs patients en hospitalisation à domicile. *Santé Publique*, *28*(4), 499-504. https://doi.org/10.3917/spub.164.0499

Levenson, R., Dewar, S., & Shepherd, S. (2008). *Understanding doctors : Harnessing professionalism*. King's Fund : Royal College of Physicians.

Lucas-Gabrielli, V., Mangeney, C., Duchaine, F., Com-Ruelle, L., Gueye, A., & Raynaud, D. (2022). Inégalités spatiales d'accessibilité aux médecins spécialistes. Proposition de méthodologie pour trois spécialités. *Working Papers*, Article DT87. https://ideas.repec.org//p/irh/wpaper/dt87.html

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579-2605.

Martin, G., Martin, P., Hankin, C., Darzi, A., & Kinross, J. (2017). Cybersecurity and healthcare : How safe are we? *BMJ*, *358*, j3179. https://doi.org/10.1136/bmj.j3179

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting Algorithms as Gradient Descent. *Advances in Neural Information Processing Systems*, *12*. https://proceedings.neurips.cc/paper_files/paper/1999/hash/96a93ba89a5b5c6c226e49b88973f46e-Abstract.html

Mauro, L. (2017). Dix ans d'hospitalisation à domicile (2006-2016). *Les dossiers de la DREES*, *Décembre 2017*(23).

Mei, K., Peng, J., Gao, L., Zheng, N., & Fan, J. (2015). Hierarchical classification of large-scale patient records for automatic treatment stratification. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1234-1245. https://doi.org/10.1109/JBHI.2015.2414876

Menger, V., Spruit, M., van der Klift, W., & Scheepers, F. (2019). Using Cluster Ensembles to Identify Psychiatric Patient Subgroups. In D. Riaño, S. Wilk, & A. ten Teije (Éds.), *Artificial Intelligence in Medicine* (p. 252-262). Springer International Publishing. https://doi.org/10.1007/978-3-030-21642-9_31

Mini-Mental State Examination (MMSE) – Strokengine. (s. d.). Consulté 26 avril 2023, à l'adresse https://strokengine.ca/fr/assessments/mini-mental-state-examination-mmse/

CIRCULAIRE N°DHOS/O/2004/44..., Pub. L. No. DHOS/O/2004/44 (2004). https://sante.gouv.fr/IMG/pdf/circulaire_44_040204.pdf

Mougeot, M., & Naegelen, F. (2014). La tarification à l'activité : Une réforme dénaturée du financement des hôpitaux. *Revue française d'économie*, *XXIX*(3), 111-141. https://doi.org/10.3917/rfe.143.0111

Najjar, A., Reinharz, D., Girouard, C., & Gagné, C. (2018). A two-step approach for mining patient treatment pathways in administrative healthcare databases. *Artificial Intelligence in Medicine*, *87*, 34-48. https://doi.org/10.1016/j.artmed.2018.03.004

Neill, D. B. (2013). Using Artificial Intelligence to Improve Hospital Inpatient Care. *IEEE Intelligent Systems*, *28*(2), 92-95. https://doi.org/10.1109/MIS.2013.51

Nichols, E., Steinmetz, J. D., Vollset, S. E., Fukutaki, K., Chalek, J., Abd-Allah, F., Abdoli, A., Abualhasan, A., Abu-Gharbieh, E., Akram, T. T., Hamad, H. A., Alahdab, F., Alanezi, F. M., Alipour, V., Almustanyir, S., Amu, H., Ansari, I., Arabloo, J., Ashraf, T., … Vos, T. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050 : An analysis for the Global Burden of Disease Study 2019. *The Lancet Public Health*, *7*(2), e105-e125. https://doi.org/10.1016/S2468-2667(21)00249-8

Oh, K.-S., & Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, *37*(6), 1311-1314. https://doi.org/10.1016/j.patcog.2004.01.013

Padovan, P. H., Martins, C. M., & Reed, C. (2023). Black is the new orange : How to determine AI liability. *Artificial Intelligence and Law*, *31*(1), 133-167. https://doi.org/10.1007/s10506-022-09308-9

Paul, M., Maglaras, L., Ferrag, M. A., & Almomani, I. (2023). Digitization of healthcare sector : A study on privacy and security concerns. *ICT Express*. https://doi.org/10.1016/j.icte.2023.02.007

Pearl, J. (2009). *Causality : Models, Reasoning and Inference* (2e éd.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Petch, J., Di, S., & Nelson, W. (2022). Opening the Black Box : The Promise and Limitations of Explainable Machine Learning in Cardiology. *Canadian Journal of Cardiology*, *38*(2), 204-213. https://doi.org/10.1016/j.cjca.2021.09.004

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). *DeepCare : A Deep Dynamic Memory Model for Predictive Medicine* (arXiv:1602.00357). arXiv. https://doi.org/10.48550/arXiv.1602.00357

Raffy-Pihan, N. (1997). *L'hospitalisation à domicile : Une alternative également adaptée aux personnes âgées*. SYSTED, Chicago.

Raftery, J. (2000). Costing in economic evaluation. *BMJ : British Medical Journal*, *320*(7249), 1597.

Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. *Proceedings of the 26th Annual International Conference on Machine Learning*, 873-880. https://doi.org/10.1145/1553374.1553486

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, *169*(12), 866-872. https://doi.org/10.7326/M18-1990

Ramadhani, F., Zarlis, M., & Suwilo, S. (2020). Improve BIRCH algorithm for big data clustering. *IOP Conference Series: Materials Science and Engineering*, *725*, 012090. https://doi.org/10.1088/1757-899X/725/1/012090

Ramesh, A. N., Kambhampati, C., Monson, J. R. T., & Drew, P. J. (2004). Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England*, *86*(5), 334-338. https://doi.org/10.1308/147870804290

Rodde-Dunet, M.-H., & Mounic, V. (2016). Le parcours du patient selon la Haute Autorité de santé. *Journal des Maladies Vasculaires*, *41*(2), 100. https://doi.org/10.1016/j.jmv.2015.12.045

Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable

models instead. *Nature Machine Intelligence*, *1*(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Salter, E. K. (2015). The Re-contextualization of the Patient : What Home Health Care Can Teach Us About Medical Decision-Making. *HEC Forum: An Interdisciplinary Journal on Hospitals' Ethical and Legal Issues*, *27*(2), 143-156. https://doi.org/10.1007/s10730-015-9268-6

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients : Appropriate Use and Interpretation. *Anesthesia & Analgesia*, *126*(5), 1763. https://doi.org/10.1213/ANE.0000000000002864

Sentilhes-Monkam, A. (2005). Rétrospective de l'hospitalisation à domicile. L'histoire d'un paradoxe. *Revue française des affaires sociales*, *3*, 157-182. https://doi.org/10.3917/rfas.053.0157

Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics : Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research, an International Journal*, *8*(4), 303-320. https://doi.org/10.1016/0010-4809(75)90009-9

Silva, J. F., & Matos, S. (2021). Patient Trajectory Modelling in Longitudinal Data : A Review on Existing Solutions. *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 480-485. https://doi.org/10.1109/CBMS52027.2021.00057

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 238-241.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Sol, P., Floccia, M., Bidalon, F., Le Ru, F., & Meliani, J. (2018). Chapitre 3. Troubles neurocognitifs. In *Hypnose en pratiques gériatriques* (p. 81-150). Dunod. https://www.cairn.info/hypnose-en-pratiques-geriatriques--9782100776221-p-81.htm

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

Suter-Crazzolara, C. (2018). Better Patient Outcomes Through Mining of Biomedical Big Data. *Frontiers in ICT*, *5*. https://www.frontiersin.org/articles/10.3389/fict.2018.00030

Tchamdja, T., Balaka, A., Tchandana, M., & Agbétra, A. (2015). Cost of hospitalization by the Activity Based Costing method in the neonatal department of Principal Hospital of Dakar. *Médecine et Santé Tropicales*, *25*(4), 392-396. https://doi.org/10.1684/mst.2015.0518

Thery, C., Biernacki, C., & Loridant, G. (1982). *CorReg : Préselection de variables en régression linéaire avec fortes corrélations*. 14ème journée de statistique.

Topol, E. (2016). *The Patient Will See You Now*. Basic Books. https://www.basicbooks.com/titles/eric-topol-md/the-patient-will-see-you-now/9780465094479/

Topol, E. J. (2019). High-performance medicine : The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), Article 1. https://doi.org/10.1038/s41591-018-0300-7

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *59*(October), 433-460. https://doi.org/10.1093/mind/lix.236.433

Umargono, E., Suseno, J. E., & S. K., V. G. (2019). K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median: *Proceedings of the International Conferences on Information System and Technology*, 234-240. https://doi.org/10.5220/0009908402340240

Vairavan, S., Eshelman, L., Haider, S., Flower, A., & Seiver, A. (2012). Prediction of mortality in an intensive care unit using logistic regression and a hidden Markov model. *2012 Computing in Cardiology*, 393-396.

Vergier, N., & Chaput, H. (2017). *Déserts médicaux : Comment les définir ? Comment les mesurer ? | Direction de la recherche, des études, de l'évaluation et des statistiques* (N° 17; Les dossiers de la DREES). https://drees.solidarites-sante.gouv.fr/publications/les-dossiers-de-la-drees/deserts-medicaux-comment-les-definir-comment-les-mesurer

Vogt, V., Scholz, S. M., & Sundmacher, L. (2018). Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. *European Journal of Public Health*, *28*(2), 214-219. https://doi.org/10.1093/eurpub/ckx169

Weiss, S., Kulikowski, C. A., & Safir, A. (1978). Glaucoma consultation by computer. *Computers in Biology and Medicine*, *8*(1), 25-40. https://doi.org/10.1016/0010-4825(78)90011-2

White, S. (2014). A review of big data in healthcare : Challenges and opportunities. *Open Access Bioinformatics*, *2014*, 13-18. https://doi.org/10.2147/OAB.S50519

Widanagamaachchi, W., Livnat, Y., Bremer, P.-T., Duvall, S., & Pascucci, V. (2018). Interactive Visualization and Exploration of Patient Progression in a Hospital Setting. *AMIA Annual Symposium Proceedings*, *2017*, 1773-1782.

Wolff, J., Pauling, J., Keck, A., & Baumbach, J. (2020). The Economic Impact of Artificial Intelligence in Health Care : Systematic Review. *Journal of Medical Internet Research*, *22*(2), e16866. https://doi.org/10.2196/16866

World Health Organization. (2012). *Dementia : A public health priority*. World Health Organization. https://apps.who.int/iris/handle/10665/75263

Xu, X., Lazar, C. M., & Ruger, J. P. (2021). Micro-costing in health and medicine : A critical appraisal. *Health Economics Review*, *11*(1), 1. https://doi.org/10.1186/s13561-020-00298-5

Yan, J., Linn, K. A., Powers, B. W., Zhu, J., Jain, S. H., Kowalski, J. L., & Navathe, A. S. (2019). Applying Machine Learning Algorithms to Segment High-Cost Patient Populations. *Journal of General Internal Medicine*, *34*(2), 211-217. https://doi.org/10.1007/s11606-018-4760-8

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH : A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, *1*(2), 141-182. https://doi.org/10.1023/A:1009783824328

Table des figures

Table des tableaux

Table des annexes

Annexe 1 : Inventaire NeuroPsychiatrique NPI du CMRR de Nice.

1

INVENTAIRE NEUROPSYCHIATRIQUE NPI

Nom: Age: Date de l'évaluation

Items	NA	Absent	Fréquence	Gravité	F x G	Retentissement
Idées délirantes	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Hallucinations	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Agitation/Agressivité	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Dépression/Dysphorie	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Anxiété	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Exaltation de l'humeur/ Euphorie	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Apathie/Indifférence	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Désinhibition	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Irritabilité/Instabilité De l'humeur	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Comportement moteur aberrant	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Sommeil	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5
Appétit/Troubles de l'appétit	X	0	1 2 3 4	1 2 3	[____]	1 2 3 4 5

NA = question inadaptée (non applicable)
F x G = Fréquence x Gravité

The Neuropsychiatric Inventory: Comprehensive assessment of psychopathology in dementia, J.L. Cummings, 1994
Traduction Française P.H.Robert. *Centre Mémoire de Ressources et de Recherche - Nice – France 1996*

Annexe 2 : Grille du Mini-Zarit des hôpitaux universitaires de Genève.

GRILLE MINI – ZARIT

Evaluation de la souffrance des aidants naturels dans le maintien à domicile des personnes âgées

Patient (Nom - Prénom):

N° SS :

Aidant évalué (nom et situation vis-à-vis du patient):

Notation : 0 = jamais , ½ = parfois , 1 = souvent	**0**	**½**	**1**
1 - Le fait de vous occuper de votre parent entraîne-t-il :			
• des difficultés dans votre vie familiale ?	O	O	O
• des difficultés dans vos relations avec vos amis, vos loisirs, ou dans votre travail ?	O	O	O
• un retentissement sur votre santé (physique et/ou psychique) ?	O	O	O
2 - Avez-vous le sentiment de ne plus reconnaître votre parent ?	O	O	O
3 - Avez-vous peur pour l'avenir de votre parent ?	O	O	O
4 – Souhaitez-vous être (davantage) aidé(e) pour vous occuper de votre parent ?	O	O	O
5 – Ressentez-vous une charge en vous occupant de votre parent ?	O	O	O

Date : Age du patient : Age de l'Aidant évalué :

SCORE : ….. + ….. + ….. + ….. + ….. + ….. + ….. **=** **/ 7**

Nom, fonction, et signature de l'évaluateur :

Date : Age du patient : Age de l'Aidant évalué :

SCORE : ….. + ….. + ….. + ….. + ….. + ….. + ….. **=** **/ 7**

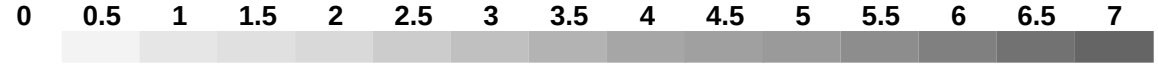
Nom, fonction, et signature de l'évaluateur :

Date : Age du patient : Age de l'Aidant évalué :

SCORE : ….. + ….. + ….. + ….. + ….. + ….. + ….. **=** **/ 7**

Nom, fonction, et signature de l'évaluateur :

Interprétation :



Fardeau absent ou léger — Fardeau léger à modéré — Fardeau modéré à sévère — Fardeau sévère

Annexe 3 : Grille d'évaluation de l'IADL de Laxton Brody publié dans « Best practices in nursing care to older adults ».

Patient Name:_____ **Date:**_____
Patient ID #

LAWTON - BRODY
INSTRUMENTAL ACTIVITIES OF DAILY LIVING SCALE (I.A.D.L.)

Scoring: For each category, circle the item description that most closely resembles the client's highest functional level (either 0 or 1).

A. Ability to Use Telephone		**E. Laundry**	
1. Operates telephone on own initiative-looks up and dials numbers, etc.	1	1. Does personal laundry completely	1
		2. Launders small items-rinses stockings, etc.	1
2. Dials a few well-known numbers	1	3. All laundry must be done by others	0
3. Answers telephone but does not dial	1		
4. Does not use telephone at all	0		
B. Shopping		**F. Mode of Transportation**	
1. Takes care of all shopping needs independently	1	1. Travels independently on public transportation or drives own car	1
2. Shops independently for small purchases	0	2. Arranges own travel via taxi, but does not otherwise use public transportation	1
3. Needs to be accompanied on any shopping trip	0	3. Travels on public transportation when accompanied by another	1
4. Completely unable to shop	0	4. Travel limited to taxi or automobile with assistance of another	0
		5. Does not travel at all	0
C. Food Preparation		**G. Responsibility for Own Medications**	
1. Plans, prepares and serves adequate meals independently	1	1. Is responsible for taking medication in correct dosages at correct time	1
2. Prepares adequate meals if supplied with ingredients	0	2. Takes responsibility if medication is prepared in advance in separate dosage	0
3. Heats, serves and prepares meals, or prepares meals, or prepares meals but does not maintain adequate diet	0	3. Is not capable of dispensing own medication	0
4. Needs to have meals prepared and served	0		
D. Housekeeping		**H. Ability to Handle Finances**	
1. Maintains house alone or with occasional assistance (e.g. "heavy work domestic help")	1	1. Manages financial matters independently (budgets, writes checks, pays rent, bills, goes to bank), collects and keeps track of income	1
2. Performs light daily tasks such as dish washing, bed making	1	2. Manages day-to-day purchases, but needs help with banking, major purchases, etc.	1
3.Performs light daily tasks but cannot maintain acceptable level of cleanliness	1	3. Incapable of handling money	0
4. Needs help with all home maintenance tasks	1		
5. Does not participate in any housekeeping tasks	0		
Score		**Score**	
		Total score_____	
A summary score ranges from 0 (low function, dependent) to 8 (high function, independent) for women and 0 through 5 for men to avoid potential gender bias.			

Source: Best Practices in Nursing Care to Older Adults, The Hartford Institute for Geriatric Nursing, New York University, College of Nursing, www.hartfordign.org

MaineHealth

Annexe 4 : Grille d'évaluation du Mini Mental State Examination (MMSE) - Version consensuelle du GRECO.

Mini Mental State Examination (MMSE) *(Version consensuelle du GRECO)*

Orientation / 10

Je vais vous poser quelques questions pour apprécier comment fonctionne votre mémoire.
Les unes sont très simples, les autres un peu moins. Vous devez répondre du mieux que vous pouvez.
Quelle est la date complète d'aujourd'hui ?

Si la réponse est incorrecte ou incomplète, posées les questions restées sans réponse, dans l'ordre suivant :

1. En quelle année sommes-nous ? ☐
2. En quelle saison ? ☐
3. En quel mois ? ☐
4. Quel jour du mois ? ☐
5. Quel jour de la semaine ? ☐

Je vais vous poser maintenant quelques questions sur l'endroit où nous trouvons.

6. Quel est le nom de l'hôpital où nous sommes ?* ☐
7. Dans quelle ville se trouve-t-il ? ☐
8. Quel est le nom du département dans lequel est située cette ville ?** ☐
9. Dans quelle province ou région est située ce département ? ☐
10. A quel étage sommes-nous ? ☐

Apprentissage / 3

Je vais vous dire trois mots ; je vous voudrais que vous me les répétiez et que vous essayiez de les retenir car je vous les redemanderai tout à l'heure.

11. Cigare	*Citron*	*Fauteuil*	☐
12. Fleur	*Clé*	*Tulipe*	☐
13. Porte	*Ballon*	*Canard*	☐

Répéter les 3 mots.

Attention et calcul / 5

Voulez-vous compter à partir de 100 en retirant 7 à chaque fois ?*

14. 93 ☐
15. 86 ☐
16. 79 ☐
17. 72 ☐
18. 65 ☐

Pour tous les sujets, même pour ceux qui ont obtenu le maximum de points, demander :
Voulez-vous épeler le mot MONDE à l'envers ?**

Rappel / 3

Pouvez-vous me dire quels étaient les 3 mots que je vous ai demandés de répéter et de retenir tout à l'heure ?

11. Cigare	*Citron*	*Fauteuil*	☐
12. Fleur	*Clé*	*Tulipe*	☐
13. Porte	*Ballon*	*Canard*	☐

Langage / 8

Montrer un crayon. 22. Quel est le nom de cet objet ?* ☐
Montrer votre montre. 23. Quel est le nom de cet objet ?** ☐
24. Ecoutez bien et répétez après moi : « PAS DE MAIS, DE SI, NI DE ET »*** ☐

Poser une feuille de papier sur le bureau, la montrer au sujet en lui disant : « Ecoutez bien et faites ce que je vais vous dire :
25. Prenez cette feuille de papier avec votre main droite, ☐
26. Pliez-la en deux, ☐
27. Et jetez-la par terre. »**** ☐

Tendre au sujet une feuille de papier sur laquelle est écrit en gros caractère : « FERMEZ LES YEUX » et dire au sujet : 28. « Faites ce qui est écrit ». ☐

Tendre au sujet une feuille de papier et un stylo, en disant :
29. « Voulez-vous m'écrire une phrase, ce que vous voulez, mais une phrase entière. » ☐

Praxies constructives / 1

Tendre au sujet une feuille de papier et lui demander :
30. « Voulez-vous recopier ce dessin ? » ☐

Annexe 5 : Code Python - dictionnaire utilisé pour le recodage des variables du stade diagnostic et du diagnostic étiologique.

```python
stade_diag_corresp = {
    "plainte isolee": "plainte cognitive isolee",
    "trouble cognitif leger amnesique": "trouble cognitif leger",
    "demence": "trouble cognitif majeur",
    "trouble cognitif leger autre domaine": "trouble cognitif leger",
    "trouble cognitif leger multi domaine": "trouble cognitif leger",
    "trouble cognitif mineur": "trouble cognitif leger",
    "non applicable": "autre",
    "autre cas": "autre"
}

```

```python
diag_etio_recoding = {
    "maladie d'alzheimer": "ad",
    "maladie d'alzheimer a composante cerebrovasculaire": "ad",
    "maladie d'alzheimer - cliniquement probable": "ad",
    "maladie d'alzheimer - probable (avec biomarqueurs)": "ad",
    "maladie d'alzheimer - certaine (genetique)": "ad",
    "maladie d'alzheimer - dans le cadre du syndrome de down": "ad",
    "angiopathie amyloide": "related disorders",
    "aphasie progressive primaire": "related disorders",
    "atrophie corticale posterieure (benson)": "related disorders",
    "atrophie multisystematisee": "related disorders",
    "degenerescence lobaire fronto-temporale": "related disorders",
    "degenerescence cortico-basale": "related disorders",
    "maladie a corps de lewy": "related disorders",
    "maladie de parkinson": "related disorders",
    "degenerescence fronto-temporale (dft-vf)": "related disorders",
    "maladie semantique (et autres dft vt)": "related disorders",
    "demence non classee ailleurs": "related disorders",
    "paralysie supranucleaire progressive": "related disorders",
    "demence vasculaire": "vascular",
    "lesion vasculaire (avc)": "vascular",
    "pathologies vasculaires non liees a atherome": "vascular",
    "sequelles d'avc": "vascular",
    "troubles cognitifs vasculaires": "vascular",
    "autre trouble cognitif d'origine vasculaire (f01,8)": "vascular",
    "encephalopathie vasculaire": "vascular",
    "autres troubles mentaux": "psy",
    "autres troubles psychiatriques": "psy",
    "etat de stress post-traumatique": "psy",
    "schizophrenie et autres troubles psychotiques (dsm5)": "psy",
    "troubles anxieux depressifs": "psy",
    "trouble anxieux (dsm 5)": "psy",
    "troubles bipolaires": "psy",
    "trouble bipolaire (dsm5)": "psy",
    "trouble bipolaire (dms5)": "psy",
    "trouble depressif (dsm 5)": "psy",
    "troubles depressifs isoles": "psy",
    "troubles depressifs recurrents": "psy",
    "troubles neurodeveloppementaux (dsm5)": "psy",
    "troubles obsessionnels-compulsifs et apparentes (dsm 5)": "psy",
    "autre trouble d'origine organique": "other neuro",
    "autres troubles neurologiques (tumeurs, anevrysmes, post-chirurgicaux)": "other neuro",
    "commotion cerebrale": "other neuro",
    "trouble cognitif lie au vih": "other neuro",
    "encephalopathie anoxique": "other neuro",
    "encephalopathie auto-immunes": "other neuro",
    "encephalopathie chronique post traumatique": "other neuro",
    "encephalopathie d'origine infectieuse": "other neuro",
    "encephalopathie d'origine metabolique": "other neuro",
    "encephalopathie ethylique": "other neuro",
```

```
    "encephalopathie toxique": "other neuro",
    "encephalite limbique/paraneoplasique": "other neuro",
    "epilepsie": "other neuro",
    "hydrocephalie chronique": "other neuro",
    "iatrogenie": "other neuro",
    "maladie a prion": "other neuro",
    "maladie de creutzfeldt jakob": "other neuro",
    "maladie de huntington": "other neuro",
    "sequelles encephaliques de traumatisme cranien": "other neuro",
    "traumatisme cranien": "other neuro",
    "sep": "other neuro",
    "syndrome des apnees du sommeil": "other neuro",
    "trouble organique cerebral directement lie a la consommation d'alcool ou d'autres toxiques": "other neuro",
    "trouble organique cerebral directement lie a une pathologie (carence metabolique...)": "other neuro",
    "tumeur intracranienne": "other neuro",
    "autres": "other neuro",
    "diagnostic en attente": "pending diagnosis",
    "diagnostic non pose": "pending diagnosis",
    "non applicable": "unknown",
    "pas de troubles": "no disorders"
}
```

Annexe 6 : Pseudo-code descriptif de la méthodologie MissForest.

```
Entrées :
    - dataframe X avec des valeurs manquantes
    - max_depth : profondeur maximale des arbres
    - n_trees : nombre d'arbres à créer

Sorties :
    - dataframe X complété avec les valeurs manquantes

Étape 1 : Création de la matrice binaire R qui indique les valeurs manquantes
    R = 1 si la valeur est manquante, 0 sinon

Étape 2 : Boucle sur chaque colonne c de X :
    2.1 : Copie de X dans une nouvelle variable X_c
    2.2 : Remplacement des valeurs manquantes dans la colonne c de X_c par des échantillons aléatoires de la même colonne c, non manquants
    2.3 : Création d'un masque binaire M qui indique les valeurs non manquantes
    2.4 : Boucle sur chaque arbre de décision i dans les n_trees :
        2.4.1 : Échantillonnage d'un sous-ensemble de X_c et R, avec remplacement
        2.4.2 : Construction de l'arbre de décision i de profondeur max_depth avec le sous-ensemble échantillonné et M comme masque de poids
    2.5 : Prédiction des valeurs manquantes dans la colonne c de X_c en utilisant la prédiction de chaque arbre de décision
    2.6 : Remplacement des valeurs manquantes de la colonne c de X par la médiane des prédictions de chaque arbre de décision

Étape 3 : Répéter les étapes 2 pour chaque colonne de X qui contient des valeurs manquantes

Étape 4 : Retourner le dataframe X complété avec les valeurs manquantes
```

Annexe 7 : Pseudo-code descriptif du fonctionnement du modèle ensembliste Gradient Boosting, *entraîné sur un jeu de données de variables (X_train, y_train) et testé sur (X_test, y_test), où X et Y sont respectivement les variables indépendantes et y la variable à prédire.*

```
# Définition des hyperparamètres
learning_rate = 0.1
n_estimators = 100
max_depth = 3

# Initialisation du modèle
model = TreeRegressor(max_depth)

# Boucle pour entraîner les modèles de prédiction
for i in range(n_estimators):

    # Prédiction avec le modèle actuel
    y_pred = model.predict(X_train)

    # Calcul des résidus
    residuals = y_train - y_pred

    # Entraînement d'un nouveau modèle pour les résidus
    weak_learner = TreeRegressor(max_depth=max_depth)
    weak_learner.fit(X_train, residuals)

    # Mise à jour du modèle actuel
    model = EnsembleRegressor(model, weak_learner, learning_rate=learning_rate)

# Prédiction finale
y_pred_final = model.predict(X_test)
```

Annexe 8 : Pseudo-code descriptif de l'entraînement d'un modèle de Support Vector Machine.

```
Entrée :
- X: un ensemble de données d'entraînement
- y: un ensemble de cibles d'entraînement
- C: paramètre de régularisation
- kernel: type de noyau à utiliser

Sortie :
- w: vecteur de poids appris
- b: biais appris

1. Initialiser le vecteur support alpha à zéro pour toutes les données d'entraînement
2. Répéter jusqu'à convergence :
   a. Pour chaque paire (x_i, y_i), (x_j, y_j) avec i != j:
      i. Calculer l'erreur E_i = f(x_i) - y_i où f(x_i) est la prédiction courante de x_i.
      ii. Calculer l'erreur E_j = f(x_j) - y_j où f(x_j) est la prédiction courante de x_j.
      iv. Calculer le noyau K(x_i, x_j).
      v. Mettre à jour alpha_j à l'aide de la formule : alpha_j := alpha_j - (y_j * (E_i - E_j) / K(x_i, x_j))
      vii. Mettre à jour alpha_i à l'aide de la formule : alpha_i := alpha_i + y_i * y_j * (alpha_j_old - alpha_j_new)
      viii. Mettre à jour le biais b en calculant :
           - b_1 = b - E_i - y_i * (alpha_i - alpha_i_old) * K(x_i, x_i) - y_j * (alpha_j - alpha_j_old) * K(x_i, x_j)
           - b_2 = b - E_j - y_i * (alpha_i - alpha_i_old) * K(x_i, x_j) - y_j * (alpha_j - alpha_j_old) * K(x_j, x_j)
           - Si 0 < alpha_i < C, alors b := b_1.
           - Sinon, si 0 < alpha_j < C, alors b := b_2.
           - Sinon, b := (b_1 + b_2) / 2.
```

```
27.    b. Si aucune mise à jour d'alpha n'a été effectuée pendant cette itération, alors
augmenter le compteur de non-amélioration de 1. Sinon, le réinitialiser à 0.
28.    c. Si le compteur de non-amélioration a atteint un certain nombre maximal (par exemple
10), alors sortir de la boucle.
29. 3. Calculer w en utilisant : w = sum(alpha_i * y_i * x_i) pour tous les alpha_i non nuls.
30. 4. Trouver un exemple de support x_i avec un alpha_i non nul et calculer le biais b en
utilisant la formule : b = y_i - sum(alpha_j * y_j * K(x_i, x_j)) pour tous les exemples de
support x_j.
31. 5. Renvoyer w et b.
```

Annexe 9 : Pseudo-code descriptif de l'entraînement d'un modèle de Random Forest.

```
1. Entrée:
 2.     - data_train: ensemble des données d'entraînement
 3.     - data_test: ensemble des données de test
 4.     - num_trees: nombre d'arbres à construire
 5.     - num_features: nombre de caractéristiques à considérer à chaque étape de construction de
l'arbre
 6.     - max_depth: profondeur maximale de chaque arbre
 7.
 8. Sortie:
 9.     - prédictions finales pour l'ensemble de données de test
10.
11. 1. Pour chaque arbre k allant de 1 à num_trees, faire :
12.     1.1 Sélection aléatoire d'un échantillon de données d'entraînement avec remplacement
13.     1.2 Sélection aléatoire de num_features caractéristiques parmi l'ensemble complet des
caractéristiques
14.     1.3 Construction d'un arbre de décision en utilisant l'échantillon de données et le sous-
ensemble de caractéristiques sélectionnés avec une profondeur maximale de max_depth
15.     1.4 Stockage de l'arbre de décision k
16.
17. 2. Pour chaque instance de test i dans data_test, faire :
18.     2.1 Pour chaque arbre k, prédire la classe de l'instance i en utilisant l'arbre k
construit à l'étape 1
19.     2.2 Stocker les prédictions pour l'instance i
20.
21. 3. Pour chaque instance de test i dans data_test, sélectionner la classe majoritaire parmi
les prédictions stockées à l'étape 2
22.
23. 4. Retourner les prédictions finales pour l'ensemble des données de test
```

Annexe 10 : CV et liste des publications

Alice Martin

Manager, Health Strategy

42 rue Césaria Evora
75019, Paris

Pro. : alice.martin@iacpartners.com
Perso. : martinalicem@gmail.com

+33 (0)6.03.03.27.94.

Career objective: combining my research interests and consulting experience to make a positive impact on health systems

Current Position

IAC Partners, Paris (France) and Cotonou (Benin)
Consulting firm

Consultant (Aug. 2018), Senior Consultant (Jan 2021), & Manager (June 2022) in Healthcare & Life Sciences with a focus on Growth Strategy

Highlighted projects:

- **Project Manager for Health Funding Reform Impact Modelling:** Leveraged expertise in analytics and stakeholder negotiations to model the impact of health funding reform for several major actors in the healthcare ecosystem and support them in discussions with government officials.
- **Project Manager for Health Digital Strategy and Transformation**: Led the definition, structuring, and implementation of a comprehensive digital health strategy for a leading In Vitro Diagnostics (IVD) company.
- **Project Manager for Innovation Strategy Projects**: Managed diverse innovation strategy projects for both public and private organizations.
- **Head of Research in AI Applied to Health**: Oversaw and launched multiple research initiatives (2 PhDs and 1 Postdoc) across various health contexts, including infectious diseases and chronic conditions such as cancer.

Skills learned:
Advanced project management. Microsoft Software Suite. Hands-on knowledge of the Healthcare Ecosystem. Technical benchmarks & market studies. Data analytics & visualization. Stakeholders' management and leadership.

DISP, Lyon
Research lab
HCL, *Lyon University Hospital*
INSA of Lyon
University

June 2019 to 2023: PhD student in Applied Computer Science at INSA of Lyon

Topic:

- Predict the patient journey, a model based on Artificial Intelligence
- Teaching activities: Data Science, AI for Health, Operations Research

Skills learned:
Machine learning. Statistics. Programming. Health Informatics.

Work Experiences

Hôpital de la Pitié-Salpêtrière, Cephepi, Paris

November 2020 to June 2021: Research Engineer and Project Manager

Cephepi (Centre de pharmaco-épidémiologie) is a mixed research structure between AP-HP, INSERM and Pierre Louis Institute. Covidom is an e-health application for COVID-19 patient monitoring. Its database provides unique research opportunities on clinical risk factors and symptoms management.

Topics:

- Providing **statistical support** for **clinical ancillary studies**.
- Following up with clinical studies. Reporting to the Scientific Committee.

Pertikos, Santiago de Chile

January to July 2018: Master's thesis in pediatric hospital Exequiel Gonzalez Cortes

Pertikos is a start-up funded by academics of the University of Chile.

Topic:

- Conception & deployment of a machine learning **algorithm suggesting treatments based on diagnoses** and clinical history.
- Support to the **implementation of an Electronic Health Record System**.

Education

INSA of Lyon
Top 10 Engineering Schools in France

2018: Master's degree in Industrial Engineering
Highlighted courses:
- Operations Research. Object Oriented Programming. Flows Simulation. Petri Net. Databases. Data analytics. Statistics.
- Industrial Fabrication Processes. Production optimization. Materials. Automation. Design of a production system. Quality and maintenance processes.

Universidad de Chile, FCFM
Best Engineering School in Chile

2018: Exchange program in Industrial Engineering
Highlighted courses:
- Modeling & optimization (linear and non-linear programming) – Top of class.
- Marketing.
- Strategic Management.

Experiences:
- Research Assistantship on Modeling and Optimization.

Hoche High School, Versailles
Top 3 Preparatory Class in France

2011 to 2012: Preparatory Class – Physics & Engineering Sciences
2011: High School Diploma with Highest Honours
Scientific Diploma major in Mathematics

Languages

Mother tongue

C2 level – fluent

C2 level – fluent
TOEIC 970/990

Skills and Interests

Langages: Python - Julia – Java – VBA
Packages: Scikit-learn – Pandas & NumPy - Matplotlib
Softwares: Jupyter lab – Atom – MS Office

Hiking
People & Learning

Publications

Alice Martin, A. Guinet, Julien Fondrevelle, Jean-Baptiste Guillaume, Eric Dubost. *Implémentation de protocoles générés par apprentissage automatique dans une structure d'hospitalisation à domicile : état des lieux et mise en perspective*. GISEH 2020, 10ème Conférence Francophone en Gestion et Ingénierie des Systèmes Hospitaliers, Apr 2020, Valenciennes

A Martin, JB Guillaume, A Guinet, J Fondrevelle. *Treatment Protocols Generated by Machine Learning: Putting a Case Study of Hospitalization at Home into Perspective*. Healthcare Systems: Challenges and Opportunities, 2022

Alice Martin, Victor Manach, Virginie Dauphinot, Pauline Desnavailles, Frédéric Gervais, Antoine Garnier-Crussard, Jean-Baptiste Guillaume, Julien Fondrevelle, Alain Guinet, Pierre Krolak Salmon. *Drivers of costs of care and disease trajectory for patients with neurocognitive disorders: a machine learning approach*. ORAHS 2022, 48th EURO Working Group on Operational Research Applied to Health Services, July 2022, Bergamo, Italy.

Submitted, under review (Annals of Operations Research) : Alice Martin, Victor Manach, Virginie Dauphinot, Pauline Desnavailles, Frédéric Gervais, Antoine Garnier-Crussard, Jean-Baptiste Guillaume, Julien Fondrevelle, Alain Guinet, Pierre Krolak Salmon. *Clustering and Markov-Chain interpretable modelling to predict disease trajectory and drivers of costs for neurocognitive disorders.*

Alice Martin, Jean-Baptiste Guillaume, Julien Fondrevelle, Alain Guinet. *Patient journey and care trajectory prediction from a medico-economic perspective: an approach based on artificial intelligence*. HCSE 2023, Portugal.