HAL
open science

Explorer les trajectoires de patients via les bases médico-économiques : application à l'infarctus du myocarde

Jessica Pinaire

To cite this version:

Jessica Pinaire. Explorer les trajectoires de patients via les bases médico-économiques : application à l'infarctus du myocarde. Médecine humaine et pathologie. Université Montpellier, 2017. Français. NNT : 2017MONTS020 . tel-01903477

HAL Id: tel-01903477

https://theses.hal.science/tel-01903477

Submitted on 24 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de Docteur

Délivré par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **I2S***
Et des unités de recherche **UPRES EA2415, UMR 5506**
et du Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier

Spécialité: **Biostatistique**

Présentée par **Mme Jessica PINAIRE**
jessica.pinaire@chu-nimes.fr

Explorer des trajectoires de patients via les bases médico-économiques : Application à l'infarctus du myocarde

Soutenue le 17/10/2017 devant le jury composé de

Jean-Pierre Daurès	Pr. émérite	Université de Montpellier	Président
Jean Charlet	CR	AP-HP & INSERM/U1142	Rapporteur
Philippe Lenca	Pr.	IMT Atlantique	Rapporteur
Régis Beuscart	Pr. émérite	Université de Lille 2	Examinateur
Sarah Cohen-Boulakia	Pr.	Université Paris-Sud	Examinateur
Paul Landais	PU-PH	Université de Montpellier	Directeur
Jérôme Azé	Pr.	Université de Montpellier	Co-directeur
Sandra Bringay	Pr.	Université Paul-Valéry	Encadrante

* **I2S :** Information, Structures Systèmes.

Résumé

Avec environ 120 000 personnes atteintes chaque année, 12 000 décès suite à la première crise et 18 000 décès après une année, l'infarctus du myocarde est un enjeu majeur de santé publique. Cette pathologie nécessite une hospitalisation et une prise en charge dans une unité de soins intensifs de cardiologie. Pour étudier cette pathologie, nous nous sommes orientés vers les bases hospitalières nationales du Programme de Médicalisation du Système d'Information (PMSI).
La collecte des données hospitalières dans le cadre du PMSI génère sur le plan national des bases de données de l'ordre de 25 millions d'enregistrements par an. Ces données, qui sont initialement recueillies à des fins médico-économiques, contiennent des informations qui peuvent avoir d'autres finalités : amélioration de la prise en charge du patient, prédiction de l'évolution des soins, planification de leurs coûts, *etc.*
Ainsi émerge un autre enjeu : celui de fournir des outils d'explorations des trajectoires hospitalières des patients à partir des données issues du PMSI. Par le biais de plusieurs objectifs, les travaux menés dans le cadre de cette thèse ont pour vocation de proposer des outils combinant des méthodes issues de trois disciplines : informatique médicale, fouille de données et biostatistique.
Nous apportons quatre contributions. La première contribution concerne la constitution d'une base de données de qualité pour analyser les trajectoires de patients. La deuxième contribution est une méthode semi-automatique pour la revue systématique de la littérature. Cette partie des travaux délimite les contours du concept de trajectoire dans le domaine biomédical. La troisième contribution est l'identification des parcours à risque dans la prédiction du décès intra-hospitalier. Notre stratégie de recherche s'articule en deux phases : 1) Identification de trajectoires types de patients à l'aide d'outils issus de la fouille de données ; 2) Construction d'un modèle de prédiction à partir de ces trajectoires afin de prédire le décès. Enfin, la dernière contribution est la caractérisation des flux de patients à travers les différents évènements hospitaliers mais aussi en termes de délais d'occurrences et de coûts de ces évènements. Dans cette partie, nous proposons à nouveau une alliance entre une méthode de fouille de données et de classification de données longitudinales.

Mots clés : Trajectoires hospitalières, PMSI, Infarctus du myocarde, Fouille de données, Grandes bases, Prédiction, Machine learning, Flux de patients, Données longitudinales, Classification.

Abstract

With approximately 120,000 people affected each year, 12,000 deaths from the first crisis and 18,000 deaths after one year, myocardial infarction is a major public health issue. This pathology requires hospitalization and management in an intensive care cardiology unit. We study this pathology using the French national Prospective Paiement System (PPS) databases.
The collection of national hospital data within the framework of the PPS generates about 25 million records per year. These data, which are initially collected for medico-economic purposes, contain information that may have other purposes : improving patient care, predicting the evolution of care, planning their costs, *etc.*
Another emerging issue is that of providing tools for exploring patients' hospital trajectories using data from the PPS. Through several objectives, this thesis aims to suggest tools combining methods from three disciplines : medical computing, data mining and biostatistics.
We make four contributions. The first contribution concerns the constitution of a quality database to analyze patient trajectories. The second contribution is a semi-automatic method for the systematic review of the literature. This part of the work delineates the contours of the trajectory concept in the biomedical field. The third contribution is the identification of care trajectories in the prediction of intra-hospital death. Our research strategy is divided into two phases : 1) Identification of typical patient trajectories using data mining tools ; 2) Construction of a prediction model from these trajectories to predict death. Finally, the last contribution is the characterization of patient flows through the various hospital events, also considering of delays and costs. In this contribution, we propose a combined-data mining and a longitudinal data clustering technique.

Keywords : Healthcare trajectories, PPS, Myocardial infarction, Data Mining, Large databases, Prediction, Machine learning, Patient flows, Longitudinal data, Clustering.

Remerciements

Parce que sa rencontre a orienté ma trajectoire professionnelle (là où elle m'a conduite aujourd'hui) et parce que son soutien a contribué à l'initialisation de cette thèse, c'est naturellement le Pr. Jean-Pierre Daurès que je remercie en tout premier. J'ai été très honorée qu'il préside mon jury.

Soutenue et dirigée par mon chef de service, Paul Landais, ces trois années ont été riches d'enseignements. Je le remercie de m'avoir permis de réaliser ces travaux, d'en avoir été le directeur d'encadrement et de m'avoir accordé sa confiance dans l'accomplissement de cette tâche. Sa patience, sa disponibilité (malgré un emploi du temps bien rempli) et sa sagesse ont été l'équivalent de la ciselure dans un travail d'orfèvrerie.

Venant d'un horizon assez différent, j'ai eu la chance de bénéficier en termes d'encadrement d'une vision plus orientée informatique médicale et fouille de données grâce à Jérôme Azé et Sandra Bringay. C'est dans ce cocktail de savoirs et d'expériences diverses qu'ont pu se développer ces travaux de recherche (et qui d'ailleurs se poursuivront). Je suis très reconnaissante envers Jérôme et Sandra qui m'ont encadrée, soutenue (en toutes circonstances) et ont éclairé le chemin à parcourir tout au long de ces trois années. De plus, leur positivisme à toutes épreuves a été pour moi un vrai bol d'oxygène roboratif.

Je remercie l'ensemble de l'équipe Advanse pour son accueil, sa chaleur humaine, sa disponibilité et tous ses conseils avisés. Ces trois années passées au sein de l'équipe m'ont beaucoup apportée à tout point de vue, j'y ai appris de nouvelles techniques et une façon de travailler qui me sera très précieuse pour la suite de ma carrière. Merci à Arnaud pour m'avoir fait partagé un peu de son expertise dans le domaine de la visualisation. J'ai pu y voir l'immense potentiel de cette discipline. Merci à Dino pour sa gentillesse, son écoute, ses conseils, les soirées et aussi le panorama au pont du Gard. Merci à Maguelonne d'avoir suivi ces travaux depuis le début, pour son soutien et sa perspicacité. Merci à Pascal, pour sa disponibilité, ses bons conseils et aussi pour les quelques soirées partagées avec les doctorants.

Pour les belles sorties, les discussions, les soirées et la bonne humeur, merci à tous les doctorants et post-doctorants Antonio, Vijay, Amine, Mike, Sarah, Erick, Samiha, Lynda, Bilel, Andon. Une mention spéciale pour Mike avec qui cela a été trois ans de soutien mutuel et de partage, mais aussi de belles soirées animées, de longues conversations et de rigolades bien entendu... Et surtout, le début d'une belle amitié aussi !

Je remercie également l'ensemble du jury pour l'évaluation de ces travaux. Merci aux rapporteurs qui ont permis d'apporter des améliorations au manuscrit tant esthétiques, qu'informatives.

Et pour finir, merci à ma famille pour son soutien inconditionnel dans les bons comme dans les mauvais moments. Ma profonde gratitude s'adresse plus particulièrement à mes parents, sans lesquels tout ceci n'existerait pas. Merci pour avoir relu encore et encore mes chapitres de thèse et pour leur aide précieuse dans les préparatifs de la soutenance. Qu'ils voient dans ce travail la concrétisation de leurs efforts.

Glossaire

ACP	Analyse en Composantes Principales
ACM	Analyse en Composantes Multiples
ADN	Acide DésoxyriboNucléique
AFC	Analyse Factorielle des Correspondances
AFCM	Analyse Factorielle des Correspondances Multiples
AIC	Akaike Information Criterion
AMI	Acute Myocardial Infarction
ANAES	Agence Nationale d'Accréditation et d'Évaluation en Santé
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
ARS	Agence Régionale de Santé
ATIH	Agence Technique de l'Information sur l'Hospitalisation
AVC	Accident Vasculaire Cérébral
CALIBER	CArdiovascular disease research using LInked Bespoke studies and Electronic health Records
CCAM	Classification Commune des Actes Médicaux
CeNGEPS	Centre National de Gestion des Essais de Produits de Santé
CépiDC	Centre d'épidémiologie sur les causes médicales de DéCès
CFPM	Contextual Frequent Pattern Mining
CIM-10	Classification Internationale des Maladies, dixième révision
CHU	Centre Hospitalier Universitaire
CMA	Complications ou Morbidités Associées
CMD	Catégorie Majeure de Diagnostics
CoPaM	Contextual discriminant Pattern Mining
DAS	Diagnostics Associés
DIM	Département d'Information Médical
DMI	Dispositif Médical Implantable
DP	Diagnostic Principal
DR	Diagnostic Relié
DRG	Diagnosis Related Groups
ENC	Échelle Nationale des Coûts
FINESS	FIchier National des Établissements Sanitaires et Sociaux

GAM	Generalized Adaptative Models
GHM	Groupe Homogène de Malades
GHS	Groupe Homogène de Séjours
GHT	Groupe Hospitalier de Territoire
HAD	Hospitalisation à Domicile
HAS	Haute Autorité de Santé (anciennement ANAES)
HR	Hazard Ratio
IC 95%	Intervalle de Confiance à 95%
ICD	International Classification Diseases
ICP	Intervention Coronaire Percutanée
IM	Infarctus du Myocarde
Insee	Institut national de la statistique et des études économiques
InVS	Institut de Veille Sanitaire
IRaMuteQ	Interface de R pour l'analyse Multidimensionnelle de textes et de Questionnaires
IRDES	Institut de Recherche et Documentation en Économie de Santé
KNN	modèle des k plus proches voisins
LCM	Linear time Closed itemset Miner
LCS	Longest Common Substring
LDA	Latent Dirichlet Allocation
MARS	Multivariate Adaptative Regression Spline Models
MCO	Médecine, Chirurgie, Obstétrique et odontologie
MINAP	Myocardial Ischaemia National Audit Project
MONICA	MONItoring of trends and determinants of CArdiovascular disease
NB	modèle Naïf de Bayes
NICOR	National Institute for Cardiovascular Outcomes Research
OCDE	Organisation de Coopération et de Développement Économiques
OMA	Optimal Matching Analysis
OMS	Organisation Mondiale de la Santé
OR	Odds-Ratio
PACA	Provence Alpes Côte d'Azur
PMSI	Programme de Médicalisation des Systèmes d'Information
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
RAUC	ROC Area Under Curve
RIM-P	Recueil d'Information Médicalisé pour la Psychiatrie
RIKS-HIA	Register of Information and Knowledge about Swedish Heart Intensive care Admissions
RL	Régression Logistique
ROC	Receiver Operating Characteristic
RSA	Résumé de Sortie Anonyme

RSS	Résumé de Sortie Standardisé
RUM	Résumé d'Unité Médicale
SAMU	Service d'Aide Médicale Urgente
SARIMA	Seasonnal AutoRegressive Moving Average
SCA	Syndrome Coronarien Aigu
SAPPHIRE	Stenting and Angioplasty with Protection of Patients with High Risk for Endarterectomy
SNIIRAM	Système National d'Information Inter-Régimes de l'Assurance Maladie
SQL	Structured Query Language
SSR	Soins de Suite et de Réadaptation
SVM	Support Vector Machine
SWEDEHEART	Swedish Web-system for Enhancement and Development of Evidence-based care in Heart disease Evaluated Accorded to Recommended Therapies
TAL	Temporal Action Logics
T2A	Tarification à l'Activité
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
USIC	Unité de Soins Intensifs de Cardiologie
VIH	Virus de l'Immunodéficience Humaine
VPP	Valeur Prédictive Positive
WHO	World Health Organization

Sommaire

1 Introduction **1**
1.1 Contexte . . . 1
1.2 Objectifs et hypothèses de recherche . . . 2
1.3 Résumé des contributions . . . 4
1.4 Organisation du manuscrit . . . 6
1.5 Production scientifique . . . 6
1.5.1 Revues internationales avec comité de lecture . . . 6
1.5.2 Conférences et ateliers nationaux avec comité de lecture . . . 7

I Les données **9**

2 Le PMSI dans tous ses états **13**
2.1 Historique . . . 14
2.2 Principes généraux . . . 15
2.2.1 Hospitalisation : production d'un Résumé de Sortie Standardisé 15
2.2.2 Classification des séjours : les Groupes Homogènes de Malades 17
2.2.3 Anonymisation : production du Résumé de Sortie Anonymisé . 17
2.3 Reconstitution du parcours hospitalier . . . 18
2.3.1 Identifiant patient : processus de création . . . 18
2.3.2 Anomalies dans le processus d'anonymisation . . . 19
2.3.3 Fiabilité des identifiants patients . . . 21
2.4 Épidémiologie et PMSI . . . 23
2.4.1 En pratique : des exemples . . . 23
2.4.2 Les limitations rencontrées . . . 26
2.5 Conclusion . . . 27

3 L'infarctus du myocarde à l'hôpital **29**
3.1 Protocole d'analyses descriptives statistiques . . . 30
3.1.1 Étape 1. Constitution de la base de données . . . 30
3.1.2 Étape 2. Hospitalisations avec IM . . . 31
3.1.3 Étape 3. Réadmissions pour IM . . . 31
3.1.4 Étape 4. Décès et facteurs de risque . . . 32
3.2 Description des résultats . . . 32
3.2.1 Étape 2. Hospitalisation avec IM . . . 33
3.2.2 Étape 3. Réadmissions pour IM . . . 41
3.2.3 Étape 4. Décès et facteurs de risque . . . 46

3.3 Commentaires . . . 49
3.3.1 Hospitalisations avec IM . . . 49
3.3.2 Ré-hospitalisations pour IM . . . 51
3.3.3 Décès et facteurs de risque . . . 52
3.3.4 Exhaustivité des données : les limites . . . 53
3.4 Conclusion . . . 54

II Les trajectoires dans la littérature 57

4 Les trajectoires dans la littérature 61
4.1 Analyse semi-automatique de la littérature . . . 62
4.1.1 Questions de recherche . . . 63
4.1.2 Processus de revue semi-automatique . . . 63
4.2 Expérimentations . . . 66
4.2.1 Étape 2. Première approche par fouille de textes . . . 66
4.2.2 Étape 3. Analyse manuelle des articles sélectionnés . . . 71
4.3 Discussion . . . 73
4.3.1 Réponses aux questions de recherche . . . 73
4.3.2 Avantages et limites . . . 76
4.4 Conclusion . . . 77

III De la trajectoire du patient à la prédiction 79

5 Motifs contextuels 83
5.1 Motifs séquentiels contextuels . . . 84
5.1.1 Définitions préliminaires . . . 85
5.1.2 Description du processus de fouille . . . 89
5.2 Expérimentations . . . 92
5.2.1 Motifs fréquents . . . 93
5.2.2 Motifs discriminants . . . 97
5.3 Discussion . . . 98
5.4 Conclusion . . . 100

6 Prédire le décès 103
6.1 Protocole de prédiction . . . 106
6.1.1 Étape 1. Tri des motifs . . . 106
6.1.2 Étape 2. Constitution de la base de données . . . 107
6.1.3 Étape 3. Modélisation par contexte . . . 107
6.1.4 Étape 4. Validation externe . . . 111
6.2 Expérimentations . . . 112
6.2.1 Étape 3.e. Choix du modèle . . . 113
6.2.2 Étape 5. Validation externe . . . 114
6.2.3 Identification des parcours à risque . . . 115
6.3 Discussion . . . 117
6.3.1 Détermination du meilleur couple *(modèle, score)* . . . 117
6.3.2 Identification des parcours à risque . . . 118

6.3.3 Compétitivité de notre approche 119
6.4 Conclusion 120

IV Des trajectoires de patients à la planification sanitaire 123

7 Extraction de motifs spatio-temporels 127
7.1 Motifs spatio-temporels 129
7.1.1 Définitions préliminaires 130
7.1.2 Description du processus de fouille 132
7.2 Expérimentations 133
7.2.1 Extraction de motifs 134
7.2.2 Visualisation des flux de patients 137
7.3 Discussion 141
7.4 Conclusion 144

8 Profils de délais et de tarifs 147
8.1 Classification de données longitudinales quantitatives 148
8.1.1 Définitions préliminaires 149
8.1.2 Description du processus de classification 152
8.2 Expérimentations 154
8.2.1 Trajectoires de délais inter-séjours 154
8.2.2 Trajectoires de tarifs de séjours 157
8.3 Discussion 159
8.3.1 Analyse des résultats 159
8.3.2 Limites de cette étude 161
8.4 Conclusion 162

V Conclusion générale et perspectives 165

9 Conclusions et perspectives 169
9.1 Bilan 170
9.2 Perspectives 171
9.2.1 Méthodes 172
9.2.2 Applications au domaine de la santé 175

VI Annexes 180

A Données 181

B Trajectoires dans la littérature 184

C Extraction de motifs séquentiels contextuels 191

D Prédire le décès **194**
D.1 Distances de similarités entre chaînes de caractères 194
D.1.1 Distances basées sur des opérations d'édition 194
D.1.2 Mesures basées sur les q-grammes 195
D.1.3 Mesures heuristiques 195
D.2 Modèle par arbre à inférence conditionnelle 196
D.3 Mesures de performances d'un modèle 197
D.4 Vecteur maximum 199

E Extraction de motifs spatio-temporels **201**

Références **208**

Table des figures

1.1 Schéma de la thèse. 5

2.1 Principe du PMSI MCO schématisé : de l'hospitalisation à l'anonymisation des données. 16
2.2 Chronologie normale des séjours hospitaliers. 22
2.3 Algorithme de détermination du niveau de fiabilité de l'identifiant patient. 22

3.1 Nombre de séjours et nombre de patients entre 20 et 99 ans hospitalisés pour IM par année. 33
3.2 Répartition des séjours hospitaliers pour IM par sexe et classe d'âge. . 34
3.3 Taux des hospitalisations pour IM (pour 10 000 habitants). 36
3.4 Taux d'hospitalisation avec IM par sexe (trait plein pour les hommes et pointillés pour les femmes) et classe d'âge pour 10 000 habitants. . 36
3.5 Taux standardisés d'hospitalisation avec IM pour 10 000 hommes par région en 2009 et 2014 avec représentation des variations régionales. . 37
3.6 Taux standardisés d'hospitalisation avec IM pour 10 000 femmes par région en 2009 et 2014 avec représentation des variations régionales. . 38
3.7 Part des séjours et de patients hospitalisés pour IM en dehors de leur région d'habitation par année. 39
3.8 Répartition de la durée de séjour hospitalier selon la classe d'âge. . . 40
3.9 Nombre de séjours hospitaliers suivant le mois de l'année. 41
3.10 Taux de réadmissions pour IM selon le nombre de jours écoulés après une hospitalisation index, par sexe. 43
3.11 Taux de réadmissions pour IM selon le nombre de jours écoulés après une hospitalisation index, par classe d'âge. 43
3.12 Taux de réadmissions pour IM à 3 mois, 9 mois et 12 mois après une hospitalisation index, par sexe et par région. 44
3.13 Proportion de décès par classe d'âge suivant le sexe et l'année. 47
3.14 Taux de létalité hospitalier par sexe (trait plein pour les hommes et pointillés pour les femmes) et par classe d'âge. 48

4.1 Étapes du processus de revue semi-automatique. 63
4.2 Nuages de mots : Trajectoire, PMSI, IM. 66
4.3 Analyse de similitude et représentation des communautés pour le thème T1. 68
4.4 Résultats d'une classification des articles pour le corpus T1 : Trajectoire. 69

4.5 Analyse de similitude sur les classes 3 et 4 issues de la première classification de T1. 70
4.6 Base de données et méthodes utilisées dans les études de trajectoires. 72

5.1 Hiérarchie des contextes de l'exemple détaillé dans le tableau 5.3 . . . 87
5.2 Étapes du processus de fouille. 89
5.3 Hiérarchie des contextes. 91
5.4 Construction de la trajectoire du patient. 92

6.1 Étapes du processus de modélisation. 106
6.2 Découpage de la base de données suivant trois échantillons : apprentissage, test et validation. 108

7.1 Visualisation des trajectoires de patients. Le temps est lié à la survenue d'une hospitalisation. 131
7.2 Étapes du processus de fouille. 132
7.3 Recalage des trajectoires de patients. 133
7.4 Essaims clos dans les trajectoires de GHM. 135
7.5 Essaims clos dans les trajectoires de DP. 136
7.6 Flux de patients dans les trajectoires de GHM. 139
7.7 Flux de patients dans les trajectoires de DP. 140

8.1 Impact du changement d'échelle dans le calcul de la distance de Fréchet. 150
8.2 Exemple de comparaison entre la moyenne de Fréchet et la distance euclidienne. 151
8.3 Étapes du processus de classification. 152
8.4 Schéma de classification et de répartition des groupes dans les classes. 153
8.5 Classes de délais à droite et mesures de dispersion à chaque estampille de temps par classe selon le sexe. 156
8.6 Classes de tarifs à droite et mesures de dispersion à chaque estampille de temps par classe selon le sexe. 158

C.1 Capture d'écran de l'interface graphique créée pour analyser et explorer les données. 193

D.1 Exemple de courbe ROC. 199

Liste des tableaux

3.1 Répartition du nombre d'hospitalisations pour IM selon le sexe. . . . 42
3.2 Distribution du délai entre deux séjours pour IM selon le sexe. 45
3.3 Risque de ré-hospitalisation pour IM selon le sexe, la classe d'âge et la présence d'un facteur de risque. 46
3.4 Répartition des décès selon le numéro de séjour d'hospitalisation. . . 46
3.5 Nombre de décès et taux de létalité pour 10 000 hospitalisations. . . . 47
3.6 Identification des facteurs de risque du décès hospitalier chez les +80 ans. 49

4.1 Questions de recherche. 63
4.2 Mots clés utilisés pour la recherche documentaire. 64

5.1 Base séquentielle de GHM. 86
5.2 Mise en valeur du motif ⟨(05M13)(05M06)⟩ (en gras) soit le GHM 05M13 suivi du GHM 05M06. Ce motif est fréquent dans la base pour un support minimum de 50%. 86
5.3 Mise en valeur du motif ⟨(05M13)(05M06)⟩ (en gras) avec les informations contextuelles sur l'âge et le sexe. Ce motif est spécifique aux personnes âgées. Une seule personne jeune est concernée. 88
5.4 Détail des contextes étudiés. 93
5.5 Extrait de motifs séquentiels contextuels fréquents pour les trajectoires de GHM au seuil de 1%. 94
5.6 Extrait de motifs séquentiels contextuels fréquents pour les trajectoires de DP au seuil de 1%. 96
5.7 Extrait de motifs séquentiels contextuels discriminants pour les trajectoires de DP au seuil de 5% et taux de croissance minimal de 2. . . 98

6.1 Récapitulatif des modèles et mesures de similarités utilisés pour la modélisation. 112
6.2 Performances des meilleurs modèles retenus pour les contextes 3-5 séjours et 5-60 séjours. 113
6.3 Validation externe des modèles retenus dans l'étape 4. 114
6.4 Modèles logistiques pour les trajectoires de GHM. 115
6.5 Modèles logistiques pour les trajectoires de DP. 116

7.1 Trajectoires de patients. Le temps est lié à la survenue d'une hospitalisation. 131

8.1 Répartition en pourcentage des hommes et des femmes (entre parenthèses) de chaque groupe dans les classes de délais. 157
8.2 Répartition en pourcentage des hommes et des femmes (entre parenthèses) de chaque groupe selon la classe de tarif. 159

A.1 Codes CIM-10 et codes CCAM pour l'algorithme de repérage d'un IM dans les bases médico-administratives. 181
A.2 Taux standardisés d'hospitalisation avec IM et pourcentage de variation annuelle des taux par région chez les hommes. 182
A.3 Taux standardisés d'hospitalisation avec IM et pourcentage de variation annuelle des taux par région chez les femmes. 183

B.1 Description des items et catégories observés 184
B.2 Références sources par item étudié 186

C.1 Codes CIM-10 retenus pour la sélection des séjours dans les trajectoires de patients. 191
C.2 Codes GHM présents dans le manuscrit. 192
C.3 Codes CIM-10 présents dans le manuscrit. 192

D.1 Matrice de confusion 198
D.2 Liste des hôtels. 200
D.3 Classement des hôtels selon leurs critères. 200

E.1 Regroupements des codes GHM pour la visualisation des flux de patients 201
E.2 Regroupements des codes CIM-10 pour la visualisation des flux de patients 205

Chapitre

1 Introduction

1.1 Contexte

L'infarctus du myocarde (IM), communément appelé « crise cardiaque » est la destruction partielle du muscle cardiaque. Le plus souvent, il est déclenché par l'occlusion d'une artère coronaire qui alimente le cœur en sang et donc en oxygène. En l'absence d'oxygène, les cellules du cœur meurent : c'est la nécrose. Cette dernière est plus ou moins étendue. Elle laisse une cicatrice sur le cœur et réduit sa capacité à se contracter. Il en résulte des troubles du rythme, une insuffisance cardiaque voire l'arrêt du cœur [Thygesen *et al.*, 2012].

L'IM est généralement dû à la présence de plusieurs facteurs [Dujardin et Cambou, 2005]. Ces facteurs sont appelés facteurs de risque de la maladie cardiovasculaire. Ils se divisent en deux catégories :

— les facteurs non modifiables tels que l'âge, l'ethnie, le sexe et l'histoire familiale (la prédisposition génétique) ;

— les facteurs modifiables, principalement influencés par les comportements de santé, tels que le diabète, la dyslipidémie, l'hypertension artérielle, le surpoids et l'obésité, le tabagisme, la consommation abusive d'alcool et les facteurs psychosociaux.

Ces facteurs de risque modifiables peuvent être évalués dans les établissements de soins de santé primaires et ils sont le signe d'un risque accru d'infarctus, d'accident vasculaire cérébral (AVC), de défaillance cardiaque et d'autres complications. Par ailleurs, de nombreuses campagnes de santé ont été développées pour lutter contre ces facteurs de risque. Ces campagnes, mises en œuvre par le ministre de la Santé, comprennent essentiellement les quatre actions suivantes : la lutte contre le tabagisme (notamment avec les prises de mesures d'interdiction de fumer dans les lieux publics), les actions sur les comportements alimentaires, la réduction de la consommation de sel et l'enrichissement des farines en vitamines B9. Elles ont pour but d'inciter les populations à faire les bons choix dans leurs comportement de santé.

L'IM est une maladie avec un risque fatal à court et à moyen terme. En 20 ans, le taux de mortalité est passé de 20 à 6% [Ghannem *et al.*, 2015]. Des progrès ont été réalisés avec la création des unités de soins intensifs cardiologiques (USIC), une prise en charge dès l'arrivée du Service d'Aide Médicale Urgente (SAMU) avec la prescription d'une thrombolyse pré-hospitalière, le développement de nouveaux outils (stents, antiagrégants...) et la création de centres et de programmes de réadaptation cardiaque. Cependant, son pronostic reste grave, en France, elle est à l'origine de 10 à 12% de la mortalité totale annuelle chez l'adulte. Par ailleurs, l'étude européenne MONICA (MONItoring of trends and determinants of CArdiovascular disease) [Investigators *et al.*, 1988] a montré qu'il y avait un gradient Nord-Sud dans la mortalité coronaire. De plus, elle confirme que pour les 35-64 ans les taux de mortalité chez la femme sont inférieurs aux taux de mortalité chez l'homme. Toutefois, on constate que, contrairement aux préjugés, les maladies cardiovasculaires ne sont plus réservées aux hommes. Elles progressent fortement chez les femmes de moins de 50 ans. En parallèle, on constate une augmentation de la prévalence de facteurs de risque tels que l'obésité [HAS, 2011], le tabagisme féminin [Beck *et al.*, 2010], la sédentarité [Le Quellec-Nathan, 2002], le stress au travail [Bureau International du Travail, 2003] et la pollution de l'air [Laaidi *et al.*, 2002, Pascal, 2009]. On peut craindre alors, un recul de cette tendance à la baisse de mortalité et une évolution à la hausse de cette mortalité dans les années à venir.

Finalement, dans ce travail de thèse nous nous focalisons sur deux enjeux liés à l'étude de cette pathologie : 1) la continuité de la lutte contre la mortalité ; 2) l'amélioration de la planification sanitaire (traitement, suivi du patient, gestion des flux de patients, prévoir les dispositifs pour les années à venir...) et dans le cadre d'une politique de réduction des dépenses de santé.

1.2 Objectifs et hypothèses de recherche

L'IM est une pathologie grave nécessitant une hospitalisation et une prise en charge dans une USIC. Malgré les limites en termes de couverture, qualité et validité des données, les bases nationales de données médico-économiques constituent une source d'informations considérable sur l'hospitalisation en France [Goldberg *et al.*, 2008]. De plus, grâce à l'identifiant anonyme de patient, il est possible de retracer le parcours d'un patient sur tout le territoire national quel que soit l'établissement de soins fréquenté [ATIH, 2014]. Ces bases de données représentent alors l'opportunité de pouvoir étudier des maladies nécessitant des soins dans un établissement de santé telles que l'IM. Nous avons donc choisi de mener nos investigations à partir de ces bases de données.

De nombreux travaux traitent de la prise en charge de l'IM, de l'apparition des premiers symptômes aux premiers soins dans une unité de cardiologie, en passant par l'intervention d'un service d'urgence comme le SAMU [Miller *et al.*, 2017]. Ces travaux concernent le suivi du patient à court et à moyen terme. Les études sur le long terme sont plus rares. Elles portent généralement sur l'estimation des tendances

des taux d'incidence [Parikh *et al.*, 2009] et de mortalité ou encore de l'apparition de complications particulières de l'IM comme l'insuffisance cardiaque [Gott *et al.*, 2007]. Dans notre cas, nous avons souhaité étudier cette pathologie non pas dans sa prise en charge immédiate, ni dans celle du suivi à court terme, mais dans celle du suivi sur une longue période sans *a priori* sur une complication en particulier. Or, le suivi du patient sur le long terme dans son parcours hospitalier est en réalité une notion intrinsèque au concept de trajectoire. Nous voici face à une prérogative qui est celle de l'exploration des trajectoires de patients via les données médico-administratives. Cette prérogative va s'accompagner d'un certain nombre de questions que nous allons résoudre tout au long de ce manuscrit par le biais d'objectifs à atteindre.

Ainsi, le premier objectif de cette thèse est de constituer une base de données de l'IM permettant d'étudier les trajectoires de ces patients. Née des sciences sociales, dans l'analyse des parcours de vie, la trajectoire est un sujet émergent dans la recherche biomédicale depuis ces 15 dernières années. Toutefois, elle recouvre diverses formes. C'est ainsi que nous avons pour deuxième objectif de déterminer de quelle façon est défini et étudié le concept de trajectoire dans la littérature biomédicale.

Nous contribuerons ensuite aux travaux existants avec notre propre conception de la trajectoire du patient via les bases médico-économiques. Dans le contexte de continuité de lutte contre la mortalité, nous nous fixons comme objectif de prédire le décès intra-hospitalier et d'identifier les parcours les plus à risque. Usuellement, ce type de modélisation prend en compte des paramètres cliniques. Nous faisons l'hypothèse que la trajectoire du patient peut-être un prédicteur du décès. Partant de cette hypothèse, nous envisageons de mettre en évidence les parties de parcours qui sont caractéristiques de sous-populations. Puis, dans un deuxième temps, il nous faudra mesurer la ressemblance entre une trajectoire de patient et ces parcours types afin d'établir s'il y a un lien de causalité avec le décès. De la sorte, nous décrirons les parcours pronostiques du décès.

Pour poursuivre ces analyses sur les trajectoires de patients, nous nous plaçons dans le contexte de la planification sanitaire. Nous nous fixons comme objectif suivant de caractériser les flux de patients. La cardiologie a d'ores et déjà établi les complications liées à l'IM. Néanmoins, nous ne savons pas dans quelles proportions les ré-hospitalisations se répartissent selon ces différentes complications, ni à quel moment du parcours elles surviennent et dans quels délais. Répondre à ces questions revient alors à déterminer si l'on peut détecter des phénomènes de groupes. De surcroît, nous rappelons que nous sommes également dans un contexte de réduction de coûts de santé. Il serait alors intéressant de caractériser ces flux de patients selon cet indicateur. Or, la plupart des méthodes qui sont mises en œuvre pour l'étude des coûts déterminent un coût moyen de coûts cumulés (prise en charge en urgence, hospitalisation, médicaments, suivi par un médecin de ville...). La comparaison porte ensuite sur ces coûts moyens qui ne permettent pas de mettre en évidence des tendances évolutives du coût. Par exemple, il n'est pas possible de déceler des patients qui auraient des parcours similaires mais avec des tendances évolutives des coûts différentes.

À l'issue de cette section, nous avons listé les objectifs à atteindre dans ce mémoire et nous avons fait émerger quelques pistes de réflexion pour y parvenir. Dans la section suivante, nous résumons des contributions.

1.3 Résumé des contributions

Ce travail de thèse se situe à l'intersection de trois disciplines qui sont la biostatistique, la fouille de données et l'informatique médicale. La combinaison de diverses méthodes issues de ces disciplines va nous permettre de construire des processus de recueil et d'analyse des données mais aussi d'extraire des connaissances. Dans l'optique de répondre aux attentes évoquées dans la section précédente, les contributions de cette thèse constituent 4 parties distinctes schématisées dans la figure 1.1.

La première contribution est la constitution de la base de données à partir des données du PMSI (Programme de Médicalisation des Systèmes d'Information). Cette contribution est centrale, puisque les autres contributions dépendent entièrement de cette dernière. Elle repose sur deux points clés : sélection de données de chaînage de qualité et repérage de la pathologie dans les bases de données. Comme évoqué au préalable, l'utilisation des bases médico-économiques présentent des inconvénients notamment dans la comparaison d'analyses longitudinales du fait de la qualité des données de chaînage. Pour remédier à cela, nous proposons une méthode de sélection des patients selon un processus de contrôle de cohérence sur un lot de séjours. Ensuite, la sélection des séjours avec IM s'effectue à l'aide d'un algorithme de repérage de la pathologie ayant une bonne sensibilité. Toutefois, il ne permet pas de récupérer les séjours pour un autre motif durant lequel le patient aurait présenté un IM. Ainsi, nous agrémentons cet algorithme avec des critères supplémentaires propres à cette pathologie : des actes de cardiologie interventionnelle.

La deuxième contribution est une étape préalable à l'analyse des trajectoires de patients dans les bases PMSI. Elle a pour but de définir les contours du concept même de trajectoire dans la littérature biomédicale. Pour atteindre cet objectif, il nous faut étudier/explorer de nombreux documents. Se pose alors la question de savoir comment traiter un grand volume de documents en un temps raisonnable ? De nombreux chercheurs utilisent d'ores et déjà des techniques issues de la fouille de textes pour faciliter la revue systématique de la littérature. C'est donc vers une approche semi-automatisée que nous nous orientons. L'approche que nous proposons débute par la combinaison de trois outils de fouille de textes (nuage de mots, analyse de similarités et classification de textes). Notre méthode se termine par une étape plus classique de revue systématique manuelle possible, grâce au filtre automatique établi par l'approche de fouille de textes.

La troisième contribution propose une modélisation du décès hospitalier suivant la trajectoire de soins et détermine ainsi les parcours les plus à risque. La modélisation du décès se fait dans l'état de l'art à l'aide d'indicateurs cliniques, parfois à l'aide d'indicateurs qualitatifs (le patient a de l'hypertension, le diabète, déjà fait un IM...) mais peu de modèles considèrent les données médico-économiques. De plus, il est

déjà possible de prédire le décès par l'association de comorbidités. Nous proposons de prendre en compte l'aspect séquentiel des données pour améliorer et expliciter cette prédiction. Pour faire cela, nous associons des méthodes issues à la fois de la fouille de données et de la biostatistique. Cette modélisation se décompose en deux phases. Une première phase consiste à extraire les parcours caractéristiques (fréquents) suivant le type de population à l'aide de motifs séquentiels. Ensuite, dans une seconde phase, ces motifs sont introduits dans un modèle prédictif du décès hospitalier comme descripteurs.

Enfin, la quatrième contribution propose une modélisation des flux de patients afin de les caractériser au travers des différents évènements hospitaliers mais aussi selon les délais entre ces évènements et leurs coûts. Cette contribution s'inscrit dans le cadre de la création d'outils de gestion sanitaire. Planifier consiste à organiser au mieux pour le plus grand nombre. Ainsi, contrairement à la contribution précédente, nous nous focalisons ici non pas sur le patient dans son individualité mais sur les ensembles de patients. Pour cela, nous avons étudié les phénomènes de groupes en combinant une nouvelle fois des méthodes issues de disciplines différentes. Nous proposons un processus en deux phases. La première phase caractérise les flux de patients à l'aide des motifs spatio-temporels. Dans le cas présent, les notions de spatialité et de temps seront redéfinies pour les besoins de notre étude. La seconde phase catégorise les patients dans des groupes de comportement type en termes de délais inter-hospitaliers et d'évolution de coûts par une méthode de classification de données longitudinales quantitatives.

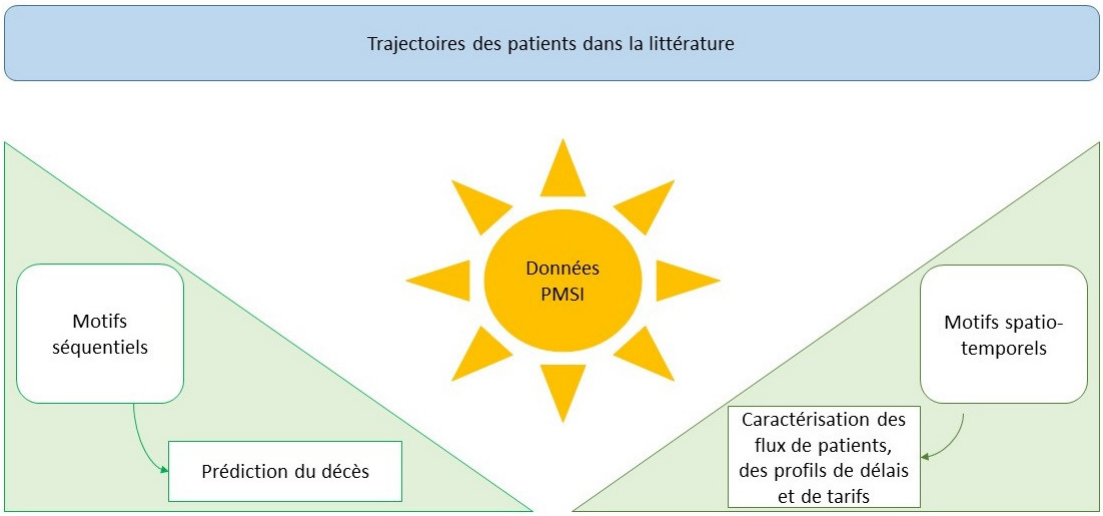


Figure 1.1 – Schéma de la thèse.

1.4 Organisation du manuscrit

Le plan de ce manuscrit est composé de cinq parties :

Partie I. Nous commençons par présenter les données du PMSI utilisées dans cette thèse. Cette partie consiste à introduire le vocabulaire spécifique du domaine dans le chapitre 2 mais aussi à faire le tour des avantages et des limites de l'usage de ces données dans les études épidémiologiques. Ensuite, dans le chapitre 3, nous décrivons plus particulièrement les données concernant l'IM.

Partie II. Nous nous intéressons à l'état de l'art sur les trajectoires de patients. Nous menons une revue systématique de la littérature à l'aide d'un processus semi-automatique afin de caractériser les concepts de la trajectoire du patient dans le domaine biomédical.

Partie III. Nous étudions la trajectoire du patient dans un but de prédiction. Le chapitre 5 introduit les concepts de la fouille de motifs séquentiels puis l'applique à nos données. Ensuite, dans le chapitre 6, les résultats issus de cette fouille sont intégrés dans un modèle de prédiction de décès hospitalier afin d'identifier les parcours hospitaliers les plus à risque.

Partie IV. Nous abordons ici la trajectoire du patient selon une autre perspective, dans un but de planification sanitaire. Le chapitre 7 introduit les concepts de la fouille de motifs spatio-temporels puis l'applique à nos données. Le chapitre 8 présente une méthode permettant de classer ces données longitudinales et l'applique sur les trajectoires de délais et de coûts hospitaliers afin de mettre en évidence au sein des flux de patients, identifiés préalablement, des profils de groupes en termes d'occurrences d'hospitalisations et de coûts.

Partie V. Enfin, ce manuscrit se termine par un bilan général des différentes contributions et nous proposons des perspectives de recherche associées.

1.5 Production scientifique

1.5.1 Revues internationales avec comité de lecture

- **Pinaire, J.**, Azé, J., Bringay, S., and Landais, P. Patient Healthcare Trajectory – an Essential Monitoring Tool : A systematic review. *Health Information Science and Systems*, April 2017.

1.5.2 Conférences et ateliers nationaux avec comité de lecture

1.5.2.1 Conférences

- **Pinaire, J.**, Azé, J., Bringay, S., Landais, P. Infarctus du myocarde : quelles sont les trajectoires de soins pronostiques du décès à l'hôpital ? *In IC2017 28es Journées francophones d'Ingénierie des Connaissances*. Caen, France, July 2017b.
- **Pinaire, J.**, Azé, J., Bringay, S., Landais, P. Extraire semi-automatiquement des connaissances dans la littérature biomédicale. *In IC2016 27es Journées francophones d'Ingénierie des Connaissances*. Montpellier, France, June 2016b.
- **Pinaire, J.**, Azé, J., Bringay, S., Landais, P. Recherche et visualisation de patterns dans les parcours de soins des patients ayant eu un infarctus du myocarde. *In JGS : 8ième Journées Grand Sud de l'Information Médicale*. Nîmes, France, June 2015b.

1.5.2.2 Ateliers

- **Pinaire, J.**, Azé, J., Bringay, S., Landais, P. Extraire semi-automatiquement des connaissances dans les archives ouvertes et les bases de publications. *In DATA4IST : Exploration et analyse des sources IST pour la recherche et ses environnements*. Paris, France, May 2016a.
- Mercadier, Y., **Pinaire, J.**, Azé, J., Bringay, S., and Teisseire, M. La confiance est dans l'air ! Application à l'identification des parcours hospitaliers. *In GAST : Gestion et Analyse des données Spatiales et Temporelles*. Reims, France, January 2016b, IRISA.
- **Pinaire, J.**, Rabatel, J., Azé, J., Bringay, S., Landais, P. Recherche et visualisation de trajectoires dans les parcours de soins des patients ayant eu un infarctus du myocarde. *In SIIM : Symposium Ingénierie de l'Information Médicale*. Rennes, France, June 2015c.

1.5.2.3 Posters/Démonstrations

- Mercadier, Y. **Pinaire, J.**, Azé, J., Bringay, S., and Teisseire, M. Manipulation interactive d'ensemble de motifs : application aux parcours hospitaliers. *In EGC : Extraction et Gestion des Connaissances*. Reims, France, January 2016a.
- **Pinaire, J.**, Ben Alouane, S., Azé, J. Bringay, S., Landais, P., and Sallaberry, Visualisation interactive de trajectoires de patients *In IC : Ingénierie des Connaissances*. Rennes, France, June 2015a.

Partie I

Les données

La vérité de demain se nourrit de l'erreur d'hier.

Antoine de Saint-Exupéry.

Table des matières

2 Le PMSI dans tous ses états **13**
2.1 Historique 14
2.2 Principes généraux 15
2.2.1 Hospitalisation : production d'un Résumé de Sortie Standardisé 15
2.2.2 Classification des séjours : les Groupes Homogènes de Malades 17
2.2.3 Anonymisation : production du Résumé de Sortie Anonymisé . 17
2.3 Reconstitution du parcours hospitalier 18
2.3.1 Identifiant patient : processus de création 18
2.3.2 Anomalies dans le processus d'anonymisation 19
2.3.3 Fiabilité des identifiants patients 21
2.4 Épidémiologie et PMSI 23
2.4.1 En pratique : des exemples 23
2.4.2 Les limitations rencontrées 26
2.5 Conclusion 27

3 L'infarctus du myocarde à l'hôpital **29**
3.1 Protocole d'analyses descriptives statistiques 30
3.1.1 Étape 1. Constitution de la base de données 30
3.1.2 Étape 2. Hospitalisations avec IM 31
3.1.3 Étape 3. Réadmissions pour IM 31
3.1.4 Étape 4. Décès et facteurs de risque 32
3.2 Description des résultats 32
3.2.1 Étape 2. Hospitalisation avec IM 33
3.2.2 Étape 3. Réadmissions pour IM 41
3.2.3 Étape 4. Décès et facteurs de risque 46
3.3 Commentaires 49
3.3.1 Hospitalisations avec IM 49
3.3.2 Ré-hospitalisations pour IM 51
3.3.3 Décès et facteurs de risque 52
3.3.4 Exhaustivité des données : les limites 53
3.4 Conclusion 54

Chapitre

2

Le PMSI dans tous ses états

Le premier système de paiement prospectif, basé sur des groupes homogènes de malades [1], a été établi aux États-Unis en 1983. L'objectif de ce système était de contrôler les dépenses des établissements de soins de santé et de rationaliser les coûts [Grant *et al.*, 1996]. Par la suite, des systèmes d'information médicale similaires ont été adoptés dans de nombreux autres pays industrialisés notamment, en France.

Le Programme de Médicalisation du Système d'Information (PMSI) a été progressivement mis en place en France depuis le début des années 90. Il concerne tous les établissements de soins de courte durée en médecine, chirurgie, obstétrique et odontologie (MCO). Son objectif principal est de mettre en relation l'activité médicale (pathologies et modes de prise en charge) et les moyens de fonctionnement des établissements. Les informations ainsi recueillies sont utilisées essentiellement pour la tarification à l'activité (T2A) et pour la planification de l'offre de soins. Néanmoins, malgré les limites inhérentes aux bases de données médico-administratives, elles sont de plus en plus utilisées pour améliorer la connaissance épidémiologique du recours à l'hospitalisation pour certaines affections et contribuer à la surveillance en santé publique [Schott *et al.*, 2002]. En outre, ces études ont été rendues possibles notamment grâce à l'adoption d'un numéro anonyme de patient dans les bases PMSI. Ce numéro relie les hospitalisations d'un même patient. Cependant, les anomalies de codage peuvent parfois impacter la qualité de cet identifiant et rendre son usage impossible [Trombert-Paviot *et al.*, 2008].

Ce chapitre a pour vocation de présenter les définitions classiques issues du domaine de l'information médicale dans le cadre du milieu hospitalier. Par ailleurs, nous présenterons une solution au problème d'exploitabilité des données de chaînage du fait de leur manque de qualité évoqué précédemment. À cet effet, nous

1. Traduit de l'anglais : Diagnosis Related Groups (DRG).

avons conçu un algorithme permettant d'évaluer la qualité d'un identifiant patient et de sélectionner ceux dont la fiabilité est la plus forte pour l'analyse des trajectoires de patients.

Dans la section 2.1, nous présentons un bref historique du PMSI. Ensuite, dans la section 2.2, nous expliquons le fonctionnement général pour le secteur d'activité de court séjour : du recueil des données à leur consolidation avant leur transmission aux instances réglementaires. Puis, dans la section 2.3, nous détaillons le processus de chaînage des séjours, ses avantages et ses faiblesses. Nous proposons également une solution pour générer des données exploitables dans le cas de l'analyse des trajectoires des patients. Enfin, nous concluons, dans la section 2.4, sur l'exploitabilité de ces données dans le cas de la recherche épidémiologique avec quelques exemples concrets issus de la littérature.

2.1 Historique

En 1980, [Fetter *et al.*, 1980] ont mené une étude à grande échelle visant à analyser le coût de l'hospitalisation avec son contenu médical. Leur objectif était de classer les séjours suivant une logique à la fois médicale et économique : les Diagnosis Related Groups (DRG). L'homogénéité de ces groupes est d'abord économique. La logique médicale ne vient que secondairement. [Fetter *et al.*, 1980] ont montré, dans sa première analyse portant sur plusieurs millions de dossiers, que cette logique économique est fortement corrélée à la durée de séjour. Le premier système de paiement prospectif a été créé à l'issu de ce constat. La classification des DRG, conçue initialement comme outil d'analyse de l'activité, a été utilisée dès 1982 par l'administration américaine pour procéder au paiement forfaitaire des séjours hospitaliers des personnes âgées et handicapées, prises en charge par le programme fédéral Medicare. Il s'est, par la suite, étendu à la majorité des centres hospitaliers [Fetter et Freeman, 1986]. La mise en place de ce système de paiement a été faite progressivement sur une période de 5 années.

Ensuite, ce système a été adopté et adapté par de nombreux pays. La France l'a nommé PMSI. Ce système s'est développé progressivement aussi bien au niveau public que privé. Cette démarche s'est d'ailleurs intégrée dans un contexte plus général d'organisation de la santé et de planification sanitaire. À partir de la loi 70-1318 du 31/12/1970, un certain nombre de réformes ont façonné le « nouveau » paysage de la santé :

— En 1991, la loi 91-178 du 31 juillet 1991, oblige les établissements publics et privés à effectuer une évaluation et une analyse de l'activité. C'est la mise en œuvre du recueil des pathologies et des modes de prise en charge qui constitue le cœur même du PMSI. Il a d'abord concerné les activités de court séjour ou MCO ;

— Ensuite, l'arrêté du 22 juillet 1996, oblige tous les établissements à mettre en place définitivement et en action le PMSI. Il a ensuite été étendu aux Soins de Suite et de Réadaptation (SSR) en 1998 et rendu obligatoire à tous les établissements en 2003 ;

— L'ordonnance n°2005-406 du 2 mai 2005, instaure une nouvelle tarification à l'activité qui privilégie les recettes sur les dépenses et les résultats sur les moyens. Elle est appliquée sur les secteurs MCO et de l'Hospitalisation à Domicile (HAD). Ce nouveau système de financement est basé sur l'activité des hôpitaux. Ainsi, la valorisation de leur activité dans le cadre du PMSI permet de rémunérer cette activité en conséquence. Notons que ce mode de financement n'est d'ailleurs pas une invention française. En effet, il peut trouver ses racines dans ce que l'on appelle communément la « tarification à la pathologie », également initiée aux États-Unis dans le cadre du programme Medicare puis reprise notamment dans plusieurs pays européens dont le Royaume-Uni ou encore l'Allemagne ;

— L'arrêté du 29 juin 2006, relatif à la généralisation du PMSI en psychiatrie instaure le Recueil d'Information Médicalisé en Psychiatrie (RIM-P). Ce recueil complète ainsi le dispositif.

2.2 Principes généraux

Dans cette section, nous expliquons le principe de fonctionnement du PMSI. Nous nous intéressons uniquement au MCO, puisque notre étude est basée sur ces données. La figure 2.1 schématise ce principe. Dans ce système, l'unité de facturation est le séjour. Chaque séjour génère un recueil de données (voir section 2.2.1). Chaque séjour est, ensuite, classé selon un Groupe Homogène de Malades (GHM) dont le critère est l'homogénéité de traitements pour des pathologies proches (voir section 2.2.2). Pour finaliser le processus, ces données sont anonymisées (voir section 2.2.3) pour la transmission à l'Agence Technique de l'Information sur l'Hospitalisation (ATIH).

2.2.1 Hospitalisation : production d'un Résumé de Sortie Standardisé

À la fin de tout séjour dans un établissement de santé, il y a production d'un Résumé de Sortie Standardisé (RSS) constitué d'un ou de plusieurs Résumés d'Unité Médicale (RUM). Dans chaque unité de soins où a été hospitalisé le patient, le médecin responsable de l'unité de soins produit un RUM qui contient des informations d'ordre administratif et médical.

1. Les informations administratives comprennent les identifiants du patient et de son séjour : sexe, date de naissance, code postal du lieu de résidence, numéro FINESS [2] de l'établissement, numéro de séjour, numéro de l'unité médicale, dates et modes d'entrée et de sortie de l'unité médicale, provenance et destination.
2. Les informations médicales sont les diagnostics ainsi que les actes médicaux réalisés pendant le séjour. Il contient également d'autres informations spécifiques [3]. Les diagnostics sont organisés de la façon suivante :

2. Le Fichier national des établissements sanitaires et sociaux assure l'immatriculation des établissements et entités juridiques porteurs d'une autorisation ou d'un agrément.

3. Comme le poids à l'entrée dans l'unité pour les nouveaux nés, la date des dernières règles pour les femmes enceintes, l'indice de gravité simplifié pour les patients en unité de réanimation...

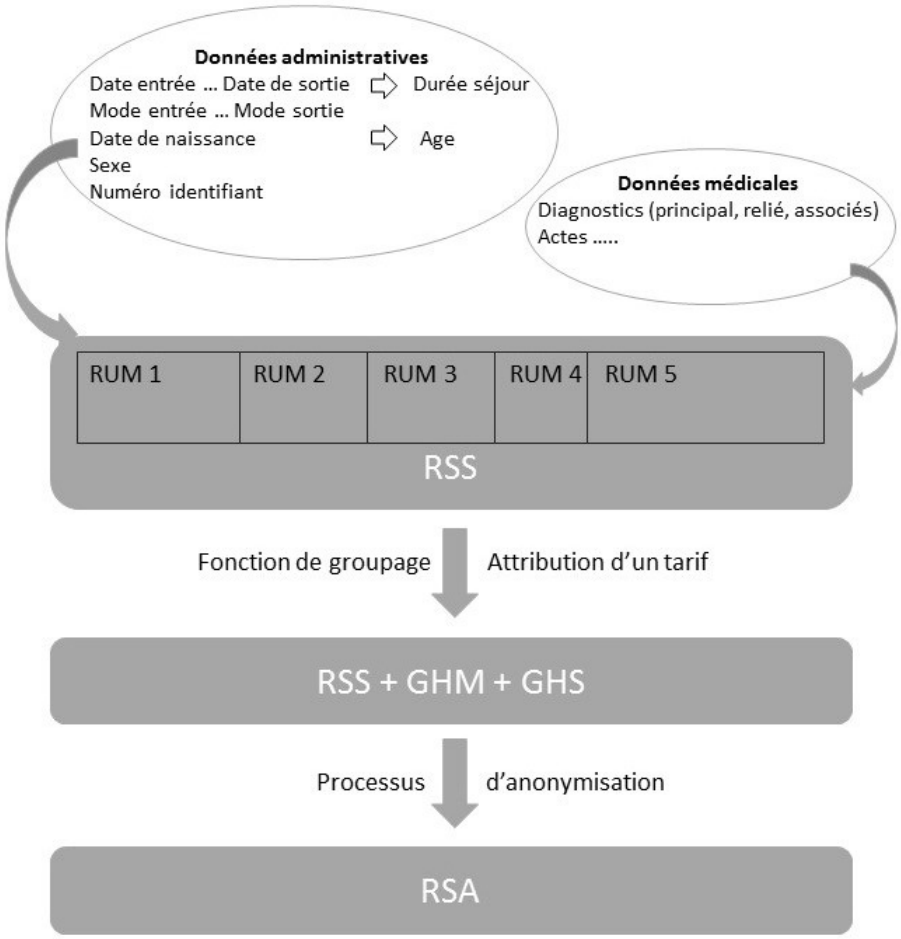


Figure 2.1 – Principe du PMSI MCO schématisé : de l'hospitalisation à l'anonymisation des données.

— le Diagnostic Principal (DP) : il s'agit de l'affection qui a motivé le séjour dans l'unité médicale ;

— le Diagnostic Relié (DR) : il complète le DP lorsque celui-ci est un motif de recours aux soins ;

— le(s) Diagnostic(s) Associé(s) (DAS) : il s'agit de toutes les autres prises en charge (surveillance, traitement, découverte de diagnostic).

Les diagnostics sont codés en référence à la 10ième révision de la Classification Internationale des Maladies (CIM-10) de l'Organisation Mondiale de la Santé (OMS) et les actes médicaux selon la Classification Commune des Actes Médicaux (CCAM). Ne doivent figurer dans le RUM que les problèmes de santé présents, actifs, diagnostiqués ou traités au moment de l'hospitalisation ou lors du passage par la structure d'accueil des urgences de l'établissement.

Ces informations sont ensuite traitées automatiquement par un algorithme de groupage permettant de décrire la prise en charge d'un séjour sous la forme d'un Groupe Homogène de Malade (GHM).

2.2.2 Classification des séjours : les Groupes Homogènes de Malades

Tout RSS est classé dans un GHM. Cette classification est dérivée de celle des DRG. Elle est effectuée par un algorithme de groupage et résulte d'une série de tests réalisés sur les informations contenues dans le RSS. La numérotation des GHM est structurée en 4 parties déterminées au cours des différentes phases de l'algorithme. Le déroulement de l'algorithme est le suivant :

1. **Premier niveau de classement : les Catégories Majeures de Diagnostics (CMD).** Généralement, c'est le DP qui détermine la CMD du séjour [4]. Elle est codée sur 2 chiffres. En cas de séjours multi-RUM [5], la fonction de groupage tient compte d'informations supplémentaires pour déterminer le DP [ATIH, 2015].
2. **Deuxième niveau de classement : type de GHM.** Après les chiffres de la CMD, il suit une lettre codant le type de GHM. L'algorithme recherche la présence d'un acte classant opératoire dans le RSS. Si cette recherche est positive, le séjour est classé dans le groupe chirurgical C. Sinon, le séjour est classé dans l'un des groupes suivants : le groupe K (s'il y a un acte classant non opératoire), le groupe médical M s'il n'y a pas d'acte classant ou le groupe indifférencié Z avec ou sans acte classant opératoire.
3. **Troisième niveau de classement : racine du GHM.** Les 2 chiffres qui suivent donnent le niveau du GHM dans l'arbre de classification. À partir de cette étape, nous obtenons ce que l'on appelle la racine du GHM.
4. **Quatrième niveau de classement : sévérité du GHM.** Les Complications ou Morbidités Associées (CMA) déterminent le niveau de sévérité sur une échelle variant de 1 (sans CMA) à 4 (le plus haut niveau de sévérité). Ces CMA sont déterminées par les DAS.

Une fois que le RSS est associé à un GHM, ce dernier va alors déterminer le tarif associé au séjour. L'étape suivante est l'affectation à un Groupe Homogène de Séjours (GHS). À chaque GHM est associé un seul GHS [6], identifié par un code numérique, renfermant le tarif du séjour. Il s'agit d'un tarif forfaitaire, tout compris, servant de base à la facturation dans le cadre d'une hospitalisation. Il est défini par l'Assurance Maladie.

2.2.3 Anonymisation : production du Résumé de Sortie Anonymisé

Le RSS est rendu anonyme et transformé en Résumé de Sortie Anonyme (RSA) avant la transmission à l'Agence Régionale de Santé (ARS). Un séjour hospitalier donne lieu à un RSA unique. Dans les RSA, la date de naissance est remplacée par l'âge calculé à la date d'entrée, le code postal de résidence par le code géographique de l'Institut national de la statistique et des études économiques (Insee), les dates

4. Exemple : CMD 02 : Affections de l'œil, CDM 04 : Affections de l'appareil respiratoire...
5. Le patient a séjourné dans différentes unités médicales.
6. À quelques exceptions près.

d'entrée et de sortie par la durée de séjour, le mois et l'année de sortie, la date de réalisation des actes par le délai en jours par rapport à la date d'entrée. Les bases régionales de RSA constituent la base nationale des RSA centralisée auprès de l'ATIH.

Depuis quelques années, le Centre Hospitalier Universitaire (CHU) de Nîmes, a investi dans un centre d'information contenant les bases nationales PMSI, lui permettant d'investiguer des champs de recherche, notamment celui de l'IM. C'est à partir de cette base de données que nous avons reconstitué le parcours du patient grâce au numéro anonyme de patient ou à l'identifiant patient. Le détail de ce processus fait l'objet de la section suivante.

2.3 Reconstitution du parcours hospitalier

Depuis 2001, une procédure de chaînage des résumés de séjours permet de relier les différentes hospitalisations d'un même patient grâce à son identifiant patient. Ce numéro est généré de façon automatique à partir du numéro d'assuré social, de la date de naissance et du sexe du patient. L'objectif de ce chaînage est de créer un identifiant patient commun à toutes les hospitalisations d'un même patient, quel que soit l'établissement (public/privé) et le secteur d'activités (MCO, SSR, psychiatrie ou encore HAD). Les différents épisodes d'hospitalisation d'un même patient peuvent ainsi être identifiés et reliés entre eux. Il devient alors possible de reconstituer le parcours de soins d'un patient, dans un établissement de santé, dans l'ordre chronologique des évènements.

Dans la section 2.3.1, nous présentons succinctement, la procédure de création de l'identifiant patient. Nous verrons, dans la section 2.3.2, que ce processus ne se fait pas toujours sans écueil. Enfin, nous proposons, dans la section 2.3.3, un algorithme d'évaluation de la fiabilité de l'identifiant patient, qui sera essentiel pour la sélection de données décrite dans le chapitre 3.

2.3.1 Identifiant patient : processus de création

Le processus de création s'effectue en deux étapes décrites ci-dessous [ATIH, 2014].

Étape 1 : au sein de l'établissement. Une première phase de cryptage [El Kalam *et al.*, 2004] est réalisée par la combinaison du numéro de sécurité sociale, de la date de naissance et du sexe à une clé de hachage. Ainsi, un premier identifiant anonyme est créé. Lors de cette phase, des contrôles sont réalisés sur les données fournies en entrée, le numéro de séjour et le rapprochement entre les numéros anonymes générés et les RSS. De plus, un « numéro de séjour » est élaboré : pour chaque identifiant une date de référence fictive est fixée. Le « numéro de séjour » correspond au nombre de jours écoulés entre cette date de référence et la date d'entrée du séjour. Ceci permet, par la suite, de calculer les délais inter-séjours.

Étape 2 : à l'ATIH. Une deuxième phase de cryptage est réalisée à l'aide du premier identifiant anonyme et d'une deuxième clé de hachage générant alors l'identifiant patient. Les contrôles effectués lors de l'étape 1 sont conservés. Par conséquent, à tout identifiant patient est associé une série de codes retours résultant de ces contrôles et permettant d'écarter les identifiants présentant des erreurs lors de leur utilisation.

Ce processus de génération de l'identifiant patient confère à ce dernier la particularité d'être :

1. **Reproductible** : les mêmes données d'identification produisent le même identifiant patient (il est propre à chaque patient) ;
2. **Irréversible** : il est impossible de retrouver les données permettant de générer l'identifiant (il est impossible d'identifier la personne à partir de son numéro de chaînage) ;
3. **Discriminant** : si les traits d'identification de deux personnes sont approchants on obtient deux numéros différents ;
4. **Spécifique** : il est propre à l'individu.

Toutefois, ce processus présente des limites. Ces limites sont, d'une part, structurelles, liées à la méthode : elles découlent de l'utilisation du numéro de sécurité sociale qui est propre à l'individu. Dès lors qu'un changement de situation (professionnel/social) se produira, il conduira à la création d'un nouveau numéro de sécurité sociale, et donc à la génération d'un identifiant patient différent pour un même patient. Notons le cas particulier des jumeaux de même sexe, où il peut y avoir une double erreur dans la mesure où le code géographique domicile et la durée de séjour sont identiques. D'autre part, ces limites sont aussi liées à un défaut de production des informations sources, comme par exemple la qualité du numéro de sécurité sociale ou la saisie de la date de naissance (mauvaise saisie, communication d'un faux ou utilisation d'un numéro générique). De plus, concernant le numéro administratif de séjour, il existe d'autres failles : par exemple pour les nouveau-nés restant à la maternité, on ne crée pas toujours un numéro administratif de séjour. Le DIM (Département d'information médical) crée alors un numéro administratif fictif non connu du système administratif.

Par ailleurs, dans les différents contrôles effectués au cours de l'étape 1, il n'y a pas de vérification de cohérence d'un RSS à l'autre. Ainsi, il est possible d'y trouver des incohérences sur le sexe et/ou la date de naissance pour un même identifiant, puisque cela n'est pas dépisté.

2.3.2 Anomalies dans le processus d'anonymisation

Le processus d'anonymisation en lui-même peut générer des anomalies dues à un phénomène de collision (*i.e.* la production d'un identifiant patient identique pour deux patients différents) ou à un phénomène de fusion (*i.e.* la production d'identifiants patients différents pour un même patient). Cependant, ces phénomènes demeurent marginaux avec une probabilité inférieure à 10^{-48} [Quantin *et al.*, 2005].

Des études ont été effectuées sur les données de chaînage et ont mis en évidence un certain nombre d'incohérences. [Tardif, 2007] est parmi les premiers à analyser la qualité des données de chaînage du PMSI. Son étude porte sur l'année 2004. Il met en évidence des incohérences sur l'âge, le sexe ou encore la chronologie des séjours pour un même identifiant patient. Ces résultats sont, par la suite, confirmés par [Le Bihan-Benjamin, 2011] qui complète ces informations avec des analyses portant sur le nombre de décès ou d'accouchements associés à un même identifiant patient. Elle souligne l'absurdité de certains chiffres comme plusieurs décès au cours d'une année pour un même patient ou encore un nombre d'accouchements dépassant les limites des capacités humaines. À travers cette étude, elle détaille également les pratiques de certains établissements dans le cas particulier de l'avortement. Il apparaît que, dans ce cas de figure, certains établissements utilisent un même numéro de sécurité sociale. Ceci explique que l'on puisse retrouver des identifiants patients avec un nombre de séjours excédant le nombre de jours contenus dans une année...

D'autres études ont été menées sur la qualité des données dans le cas de l'utilisation du chaînage. [Bocquier *et al.*, 2011] ont évalué la qualité du chaînage notamment sur la cohérence de l'âge et du sexe pour trois régions françaises (Picardie, Bretagne, PACA), pour les maladies de type cancer et asthme. Ils ont établi que les sources d'erreurs variaient selon la région. Bien qu'ils aient estimé une bonne qualité du chaînage dans l'ensemble, ils ont souligné les mêmes incohérences évoquées dans le paragraphe précédent. [Beyeme-Ondoua, 2007] a évalué la qualité des données chaînées nationales dans le cadre du cancer colorectal concernant l'année 2003. Il a mis en évidence des disparités en termes de qualité de codage des séjours (information manquante telles que le DP, la durée de séjour...) conduisant à des erreurs de groupage des séjours [7] ou à un identifiant de patient manquant [8]. [Michel *et al.*, 2008] ont analysé la fiabilité des causes de décès en PACA. Leur étude fait ressortir un manque de précision dans le codage amenant à des aberrations telles qu'évoquées dans [Le Bihan-Benjamin, 2011]. Une étude plus récente [Le Bihan-Benjamin *et al.*, 2013] met en exergue une nette amélioration de la qualité du codage ces dernières années, en corrélation avec la montée en charge de la T2A. Selon cette étude, en 2009, la proportion de séjours sans numéro de chaînage ou avec anomalies sur le numéro de chaînage était égale à 2,2% contre 2,6% en 2007. Par conséquent, la qualité des données de chaînage s'est améliorée ouvrant la voie à de nouvelles perspectives d'exploitation de ces données pour la recherche épidémiologique par exemple.

Néanmoins, le chercheur souhaitant exploiter les données de chaînage dans le cas d'une étude des parcours de soins se doit de consolider ses données en tenant compte des éléments évoqués dans cette section. Pour ce faire, nous proposons une solution dans la section suivante.

7. Ce séjour n'est alors pas chaînable avec les autres dans la mesure où des informations sont manquantes : on perd une partie de l'histoire du parcours du patient.

8. En effet, s'il manque une des trois informations nécessaires à la création de l'identifiant ce dernier n'est alors pas créé.

2.3.3 Fiabilité des identifiants patients

À la lumière de ces limitations, nous avons nettoyé et/ou filtré les données de sorte à obtenir une base de données permettant de mener des analyses de données longitudinales dans de bonnes conditions. Sur la base des anomalies mises en évidence dans la section 2.3.2, nous avons participé à l'élaboration d'un algorithme [Boudemaghe, 2016] permettant d'évaluer la fiabilité d'un identifiant patient sur une échelle de valeurs allant de 0 (identifiant non exploitable pour le chaînage) à 5 (identifiant de haute qualité). Nous avons défini cinq tests de contrôle qualité :

Test 1 : Génération de l'identifiant sans écueil. S'il n'y a pas eu d'erreur lors du processus de génération de l'identifiant patient alors celui-ci a une fiabilité de niveau 1, 0 sinon. Ce test consiste à vérifier les codes retours évoqués dans la section 2.3.1. Si l'identifiant passe le test, il est alors au niveau 1 sur l'échelle de fiabilité. Nous procédons ensuite au test suivant.

Test 2 : Cohérence du sexe. Ce test consiste à vérifier que le sexe est identique pour tous les séjours associés à un même identifiant. Dans le cas où le test est positif, le niveau de fiabilité de l'identifiant monte à 2 et nous procédons au test suivant.

Test 3 : Cohérence de l'âge. Ce test se base sur le principe suivant : la différence d'âge entre deux séjours consécutifs au cours de la même année ne doit pas être supérieure à une année. Dans le cas où le test est positif, le niveau de fiabilité de l'identifiant monte à 3 et nous procédons au test suivant.

Test 4 : Cohérence de la chronologie. Ce test s'effectue en deux parties. La première partie vérifie qu'un séjour postérieur à un autre, relativement à son numéro de séjour, n'a pas une année et un mois de sortie antérieurs au séjour précédent. La deuxième partie vérifie qu'un séjour contenu à l'intérieur d'un autre[9] se termine au plus tard au même moment. La figure 2.2 schématise les différents cas de figure évoqués. S'il n'y a pas d'incohérence de chronologie des séjours, alors, l'identifiant passe au niveau de fiabilité 4 et nous effectuons le dernier test.

Test 5 : Cohérence médicale. Ce test vérifie les éléments suivants :

- Un patient ne peut être décédé qu'une seule fois et ceci lors du dernier séjour ;
- Une femme ne peut accoucher plus de deux fois au cours d'une même année ;
- Une femme ne peut avoir plus de 6 interruptions de grossesses au cours d'une même année.

Dans le cas où le test est positif, le niveau de fiabilité de l'identifiant monte à 5. Il est alors considéré comme étant de haute qualité, puisque de fiabilité maximale selon les critères définis ci-dessus. Les différentes étapes de l'algorithme sont résumées dans la figure 2.3.

9. Comme c'est le cas pour une prestation inter-établissement.

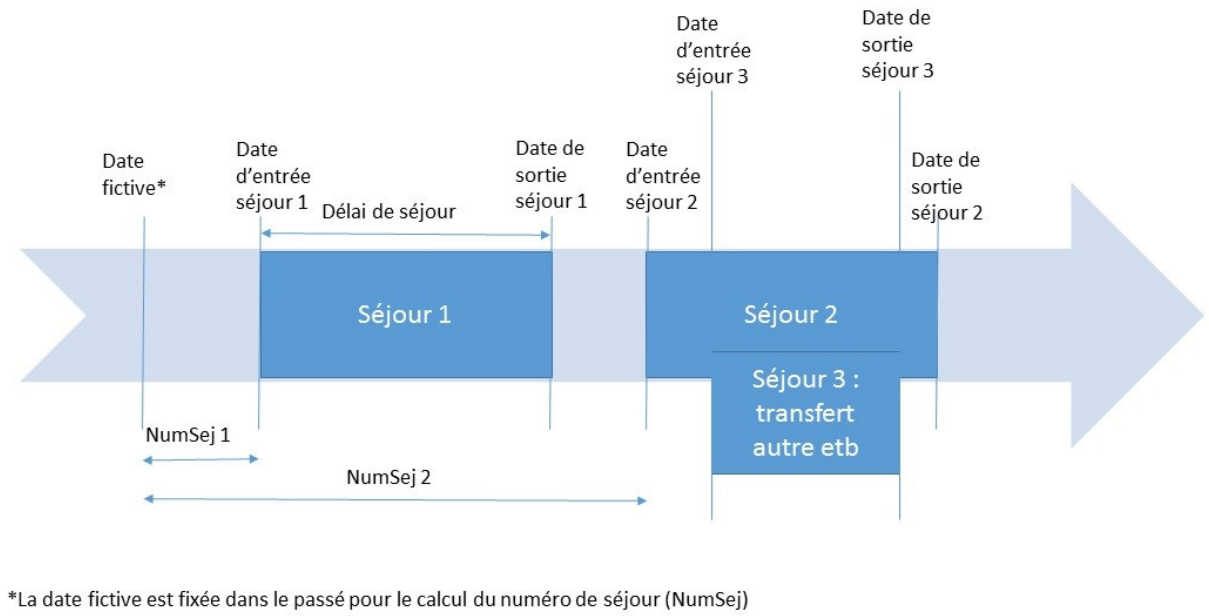


Figure 2.2 – Chronologie normale des séjours hospitaliers.

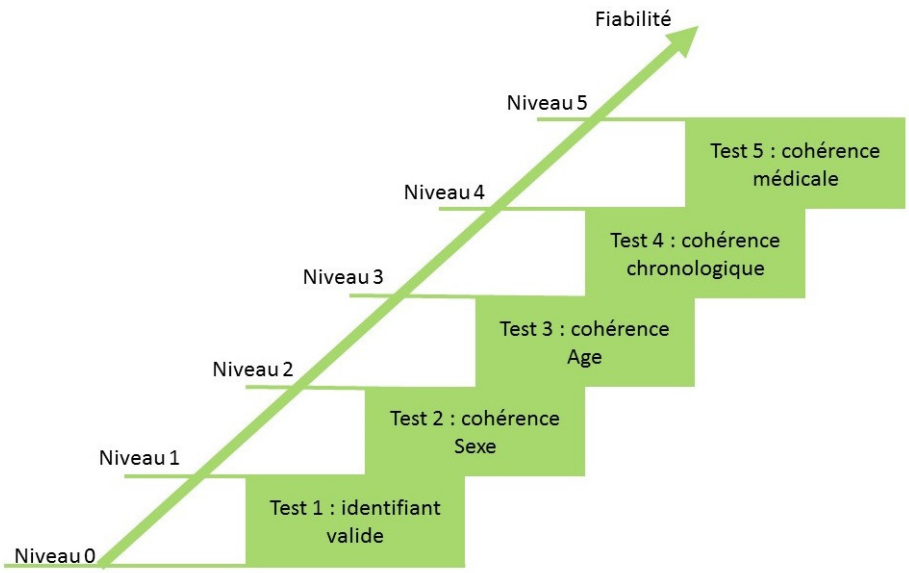


Figure 2.3 – Algorithme de détermination du niveau de fiabilité de l'identifiant patient.

Nous avons, ensuite, procédé à des analyses statistiques concernant la répartition des niveaux de fiabilité selon le nombre de séjours [10]. [Boudemaghe, 2016] a établi que le taux d'identifiants avec un niveau de fiabilité égal à 5 avoisine les 95% quel que soit le nombre de séjours. Ce résultat est satisfaisant et nous permet d'aborder la question de la sélection d'une population représentative pour l'étude des trajectoires.

Notre méthode a été utilisée pour sélectionner des séjours dans le cadre d'une étude portant sur la comparaison de la survenue de complications post-opératoires et de la mortalité hospitalière entre patients opérés pour fracture de hanche et patients opérés pour chirurgie programmée de remplacement de la hanche [Le Manach *et al.*, 2015]. Cette sélection a donné une base de données comportant 700 000 séjours. Elle a également été utilisée dans une étude d'addictologie dont l'objectif était d'élaborer un modèle prédictif d'évènement intra-hospitalier pour les patients admis pour dépendance ou abus de drogue [Nguyen *et al.*, 2017] et a fourni une base de 66 000 séjours.

Nous sommes donc capables d'analyser les données nationales et celles de leur chaînage dans des conditions de qualité, définies ci-dessus. En outre, les chercheurs se sont d'ores et déjà intéressés aux données nationales, à diverses fins, comme nous allons le voir dans la section qui suit.

2.4 Épidémiologie et PMSI

Dans cette section, nous présentons des cas concrets d'applications pratiques de l'utilisation des bases médico-administratives mais aussi, les différentes limitations rencontrées.

2.4.1 En pratique : des exemples

Nous avons vu que grâce à l'identifiant patient (voir section 2.3), les données nationales peuvent être reliées. Cet historique permet de retracer les différentes étapes qui ont ponctué le parcours de soins d'un patient. Cet historique offre de nouvelles perspectives d'analyses dans divers domaines : épidémiologie, économie médicale ou organisation des soins. La littérature sur ces sujets est pléthorique, nous présentons dans la suite un aperçu des différents types d'études menées à partir des bases de données du PMSI, au travers de quelques exemples.

De récentes publications ont porté sur l'estimation de la prévalence hospitalière ou de son incidence. [Remontet *et al.*, 2008] ont proposé un algorithme de sélection des séjours afin d'estimer le taux d'incidence du cancer dans douze départements français. Ils ont validé leur méthode en la comparant aux résultats obtenus à partir des registres cancers. [Colonna *et al.*, 2012] ont estimé la prévalence du cancer du sein chez la femme. Leur méthode s'appuie sur un processus spécifique d'extraction de données à partir de bases de données médico-administratives et une modélisation épidémiologique des données. [de Peretti *et al.*, 2012] ont analysé les évolutions des

10. Sur la période 2009 à 2014.

taux de patients hospitalisés et de la prise en charge en unités neuro-vasculaires depuis 2008. [Coloma *et al.*, 2013] ont élaboré une méthode de sélection des séjours pour déterminer l'incidence de l'IM aigu à partir des bases médico-administratives. Ils l'ont construite aussi bien à partir des codes de la CIM-9 que ceux de la CIM-10. Ils ont comparé leurs résultats à des registres de cardiologie et ont obtenu une valeur prédictive positive (VPP) de l'ordre de 75%. Par ailleurs, un grand nombre de ces études ont été à la base d'un guide méthodologique d'algorithmes de repérage des pathologies dans les bases médico-administratives [Quantin et Cnamts, 2015].

D'autres études ont porté sur le suivi de cohortes hospitalières pour la recherche d'associations de pathologies. Par exemple, [Dalichampt *et al.*, 2014] ont étudié les risques d'embolie pulmonaire, d'AVC ischémique et d'IM chez les femmes sous contraceptif oral. Ils ont montré qu'il y avait une augmentation des risques en association avec la prise du contraceptif. [Bernier *et al.*, 2012] ont étudié le risque de cancer radio-induit après exposition dans l'enfance à des examens scanographiques à partir de la cohorte Enfant Scanner. Ils ont estimé le pourcentage d'enfants de la cohorte identifiés dans la base PMSI et ont construit un algorithme pour individualiser les enfants présentant un cancer ou une pathologie à risque de cancer à partir des diagnostics cliniques de la base PMSI. [Lorgis *et al.*, 2013] ont évalué la mortalité intra-hospitalière et le pronostic à 1 an chez les patients infectés par le VIH ayant un IM aigu. Ils ont montré qu'après un IM aigu, le statut du VIH influence le risque à long terme, bien que le risque à court terme chez les patients atteints du VIH soit comparable à celui des patients non infectés. [Tzoulaki *et al.*, 2009] ont examiné toutes les causes de mortalité associées à la prescription de médicaments antidiabétiques oraux, mais aussi le risque d'IM et le risque d'insuffisance cardiaque congestive. Les résultats de leur étude suggèrent que certains médicaments comparés à d'autres augmentent les risques. Par exemple, ils ont établi que les sulfonylurées ont un profil de risque défavorable par rapport à la metformine.

Dans un aspect plus médico-économique, des études ont porté sur l'analyse des coûts de prise en charge ou encore de la planification sanitaire. [Schmidt *et al.*, 2015] se sont intéressés à la prise en charge de l'AVC ischémique aigu et aux coûts associés sur une période d'un an. Leur étude souligne l'importance que représente la prise en charge de cette pathologie, tant au niveau médical qu'au niveau économique. Elle suggère également que des changements dans la prise en charge thrombolytique pourraient impacter les parcours de patients et par voie de conséquence permettre des économies substantielles. [Colin *et al.*, 2007] ont comparé les coûts de prises en charge des patients diabétiques avec ceux des non diabétiques dans le cas de complications cardiovasculaires telles que l'AVC, l'IM, l'angine instable, l'arrêt cardiaque et la revascularisation coronarienne. Leur étude atteste que la prise en charge pour les patients diabétiques est nettement plus élevée que pour les autres patients. [Boddaert *et al.*, 2015] ont évalué l'impact médico-économique de la prise en charge des fractures de hanche dans une unité périopératoire gériatrique. Cet impact s'est avéré positif tant sur le plan économique que sur la réduction du risque de mortalité à long terme. [Jay *et al.*, 2013] ont étudié les coûts cumulés des trajectoires de patientes ayant subi une intervention chirurgicale pour un cancer du sein. Ils ont constitué des groupes de patientes pertinents à la fois sur le plan économique et médical, en

se basant sur l'analyse de concept formel. Leur étude offre des perspectives d'applications dans le cadre de la planification sanitaire des maladies chroniques afin de mieux répartir les ressources et de rendre la prise en charge efficace.

Enfin, des études ont identifié les facteurs prédictifs de ré-hospitalisation. [Gusmano *et al.*, 2015] ont comparé les taux de ré-hospitalisation, toutes causes confondues, à 30 jours en France et aux États-Unis chez les patients âgés de 65 ans. Ils ont observé des taux de ré-hospitalisation plus faibles pour la France et ont attribué cette différence à une combinaison de facteurs : un meilleur accès aux soins primaires, une meilleure santé chez les personnes âgées, une plus longue durée de séjour dans les hôpitaux français... [Delmas *et al.*, 2011] ont établi que le taux de réadmission des patients atteints d'asthme est un indicateur pertinent de la surveillance de l'asthme et plus particulièrement de sa prise en charge. Ce dernier associé aux données sur les médicaments représente une source précieuse d'informations pour améliorer la compréhension de l'asthme sévère en France. [Béjot *et al.*, 2011] ont analysé les données nationales concernant les patients hospitalisés soit pour un syndrome coronarien aigu (SCA) ou pour un syndrome cérébrovasculaire. Ils en ont déduit que les patients ayant un profil de risque vasculaire élevé sont à risque de réadmission, mais surtout à risque de mortalité précoce. [Lainay *et al.*, 2015] ont étudié les réadmissions non programmées au cours de la première année qui a suivi un AVC. Ils ont mis en évidence le fait que les patients victimes d'un AVC, ont un risque élevé de ré-hospitalisation associé à un risque élevé de décès. Leur étude souligne ainsi la nécessité d'un suivi d'intervention pour prévenir la réadmission.

L'identifiant patient est également présent dans les bases de données du Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM). Ainsi, les données du PMSI peuvent être croisées avec celles du SNIIRAM offrant la possibilité de relier la consommation d'un médicament avec les séjours hospitaliers et les diagnostics répertoriés dans ces séjours. [Conte *et al.*, 2016] ont mesuré l'incidence de la consommation de drogues psychotropes, pendant le diagnostic et la phase de traitement, des patients atteints de lymphomes non-Hodgkiniens par rapport aux témoins de la population générale et ont identifié les facteurs associés à cette utilisation. Ils ont révélé qu'un taux élevé de prise de médicaments psychotropes était associé à un niveau élevé d'anxiété à la phase initiale du développement de la maladie. [Perlbarg *et al.*, 2014] ont créé un algorithme de sélection des patients diabétiques et atteints d'hypertension en appariant les bases de données SNIIRAM à celles du PMSI. Leur méthode ouvre la voie à de nouvelles exploitations sur l'analyse des comorbidités, des pratiques et des parcours de soins. Toutefois, ces bases de données sont complexes de part leur architecture et volumineuses, car elles contiennent de nombreuses informations (données sur les patients, données sur les prestations remboursées dans le cadre des soins réalisés en médecine de ville...). Ainsi, leur exploitation nécessite avant tout une préparation technique et méthodologique de contrôle, de validation, mais également une bonne connaissance de ces données pour leur interprétation [Goldberg *et al.*, 2016].

Bien que l'on trouve de plus en plus d'applications et que les applications soient diverses, les auteurs ont été confrontés à certaines limitations comme nous allons le voir dans la section suivante.

2.4.2 Les limitations rencontrées

Le PMSI est un outil d'allocation budgétaire, mais il présente des limites dans le domaine épidémiologique car il expose à des imprécisions et à des erreurs. En effet, un certain nombre de critiques ont été soulevées comme : 1) l'exhaustivité du codage ; 2) la qualité du codage ; 3) les difficultés du codage. Par conséquent, la fiabilité du codage des séjours via les bases médico-administratives est controversée [Bayat *et al.*, 2001, Goldberg *et al.*, 2008].

Exhaustivité du codage. La nature de l'utilisation des informations du PMSI, fait que l'ajout d'informations supplémentaires ne se fait que dans le but de valoriser le séjour et ainsi d'être au plus proche des dépenses encourues par l'établissement de soins. Ainsi, si ces informations supplémentaires ne rapportent pas de supplément financier à l'hôpital, alors ces informations n'ont pas nécessité à être saisies. Ces informations ne sont donc pas toujours adaptées pour certaines analyses [Landais *et al.*, 1998, Lombrail *et al.*, 1994] lorsque les résultats des études sont liés à l'exhaustivité du codage. Dans le chapitre 3, nous montrerons d'autres exemples de l'insuffisance de ces données plus précisément dans le cas des DAS.

Qualité du codage. L'étude de [Aboa-Eboulé *et al.*, 2013] fait valoir que la précision du recueil de l'algorithme de repérage d'une pathologie repose sur la qualité des données. Or, cette dernière est dépendante de la qualité de documentation des dossiers médicaux, mais aussi de l'expérience et l'expertise du codeur. Toutefois, le codage n'est pas toujours réalisé par le spécialiste concerné, ce qui peut conduire à des erreurs d'appréciation et induire des erreurs de codage. Par ailleurs, les séjours sont contrôlés par les TIM (Technicien d'Information Médicale) et peuvent être recoder *a posteriori* sous le contrôle du médecin DIM. Toutefois, ces interventions ont essentiellement pour objectif d'enrichir les CMA afin de valoriser le séjour [11]. De plus, le PMSI au quotidien n'est jamais un langage tout à fait stabilisé, mais en cours permanent de stabilisation. En effet, la traduction des données médicales, contenues dans le dossier patient, en information médicale est propre au codeur ce qui induit une hétérogénéité du codage [Chantry *et al.*, 2012, Grammatico, 2014]. De plus, les versions de GHM et de GHS évoluent ce qui peut rendre délicates les comparaisons longitudinales.

Difficultés de codage. Les nomenclatures ont été créées dans un but de codification de l'activité pour la tarification. Ces nomenclatures se sont parfois avérées inappropriées. En effet, elles ne permettent pas toujours de retranscrire la véritable description de la pathologie dont souffre le patient car cette dernière n'existe pas

11. Notons que dans notre domaine d'étude : la cardiologie, les erreurs sur le diagnostic principal sont marginales.

dans les nomenclatures [Boutault *et al.*, 1999]. Coder implique une conception singulière du soin qui est ici gestionnaire. Il émane, alors, de ce processus de traduction, un ensemble d'incertitudes, notamment médicales et économiques [Juven, 2013]. Ainsi, l'adéquation entre la réalité observée chez le patient et la codification de la maladie n'est pas toujours patente.

Exhaustivité des données. Par ailleurs, il peut y avoir une autre limitation dans l'utilisation de ces données : il s'agit non pas de l'exhaustivité du codage, mais de celle des patients. En effet, pour diverses raisons, certains patients se privent du recours aux soins. Une enquête de santé et de protection sociale menée en 2012 [Célant *et al.*, 2014], rapporte que 5% des personnes sous le régime de l'assurance maladie ont renoncé à consulter le médecin, 4% ont renoncé à un autre type de soins, en dehors des soins dentaires et optiques, pour des raisons financières. Ce renoncement était lié au délai d'attente pour obtenir un rendez-vous dans 17% des cas, mais aussi à l'éloignement du cabinet ou à des difficultés de transports dans 3% des cas.

Malgré toutes ces limitations évoquées, les bases médico-administratives représentent indéniablement une source importante d'informations. Elles couvrent la majorité des établissements de santé privés et publics sur le plan national et colligent des données médicales sur tous les séjours hospitaliers. Une alternative pour réduire ce biais intrinsèque au codage est par exemple l'appariement des informations avec les bases de données du SNIIRAM [Goldberg *et al.*, 2016].

2.5 Conclusion

Dans ce chapitre, après avoir introduit un bref historique du PMSI, nous avons présenté son principe de fonctionnement pour le secteur d'activité de court séjour. Nous avons ensuite expliqué comment reconstituer la trajectoire de soins d'un patient à partir des bases anonymisées nationales. La littérature ayant souligné des faiblesses dans la qualité du chaînage, nous avons participé à la conception d'un algorithme permettant de filtrer les données et d'assurer une certaine qualité des informations de sorte à pouvoir mener des analyses sur les trajectoires de patients.

Les bases médico-administratives sont depuis plusieurs années explorées, et ont servi de support dans de nombreuses études épidémiologiques, économiques et aussi de planification sanitaire. Beaucoup d'entre-elles sont à la base d'une méthodologie [Quantin et Cnamts, 2015] facilitant le repérage de(s) pathologie(s) dans les bases PMSI. Toutefois, un certain nombre de ces auteurs met en garde le chercheur, qui voudrait mener ses propres analyses, sur la qualité des données et recommande la plus grande prudence quant à leur interprétation.

Ces bases de données constituent une mine d'informations que nous allons explorer, dans la suite de cette thèse, pour le cas particulier de l'IM. Dotés d'une méthode de sélection des données longitudinales, nous allons construire la base de données, dans le chapitre 3, à partir de laquelle nous allons mener nos analyses de trajectoires de patients dans les parties III et IV.

Chapitre

3 L'infarctus du myocarde à l'hôpital

Dans le chapitre 2, nous avons présenté brièvement le principe de création des bases PMSI, leur usage possible pour la recherche et les limites dans leur utilisation pour des études épidémiologiques. Partant de là, nous allons maintenant construire la base de données des IM en France à partir de laquelle nous procéderons aux analyses de trajectoires.

En 2012, une étude suédoise [Schmidt *et al.*, 2012] présente un suivi des patients atteints d'IM sur 25 ans. Les auteurs ont établi que, quel que soit le sexe et les comorbidités associées, le risque de mortalité à court et long termes a diminué de moitié au cours de la période d'observation. Toutefois, ils ont souligné que les comorbidités demeurent un fort facteur de risque alors que le sexe ne l'est plus. On y voit un changement dans les tendances établies jusqu'à présent. Quelques années plus tard, en 2015, dans une étude similaire, [Dégano *et al.*, 2015] ont comparé les taux d'incidence et de mortalité entre 6 pays européens. Les auteurs ont également mis en évidence une baisse de l'incidence de l'IM et des taux de mortalité. Par ailleurs, ils ont constaté des taux de mortalité plus élevés chez les femmes et ont également souligné un changement des tendances selon le sexe. D'autres études [Gillum, 1994, Yeh *et al.*, 2010, Singh *et al.*, 2014, Journath *et al.*, 2015] abondent dans ce sens montrant une baisse à la fois de l'incidence de l'IM et des taux de mortalité. De plus, certaines d'entre-elles [Gillum, 1994, Singh *et al.*, 2014] se sont plus spécifiquement intéressées à la comparaison de sous-populations suivant l'origine ethnique et le sexe. Les auteurs de ces deux études en arrivent aux mêmes conclusions : à savoir des taux d'incidence et de mortalité plus élevés pour les personnes de couleurs mais aussi pour les femmes.

Ces études témoignent des progrès de la médecine dans la diminution de la mortalité liée aux maladies cardio-vasculaires au cours des dernières décennies, mais aussi l'impact positif des campagnes de santé visant à agir sur les facteurs de risque tels que la lutte contre le tabagisme et les prises de mesures d'interdiction de fumer dans les lieux publics [Vacheron, 2010], les actions sur les comportements alimentaires,

la réduction de la consommation de sel et l'enrichissement des farines en vitamines B9[1]. Toutefois, elles témoignent également d'un changement sociétal plaçant notamment les femmes sur un pied d'égalité avec les hommes en termes de risque de survenue de l'IM et de mortalité.

En France, des analyses locales ou régionales, voire nationales des IM ont été menées à partir des données du PMSI [Gabet *et al.*, 2016, De Peretti et Bonaldi, 2010, Tuppin *et al.*, 2010]. L'Institut de Veille sanitaire (InVs) a récemment publié les premières données nationales concernant les hospitalisations pour IM à partir du PMSI pour les années 2002 à 2008 [De Peretti *et al.*, 2012]. Ces études reflètent des résultats similaires à ceux évoqués plus haut. Avec l'amélioration de la qualité du codage et du chaînage (évoqués dans le chapitre 2), nous nous sommes posés la question de l'actualisation de ces tendances à partir des données du PMSI-MCO. Dans ce chapitre, nous présentons une étude qui a pour but de compléter ces données d'hospitalisation avec celles des années suivantes jusqu'en 2014, d'évaluer le risque de réadmission et de déterminer les facteurs de risque pour les populations à fort taux de mortalité hospitalière. Cette analyse est importante pour mieux comprendre les données sur lesquelles nous allons travailler notamment pour cerner les caractéristiques importantes de la maladie étudiée et pour vérifier les connaissances extraites via les motifs dans les chapitres 5 et 7.

Dans la section 3.1, nous décrivons les étapes du protocole d'analyses descriptives de l'IM au travers des données du PMSI-MCO. Ensuite, dans la section 3.2, nous détaillons les différents résultats de ces analyses de l'hospitalisation au décès en passant par la récidive. Puis, dans la section 3.3, nous commentons ces résultats et posons un regard critique quant à la complétude des informations présentes dans les bases médico-administratives. Enfin, dans la section 3.4, nous concluons par des perspectives.

3.1 Protocole d'analyses descriptives statistiques

Le protocole d'analyses commence par la constitution de la base de données et se décompose ensuite en trois étapes. Nous considérerons trois axes d'analyse descriptive des données : les hospitalisations, les réadmissions pour IM (les récidives) et les décès. Ces étapes sont décrites dans les sections qui suivent. Toutes les analyses ont été réalisées à l'aide des logiciels R [R Core Team, 2016] et Microsoft SQL Server Management Studio, version 10.0.1600.22.

3.1.1 Étape 1. Constitution de la base de données

La constitution de la base de données repose sur deux points décrits ci-dessous.

1. **Repérage de la pathologie dans les bases PMSI.** Selon le guide [Quantin et Cnamts, 2015] de l'Assurance Maladie, l'IM est défini dans les bases par le repérage d'un syndrome coronarien aigu (SCA). Les personnes hospitalisées l'année n pour cardiopathie ischémique aiguë, se repèrent à l'aide du DP d'un

1. www.sante.gouv.fr.

des RUM. De nombreuses études [Chevreul *et al.*, 2012, Coloma *et al.*, 2013, Haesebaert *et al.*, 2013, Aboa-Eboulé *et al.*, 2013] utilisent le DP comme critère pour repérer ces séjours avec un taux de sensibilité de 76% [De Peretti et Bonaldi, 2010]. Cependant, cela ne permet pas de repérer les IM lors d'une hospitalisation pour un autre motif. Nous avons donc ajouté un autre critère portant sur les actes : les actes de cardiologie interventionnelle. La liste des codes CIM-10 et CCAM, servant au repérage de la pathologie, est référencée dans le tableau A.1 situé en annexe.

2. **Critères d'inclusion.** Nous avons retenu tous les séjours avec IM (selon les critères définis ci-dessus), entre le 1er mars 2009 et le 31 décembre 2014 dans les établissements MCO publics et privés de France métropolitaine, concernant les personnes entre 20 et 99 ans domiciliées en France métropolitaine. Nous avons filtré cette sélection avec l'algorithme évaluant la fiabilité des identifiants patients (décrit dans la section 2.3.3 du chapitre 2) pour ne retenir que les patients ayant un identifiant de fiabilité de niveau 5.

3.1.2 Étape 2. Hospitalisations avec IM

Cette étape de description des séjours se décompose en deux parties :

1. **Nombres de séjours et taux d'hospitalisation annuels.** Cette analyse correspond à l'ensemble des IM hospitalisés sans différencier les IM isolés des récidives. Nous avons étudié les données suivant des critères démographiques (âge et sexe). Nous avons déterminé les taux spécifiques d'hospitalisation annuels, par classe d'âge et par sexe, en rapportant le nombre d'hospitalisations dans une année donnée à la population moyenne par classe d'âge et par sexe de cette même année. Les populations moyennes ont été calculées pour chaque année à partir des données de population au 1er janvier fournies par l'Insee. Nous avons ensuite déterminé les taux standardisés sur l'âge et le sexe par la méthode de standardisation directe. La structure de la population française métropolitaine de l'année en cours constitue la référence. Les taux sont déterminés par région d'habitation. Nous avons complété ces informations par les proportions d'hospitalisations en dehors de la région de résidence.
2. **Caractérisation des séjours.** Cette analyse consiste à décrire les séjours selon les modalités de sortie, la durée de séjour, le mois de survenue de l'IM mais aussi les diagnostics d'hospitalisation. Ces données sont également étudiées selon des critères démographiques.

3.1.3 Étape 3. Réadmissions pour IM

Ensuite, nous avons chaîné les séjours des patients qui ont eu plusieurs hospitalisations avec IM au cours des six années d'observation (2009-2014). Nous avons étudié la réadmission suite à un évènement initial. Ce dernier est défini comme la première hospitalisation avec IM durant les six années d'observation, que nous avons appelé *hospitalisation index*. L'événement d'intérêt est alors la première réadmission avec IM. Notre plan d'analyse des réadmissions se décompose en deux parties :

1. **Nombre de patients, taux et délai de réadmission.** Nous avons décrit la répartition des patients selon le nombre d'hospitalisations suivant des critères démographiques. Puis, nous avons déterminé les taux de réadmission par la méthode de Kaplan-Meier [Alberti *et al.*, 2005]. Enfin, nous avons déterminé le délai entre deux ré-hospitalisations. Ce dernier a été calculé à l'aide de la variable « numéro de séjour » (présentée dans la section 2.3.1 du chapitre 2).
2. **Évaluation du risque de réadmission.** Nous avons comparé le risque de réadmission selon les caractéristiques des patients à l'aide d'un modèle de Cox [Timsit *et al.*, 2005].

Nota Bene : Prendre en compte tous les séjours surestime les chiffres en termes de nombre de séjours. Or, nous avons souhaité mettre en évidence la consommation réelle de soins, du point de vue de la production de séjours, générée par cette pathologie. Pour le calcul de moyenne, et la répartition des délais de réadmission, nous n'avons pas pris en compte les prestations inter-établissements (voir chapitre 2 section 2.3.3).

3.1.4 Étape 4. Décès et facteurs de risque

L'analyse du décès se déroule en deux parties :

1. **Nombres de décès et taux de létalité.** Il s'agit de décrire à quel moment se produit le décès (quel séjour), pour quelle population (âge et sexe). Nous avons calculé les effectifs puis les taux de létalité (bruts et standardisés) par rapport à la population cible.
2. **Facteurs de risque.** Pour des populations cibles, nous cherchons à déterminer les facteurs de risque du décès hospitalier. Nous avons utilisé les comorbidités présentes dans le calcul du score de [Charlson *et al.*, 1987]. Il s'agit d'un index pondéré de comorbidités construit pour prédire la mortalité à un an. Cet index référence 17 comorbidités : *insuffisance myocardique, insuffisance vasculaire périphérique, maladie cérébrovasulaire, démence, maladie pulmonaire chronique, maladie du tissu conjonctif, maladie ulcéreuse, hépatopathie, diabète, hémiplégie, maladie rénale modérée à sévère, diabète avec lésions organiques, tumeurs de toute origine, leucémies, lymphome, hépatopathie modérée à sévère, tumeurs solides métastatiques et sida.* Nous avons repéré ces comorbidités à l'aide des codes CIM-10 fournit dans [Quan *et al.*, 2005], présents dans les DP, DR et DAS des séjours de chaque patient. Nous avons modélisé le décès par la régression logistique (RL).

3.2 Description des résultats

À l'issue de la première étape, notre base de données comprend un total de 900 121 séjours pour 678 021 patients. Dans la suite de cette section, nous allons décrire les résultats des analyses menées pour les étapes 2 à 4.

3.2.1 Étape 2. Hospitalisation avec IM

3.2.1.1 Nombres de séjours et taux d'hospitalisation

Nombres de séjours et de patients. Nous avons analysé :

a) le cas général : le nombre de séjours et le nombre de patients âgés de 20 ans et plus hospitalisés pour IM dans les établissements publics et privés MCO en France métropolitaine sont présentés par année de 2009 à 2014 dans la figure 3.1. Durant cette période, entre 120 000 et 169 000 hospitalisations annuelles pour IM ont lieu en France métropolitaine. Ces hospitalisations concernent entre 101 000 et 140 000 patients par an. L'évolution du nombre de patients et du nombre de séjours a suivi des courbes parallèles. Elle est marquée par une forte augmentation entre 2009 et 2010. Elle continue ensuite d'augmenter progressivement d'année en année. À noter qu'environ 63% des hospitalisations ont lieu dans des établissements publics.

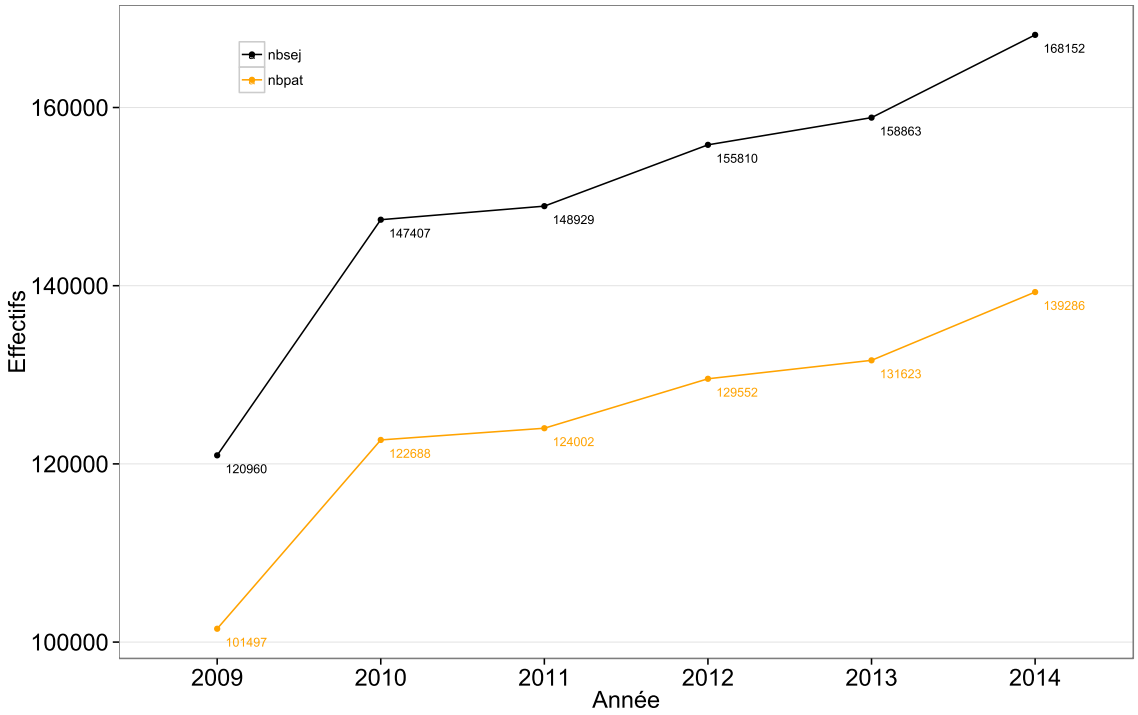


Figure 3.1 – Nombre de séjours et nombre de patients entre 20 et 99 ans hospitalisés pour IM par année.

b) le sexe : la répartition par année du nombre d'hospitalisations avec IM selon le sexe et la part de séjours des hommes. Entre 2009 et 2014, le nombre total d'hospitalisations chez les hommes pour IM est de 660 885, celui des femmes est de 239 236. La répartition selon le sexe est globalement de 2,7 hommes pour 1 femme. Au cours des années, la part des séjours concernant les hommes est relativement constante : elle représente 72,9% (resp. 73,7%) de l'ensemble des séjours pour IM en 2009 (resp. 2014).

c) le sexe et l'âge : si, tous âges confondus, les séjours des hommes représentent environ les trois-quarts de la totalité des hospitalisations avec IM, la répartition par genre est différente selon les classes d'âge (voir figure 3.2). En effet, chez les 25-64 ans, les hospitalisations avec IM concernent 4,5 fois plus souvent les hommes que les femmes. La différence dans la répartition selon le sexe tend ensuite à diminuer pour atteindre 42% de femmes et 58% d'hommes chez les 80-84 ans. À partir de 85 ans, les séjours hospitaliers pour IM concernent autant les femmes que les hommes.

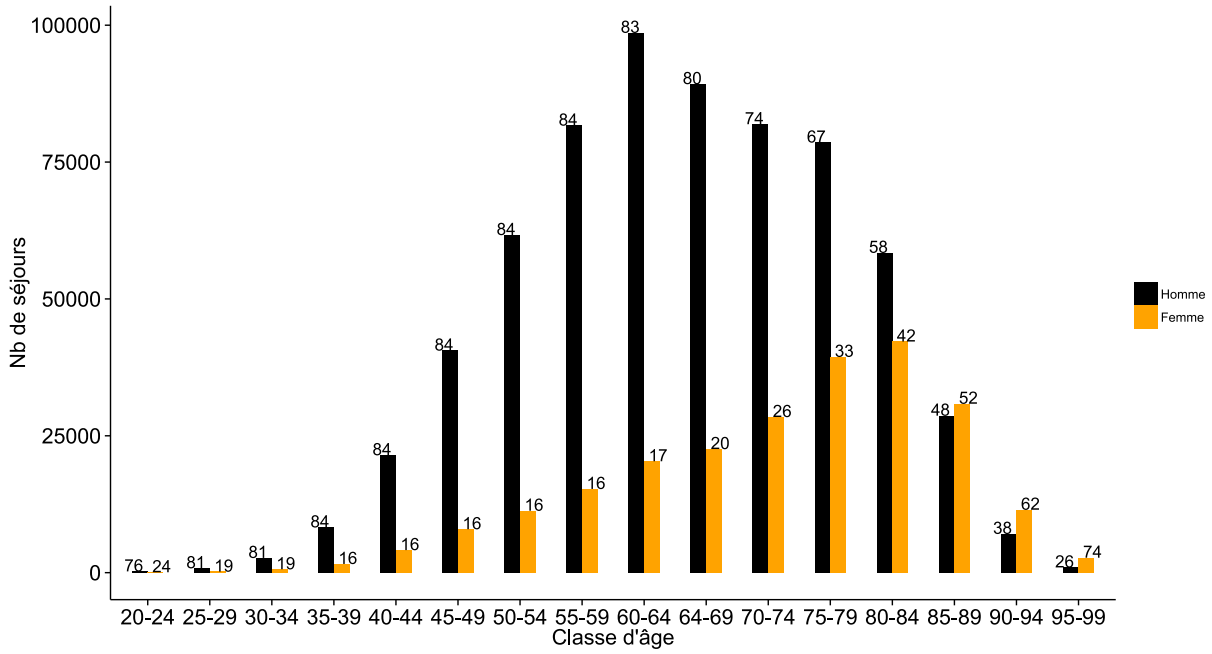


Figure 3.2 – Répartition des séjours hospitaliers pour IM par sexe et classe d'âge.

Taux d'hospitalisation. Nous avons calculé les taux selon les critères suivants :

a) le sexe : la figure 3.3 présente les taux bruts par sexe et pour l'ensemble de la population. Les taux diffèrent selon le sexe : 35,9 pour 10 000 hommes et 12,2 pour 10 000 femmes. La tendance évolutive est à l'augmentation : 29,2 pour 10 000 en 2009 à 39,9 pour 10 000 en 2014 chez les hommes et 10,2 pour 10 000 à 13,4 pour 10 000 chez les femmes. Le taux de séjours hospitaliers pour IM est en moyenne de 23,7 pour 10 000 habitants par an sur l'ensemble de la période d'étude.

b) le sexe et l'âge : lorsque l'âge est pris en compte, on observe des différences selon le sexe (voir figure 3.4). Chez les femmes les taux sont systématiquement inférieurs aux taux des hommes à partir de 30 ans. Chez les hommes, les taux de séjours hospitaliers pour IM augmentent avec l'âge pour atteindre des maxima de 8 et 7 pour 10 000 respectivement dans les classes d'âge des 60-64 ans et 75-79 ans. Entre 60 et 70 ans, les taux sont systématiquement en baisse,

sauf en 2009, où le taux est constant. Au contraire, chez les femmes, les taux hospitaliers pour IM augmentent également avec l'âge avec un maximum à 3 pour 10 000 pour les 75-84 ans. Ensuite, les taux d'hospitalisation pour IM diminuent au-delà de 90 ans.

Taux standardisés d'hospitalisation selon la région. Nous avons avons analysé les données selon le sexe en comparant les années 2009 et 2014.

a) Chez les hommes : les résultats sont représentés dans la figure 3.5 et résumés dans le tableau A.2 de l'annexe. Les taux standardisés nationaux sont respectivement de 39,4 et 52,4 pour 10 000 hommes en 2009 et 2014. Les taux varient en 2009 de 27,6 pour 10 000 dans les Pays de la Loire (resp. 37,1 pour 10 000 en Picardie en 2014) à 49,2 pour 10 000 en PACA (resp. 69,7 pour 10 000 en Corse en 2014). L'évolution annuelle est la plus importante pour la Corse, mais c'est également dans cette région que l'on observe l'une des variations la plus importante de la population [2]. *A contrario*, l'évolution de la population la plus faible concerne la Bourgogne. Or c'est le deuxième pourcentage le plus élevé avec 9,6% en moyenne par an.

b) Chez les femmes : les résultats sont représentés dans la figure 3.6 et résumés dans le tableau A.3 de l'annexe. Les taux standardisés nationaux sont respectivement de 11,5 et 15,2 pour 10 000 femmes en 2009 et 2014. Les taux varient en 2009 de 8,6 pour 10 000 en Auvergne (resp. 10,8 pour 10 000 en Alsace en 2014) à 12,9 pour 10 000 en Alsace (resp. 17,7 pour 10 000 dans les Pays de la Loire en 2014). Les évolutions des pourcentages des taux standardisés de séjours hospitaliers pour IM entre 2009 et 2014 sont les plus importantes pour les régions de Bourgogne et du Limousin, alors que le taux de croissance de la population entre 2009 et 2014 est respectivement de 0,3% et -1,1%. En revanche, le taux de variation le plus faible est de -0,3% pour la Franche-Comté, dont la population augmente de 1,30% sur six ans.

2. La population de cette région augmente de 5,6%.

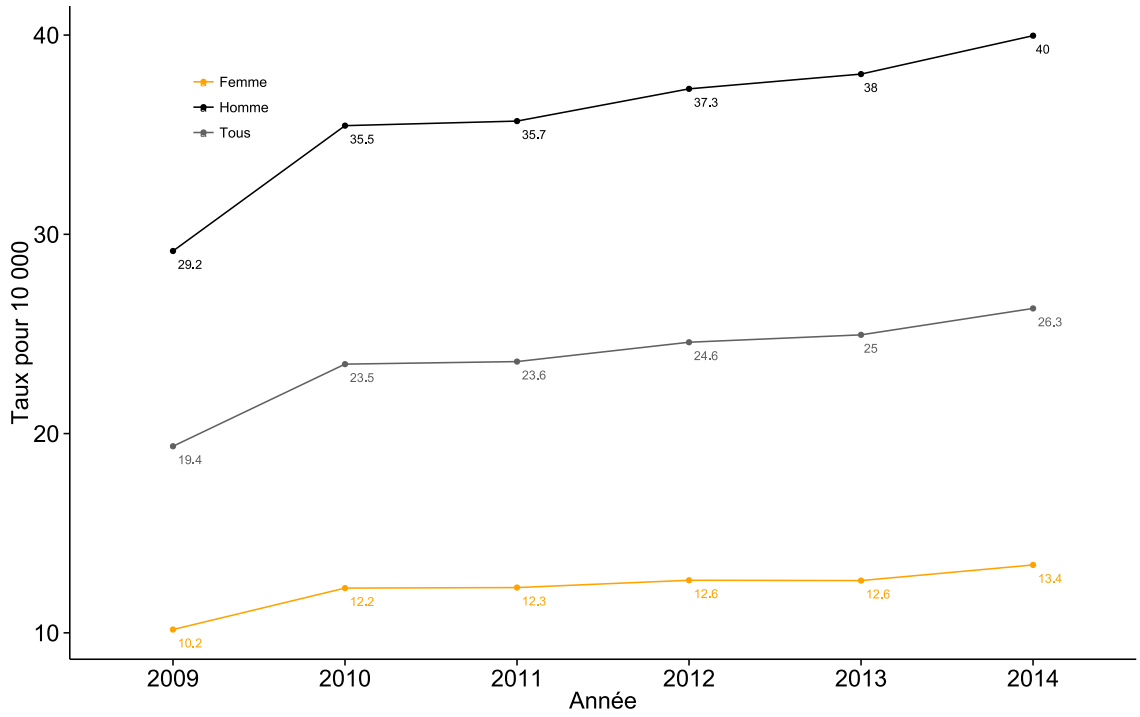


Figure 3.3 – Taux des hospitalisations pour IM (pour 10 000 habitants).

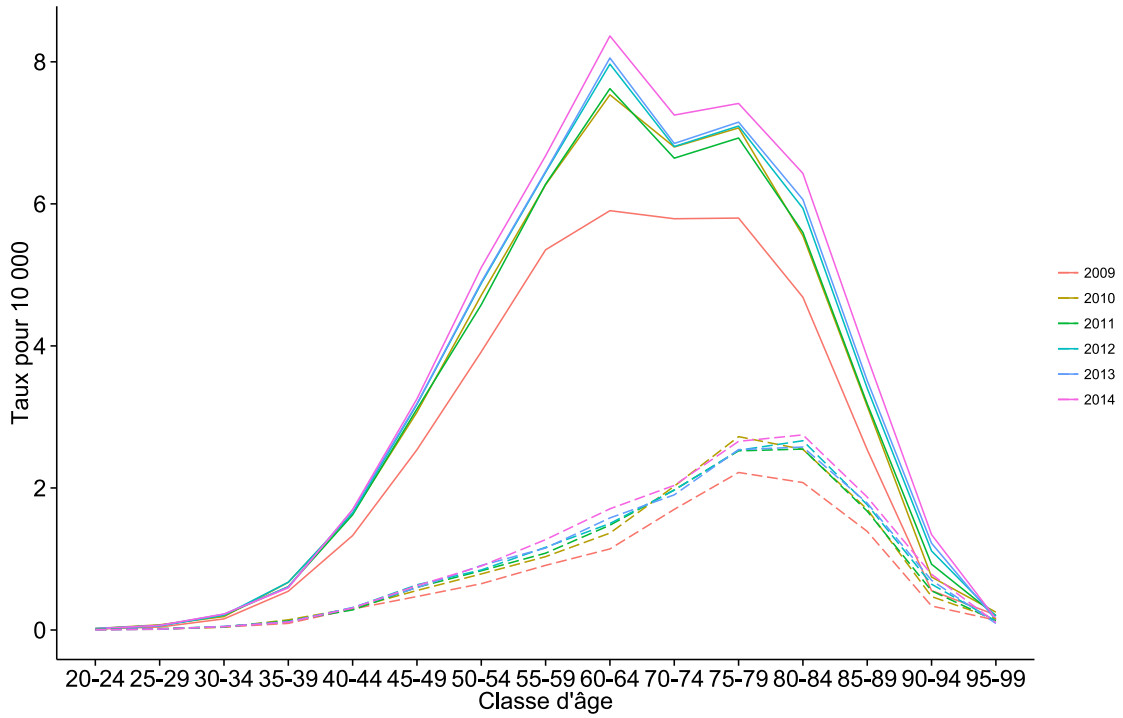


Figure 3.4 – Taux d'hospitalisation avec IM par sexe (trait plein pour les hommes et pointillés pour les femmes) et classe d'âge pour 10 000 habitants.

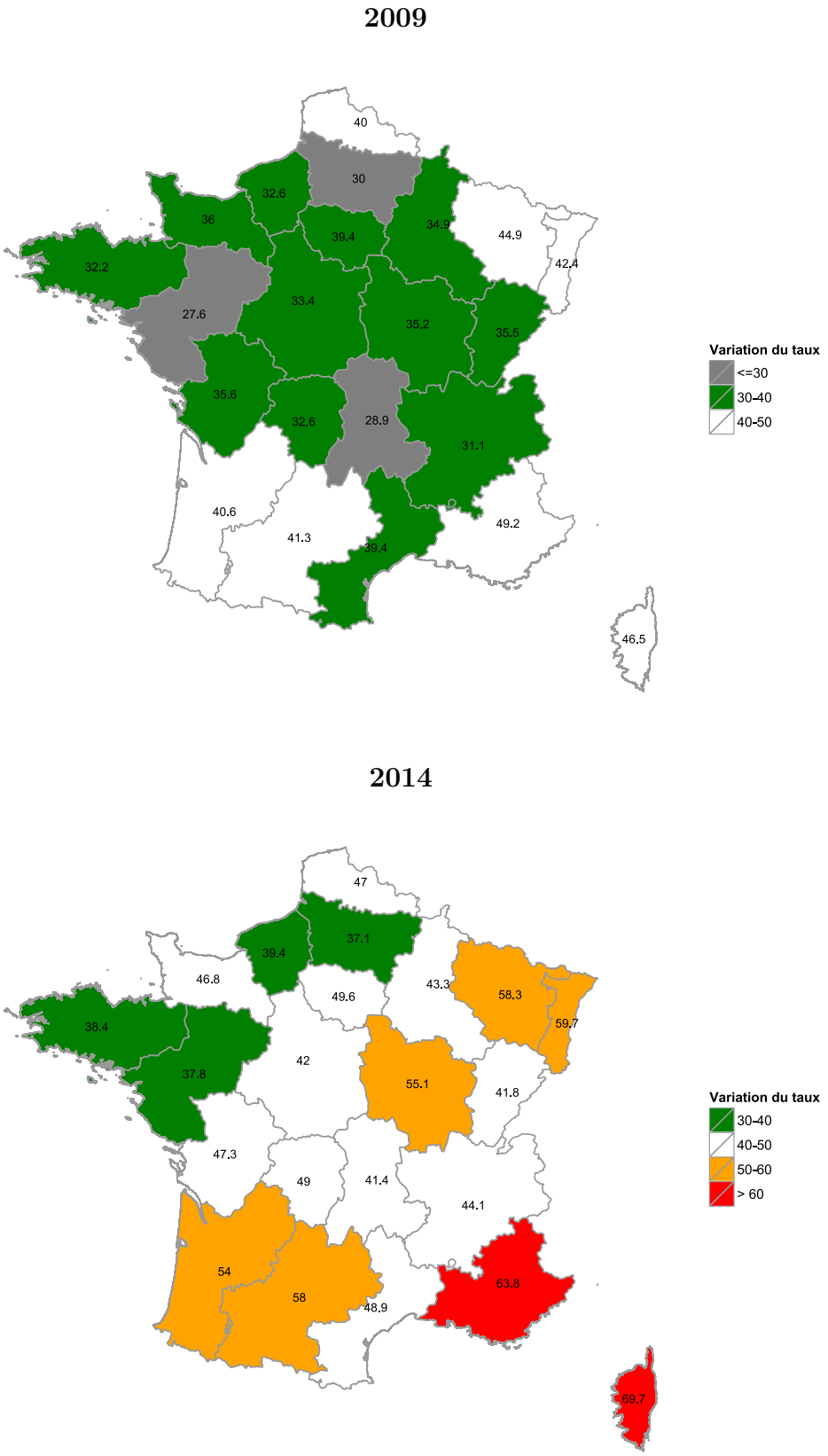


Figure 3.5 – Taux standardisés d'hospitalisation avec IM pour 10 000 hommes par région en 2009 et 2014 avec représentation des variations régionales.

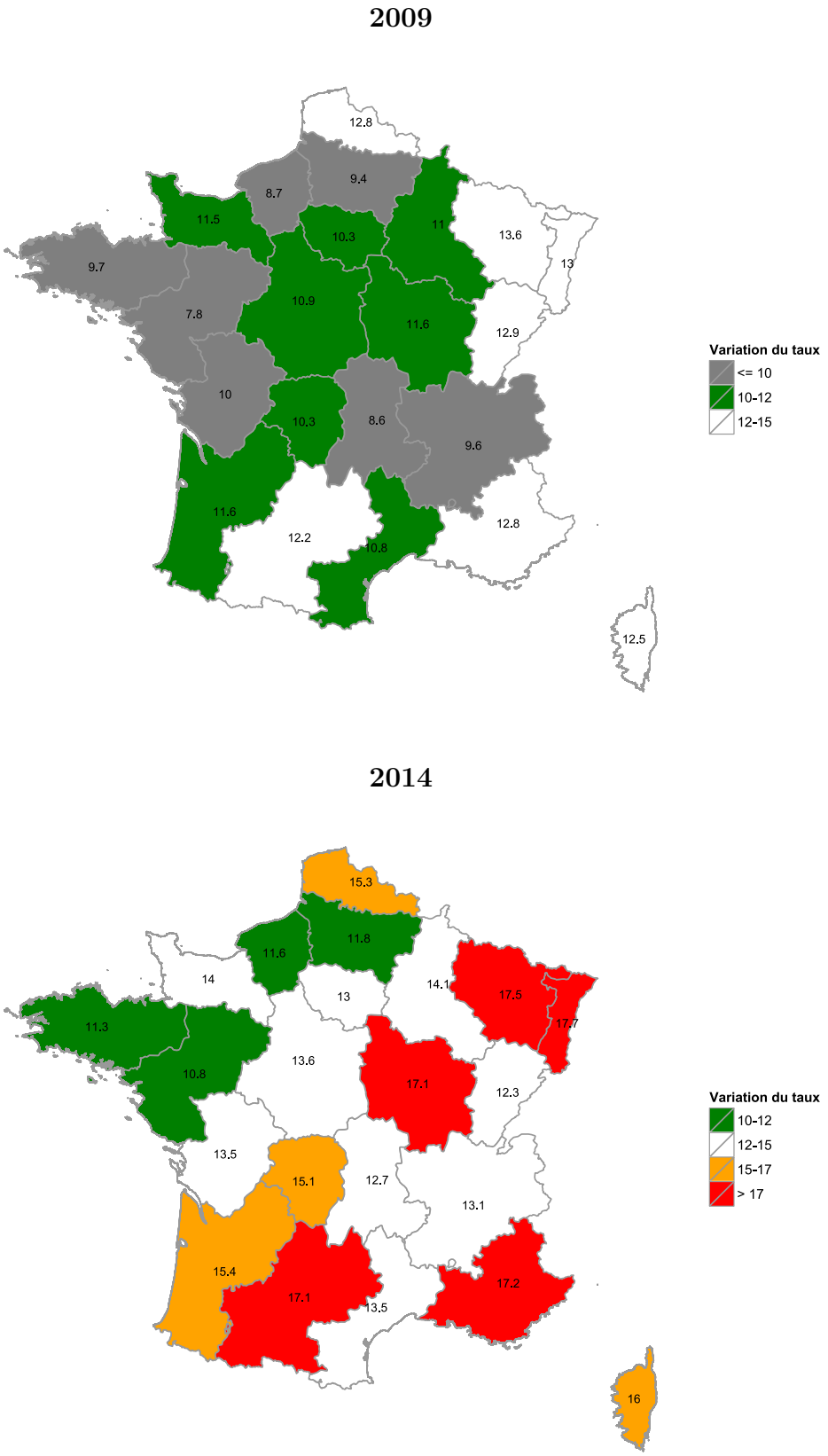


Figure 3.6 – Taux standardisés d'hospitalisation avec IM pour 10 000 femmes par région en 2009 et 2014 avec représentation des variations régionales.

Hospitalisation hors de la région d'habitation. Sur six ans, de 2009 à 2014, 7,6% des séjours pour IM se sont déroulés en dehors de la région d'origine. Parmi ces séjours, 76,3% concernent des hommes et 23,7% concernent des femmes. La figure 3.7 décrit l'évolution de la part de séjours et de patients au cours des différentes années. Quel que soit le sexe, ces courbes suivent la même évolution. Elles fluctuent entre 8,5% et 7,6% de séjours pour les hommes, et 6,9% et 6,7% pour les femmes avec une petite inflation en 2011.

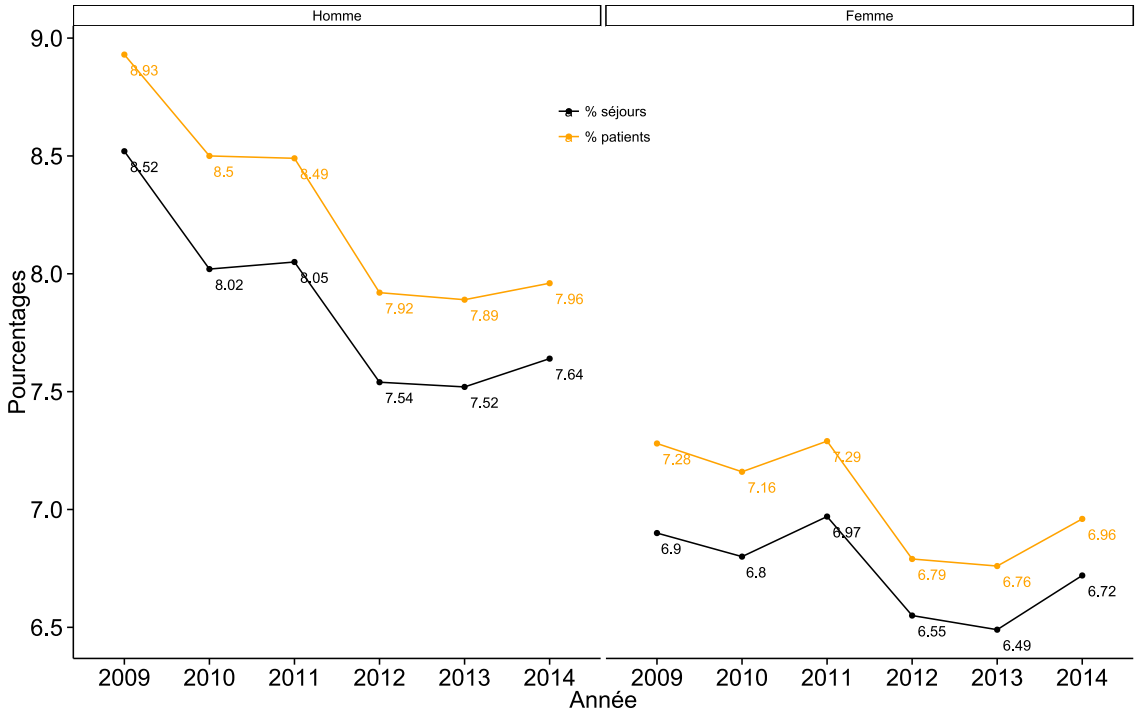


Figure 3.7 – Part des séjours et de patients hospitalisés pour IM en dehors de leur région d'habitation par année.

3.2.1.2 Caractérisation des séjours

Modalités de sortie. Après hospitalisation, un retour à domicile est observé dans 78,5% des cas et un transfert vers une autre structure de soins dans 17,4% des cas. Dans 3,3% des cas, les patients sont décédés au cours de l'hospitalisation. L'analyse par genre montre des différences : les taux de transfert (22,2% vs 15,7%) et de décès (5,4% vs 2,5%) sont plus importants pour les femmes tandis que le taux de retour à domicile est plus élevé chez les hommes (81,3% vs 70,7%).

Durée de séjour. La grande majorité des hospitalisations avec IM (74%) a duré 5 jours ou moins. Nous observons que, 3,7% des séjours ont duré moins d'une journée, 9,4% des séjours n'ont duré qu'un jour tandis que les séjours de 2 jours et ceux de 3-5 jours ont concerné respectivement 27,4% et 33,5% des hospitalisations avec IM. Les séjours de +15 jours n'ont concerné que 3,5% de l'ensemble des hospitalisations. Les durées d'hospitalisation pour IM n'ont pas varié entre 2009 et 2014. Nous présentons les résultats selon différents critères :

a) le sexe : la proportion d'hospitalisations pour moins d'une journée est plus importante chez les femmes que chez les hommes (resp. 8,3% et 3,6%). Les proportions des séjours de durée égale à 1 jour ne sont en revanche pas différentes selon le sexe (9%).

b) l'âge : la durée des séjours pour IM est différente selon l'âge des patients. Les durées moyennes de séjour sont de 5 jours chez les 20-24 ans, 3 jours chez les 55-59 ans et de 7 jours chez les +90 ans. La figure 3.8 montre que la durée d'hospitalisation augmente avec l'âge. Les séjours de 11-15 jours représentent 4% des séjours pour les 20-25 ans et 15% des séjours des 95-99 ans. Notons que pour ces deux classes d'âge, les proportions de séjours d'une journée sont identiques.

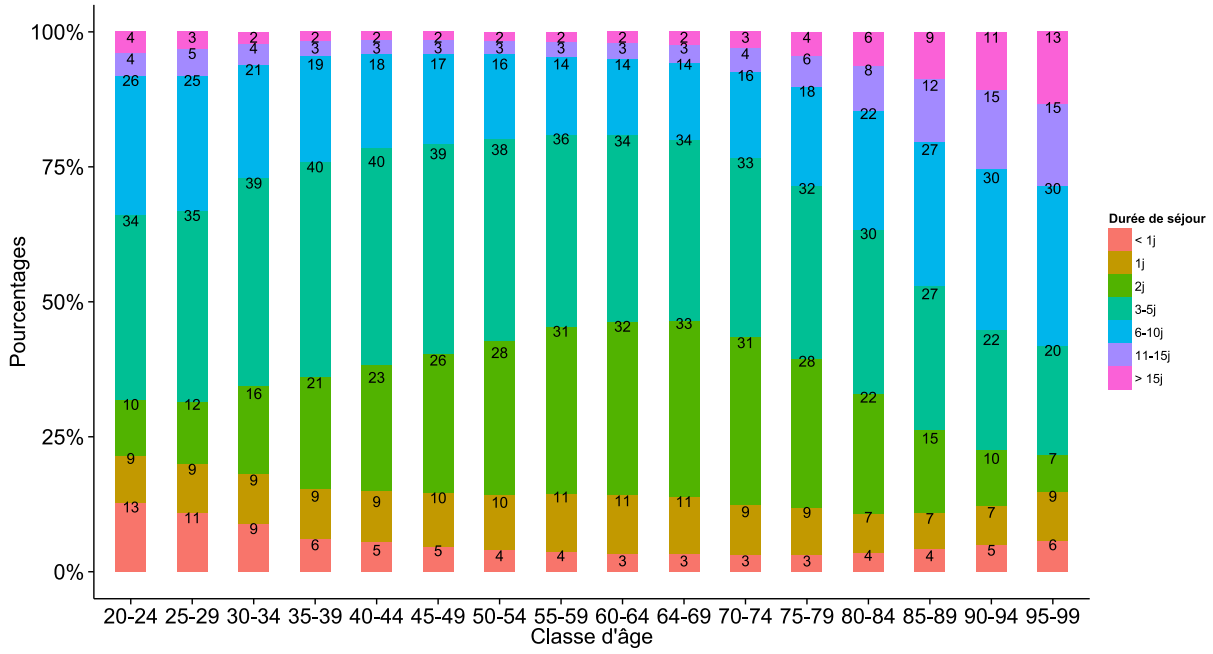


Figure 3.8 – Répartition de la durée de séjour hospitalier selon la classe d'âge.

Mois de survenue de l'IM. La survenue des hospitalisations avec IM est fluctuante selon les mois et les années. La figure 3.9 montre que le nombre de séjours hospitaliers pour IM est constamment moins élevé en juillet, août et septembre par opposition au nombre de séjours des mois de mars, octobre (sauf en 2010) et décembre. L'analyse par sexe donne des résultats identiques à ceux présentés dans la figure 3.9.

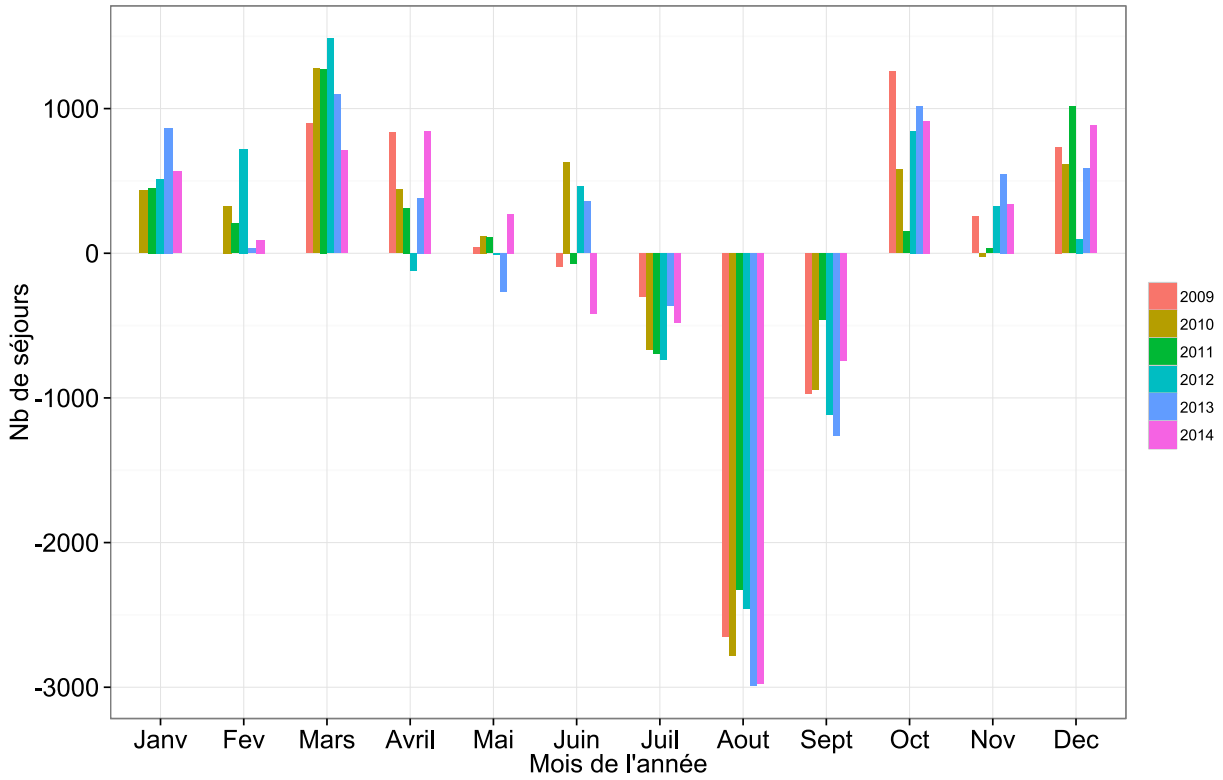


Figure 3.9 – Nombre de séjours hospitaliers suivant le mois de l'année.

Diagnostic d'hospitalisation. Quelle que soit l'année considérée, la fréquence relative des diagnostics d'hospitalisation pour les séjours IM n'est pas très différente. L'IM est le DP de loin le plus fréquent et concerne 65% des IM hospitalisées, soit entre 200 225 et 238 900 séjours hospitaliers par an entre 2009 et 2014 (dont 70% concerne les hommes). La cardiopathie ischémique chronique est le deuxième DP et représente 22% de l'ensemble des séjours hospitaliers pour IM. Le troisième DP est l'angine de poitrine, il représente 8% de l'ensemble des séjours. L'analyse selon le sexe ne montre pas de différence. En revanche, si l'on prend en compte l'âge, la fréquence des séjours en fonction du DP n'est pas la même. La part relative des hospitalisations avec un DP IM augmente avec l'âge chez les femmes, avec un pic à 80-84 ans, alors que l'on constate une baisse chez les hommes. Chez les hommes, l'augmentation se fait plus tôt avec un pic pour les 60-64 ans.

3.2.2 Étape 3. Réadmissions pour IM

3.2.2.1 Nombre de patients, taux et délai de réadmission

Nombre de patients. Le tableau 3.1 résume la répartition des patients selon le nombre d'hospitalisations. Dans le cas général, 74,8% (n = 507 120) des patients ont été hospitalisés une seule fois pour IM et 25,2% d'entre eux (n = 170 901) pour plusieurs séjours.

Nous présentons les résultats des analyses par :

a) sexe : les hommes ré-hospitalisés (n=491 769, 72,5%) sont plus nombreux que les femmes (n=186 252, 27,5%). Parmi l'ensemble des hommes hospitalisés pour IM, 26,4% ont eu plusieurs séjours pour IM. Ce pourcentage est de 24,1% parmi les femmes.

a) sexe et âge : le pourcentage de ré-hospitalisations pour IM est plus important chez les hommes excepté chez les 20-24 ans. La différence entre les pourcentages de ré-hospitalisations pour IM selon le sexe est nulle chez les 25-29 ans. Pour les autres classes d'âge, les taux varient très peu, avec un pourcentage toujours plus élevé chez les hommes.

Tableau 3.1 – Répartition du nombre d'hospitalisations pour IM selon le sexe.

	Hommes		**Femmes**		**Total**	
Nb hosp	Effectif	%	Effectif	%	Effectif	%
1	362 092	73,63	145 028	77,87	507 120	74,79
2	100 098	20,35	32 268	17,32	132 366	19,52
3	22 540	4,58	6 971	3,74	29 511	4,35
4	5 174	1,05	1 442	0,77	6 616	0,98
5	1 293	0,26	377	0,2	1 670	0,25
6	360	0,07	115	0,06	475	0,07
≥ 7	212	0,04	51	0,03	263	0,04
Total	491 769	100	186 252	100	678 021	100

Taux de réadmissions après une hospitalisation index. Nous présentons les résultats des analyses par :

a) sexe : les taux de réadmissions sont représentés dans la figure 3.10. Après une hospitalisation index[3], le taux d'une première ré-hospitalisation pour IM est globalement de 22,6% à un mois, 47,7% à 3 mois, 69,5% à 6 mois, 86,8% à 12 mois, 95,8% à 2 ans. Il n'y a pas de différence significative entre les hommes et les femmes.

b) âge : les résultats sont présentés dans la figure 3.11. À un mois, la réadmission a concerné 20,5% des -45 ans, 21,9% des 45-65 ans, 22,8% des 65-85 ans et 17,9% des +85 ans. Jusqu'à 1 an et 4 mois, le taux de réadmission le plus bas concerne les +85 ans par opposition aux -45 ans qui ont le taux de réadmission le plus élevé. Au delà de cette période, les +85 ans ont des taux surpassant toutes les autres tranches d'âge. Au delà de 2 ans et 3 mois, les taux se rejoignent toutes classes d'âge confondues.

c) la région : l'analyse des taux de réadmissions pour IM selon les régions (figure 3.12) a été effectuée sur un intervalle assez court au vue de la période d'observation. Ainsi, les taux comparés concernent les taux de réadmissions à 3 mois, 9 mois et un an. Tout d'abord, nous observons un contraste Nord-Sud pour les taux de ré-hospitalisations chez les hommes et une dissymétrie dans la représentation graphique. Les taux de ré-hospitalisations selon la région ne sont pas

3. C'est-à-dire la première hospitalisation pour IM (voir définition section 3.1.3).

les mêmes suivant le sexe. En effet, une grande différence est observée entre la Basse Normandie et la Bourgogne, où les hommes sont plus ré-hospitalisés que les femmes, puis entre Les Pays de la Loire et la Corse, où *a contrario*, les femmes sont plus ré-hospitalisées que les hommes.

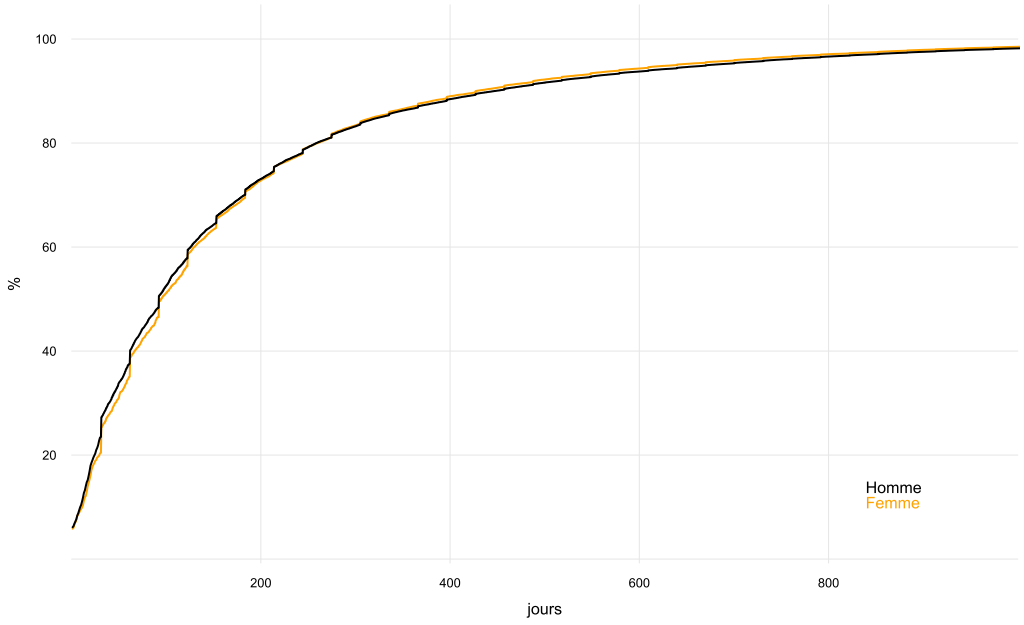


Figure 3.10 – Taux de réadmissions pour IM selon le nombre de jours écoulés après une hospitalisation index, par sexe.

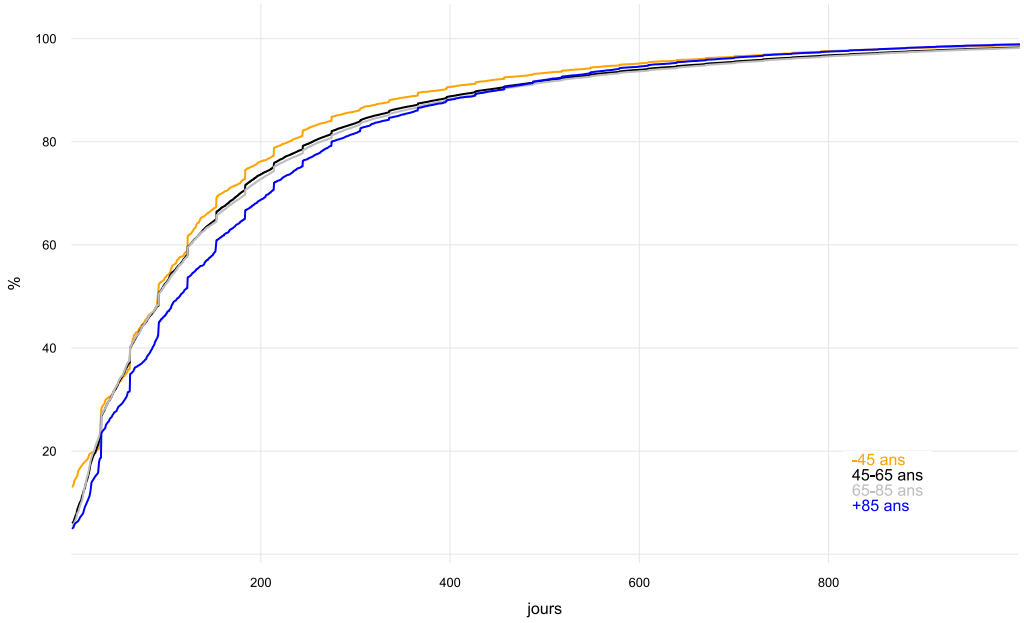


Figure 3.11 – Taux de réadmissions pour IM selon le nombre de jours écoulés après une hospitalisation index, par classe d'âge.

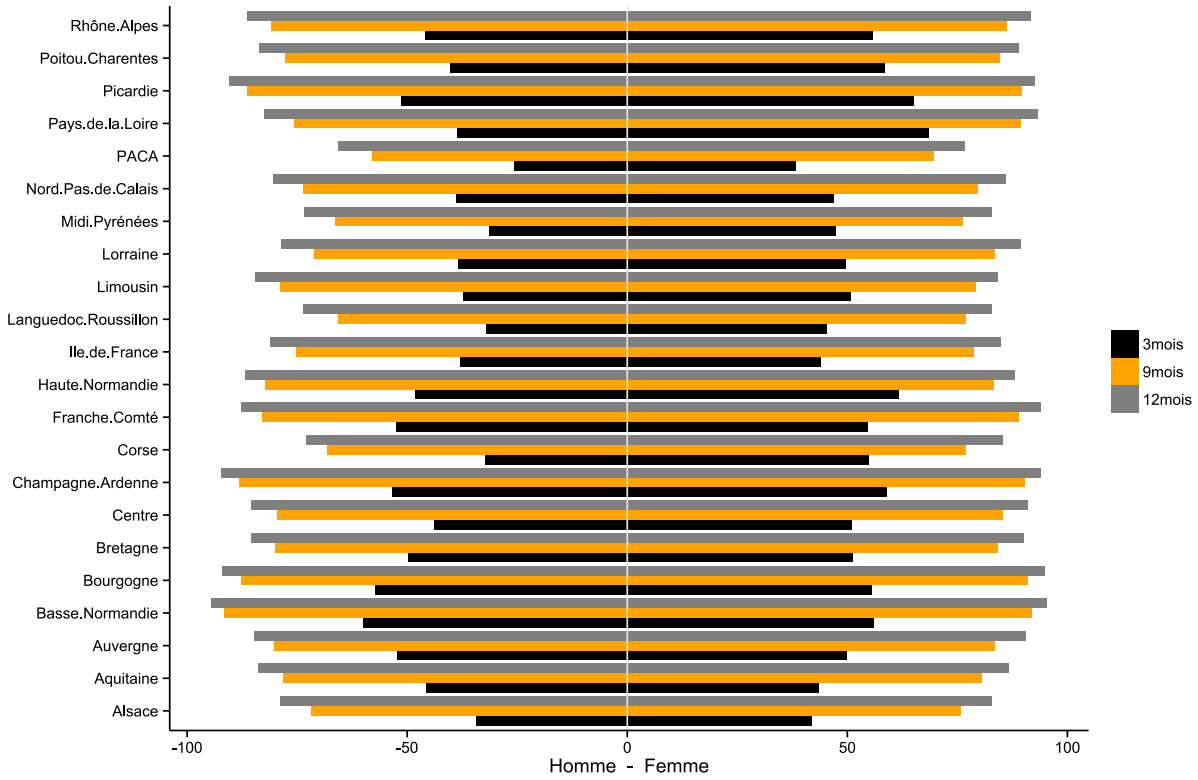


Figure 3.12 – Taux de réadmissions pour IM à 3 mois, 9 mois et 12 mois après une hospitalisation index, par sexe et par région.

Délai entre deux séjours pour IM. Nous détaillons les résultats de la façon suivante :

a) Nombre moyen de séjours : lorsqu'il y a réadmission pour IM, le nombre moyen de séjours hospitaliers entre 2009 et 2014 est de 2,3 (IC 95% [4] : 2,29-2,3) avec une différence significative ($p < 2.2e-16$) entre les hommes et les femmes. Les 41 224 femmes et les 129 677 hommes qui ont séjourné plusieurs fois à l'hôpital pour IM entre 2009 et 2014 ont eu un total de 353 970 séjours (85 867 pour les femmes et 268 112 pour les hommes).

b) Délais moyen et médian en jours : le délai moyen entre deux séjours pour IM est estimé à 275 jours (IC 95% : 273-277). Le délai médian est de 91 jours : lorsqu'il y a eu ré-hospitalisation pour une récidive, une personne sur deux l'a été dans les 3 mois. Le délai de ré-hospitalisation est significativement différent selon le sexe : la récidive survient en moyenne onze jours plus tard chez les hommes. Le délai moyen de ré-hospitalisation est de 277 jours (IC 95% : 275-279) chez les hommes contre 266 jours (IC 95% : 262-270) chez les femmes et le délai médian est de 3 mois pour les deux sexes.

4. Intervalle de confiance à 95%.

c) Distribution des patients selon le délai : les résultats sont détaillés dans le tableau 3.2. Lorsqu'il y a eu ré-hospitalisation pour IM, celle-ci a eu lieu dans les 48 heures qui a suivi la précédente hospitalisation pour IM, plus souvent pour les femmes que pour les hommes (16,3% parmi les hommes *vs* 18,4% parmi les femmes). De même, la réadmission a lieu au delà de 6 mois dans 38,2% des cas (38,4% parmi les hommes *vs* 37,6% parmi les femmes).

Tableau 3.2 – Distribution du délai entre deux séjours pour IM selon le sexe.

	Hommes		**Femmes**		**Total**	
	Effectif	%	Effectif	%	Effectif	%
-24h	1 351	1,13	562	1,82	1 913	1,28
24h	623	0,52	176	0,57	799	0,53
24-48h	1 095	0,92	294	0,95	1 389	0,93
2-3j	1 380	1,16	360	1,17	1 740	1,16
3-7j	7 322	6,14	1 691	5,49	9 013	6,01
7-15j	12 524	10,51	3 074	9,97	15 598	10,4
15j-1 mois	17 527	14,71	4 289	13,91	21 816	14,55
1-3 mois	17 795	14,93	4 914	15,94	22 709	15,14
3-6 mois	13 789	11,57	3 884	12,6	17 673	11,78
6 mois - 1 an	15 743	13,21	4 174	13,54	19 917	13,28
+1 an	30 012	25,19	7 406	24,03	37 418	24,95
Total	119 161	100	30 824	100	149 985	100

3.2.2.2 Évaluation du risque de réadmission

Nous avons modélisé le risque de réadmission pour un IM en prenant en compte, le sexe, l'âge discrétisé en 4 classes (-45 ans, 45-65 ans, 65-85 ans et +85 ans) et les facteurs de risque (décrits dans les DAS). Dans le modèle final, après une sélection pas à pas des variables, nous retenons : le sexe, l'âge puis le diabète, les dyslipidémies, le sepsis, l'insuffisance respiratoire (IR), la consommation de substances psycho-actives (TabacAutres) telles que l'alcool, le tabac ou encore le cannabis.

Les résultats sont présentés dans le tableau 3.3. En prenant pour référence les femmes, le risque de réadmission pour les hommes diminue de 13%. Comme pour l'âge, en prenant pour référence la classe -45 ans, le risque de réadmission diminue pour les autres classes d'âge. Notons que seule la consommation de substances psycho-actives est un facteur de risque de la réadmission. Ce dernier augmente alors de 13%.

Tableau 3.3 – Risque de ré-hospitalisation pour IM selon le sexe, la classe d'âge et la présence d'un facteur de risque.

	HR	**IC 95%**	**p.value**
Sexe			
Femme	1		
Homme	0,87	0,78 à 0,97	9,8e-03
Classe d'âge			
-45 ans	1		
45-65 ans	0,66	0,53 à 0,83	3,92e-04
65-85 ans	0,64	0,51 à 0,81	1,59e-04
+85 ans	0,71	0,53 à 0,94	0,02
Diabète	0,85	0,77 à 0,93	8,16e-04
Dyslipidémies	0,84	0,77 à 0,91	7,51e-05
Sepsis	0,49	0,24 à 0,98	0,04
IR	0,78	0,60 à 0,99	0,05
TabacAutres	1,13	1,01 à 1,27	0,03

HR : hazard ratio ; **IC 95%** : intervalle de confiance à 95%.

3.2.3 Étape 4. Décès et facteurs de risque

3.2.3.1 Nombres de patients et taux de létalité

Nombres de patients. Nous nous sommes intéressés à plusieurs critères :

a) le numéro de séjour : le tableau 3.4 détaille la répartition du nombre de décès selon l'année suivant le numéro de séjour de l'année concernée. La numérotation des séjours est remise à 1 à chaque changement d'année. En dépit d'une augmentation du nombre de personnes décédées au cours de ces six années, la répartition suivant le numéro de séjour, par année, n'est pas significativement différente. Cette répartition montre que le décès se produit majoritairement lors du premier séjour.

Tableau 3.4 – Répartition des décès selon le numéro de séjour d'hospitalisation.

	2009	**2010**	**2011**	**2012**	**2013**	**2014**
	n=4 212	n=5 252	n=5 034	n=5 127	n=5 085	n=5 018
1	3 798	4 700	4 539	4 627	4 579	4 543
2	360	480	422	438	437	414
3	44	66	64	50	55	48
≥ 4	10	6	9	12	14	13
NbM	1,11	1,12	1,11	1,11	1,12	1,11

NbM : représente le nombre moyen de séjours pour IM au décès

b) l'âge et le sexe : la figure 3.13 décrit parmi les décès, la proportion par classe d'âge concernée, suivant le sexe et l'année [5]. Sur la période des six ans, quel que soit le sexe, les décès touchent en majorité les 80-89 ans avec un taux de 20%. Chez les femmes, la proportion de décès est plus importante pour les 80-94 ans avec des pourcentages allant de 19% à 28% selon les années. Tandis que, chez les hommes la proportion augmente avec l'âge et diminue ensuite pour les +90 ans.

5. Pour des questions de lisibilité, l'affichage des proportions n'est fait que pour les valeurs supérieures ou égales à 5%.

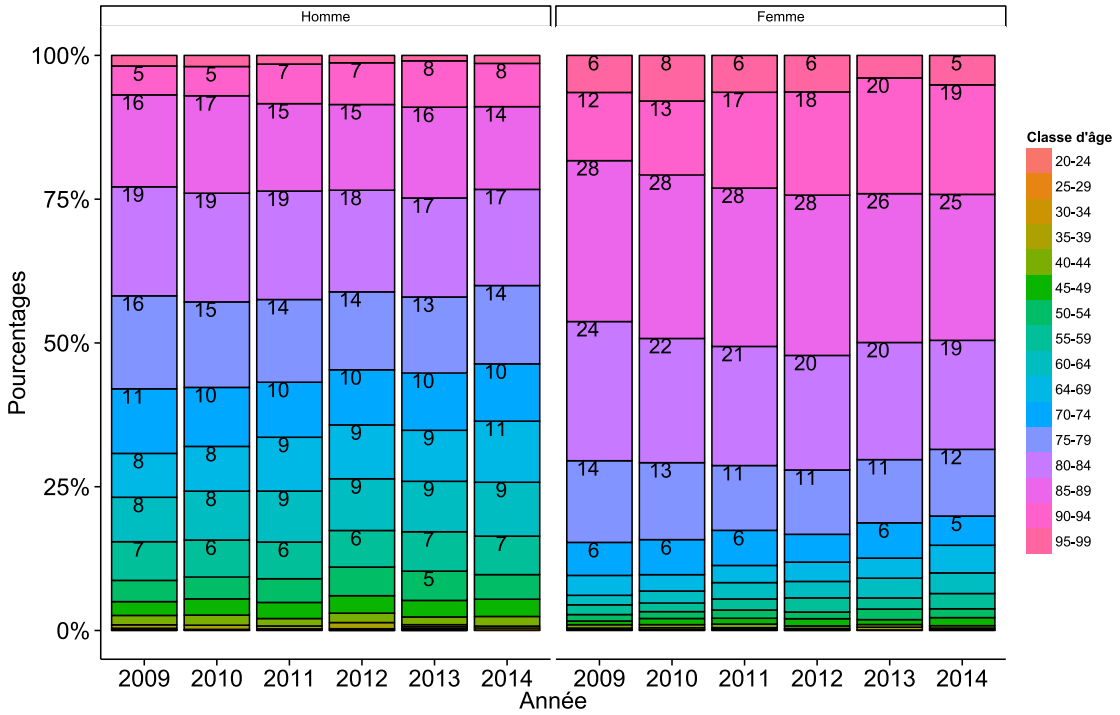


Figure 3.13 – Proportion de décès par classe d'âge suivant le sexe et l'année.

Taux de létalité. Les données ont été analysées suivant plusieurs critères :

a) le sexe : le tableau 3.5 résume les taux de létalité suivant le sexe. Au cours des années 2009 à 2014, 3,3% (n=29 728) des patients hospitalisés pour IM sont décédés avec un taux de létalité variant entre 80 et 95 pour 10 000 séjours hospitaliers pour IM par an. Ainsi, chaque année, entre 4 212 et 5 252 personnes admises pour IM sont décédées au cours de leur séjour à l'hôpital. Quelle que soit l'année considérée, environ 14,2% des décès sont survenus au cours de la première hospitalisation. Le taux de létalité est plus important chez les femmes (entre 0,013 et 0,022 pour 10 000 hospitalisations) que chez les hommes (0,003 pour 10 000 hospitalisations).

Tableau 3.5 – Nombre de décès et taux de létalité pour 10 000 hospitalisations.

	2009	**2010**	**2011**	**2012**	**2013**	**2014**
Nb décès Hommes	2 332	2 905	2 789	2 892	2 964	2 871
Taux de létalité	0.003	0.003	0.003	0.003	0.003	0.002
Nb décès Femmes	1 880	2 347	2 245	2 235	2 121	2 147
Taux de létalité	0.021	0.018	0.017	0.015	0.015	0.013
Nb décès totaux	4 212	5 252	5 034	5 127	5 085	5 018
Taux de létalité	0.003	0.003	0.003	0.002	0.002	0.002

b) le sexe et l'âge : la figure 3.14 représente les taux bruts de létalité suivant la classe d'âge, le sexe et l'année. Elle montre que les taux sont importants pour les -35 ans quel que soit le sexe. Ensuite, ils décroissent, jusqu'à 40 ans et augmentent de façon exponentielle avec l'âge, à partir de la classe d'âge 75-79 ans.

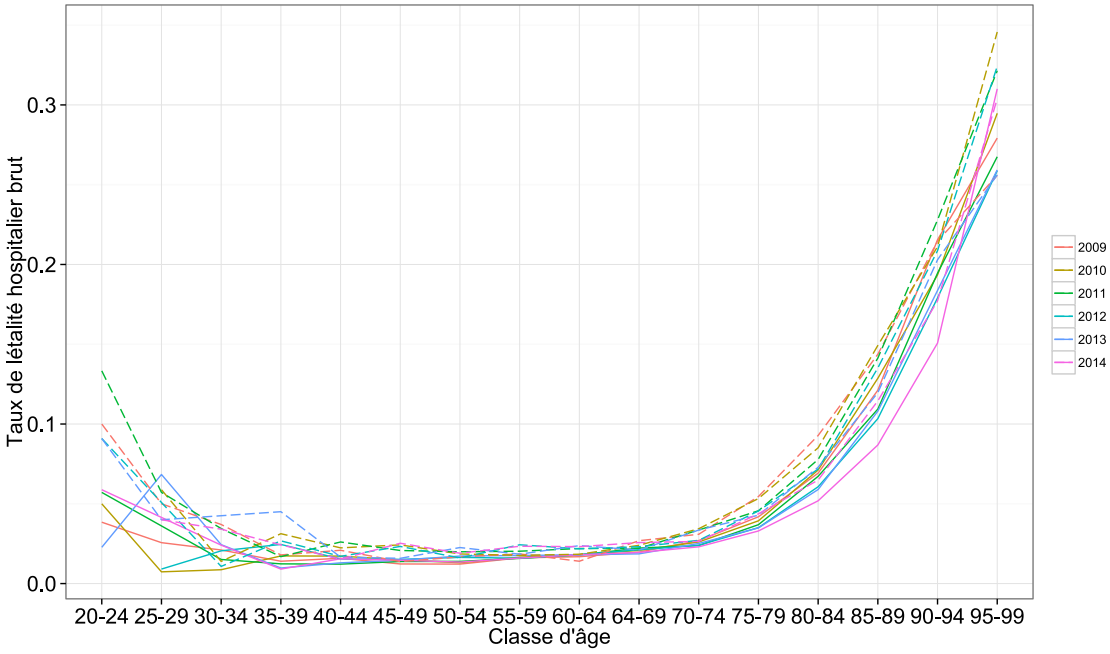


Figure 3.14 – Taux de létalité hospitalier par sexe (trait plein pour les hommes et pointillés pour les femmes) et par classe d'âge.

3.2.3.2 Facteurs de risque

Pour aller plus loin dans notre analyse, nous avons étudié les deux populations dont le taux de létalité est plus élevé que la moyenne, c'est-à-dire pour les -40 ans et les +80 ans. Nous avons cherché à identifier les facteurs de risque pour ces deux populations ayant de taux de létalité importants à l'aide de la régression logistique. Dans le cas des -40 ans, les modèles ne permettent pas d'établir de façon probante les facteurs de risque. Nous ne détaillerons donc pas le résultat de la modélisation. Dans le cas des +80 ans, les résultats sont présentés dans le tableau 3.6. Après une sélection pas à pas des variables, sont retenues dans le modèle : le sexe, l'insuffisance myocardique (InsufMyo), la maladie rénale chronique modérée à sévère (RMS), les tumeurs métastatiques et solides (MetaTumeur), l'hépatopathie modérée à sévère (HMS), l'hémiplégie, la démence et les leucémies.

Tableau 3.6 – Identification des facteurs de risque du décès hospitalier chez les +80 ans.

Variables	**Coeff**	**OR**	**IC 95%**
Constante	-0,58***		
Sexe			
Femme	0,11**	1,12	1.04 à 1.2
InsufMyo	0,53***	1,71	1,58 à 1,83
RMS	0,49***	1,63	1,51 à 1,75
MetaTumeur	0,77***	2,15	1,79 à 2.6
HMS	1,12***	3,07	2,18 à 4,42
Hémiplégie	0,47***	1,6	1,34 à 1,9
Démence	0,18***	1,20	1,09 à 1,33
Leucémies	0,63***	1,89	1,31 à 2,74

Coeff : valeur des β_i du modèle
et test de nullité des β_i avec *p<0,05 ; **p<0,01 ;
***p<0,001 ;**OR** : odds-ratio ;
IC 95% : intervalle de confiance à 95% des OR.

Ces résultats nous indiquent que par rapport aux hommes, le risque chez les femmes est multiplié par 1,12. Les facteurs de risque les plus importants sont les tumeurs métastatiques et l'hépatopathie modérée à sévère pour lesquelles le risque est multiplié respectivement par 2,15 et par 3.

3.3 Commentaires

Dans cette section, nous commentons les résultats de nos analyses concernant les étapes 2 à 4, dans les sections 3.3.1 à 3.3.3. Puis, nous évoquons les limites de cette étude en matière d'exhaustivité des données dans la section 3.3.4.

3.3.1 Hospitalisations avec IM

La répartition du nombre de patients et de séjours par an permet d'établir que l'IM représente une moyenne de 150 000 séjours annuels pour une moyenne d'environ 125 000 patients par an dans les services de MCO. De plus, l'évolution annuelle des taux d'hospitalisation pour IM entre 2009 et 2014 est globalement en hausse. Dans le même temps, d'après l'ATIH [6], le nombre total de séjours hospitaliers en MCO est passé de 11,96 millions en 2009 à 11,31 millions en 2014, les séjours pour IM représentant ainsi 1% des séjours totaux en 2009 et 1,5% en 2014. Cette augmentation peut s'expliquer par :

— un plus grand recours aux soins dès les premiers symptômes d'un IM grâce aux campagnes d'informations et de sensibilisation, ou grâce à une meilleure diffusion des recommandations de la Haute Autorité de Santé (HAS), publiées depuis 2007 [HAS, 2007] suite à la conférence de consensus du 23 novembre 2006 ;

— une augmentation du nombre d'IM dans la population, bien que l'enquête de la [DREES, 2015] parue en 2015, n'indique ni une baisse, ni augmentation du nombre d'IM dans la population française sur la période de 2000 à 2010 ;

6. http://www.scansante.fr/applications/statistiques-activite-MCO-par-diagnostique-et-actes

— une augmentation des facteurs de risque. Par exemple, sur la période allant de 2005 à 2010, la prévalence du tabagisme quotidien a augmenté de 7% [Dujardin et Cambou, 2005].

Nous avons également retrouvé des résultats de la littérature internationale [Lloyd-Jones *et al.*, 2010, Campbell, 2008] concernant la prépondérance des séjours pour les hommes : ils sont 3 fois plus fréquents que chez les femmes. Toutefois, nous avons également établi une différence selon la classe d'âge et au-delà de 75 ans. Les femmes sont touchées plus tardivement par la maladie. Néanmoins ceci est à nuancer avec l'espérance de vie qui est plus élevée pour les femmes (84,4 ans *vs* 77,7 ans pour les hommes en 2009 [7]) mais aussi avec l'augmentation de la population des +90 ans entre 2009 et 2014 [8].

Lorsque l'on s'intéresse à la région de survenue des IM, elle concerne en majorité les habitants des régions du nord-est de la France mais également des régions du sud et sud-ouest. Nous expliquons les taux élevés d'hospitalisation pour les régions du sud du fait de l'accroissement démographique de la population des personnes âgées en raison de l'héliotropisme. Par ailleurs, ces résultats ont déjà été constatés dans l'étude épidémiologique de 1995 [Cambou *et al.*, 1997] indiquant un fort contraste régional dans la gravité des IM. De plus, à l'échelle nationale, le projet MONICA [Investigators *et al.*, 1988] a souligné une importante disparité géographique de la maladie, à l'échelle mondiale, mais également dans chaque pays avec un gradient Nord-Sud [Yusuf *et al.*, 2001].

Nous avons ensuite étudié les caractéristiques de ces séjours avec notamment la durée de séjour. La durée moyenne du séjour pour un IM dépend de l'âge de la personne atteinte et des complications de l'infarctus lui-même. La durée moyenne de séjour est de 7 jours pour les personnes très âgées et de l'ordre de trois jours pour les personnes ayant entre 55 et 59 ans. Ceci est cohérent avec la norme française. En France, la durée moyenne de séjour est de l'ordre d'une semaine. Toutefois, il est possible de faire sortir un patient plus précocement. Dans le cas d'un infarctus revascularisé non compliqué, le délai après lequel le risque vital chute de façon significative est de 3 jours [Mark et Newby, 2003]. Cette perception de la prise en charge est différente suivant les pays : au Japon elle est de l'ordre de 4 semaines [Kinjo *et al.*, 2005] ; en Norvège et au Danemark elle est de 4 jours, en Allemagne et en Grèce elle est d'environ 10 jours selon les données de l'OCDE (Organisation de Coopération et de Développement Économiques).
De même, nous avons étudié la survenue d'un épisode d'IM. Nous déduisons que, les hospitalisations pour IM sont systématiquement plus fréquentes pendant les mois d'hiver (mars, octobre et décembre), tandis qu'elles sont moins fréquentes pendant la période estivale (de juillet à septembre). Diverses études [Spencer *et al.*, 1998, Ornato *et al.*, 1996, Loughnan *et al.*, 2008, Hernández *et al.*, 2004] menées dans différents pays obtiennent des conclusions similaires sur la saisonnalité des IM. Ces études s'accordent sur la survenue des épisodes dans les périodes d'hiver dans la majeure partie des cas. Bien que ces données doivent être considérées avec précau-

7. Source Insee 2015.
8. De l'ordre de 57%, d'après les chiffres Insee de 2016.

tion, puisque la date exacte d'entrée des patients n'est pas présente dans les bases anonymisées du PMSI, le fait que la survenue des IM semble être plus fréquentes certains mois d'hiver pourrait par ailleurs constituer une aide dans l'organisation des soins.

3.3.2 Ré-hospitalisations pour IM

Après la première hospitalisation incluse dans notre étude, sur un suivi pouvant aller jusqu'à six années, il y a une ré-hospitalisation pour IM pour 30% des patients. Les ré-hospitalisations sont plus fréquentes chez les 30-49 ans, alors qu'ensuite elles diminuent avec l'âge. Ces résultats sont corroborés par ceux de la littérature [Andrés *et al.*, 2012, Dharmarajan *et al.*, 2015]. Une ré-hospitalisation est plus fréquente chez les femmes. Bien que jusqu'à présent cette maladie concerne plutôt les hommes, cette tendance a changé au cours des dernières années. D'ailleurs, certains auteurs [Dreyer *et al.*, 2015, Dreyer *et al.*, 2014, Gabizon et Lonn, 2015, Gabet *et al.*, 2016] ont établi que cette tendance serait liée à une augmentation des facteurs de risque tels que le tabagisme, l'association du tabagisme avec la prise de la pilule, l'éclampsie ou pré-éclampsie.

Notons qu'au-delà de deux ans, le taux de réadmission frôle les 100%. Autrement dit, tous les patients ont été ré-hospitalisés. Ce phénomène s'explique par des durées d'observation courtes d'une part et la longue durée d'observation pour ce type de pathologie. En effet, plus l'observation est longue plus la probabilité d'observer l'événement augmente. De plus, la modélisation choisie pour les taux de récidives par Kaplan-Meier, avec des intervalles, entraîne une réduction par moité des individus à risque à chaque intervalle suivant. Ainsi, avec de nombreux petits intervalles de temps, nous nous retrouvons assez rapidement avec peu de différence entre les patients à risque et les patients ré-hospitalisés ce qui explique ce fort taux de réadmissions.

Nous avons fait ressortir une différence régionale avec un contraste Nord-Sud pour les hommes. Les hommes sont moins ré-hospitalisés dans les régions du sud que dans celles du nord, alors que les taux d'hospitalisations sont plus élevés dans ces régions. De plus, les taux de décès hospitaliers ne montrent pas de différence entre les régions concernées et les autres. En outre, nous faisons également ressortir une différence de genre suivant certaines régions qui ont toutes des taux d'hospitalisation plus élevés pour les hommes, et des effectifs plus importants pour les femmes. Pourtant, hormis pour la Corse, il n'y a pas des taux de décès anormalement élevés qui pourraient expliquer cette différence régionale. Les différences territoriales sont également rapportées dans une étude [Marrugat *et al.*, 2004] comparant les taux de mortalité et d'incidence en Espagne, mais aussi entre divers pays par le projet WHO [9] MONICA [Thorvaldsen *et al.*, 1995].

9. World Health Organization.

Nous avons ensuite étudié le délai de réhospitalisation. Nos résultats montrent que, parmi les patients ré-hospitalisés, 23% ont eu une récidive un mois plus tard, et presque 50% trois mois plus tard, que ce soit pour les hommes ou pour les femmes. Nous obtenons des résultats très proches de l'étude [Oliver, 2014], où 20% de ces patients vont connaître une ré-hospitalisation ou le décès dans le mois qui suit l'hospitalisation pour événement cardiaque. Ainsi, l'examen des délais entre une première hospitalisation pour IM et une récidive révèle que pour 75% des patients qui ont eu une récidive, elle a eu lieu dans l'année qui suivait.

Enfin, nous terminons cette partie avec la modélisation du risque de réadmission. Les résultats viennent renforcer les éléments déjà évoqués plus haut quand à savoir un risque plus élevé pour les femmes et les -45 ans. De plus, parmi les facteurs de risque de l'IM, la consommation de substances psychoactives augmente le risque de récidive de 13%. Cependant, nous observons une diminution du risque de réadmission pour le diabète, l'insuffisance respiratoire, Sepsis et dyslipidémies. Les facteurs de risque étant recueillis dans les DAS, nous pouvons expliquer ce résultat contraire à toute attente par le manque d'exhaustivité du codage (évoqué dans la section 2.4.2 du chapitre 2). Dans la section suivante, nous allons revenir plus en détails sur ces facteurs de risque.

3.3.3 Décès et facteurs de risque

Dans de nombreux pays [Freisinger *et al.*, 2014, Chung *et al.*, 2015], les registres ou données administratives hospitalières sont utilisés pour la surveillance des IM. En Grande Bretagne, avec le NICOR [10]/MINAP [11] et le CALIBER [12], le taux de mortalité estimé des IM aigus est de 9,7%. En Suède le SWEDEHEART [13]/RIKS-HIA [14] permet d'établir ce taux à 8,4%. En Allemagne, avec les bases hospitalières, ce taux est estimé à 10,8%. Toutefois, les pratiques de codages, le repérage des séjours concernés, les systèmes de santé et de recours aux soins, ainsi que les années, sont trop différents pour comparer facilement les taux relatés par ces pays avec les résultats issus de notre étude.

Nos analyses nous ont permis d'établir que le décès, lorsqu'il se produit, arrive le plus souvent lors du premier séjour [Asaria *et al.*, 2017]. De plus, il touche majoritairement les personnes âgées de +80 ans et le plus souvent les femmes. Toutefois, ce dernier élément est à tempérer avec le fait que cette pathologie atteint les femmes plus tardivement. Elles sont donc plus âgées pour les mêmes risques [Vaccarino *et al.*, 1995].

10. National Institute for Cardiovascular Outcomes Research.
11. Myocardial Ischaemia National Audit Project.
12. CArdiovascular disease research using LInked Bespoke studies and Electronic health Records.
13. Swedish Web-system for Enhancement and Development of Evidence-based care in Heart disease Evaluated Accorded to Recommended Therapies.
14. Register of Information and Knowledge about Swedish Heart Intensive care Admissions.

Par ailleurs, nous établissons que les taux de létalité sont fortement élevés chez les -40 ans. Cependant, nous n'avons pas pu déterminer de façon probante un lien de cause à effet avec des comorbidités associées. Néanmoins, nous avançons deux explications à l'observation de ce phénomène :

— pour certaines personnes, il y a des facteurs de risque associés aggravant le contexte cardiovasculaire, mais les faiblesses de l'exhaustivité du codage ne permettent pas de les mettre en évidence avec ce type de données (voir section 2.4.2 du chapitre 2) ;

— pour d'autres personnes, il y a également des facteurs liés à l'environnement : la pollution de l'air [Laaidi *et al.*, 2002, Pascal, 2009], le stress au travail [Kivimaki *et al.*, 2012, Bureau International du Travail, 2003], l'hygiène de vie [Ruidavets *et al.*, 2010, Beck *et al.*, 2010, Kohl *et al.*, 2012] ou encore la prise de certains contraceptifs pour les femmes [Dalichampt *et al.*, 2014].

Nous remarquons que pour la population des +80 ans, nous n'avons pas retrouvé les facteurs de risque habituellement associés à cette pathologie comme le diabète. En outre, nous établissons que le décès survient par le biais d'autres pathologies telles que l'insuffisance rénale, les leucémies, la démence... D'autres investigations pourraient être menées dans l'identification de ces facteurs de risque en s'intéressant aux comorbidités introduites dans le calcul du score d'Elixhauser [Elixhauser *et al.*, 1998] par exemple. Par ailleurs, il serait également intéressant de déterminer parmi les deux scores de comorbidités (Charlson et Elixhauser), lequel est le plus prédicteur du décès dans le cas d'étude des patients atteints d'IM.

3.3.4 Exhaustivité des données : les limites

Dans le chapitre 2 section 2.4.2, nous avons détaillé les différentes limites relatives à l'utilisation des bases médico-administratives pour des études épidémiologiques. Certaines limites ont par ailleurs fait surface dans les sections précédentes, en particulier dans les études des risques de réadmissions et de décès. Dans cette section, nous allons plus spécifiquement nous intéresser à la sélection des patients et souligner des limites concernant l'exhaustivité de cette sélection. Pour évoquer cela, nous nous posons trois questions :

1. *Tous les patients ayant un IM sont-ils nécessairement hospitalisés ?* Les patients ayant eu un IM mais qui ne sont pas vus dans le système de soins (médecine de ville ou établissement de santé) regroupent deux catégories de patients :

   — ceux qui ont fait un IM « silencieux » [Cohn, 1985] n'ayant entraîné aucun recours aux soins. Le patient ignore qu'il a fait un incident cardiaque puisqu'il ne ressent pas les symptômes classiques comme la douleur thoracique.

   — ceux dont l'IM a entraîné un décès immédiat. En effet, sur les 33 435 décès par IM annuels [15], environ 5 085 décès [16] se sont produits au cours de l'hospitalisation. La grande majorité des décès par IM survient donc en dehors de l'hôpital.

15. Données 2013 CépiDC (Centre d'épidémiologie sur les causes médicales de décès).
16. Données 2013 ATIH.

2. *Les patients hospitalisés apparaissent-ils nécessairement dans la base PMSI-MCO ?* Certains patients atteints, non hospitalisés après passage dans les services d'urgences, probablement ceux considérés comme les moins graves ou *a contrario* décédés aux urgences, ne sont pas comptabilisés dans le PMSI. Ainsi, le fait de circonscrire l'analyse aux IM hospitalisés en MCO donne une image homogène mais partielle de la pathologie.
3. *Le codage de l'IM est-il bien renseigné dans les bases PMSI-MCO ?* Des études ont comparé les données de registres propres aux pathologies cardiaques et les données médico-administratives. Par exemple, l'étude [De Peretti et Bonaldi, 2010] de l'InVS d'étalonnage du PMSI MCO dans la surveillance des IM rapporte que la détection des IM, à partir des RSA à l'aide des codes CIM-10, a une sensibilité de 76% (et une VPP de 78%), quel que soit l'âge entre 35 et 74 ans, et quel que soit le genre. Ce recueil a été fait avec les mêmes contraintes que celles des registres, c'est-à-dire en ne comptabilisant qu'un seul événement par patient et par période de 28 jours. Une autre étude [Aboa-Eboulé *et al.*, 2013] basée sur des données plus récentes de 2004 à 2008 compare la précision des données PMSI par rapport au registre AVC et obtient une sensibilité de 77% et une VPP de 69%, en utilisant un algorithme de recueil s'appuyant sur les codes de la CIM-10.
   Au niveau international, il existe une enquête [Coloma *et al.*, 2013] sur la validation d'algorithmes de repérage de l'IM dans les bases de données électroniques, dans divers pays utilisant des techniques de codage différentes. Le recueil des IM par différentes bases de données (hospitalières et des bases de praticiens généralistes) pour l'Italie, le Danemark et les Pays Bas, à l'aide des codes CIM-9 et CIM-10 est évalué avec des VPP supérieures à 90%. De plus, cet article rapporte que les différentes bases de données peuvent être complémentaires et offrent diverses perspectives. En effet, les bases hospitalières contiennent l'information liée à l'événement et aux soins pratiqués ne pouvant se faire que dans le cadre d'un établissement de soins, comme par exemple une pose d'endoprothèse, ou une dilatation intraluminale. Au contraire, les bases des généralistes apportent une information sur le suivi du patient. Exploiter ces données pourrait être une perspective d'étude complémentaire.

3.4 Conclusion

Dans ce chapitre, nous avons constitué notre base de données à partir de laquelle nous allons explorer les trajectoires de patients dans les chapitres suivants. Au préalable, nous avons mené des analyses descriptives de sorte à quantifier et à caractériser cette problématique de santé publique que constitue l'IM.

Les points clés mis en évidence lors de ces analyses, en adéquation avec la littérature du domaine, sont les suivants : nous observons 1) une augmentation du taux d'IM chaque année ; 2) une augmentation alarmante des cas chez les femmes, plus précisément les femmes jeunes dont le risque surpasse celui des hommes ; 3) une régionalisation de l'IM avec un contraste Nord-Sud ; 4) un risque de récidive notoire dans les 3 mois après le 1er épisode ; 5) 90% des décès se produisent lors du 1er séjour.

Bien que les résultats soient à modérer du fait du manque d'exhaustivité des données, les bases médico-administratives peuvent servir de source de prospection, notamment dans le cadre de la surveillance de diverses pathologies car elles contiennent de nombreuses informations. Cependant, dans l'exploration de certains domaines, elles restent limitées comme nous avons pu le voir par les modèles de risque de réadmission. Dans ce cas de figure, il serait plus opportun de compléter les informations concernant les patients à l'aide des données du SNIIRAM par exemple.

Par ailleurs, des investigations pourraient être menées afin de mieux caractériser et comprendre les raisons de la hausse observée des hospitalisations pour IM, mais aussi d'expliquer les différences territoriales observées en termes de taux d'hospitalisation, de réadmission mais également de décès. De surcroît, des enquêtes de santé pourraient être conduites afin de mesurer la proportion de personnes ayant succombé avant l'intervention des secours ou ayant fait un IM silencieux et déterminer les éventuels risques associés à la survenue de l'infarctus. Selon [Valensi *et al.*, 2011], ces patients sont sujets aux mêmes facteurs de risque que les autres, d'autres analyses permettraient de réactualiser ces informations. Enfin, des études devraient être menées pour évaluer les modalités de suivi des patients afin de prévenir les récidives, en particulier auprès des personnes à risque telles que les femmes.

Nous avons maintenant à disposition une base de données nous permettant de mener des analyses sur les trajectoires de patients. De plus, ces analyses nous permettent de mieux appréhender les données et surtout de vérifier la cohérence des résultats avec ceux que nous obtiendrons lors de l'extraction de motifs dans les chapitres 5 et 7. Dans la suite de cette thèse, après avoir défini le concept de trajectoire (partie II) à travers l'étude de la littérature du domaine, nous allons étudier les trajectoires de patients dans un but de prédiction (partie III) et dans un but de planification sanitaire (partie IV). Dans le chapitre 6, nous proposons une méthode pour prédire le décès qui pourrait également être utilisée dans la prédiction de la récidive. Dans le chapitre 8, nous proposons une méthode pour mettre en lumière des profils de délais de ré-hospitalisation qui pourrait servir à caractériser plus spécifiquement le cas de la récidive.

Partie II

Les trajectoires dans la littérature

La seule chose que je sais, c'est que je ne sais rien.

Socrate.

Table des matières

4 Les trajectoires dans la littérature **61**
4.1 Analyse semi-automatique de la littérature 62
4.1.1 Questions de recherche 63
4.1.2 Processus de revue semi-automatique 63
4.2 Expérimentations 66
4.2.1 Étape 2. Première approche par fouille de textes 66
4.2.2 Étape 3. Analyse manuelle des articles sélectionnés 71
4.3 Discussion 73
4.3.1 Réponses aux questions de recherche 73
4.3.2 Avantages et limites 76
4.4 Conclusion 77

Chapitre

4 Les trajectoires dans la littérature

L'étude des trajectoires hospitalières des patients est un sujet émergent récent dans la littérature, englobant des concepts généraux. Notre recherche a porté sur les trajectoires des patients. Elle est plus précisément axée sur la gestion et les soins de la maladie, tout en tenant compte des aspects médico-économiques de la prise en charge. Nous avons abordé la trajectoire du patient au travers d'un exemple : la survenue d'un IM. Puisque le traitement de l'IM est effectué dans un établissement de santé, nous avons pu tracer les trajectoires des patients par le biais du système national de financement hospitalier, en utilisant des bases de données ou des registres complets des hôpitaux, collectés régulièrement à des fins de facturation.

Dans ce chapitre, nous étudions comment le concept de trajectoire est défini et étudié. Nous avons mené une revue de la littérature sur PubMed à l'aide de mots-clés liés à la trajectoire, aux concepts PMSI et IM. Nous procédons ensuite en deux étapes : 1) une recherche automatique sans *a priori* avec des techniques de fouille de textes ; et 2) une analyse plus classique d'une sous-sélection de documents. De telles revues systématiques de la littérature [Van Hecke *et al.*, 2015] ont été réalisées avant, mais sans utiliser de procédures automatiques. Cependant, l'intérêt de cette phase automatique est de traiter un grand nombre de documents. L'exploration automatique des textes permet de mieux cibler les articles d'intérêt et de réduire le temps de recherche [Cohen *et al.*, 2006], tout en permettant aux utilisateurs de se focaliser sur un ensemble d'intérêt.

Dans la section 4.1, nous présentons la stratégie de revue notamment les questions de recherche et les outils utilisés pour sa mise en œuvre. Puis, nous détaillons les résultats dans la section 4.2. Nous répondons aux questions de recherche dans la section 4.3. Pour conclure, nous discutons les avantages et les limites de ce travail dans la section 4.3.2.

4.1 Analyse semi-automatique de la littérature

Actuellement, les chercheurs en santé explorent la littérature manuellement et utilisent des méthodes statistiques ou des modèles nécessitant une simplification *a priori* extrême des processus. Cependant, il existe de nombreuses techniques de fouille de textes développées pour la recherche documentaire et la revue systématique de la littérature [Thomas *et al.*, 2011]. Dans les revues systématiques, les techniques de fouille de textes sont essentiellement utilisées pour :

— La reconnaissance automatique des termes pour identifier et extraire les termes automatiquement des textes [Frantzi *et al.*, 2000] ;

— La classification des documents en générant des sous-groupes de documents axés sur un sujet spécifique [Frunza *et al.*, 2011, Joachims, 1998, Sebastiani, 2002, Mo *et al.*, 2015] ;

— La classification de documents pour regrouper les documents par thème. Ceux-ci correspondent à des thèmes partagés par tous les documents du groupe et par aucun autre document de la collection [Reinert, 1983, Blei *et al.*, 2003, Bada, 2014] ;

— La rédaction des résumés en sélectionnant des phrases de chaque document en fonction de l'importance de ses termes, qui sont combinés par des techniques de classification [Bollegala *et al.*, 2010].

Toutefois, certains auteurs ont également utilisé l'exploration de textes pour d'autres finalités. Par exemple, [Lin *et al.*, 2008] ont créé des bases de données en reliant les auteurs aux abréviations de leurs noms et ont procédé à une analyse de co-auteurs. [Leitner et Valencia, 2008] ont annoté les résumés de deux façons : d'abord le gène ou la protéine d'intérêt, puis les interactions protéines et/ou les fonctions génétiques. Puis, ils ont catégorisé les documents selon ces annotations.

Ces exemples montrent que l'utilisation de méthodes de fouille de textes pour la réalisation de revues systématiques de la littérature est un sujet d'actualité [O'Mara-Eves *et al.*, 2015, Paynter *et al.*, 2016, Jonnalagadda *et al.*, 2015, Lefebvre *et al.*, 2013]. Par ailleurs, les méthodes d'exploration de données telles que la fouille de textes aboutissent à l'interprétation et l'exploration des processus sans *a priori* sur les connaissances. Nous avons donc intégré certaines de ces méthodes dans le processus classique de revue de la littérature afin d'explorer un grand volume de documents.

Nous commençons par introduire les différentes questions de recherche qui ont guidé notre revue dans la section 4.1.1. Ensuite, nous expliquons, dans la section 4.1.2, les différentes étapes de notre approche.

4.1.1 Questions de recherche

Nous avons formulé des questions pratiques, contenues dans le tableau 4.1, pour guider le processus d'analyse. Nous avons identifié sept types de questions sans *a priori*, exprimées en termes généraux qui intègrent des questions thématiques axées sur la médecine et satisfont à la fois les aspects scientifiques et médicaux pour l'expertise des professionnels de la santé. Nous avons également identifié sept autres questions spécifiques, avec *a priori*, nécessitant une analyse experte.

Tableau 4.1 – Questions de recherche.

	Questions sans *a priori*
Q1	Existe-t-il des études sur les trajectoires de patients ?
Q2	Quels sont les thèmes abordés dans ces études ? (la prise en charge, le traitement, les coûts...)
Q3	Pour quelles pathologies sont étudiées les trajectoires ?
Q4	Utilise-t-on le PMSI pour la recherche ?
Q5	Utilise-t-on le PMSI dans l'étude des trajectoires ?
Q6	Y a-t-il des études sur les trajectoires de patients ayant présenté un IM ?
Q7	Qu'étudie-t-on lien avec l'IM ?
	Questions avec *a priori*
Q8	Quels sont les différents concepts de la trajectoire ? (Comment est définie cette notion ?)
Q9	Quel est l'intérêt pour le sujet : y a-t-il beaucoup d'études sur les trajectoires de patients ?
Q10	Quels sont les pays menant des études sur les trajectoires ?
Q11	Quels sont les objectifs des études concernant les trajectoires de patients ?
Q12	Quelles sont les méthodes utilisées dans les études de trajectoires de patients ?
Q13	Quelles sont les caractéristiques de ces études : nombre de patients impliqués, temps d'observation ?
Q14	Quelles données sont utilisées dans ces études : hospitalières ou autres ?

4.1.2 Processus de revue semi-automatique

Notre méthode a été structurée en deux étapes automatiques suivie d'une analyse manuelle des articles sélectionnés. Ces deux étapes sont basées sur la récupération de documents et les techniques de fouille de textes. L'articulation des différentes étapes est schématisée dans la figure 4.1.

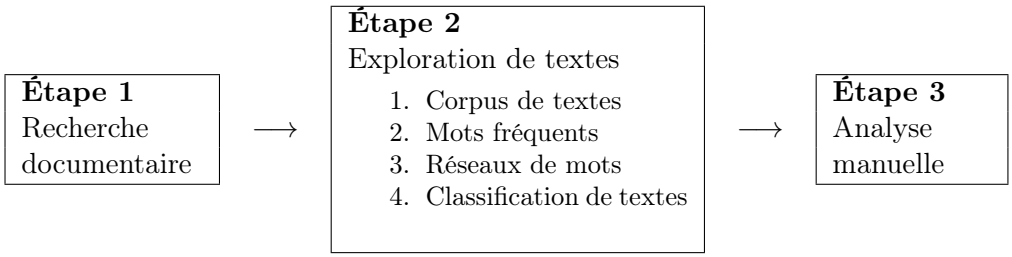


Figure 4.1 – Étapes du processus de revue semi-automatique.

Étape 1. Recherche documentaire. Nous avons mené des recherches, dans PubMed, selon les thèmes et contraintes résumés dans le tableau 4.2. Par exemple, la requête : C1 + T1 + C2 + C3, sélectionne les articles du domaine médical qui traitent des trajectoires, écrits entre le 1er janvier 2000 et le 31 octobre 2015, en anglais.

Tableau 4.2 – Mots clés utilisés pour la recherche documentaire.

Thèmes et contraintes	Mots clés
C1 : contexte médical	« health », « patient(s) »
T1 : Trajectoire	« trajectories », « trajectory », « path », « pathway(s) »
T2 : PMSI	« prospective payment system »*, « PMSI », « DRG », « ICD »**, « regional information system », « fee for service system », « registry », « Activity-based Payment »
T3 : IM	« myocardial infarction » in the title
C2 : dates	January 1st 2000 to October 31th 2015
C3 : langue	English

*est l'équivalent anglais pour désigner le PMSI.
**International Classification Diseases.

Étape 2. Première approche par fouille de textes. Cette étape se décompose en plusieurs sous-étapes :

1. À l'issue de l'étape 1, nous avons créé un corpus de textes, divisé en trois parties, T1 à T3 (correspondant aux thèmes du tableau 4.2), composé du titre et du résumé, dans lequel nous avons supprimé les mots clés (voir tableau 4.2), afin de ne conserver que les autres termes ;
2. Nous avons appliqué les prétraitements suivants :
   a) Lemmatisation des textes ;
   b) Enrichissement du dictionnaire : nous avons lemmatisé des termes non reconnus par TreeTagger [1] et ajouté des termes médicaux spécifiques et des acronymes bien connus tels que AMI [2] (Acute Myocardial Infarction). Par la suite, les analyses ont été menées avec les formes complètes (noms, adjectifs, adverbes et verbes) ;
3. Les trois parties du corpus ont été analysées séparément avec le logiciel IRaMuteQ [3]. Il s'agit d'une interface de R pour l'analyse multidimensionnelle de textes et de questionnaires [Ratinaud et Déjean, 2009], permettant de faire des analyses statistiques sur des corpus de textes [Ratinaud et Marchand, 2012]. Nous avons effectué des analyses textuelles classiques, puis des analyses de similitude et enfin des classifications à l'aide des outils suivants :

   Nuage de mots. Il s'agit d'une représentation synthétique de la distribution des termes : les mots les plus récurrents sont au centre avec une taille de texte proportionnelle au nombre d'occurrences. Ainsi, ce type de représentation symbolise, par ordre d'importance, les concepts couverts dans tous les articles. Cette méthode fournira une réponse à la question Q1.

1. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
2. La recherche se fait sur des textes en anglais.
3. http://www.iramuteq.org/

Analyse de similitude. C'est une technique, reposant sur la théorie des graphes, classiquement utilisée pour décrire des représentations sociales, sur la base de questionnaires d'enquête [Flament, 1981]. L'analyse de similitude est appliquée pour étudier la proximité et les relations entre les éléments dans un ensemble sous la forme d'arbres maximaux. L'objectif est de réduire le nombre de liens entre deux éléments, pour obtenir un graphique connecté acyclique. L'arborescence maximale est donc l'arbre créé par les arêtes les plus fortes du graphique, où la force est mesurée par la co-occurrence des termes liés. Pour chaque corpus, nous avons sélectionné la représentation décrite dans [Fruchterman et Reingold, 1991] et l'algorithme de [Brandes, 2001], pour décrire les communautés via le chemin le plus court, mettant ainsi en évidence les mots les plus fréquemment associés dans la même phrase ou texte. Le graphique génère une idée plus précise du contenu des articles concernant les concepts et les thèmes abordés en liant des termes importants. Cette méthode fournira des réponses aux questions Q3, Q5 et Q7.

Classification de textes. La classification de Reinert [Reinert, 1983] est une classification hiérarchique divisive s'effectuant en plusieurs étapes, offrant ainsi une approche globale du corpus. Après partitionnement de celui-ci, elle identifie des classes statistiquement indépendantes de mots, qui sont caractérisés par des mots spécifiques corrélés entre eux. Ce type d'analyse nous apporte une vision complémentaire de l'analyse de similitude en regroupant des articles selon des concepts, en partie identifiés par une analyse de similitude, caractérisée par des groupes de mots. Cette méthode complètera la réponse à la question Q7 et répondra aux questions Q2, Q4 et Q6.

Étape 3. Analyse manuelle des articles sélectionnés. Nous avons utilisé la sous-sélection précédente et avons croisé les thèmes : T1 et T2, noté $T1 \cap T2$, puis T1 et T3, noté $T1 \cap T3$. Cette sélection a été effectuée de la même manière que celle décrite dans le tableau 4.2. Nous avons ajouté une contrainte supplémentaire pour mieux cibler notre étude en comptant le nombre d'occurrences K du concept de trajectoire dans chaque document et en sélectionnant ceux pour lesquels : $K \geq 2$. De façon pratique, nous avons compté le nombre d'occurrences des mots « trajectories », « trajectory » ou « pathway » dans les titres et les résumés des articles.
Nous avons élaboré notre grille de lecture en nous basant sur celle décrite dans PRISMA [4] (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [Moher, 2009]. Nous avons retenu les items permettant de répondre aux questions de recherche avec *a priori* (voir tableau 4.1) : année de publication, pays d'étude, nombre de patients, période d'observation, méthodes et objectifs. Les autres items, n'étant pas pertinents pour notre étude, n'ont pas été conservés. Nous avons ajouté trois items spécifiques à notre problématique : les pathologies étudiées, les bases de données utilisées et la définition du concept de trajectoire.

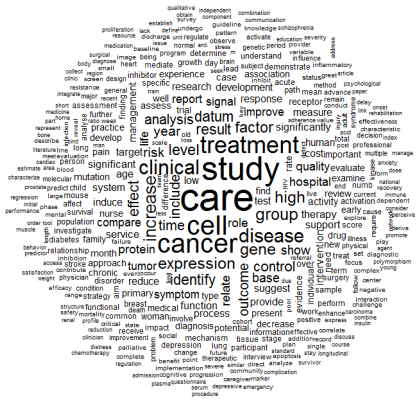
4. www.prisma-statement.org

4.2 Expérimentations

À l'issue de l'étape 1, nous avons collecté un total de 33 514 articles. Dans la section 4.2.1, nous détaillons les résultats de l'exploration des textes essentiellement pour le thème T1. Les autres résultats sont disponibles à l'adresse suivante : http://www.lirmm.fr/~pinaire/survey.html. Dans la section 4.2.2, nous faisons une synthèse de l'analyse manuelle des articles.

4.2.1 Étape 2. Première approche par fouille de textes

Nuage de mots. Pour chaque corpus, nous avons représenté les 400 premières formes (voir figure 4.2). Pour T1, les termes les plus saillants sont « care », « study », « cancer », « cell », « treatment », « increase ». Pour T2, ce sont les termes « study », « registry », « datum », « cancer ». Pour T3, ce sont les termes « AMI », « acute », « hospital ».



T1

T2

T3

Figure 4.2 – Nuages de mots : Trajectoire, PMSI, IM.

Analyse de similitude. Dans un arbre maximum, seules les arêtes les plus fortes du graphe sont conservées. Une arête symbolise la co-occurrence entre les termes, et son épaisseur représente l'importance du nombre de co-occurrences. Par exemple, dans la figure 4.3, le lien entre « care » et « study » est plus important que le lien entre « significant » et « difference ». Pour T1, la figure 4.3 comprend trois parties : une partie inférieure avec un grand réseau caractérisé par « care », puis un petit réseau contigu qui regroupe les termes « clinical » et « outcome ». Sur la partie supérieure droite, il y a un réseau englobant les termes liés à la génétique, avec « cell », « expression » et « gene », puis un réseau attenant plus petit regroupant les termes « increase », « high » et « significantly ». La partie supérieure centrale contient les termes « cancer », « diagnostic » et « treatment ». Enfin, la partie supérieure gauche comporte un grand réseau contenant « study » auquel sont rattachés plusieurs grappes plus petites, caractérisées par les termes « risk », « disease », « time », « year » et enfin « disease ». Les communautés de mots les plus étroitement liées sont « genetics » avec « cancer », « cancer » avec « study » et « study » avec « care ».

Classification de textes. À la suite de cette classification, 80% des articles de T1 ont été distribués dans onze classes disjointes, 86% pour T2 en cinq classes et 98% pour T3 dans cinq classes. Pour T1, la figure 4.4 montre, de gauche à droite, deux classes regroupant les concepts d'organisation génétique (classe 5), l'organisation du signal et la médiation cellulaire (classe 10). La classe 8 rassemble des concepts liés à la réponse du système immunitaire dans un processus inflammatoire. La classe 1, regroupe les dysfonctionnements liés au diabète et les conséquences. Les classes 2 et 7 symbolisent respectivement le temps dans l'organisation des séjours à l'hôpital et le temps dans la trajectoire. La classe 6 concerne les questionnaires et les échelles psychométriques avec dépression. La classe 11 contient des concepts liés à l'imagerie médicale. Dans la dernière branche, la classe 9, décrit la médecine dans ses aspects financiers et réglementaires. La classe 4 concerne la gestion des patients, y compris les pratiques. La classe 3 regroupe les termes relatifs à la façon dont les informations sont transmises.
Pour ces deux dernières classes, nous avons procédé à un zoom, en étudiant l'arbre des similitudes. Elles réunissent 1 645 articles à elles deux. Pour la classe 3, nous distinguons trois nœuds (voir la figure 4.5), celui de « care », qui est étroitement lié à celui de « study », qui est à son tour étroitement lié à celui du « cancer ». Pour la classe 4 (voir la figure 4.5), il n'y a qu'un seul nœud représenté par « care », à partir duquel partent plusieurs branches pour « research » et « process », puis plus haut un sous-nœud pour « clinic », relié à « trial », « datum », « bases » et « identify ».
Dans la suite, nous avons effectué une deuxième classification sur les 3 160 articles non classés lors de la première analyse, nous identifions cinq classes constituées de 99% des articles. De droite à gauche, la classe 1 regroupe les concepts méthodologiques. La classe 5 traite de la fin de vie. La classe 4 réunit le côté macroscopique des soins avec la prise en charge publique. La classe 3, traite des études à partir d'expériences faites sur des animaux. Et enfin la classe 2, évoque les mutations génétiques et les anomalies. Trois articles n'ont pas pu être classés.

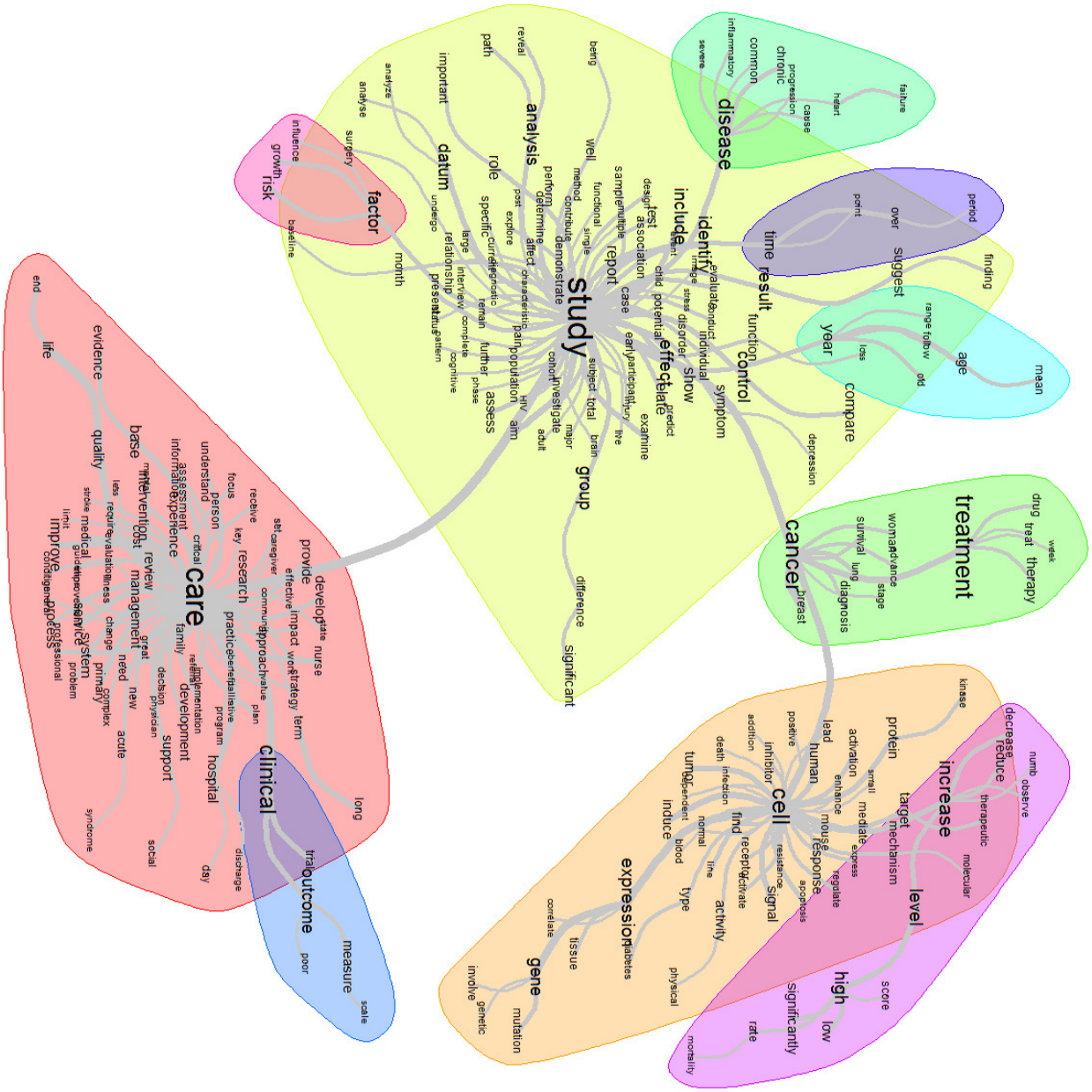


Figure 4.3 – Analyse de similitude et représentation des communautés pour le thème T1.

Figure 4.4 – Résultats d'une classification des articles pour le corpus T1 : Trajectoire.

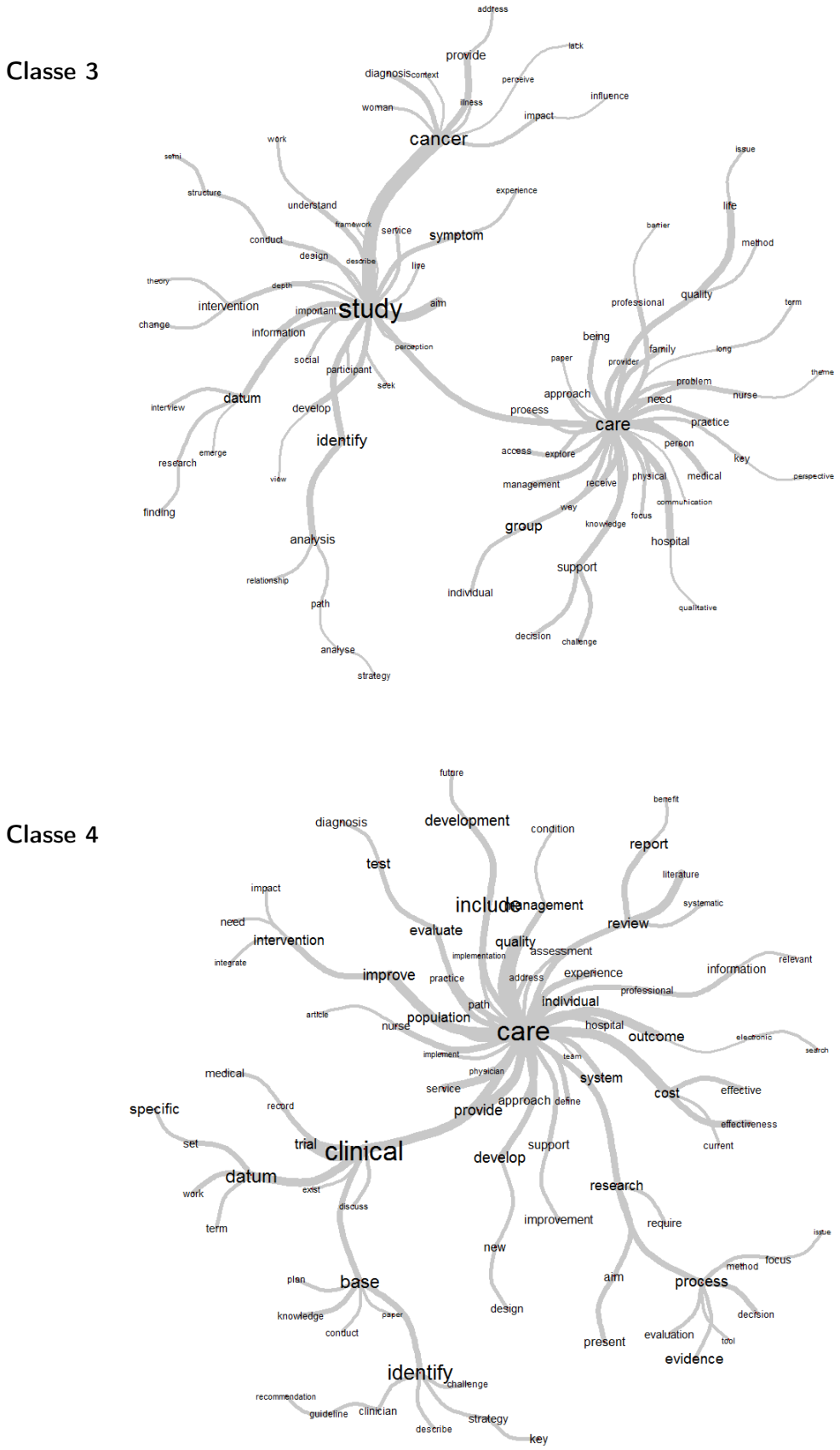


Figure 4.5 – Analyse de similitude sur les classes 3 et 4 issues de la première classification de T1.

4.2.2 Étape 3. Analyse manuelle des articles sélectionnés

Grâce au processus de filtration précédent, nous avons généré une sous-sélection de 84 articles, dont 53 pour $T1 \cap T2$ et 31 pour $T1 \cap T3$. Nous avons éliminé 8 articles pour $T1 \cap T2$ et 6 pour $T1 \cap T3$ car ces articles ne traitaient pas de la trajectoire du patient au sens qui nous intéresse. Par exemple [Suresh *et al.*, 2014] traite de la trajectoire d'un gène et [Tada *et al.*, 2015] traite de celle d'une protéine. Pour la plupart des items (objectif de l'étude, pathologies étudiées et définition du concept de trajectoire), nous avons créé des catégories, détaillées en annexe dans le tableau B.1 et les sources des références associées sont répertoriées dans le tableau B.2. Dans la suite de cette section, nous synthétisons les résultats des autres items retenus dans notre grille de lecture :

Bases de données et méthodes utilisées. Généralement, les auteurs ont utilisé plusieurs sources et méthodes dans leurs études. Les résultats concernant ces items sont représentés dans la figure 4.6 ;

Pays d'étude. Nous avons regroupé les pays par continent. Pour $T1 \cap T2$, nous avons noté une forte représentation de l'Europe (55% des articles) et des Amériques (29%). Nous dénombrons quelques études pour l'Océanie (9%) et pour l'Asie (7%). Pour $T1 \cap T3$, la distribution des articles se répartie essentiellement entre trois continents : Europe (36%), Amériques (28%) et Asie (24%). L'Australasie était marginale, avec 4% d'articles. Nous avons trouvé des études atypiques avec des données provenant de continents multiples (8%) ;

Année de publication. Pour $T1 \cap T2$, les résultats ont mis en évidence une activité qui a commencé à se développer en 2013. Alors que, pour $T1 \cap T3$, nous avons noté un pic d'activité en 2004 et une activité en augmentation en 2012 ;

Nombre de patients. Ce nombre varie de 14 à 6,2 millions pour $T1 \cap T2$ (resp. 20 à 30,20 millions pour $T1 \cap T3$), avec une médiane (Med) de 859 et un intervalle interquartile (IQ) de 3 250 (resp. Med $= 604,5$ et IQ $= 933,25$ pour $T1 \cap T3$), avec des données manquantes pour trois articles (resp. cinq pour $T1 \cap T3$) ;

Durée d'observation. La durée varie de 5 à 180 mois dans $T1 \cap T2$ (resp. 3 à 240 mois dans $T1 \cap T3$) avec une médiane de 36 mois et un IQ de 54 mois (resp. Med $= 12$ et IQ $= 99$ mois dans $T1 \cap T3$).

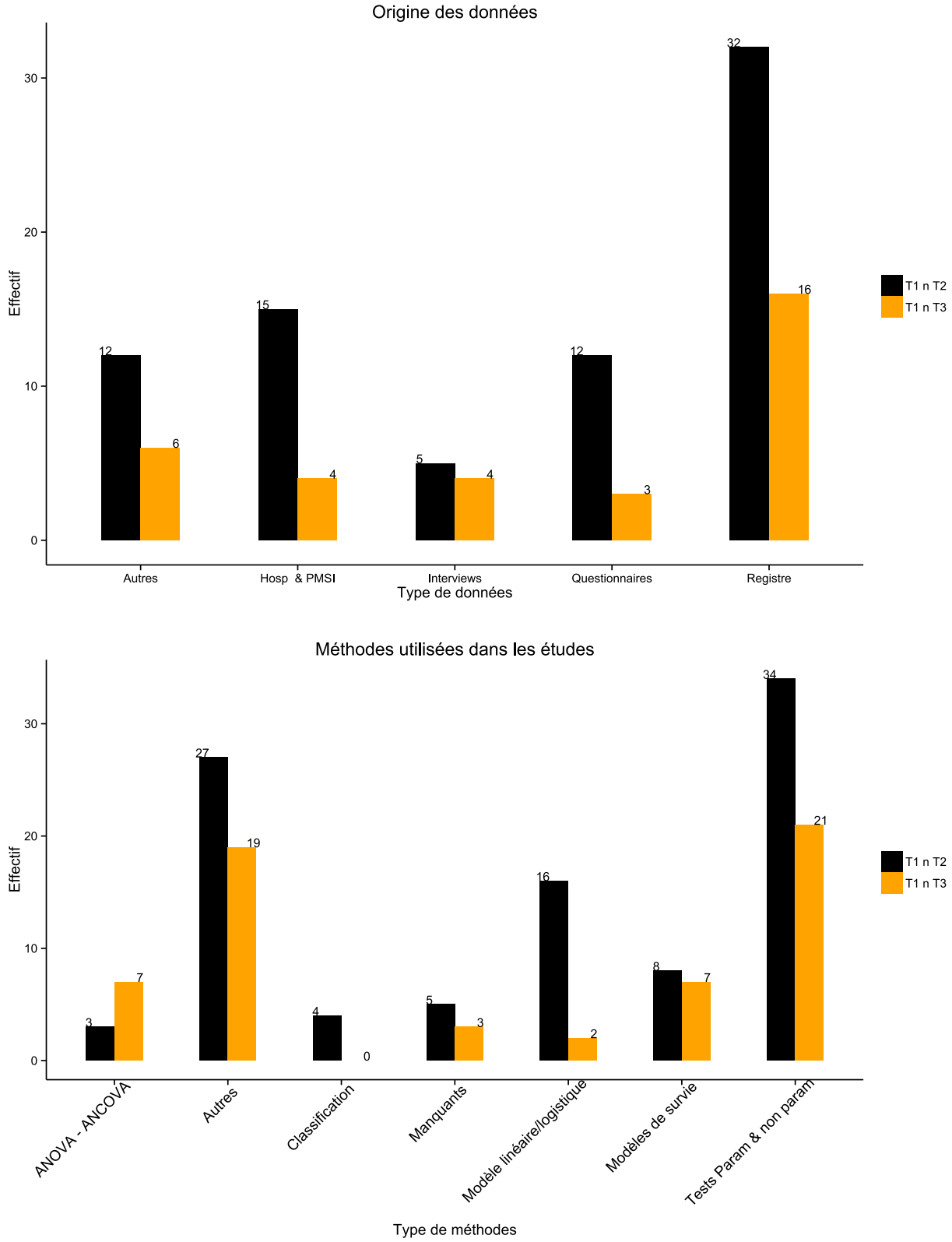


Figure 4.6 – Base de données et méthodes utilisées dans les études de trajectoires.

4.3 Discussion

Dans la suite de cette section, nous répondons aux questions de recherche (tableau 4.1) en nous appuyant sur les résultats de l'exploration de textes puis, sur ceux de l'analyse manuelle des articles. Ensuite, nous discutons des avantages et des limites des différentes étapes de notre processus de revue.

4.3.1 Réponses aux questions de recherche

Q1 : Existe-t-il des études sur les trajectoires de patients ? Dans les nuages de mots, l'apparition des termes « study » et « care » pour tous les domaines étudiés signifie que ces articles recouvrent le concept de soins et les études sur des sujets tels que les maladies ou les médicaments. Pour T1, les termes « treatment » et « increase » reflètent l'accent mis sur les trajectoires de soins des patients. Ainsi, il existe de nombreuses études sur les trajectoires des patients.

Q2 : Quels sont les thèmes abordés dans ces études ? Une synthèse des résultats de la classification de textes nous permet de répondre à la question. Le premier sujet abordé est la maladie avec, par exemple, les troubles métaboliques tels que le diabète et les complications cardiovasculaires. Certains articles ont abordé les sentiments, les angoisses et l'expérience de la maladie par les patients. Le soutien par l'environnement proche et la famille du patient a été abordé, ainsi que les dispositifs mis en place comme l'intervention d'une infirmière à domicile. Des articles abordent la fin de vie, les soins palliatifs et les processus mis en place pour gérer cette dernière étape de la maladie. Un autre sujet est la recherche clinique, impliquant le développement de cohortes, la collecte de données et les méthodes utilisées dans différentes études. Quelques études se sont concentrées sur l'organisation des hôpitaux, des services, des personnels de soins et les coûts associés. D'autres articles ont porté sur les règlements sanitaires et les recommandations de guides de bonnes pratiques.

Q3 : Pour quelles pathologies sont étudiées les trajectoires ? Dans l'analyse de la similitude, les résultats ont montré pour T1 que, les études étaient étroitement liées aux soins, à la maladie et plus spécifiquement au cancer. Les maladies étudiées sont celles qui causent des dysfonctionnements organiques (le cœur, les reins ou les poumons) sévères et chroniques. Pour T2, le cancer a été principalement étudié à partir des données du registre qui décrit l'histoire de la maladie depuis le diagnostic. Notons que nous avons retrouvé ces résultats dans l'étape 3 (voir item pathologie du tableau B.1).

Q4 : Utilise-t-on le PMSI pour la recherche ? Les deux classifications de textes de T2 nous permettent de conclure que le PMSI est utilisé dans la recherche principalement pour l'étude des maladies, parfois sur la survenue de la maladie, mais surtout dans sa prise en charge, ses coûts associés, son traitement et les complications possibles mais aussi dans son codage. Les maladies les plus étudiées via le PMSI sont les troubles neurologiques, le cancer, le rythme cardiaque irrégulier et les maladies

cardiovasculaires, les dispositifs médicaux implantables pour réguler ces anomalies, traumatismes et plaies, maladies infectieuses, transplantations d'organes, maladies génétiques et auto-immunes et enfin insuffisance rénale. La grossesse et la naissance sont également beaucoup étudiées via le PMSI.

Q5 : Utilise-t-on le PMSI dans l'étude des trajectoires ? On peut déduire de la question précédente que T2 est utilisé dans l'étude des trajectoires de patients : dans les études de maladie [Guldbrandt *et al.*, 2015, Harlos *et al.*, 2015, Jensen *et al.*, 2014, Danielsson *et al.*, 2009, Jiwa *et al.*, 2010, Schwartz *et al.*, 2013, Sieberg *et al.*, 2013, Gebregziabher *et al.*, 2010], pour comparer les soins et le codage [Palmer *et al.*, 2013b, Aeyels *et al.*, 2016, Kesavan *et al.*, 2013] mais aussi dans le suivi des patients au fil du temps et l'évolution de la survie [Harlos *et al.*, 2015, Biffl *et al.*, 2001, Diaz *et al.*, 2008, Myers *et al.*, 2014, Wang *et al.*, 2013]. L'évolution de la survie fait partie du concept de trajectoire. Ce concept de trajectoire peut également englober la notion de registre. Il s'agit là d'un concept renfermant de nombreux concepts liés à la longitudinalité [5].

Q6 : Y a-t-il des études sur les trajectoires de patients ayant présenté un IM ? Dans la classification de textes, les résultats de T3 ont montré que l'IM est étudié selon plusieurs aspects : les facteurs de risque (socio-économique, âge, hypertension, diabète), les aspects fonctionnels (biochimiques et cardio-circulatoires, mécanismes conduisant à l'IM et prédisposition génétique [Ginzburg *et al.*, 2003]), les aspects psychologiques, la prise en charge aux urgences avant l'admission à l'hôpital, y compris les transports et les premiers secours. Ensuite, il y a les soins à l'admission, la prise en charge médicamenteuse [Wang *et al.*, 2013] et les coûts associés - ici le concept de trajectoire émerge. Il y a également un aspect concernant l'efficacité des mesures mises en œuvre [Kinsman *et al.*, 2012, Kristoffersen *et al.*, 2015, Bestul *et al.*, 2004, Kucenic *et al.*, 2000, Mazzini *et al.*, 2008, Pelliccia *et al.*, 2004, Smith *et al.*, 2011] et des différents traitements [Hagiwara *et al.*, 2014, O'Donnell *et al.*, 2006, Lewis *et al.*, 2014, Rankin *et al.*, 2002]. Un autre aspect étudié est le style de vie (habitudes alimentaires, hygiène de vie [Rankin *et al.*, 2002], comorbidités (tabagisme et/ou alcool) [Dharmarajan *et al.*, 2015, Myers *et al.*, 2014]) et les facteurs environnementaux comme la pollution atmosphérique.

Q7 : Qu'étudie-t-on en lien avec l'IM ? Nous avons en partie répondu à cette question dans la question précédente. Le graphique issu de l'analyse de similitude pour T3 permet de compléter cette réponse. Il met en évidence deux points de vue concernant l'IM : 1) celui du clinicien qui étudie l'IM, ses risques et les facteurs aggravants pour mieux comprendre, prévenir et, le cas échéant, prendre en charge ces patients ; 2) celui du patient atteint de symptômes coronariens, qui pourraient évoluer vers des incidents, un IM aigu exigeant une hospitalisation et encourant un risque élevé de mortalité suivant son âge.

5. Arrêté du 6 novembre 1995 relatif au Comité national des registres.

Q8 : Quels sont les différents concepts de la trajectoire ? Les résultats de l'analyse manuelle ont montré que dans la plupart des cas, la trajectoire se caractérise par des processus de soins établis pour une maladie spécifique afin d'améliorer la prise en charge des malades, faciliter la planification sanitaire au sein des établissements, faire de la prévention, anticiper l'évolution de la pathologie et prévenir l'apparition des symptômes.

Q9 : Quel est l'intérêt pour le sujet : y a-t-il beaucoup d'études sur les trajectoires de patients ? Nous avons constaté que l'intérêt pour les études de trajectoire des patients a augmenté au cours des 5 dernières années notamment en 2013. Cela s'explique par l'amélioration de la qualité des bases de données à partir de 2009 [Le Bihan-Benjamin *et al.*, 2013], notamment en France, et par la possibilité de chaîner les séjours hospitaliers et de reconstituer le parcours de soins d'un patient sur tout le territoire.

Q10 : Quels sont les pays menant des études sur les trajectoires ? Cet intérêt pour les trajectoires provient principalement d'Europe et des Amériques avec respectivement 47% et 29% d'études. Néanmoins, le concept de PMSI conduit à inclure seulement les pays ayant une organisation similaire des systèmes de base de données santé. Il s'agit là d'une limite de notre étude, puisque les pays ayant un système d'information de santé différent du modèle américain n'ont pas été sélectionnés via ce filtre.

Q11 : Quels sont les objectifs des études concernant les trajectoires de patients ? La répartition des articles dans les six catégories que nous avons définies (voir tableau B.1) a montré que l'objectif de la plupart des études était de comparer les traitements, les techniques ou les procédures de soins. Dans chaque cas, l'objectif était de réduire les coûts tout en améliorant la qualité des soins. L'étude de la trajectoire du patient apporte un double bénéfice : 1) la trajectoire permet de mieux appréhender l'évolution de la pathologie suivant les soins prodigués tant sur le plan médical que chirurgical ; 2) la trajectoire est informative sur le plan médico-économique dans la mesure où une potentialisation du suivi évite la dispersion des soins.

Q12 : Quelles sont les méthodes utilisées dans les études de trajectoires de patients ? Les méthodes utilisées viennent étayer l'argumentaire d'études comparatives sur ces techniques de soins et de traitements. Ces méthodes classiques (Anova, tests de comparaison, modèles de survie, régression linéaire ou logistique...) sont répertoriées dans la deuxième partie de la figure 4.6.

Q13 : Quelles sont les caractéristiques de ces études : nombre de patients impliqués, temps d'observation ? Dans les études, le nombre de patients recrutés a été estimé *a priori* de sorte à mener des analyses statistiques dans de bonnes conditions avec une puissance suffisante. Cependant, nous avons identifié quelques études menées sur l'ensemble de la population, sans échantillonnage. Dans l'ensemble, le temps d'étude est de courte durée entre 3 mois et 3 trois ans, ce qui

s'explique par des considérations économiques ou par un manque de données. En effet, pour des études rétrospectives, par exemple, il est parfois difficile de remonter plusieurs années en arrière car les informations ne sont pas conservées au-delà d'un certain laps de temps.

Q14 : Quelles données sont utilisées dans ces études : hospitalières ou autres ? Pour $T1 \cap T3$, les données utilisées sont des données de registre. Pour $T1 \cap T2$, les bases de données des hôpitaux et celles de la facturation ont été utilisées. Ceci corrobore le fait que les études sont en majorité des études hospitalières. De plus, en dehors des bases de données hospitalières, certaines études ont pris en compte les sentiments des patients à l'aide de questionnaires et/ou d'entretiens. Certaines études nécessitaient des informations supplémentaires sur, par exemple, les médicaments [Harlos *et al.*, 2015, Sundberg *et al.*, 2014] par le biais de bases de données pharmaceutiques ou de données de soins pratiqués en dehors du milieu hospitalier [Couchoud *et al.*, 2015, Bossuyt *et al.*, 2015] avec des bases de données de sécurité sociale pour un suivi complet du parcours des patients.

Notre étude nous a permis de répondre à toutes les questions initiales. Dans la suite, nous listons les avantages et les inconvénients de notre approche.

4.3.2 Avantages et limites

Recherche documentaire. Nous avons choisi de concentrer notre étude sur PubMed. Les résultats de la recherche dépendent entièrement du choix des mots-clés, ce qui en fait une tâche particulièrement délicate puisque les définitions varient d'un pays à l'autre. Par exemple, nous avons rencontré cette difficulté pour le thème T2. Comme le montre le tableau 4.2, les mots-clés utilisés sont « Prospective Payment System », « PMSI », « DRG », « ICD », « regional information system », « fee for service system », « registry », « Activity-based Payment ». Cependant, certains documents ont utilisé des mots qui ne sont pas dans notre sélection initiale, comme dans [Jay *et al.*, 2013], qui utilise le terme « national case-mix system » pour désigner les bases de données du PMSI. Toutefois, notre objectif n'était pas d'être exhaustif, mais plutôt de définir une méthode d'analyse générale. Une façon d'améliorer notre approche serait de mettre en place un algorithme adaptatif pour l'enrichissement des mots clés [Rusmevichientong et Williamson, 2006].

Choix des méthodes. Bien qu'il n'y ait pas de consensus sur une méthode préconisée pour les revues systématiques de la littérature avec un grand nombre de documents, plusieurs techniques de fouille de textes ont été utilisées dans divers domaines pour explorer les données textuelles [Lebart *et al.*, 1998, Van Eck et Waltman, 2011]. Notre objectif était de conduire une analyse des documents de la littérature dans le domaine des trajectoires de patients, afin de fournir des informations générales et de répondre aux questions de recherche. Pour cela, nous avons cherché à explorer un grand volume de documents et à mieux sélectionner les publications en créant des filtres. Avec notre méthode, les recherches sont effectuées en fonction de la signification des mots et des concepts émergeant des techniques de classification et non en fonction de la simple présence d'un terme et/ou d'un concept. Ainsi, nous

avons pu explorer les textes, en mettant en évidence des mots clés, souvent utilisés dans les résumés. La représentation du nuage de mots est la plus adaptée pour cette étape car elle apporte une lecture visuelle et rapide des résultats. Toutefois, au-delà de l'aspect visuel, elle n'apporte pas beaucoup d'informations.
Une façon d'obtenir plus d'informations sur ces articles, est de mettre en évidence un univers lexical attaché à ces mots-clés. Ainsi, un même mot peut être interprété différemment selon les termes qui lui sont associés. L'analyse de similitude répond le mieux à ce problème. Sa construction en arborescence relie des réseaux de termes hautement coexistants et permet une meilleure compréhension des thèmes les plus fréquemment discutés au travers des différents éléments composant chaque corpus.
La dernière étape du processus d'exploration a consisté à déterminer s'il était possible de classer ces articles dans les thèmes mis en évidence par une analyse de similitude. Nous avons comparé ces résultats en utilisant la classification de Reinert car elle a l'avantage de respecter la construction du texte. Elle offre également plus de flexibilité que, par exemple, l'allocation latente de Dirichlet (LDA) dans laquelle le chercheur doit pré-déterminer le nombre de classes, bien que certains auteurs aient proposé des solutions pour déterminer le nombre « optimal » de classes dans la modélisation de thèmes [Greene *et al.*, 2014, Zhao *et al.*, 2015].
In fine, les méthodes de fouille de textes que nous avons sélectionnées se sont révélées efficaces pour explorer les corpus sans *a priori* avec des questions ouvertes, ce qui nous a permis d'identifier rapidement les documents hors sujets traitant de la génétique. En effet, dans l'étude des graphiques issus de l'analyse de similitude, nous avons noté que le concept de cancer était également étroitement lié à celui de la génétique. Il s'avère que, l'utilisation du mot-clé « pathway » met en évidence tous les articles relatifs à la signalisation cellulaire ou aux voies génétiques [Burke *et al.*, 2015, Cresci *et al.*, 2011, Davis *et al.*, 2014, Park *et al.*, 2015, Pedersen *et al.*, 2015, Peters *et al.*, 2011, Suresh *et al.*, 2014, Nubukpo, 2014, Tada *et al.*, 2015, Zhang *et al.*, 2014]. Cela a facilité le filtrage des articles pour appliquer des méthodes avec *a priori* afin de répondre à des questions spécifiques.

4.4 Conclusion

Dans ce chapitre, nous avons appliqué une méthodologie semi-automatique d'exploration de textes pour étudier la trajectoire du patient à partir des articles publiés dans PubMed. Les techniques de fouille de textes, nous ont permis d'analyser de grandes quantités de données textuelles, ce qui n'aurait pas été possible autrement. L'originalité de notre démarche repose sur une approche sémantique pour assister une revue systématique. Cette méthode est adaptée pour des questions complexes ou des sujets dont le contour est difficile à définir tels que ceux abordés en santé publique et plus particulièrement dans le contexte de la littérature sur la trajectoire de patients. Enfin, notre stratégie nous a permis d'explorer le concept de trajectoire dans le domaine médical.

La recherche documentaire sur les trajectoires des patients a été combinée avec deux thèmes principaux : les bases de données du PMSI et l'IM. Les résultats ont montré que ce type d'étude est un sujet d'intérêt dans la communauté biomédicale : que ce soit pour en apprendre davantage sur l'évolution d'une pathologie grâce au suivi du patient, ou pour comparer les parcours de soins afin de mettre en place des stratégies par des procédures de soins facilitant le travail des personnels de santé tout en fournissant un cadre rassurant au patient et en réduisant les coûts. Ce gain d'intérêt est relativement récent, car les bases étaient peu exploitées jusqu'alors, en témoigne le peu d'articles sélectionnés lors de la deuxième partie de notre étude. Nous retenons de cette étude que le concept de trajectoire, quel que soit sa forme est ainsi exploré, analysé et exploité, en particulier en oncologie à travers le dossier médical de communication et les réunions multidisciplinaires.

Pour compléter cette recherche, il serait intéressant d'inclure des études sur les trajectoires de patients dans les dossiers de santé électroniques. Certaines études récentes ont porté sur l'utilisation de ces nouvelles technologies afin d'offrir aux patients ayant des problèmes de mobilité des soins intégrés en regroupant les dossiers électroniques des patients avec ceux des soignants, ou des équipes de soins de santé, ainsi que ceux concernant le suivi par les médecins [Skinner *et al.*, 2014]. Cependant, la mise en œuvre de tels processus exige une organisation considérable et des ressources adéquates [Dent et Tutt, 2014] et peut conduire à des problèmes techniques d'interopérabilité [Waterson *et al.*, 2012].

Dans la suite de cette thèse, nous allons apporter notre contribution aux travaux existants en proposant deux nouvelles approches pour extraire et interpréter les trajectoires de patients concernant l'IM. Dans la partie III, nous définirons le concept de trajectoire que nous étudierons tout au long de cette thèse. Nous appliquerons des techniques de fouille de données afin de mettre en évidence des motifs fréquents dans les parcours hospitaliers, dans un but de prédiction du décès. Enfin, dans la partie IV, nous étudierons les flux de patients à l'aide des motifs spatio-temporels dans un but de caractérisation des différents parcours hospitaliers (les schémas de flux) à des fins de planification sanitaire.

Partie III

De la trajectoire du patient à la prédiction

Un homme ne peut jamais avoir de connaissances certaines.

Hérodote, Histoires, 7, 50 - Ve s. av. J.-C.

Table des matières

5 Motifs contextuels **83**
5.1 Motifs séquentiels contextuels 84
5.1.1 Définitions préliminaires 85
5.1.2 Description du processus de fouille 89
5.2 Expérimentations 92
5.2.1 Motifs fréquents 93
5.2.2 Motifs discriminants 97
5.3 Discussion 98
5.4 Conclusion 100

6 Prédire le décès **103**
6.1 Protocole de prédiction 106
6.1.1 Étape 1. Tri des motifs 106
6.1.2 Étape 2. Constitution de la base de données 107
6.1.3 Étape 3. Modélisation par contexte 107
6.1.4 Étape 4. Validation externe 111
6.2 Expérimentations 112
6.2.1 Étape 3.e. Choix du modèle 113
6.2.2 Étape 5. Validation externe 114
6.2.3 Identification des parcours à risque 115
6.3 Discussion 117
6.3.1 Détermination du meilleur couple *(modèle, score)* 117
6.3.2 Identification des parcours à risque 118
6.3.3 Compétitivité de notre approche 119
6.4 Conclusion 120

Chapitre

5 Extraction de motifs séquentiels contextuels

La collecte des données hospitalières dans le cadre du PMSI génère sur le plan national des bases de données de l'ordre de 25 millions d'enregistrements (séjours) par an[1]. Ces données recueillies à des fins médico-économiques (voir chapitre 2), peuvent *a posteriori* servir à des fins d'analyse et de recherche, pour examiner des questions médicales et épidémiologiques (voir chapitre 2, section 2.4), en suivant le parcours hospitalier d'un patient grâce au chaînage des séjours permis par le numéro anonyme de patient (voir chapitre 2, section 2.3). Les enjeux associés à l'extraction de connaissances dans ces types de données sont importants : 1) au niveau de l'individu pour prédire sa trajectoire de soins [Egho *et al.*, 2013] ; 2) au niveau de la population pour prédire l'évolution et les coûts associés à la santé des populations (notamment dans le cas des maladies chroniques) [Jensen *et al.*, 2014].

Notre objectif est d'extraire des profils de parcours de soins fréquents pour l'IM qui prennent en compte des informations contextuelles fréquemment associées aux données séquentielles. Dans un premier temps, nous avons utilisé des motifs séquentiels [Srikant et Agrawal, 1996] pour identifier des évènements chronologiques fréquents associés à des pathologies. Par exemple, dans le cas des séquences d'actes réalisées pour des patients dans un hôpital, nous allons pouvoir extraire les informations suivantes : fréquemment un patient ayant eu un IM subit une *coronarographie*, puis la *pose d'un stent*. L'extraction classique de motifs séquentiels se focalise sur les séries d'actes sans considérer des informations relatives aux patients comme leur sexe, leur âge, leur poids, *etc.* Or, en considérant le fait qu'un motif séquentiel est spécifique à un contexte donné (*e.g.* les jeunes hommes), un professionnel de santé pourra adapter sa stratégie de soin au contexte du patient et prendre les décisions adéquates. Nous avons pris en compte ces informations en utilisant la méthode d'ex-

1. Guide méthodologique de production des résumés de séjour du PMSI en médecine, chirurgie et obstétrique (fascicule spécial 2004/2 bis du Bulletin officiel) : http://www.atih.sante.fr/textes-officiels-du-pmsi-en-mco

traction de motifs contextuels définie dans [Rabatel *et al.*, 2010a]. Nous verrons dans le chapitre 6 que cette approche nous a permis d'envisager l'application suivante : prédire le décès à l'hôpital.

Dans la section 5.1, nous présentons les définitions générales du domaine puis, nous décrivons le processus d'extraction de connaissances mis en place, en particulier la transformation de la base de données et la hiérarchie des contextes utilisées en entrée de l'algorithme de fouille de données. Nous définissons une trajectoire fréquente sous la forme d'un motif séquentiel contextuel et nous évoquons les techniques d'extraction de motifs fréquents ou discriminants. Les résultats sont décrits dans la section 5.2. Enfin, nous analysons les résultats obtenus dans la section 5.3 avant de conclure par des perspectives dans la section 5.4

5.1 Motifs séquentiels contextuels

La recherche de motifs séquentiels est une discipline de fouille de données, utilisée à des fins d'identification de séries d'évènements ordonnés dans une base de données. Elle a été introduite en 1995 par [Agrawal et Srikant, 1995], pour le commerce, dans le but de prédire les produits susceptibles d'intéresser un acheteur ayant le même comportement que d'autres. Elle a été assez rapidement appliquée en médecine, comme dans [Brossette *et al.*, 1998], pour créer un système de surveillance clinique des infections à partir des bases hospitalières. D'autres systèmes de surveillance ont vu le jour, comme dans [Batal *et al.*, 2011] où les auteurs ont généré des sous-ensembles de données pertinentes dont le pouvoir prédictif permet ensuite de prévenir une allergie à un médicament. Dans le cadre de la pharmacovigilance, de nombreux travaux ont émergé ces dernières années, tels que [Norén *et al.*, 2008], pour détecter, à partir des motifs temporels, des effets indésirables, qu'ils soient éphémères ou persistants, en fonction de la maladie et des traitements administrés aux patients. Dans [Nikfarjam et Gonzalez, 2011], les auteurs ont enrichi la connaissance sur les médicaments, plus précisément sur leurs effets secondaires, à partir des réseaux sociaux en extrayant les motifs associant les termes dans un langage familier aux effets négatifs des médicaments. [Wright *et al.*, 2015] proposent un outil de prédiction des prescriptions chez le patient diabétique, ce qui ouvre des perspectives pour d'autres types de pathologies et des applications au sein des établissements de soins. Ces quelques exemples montrent l'applicabilité de l'extraction des motifs séquentiels et les avantages potentiels dans le domaine médical.
Si, comme nous venons de le montrer, les motifs séquentiels se sont avérés très utiles pour des applications médicales, d'autres variantes de motifs ont également été utilisées. Pour la détection d'effets secondaires des médicaments, [Jin *et al.*, 2008] utilisent d'autres types de motifs comme les motifs rares. [Salle *et al.*, 2009] se sont servis des motifs discriminants pour distinguer des patients sains des patients malades, à partir des données de puces à ADN. [Gotz *et al.*, 2014] ont créé une plateforme de visualisation des évènements cliniques, à partir du dossier patient, pour mieux comprendre de quelle façon des variations dans la séquence des évènements cliniques peuvent influencer les résultats médicaux.
Plusieurs algorithmes sont capables d'extraire des motifs séquentiels comme GSP

[Srikant et Agrawal, 1996], basé sur l'algorithme Apriori, PrefixSpan [Pei *et al.*, 2004], SPADE [Zaki, 2001], SPAM [Ayres *et al.*, 2002], *etc.* Dans ce chapitre, nous nous intéressons à un type particulier de motifs, **les motifs contextuels**.

Dans la suite de cette section, nous introduisons (section 5.1.1) le vocabulaire spécifique au domaine de la recherche de motifs séquentiels et l'illustrons sur un exemple d'application. Dans ce dernier, nous mettons en évidence des motifs fréquents à partir de séquences d'évènements médicaux représentant les GHM d'un patient. Puis, nous poursuivons avec le descriptif de la méthode (section 5.1.2) appliquée aux données du PMSI.

5.1.1 Définitions préliminaires

5.1.1.1 Séquences d'évènements

Définition 1 (Itemset)
Soit $I = \{i_1, i_2..., i_k\}$ l'ensemble de tous les items. Un sous-ensemble de I est appelé un itemset.

Définition 2 (Séquence d'évènements)
Une séquence $s = \langle e_1 e_2 ... e_m \rangle$ est une liste ordonnée d'itemsets, où $e_i \subseteq I$ pour $1 \leq i \leq m$.

Définition 3 (Sous-séquence d'évènements)
Une séquence $s' = \langle r_1 r_2 ... r_p \rangle$ est une sous-séquence de $s = \langle e_1 e_2 ... e_m \rangle$, s'il existe des entiers $1 \leq i_1 \leq i_2 \leq ... \leq i_p \leq m$, tels que $r_1 \subseteq e_{i_1}, r_2 \subseteq e_{i_2}, ..., r_m \subseteq e_{i_p}$.

Définition 4 (Support)
Soit une base de séquences $B = \{s_1, ..., s_n\}$, le support de la séquence s, $Freq_B(s)$, est le nombre de séquences dans B ayant s comme sous-séquence.

Exemple : Dans cet exemple, nous allons considérer les évènements (les séjours hospitaliers) de 14 patients sur une période de 4 mois. Le temps est perçu comme une variable non continue. Il est divisé en estampilles temporelles représentées par les mois. Supposons qu'à chaque estampille temporelle, il ne peut se produire qu'un seul évènement. En d'autres mots, un patient a un seul séjour par mois. Ces informations sont contenues dans la base de données présentée dans le tableau 5.1. À chaque séjour est associé un GHM (05M13 : Douleurs thoraciques ; 05M06 : Angine de poitrine ; 05M16 : Athérosclérose coronarienne ; 05M04 : IM aigu ; 05M09 : Insuffisance cardiaque). Ces données sont séquentielles car elles présentent des évènements (les GHM) disposés suivant un ordre (le temps). Par exemple, pour le patient P_{12}, le GHM 05M09 a été associé au séjour de janvier, le GHM 05M06 a été associé au séjour de février, puis le GHM 05M04 a été associé au séjour de mars, enfin le GHM 05M13 a été associé au séjour d'avril. Une sous-séquence de la séquence du patient P_{12} est par exemple, la séquence $\langle(05M09)(05M06)\rangle$. Elle est également présente dans la séquence du patient P_2 : son support est donc de 14% (2 sur 14). Il est important de constater que dans notre cas d'étude les itemsets sont réduits à un item.

Tableau 5.1 – Base séquentielle de GHM.

Patients	Janvier	Février	Mars	Avril
P_1		05M13	05M04	05M06
P_2	05M13	05M09	05M06	
P_3	05M13	05M13		05M06
P_4	05M16	05M13	05M06	05M16
P_5	05M04	05M13	05M06	05M04
P_6		05M06		05M13
P_7		05M13	05M06	05M13
P_8	05M04	05M13	05M16	05M06
P_9		05M13	05M13	05M06
P_{10}		05M06	05M16	05M04
P_{11}		05M06	05M04	05M13
P_{12}	05M09	05M06	05M04	05M13
P_{13}	05M06	05M04	05M09	
P_{14}	05M06		05M13	05M09

5.1.1.2 Motif séquentiel

Définition 5 (Motif séquentiel fréquent)
Une séquence s est fréquente et appelée motif séquentiel si son support est supérieur ou égal à un seuil appelé support minimum $k_\sigma > 0$ fixé : $Freq_B(s) \geq k_\sigma$.

Exemple : En examinant le tableau 5.2, nous constatons que le motif 05M13 suivi plus tard par 05M06 est vérifié par plus de 50% des patients (8 sur 14). En supposant que le professionnel de santé précise qu'il est intéressé par des GHM qui apparaissent dans au moins 50% des cas (support minimum) présents dans la base alors il s'avère que la sous-séquence $\langle(05M13)(05M06)\rangle$ est un motif séquentiel fréquent.

Tableau 5.2 – Mise en valeur du motif $\langle(05M13)(05M06)\rangle$ (en gras) soit le GHM 05M13 suivi du GHM 05M06. Ce motif est fréquent dans la base pour un support minimum de 50%.

Patients	Janvier	Février	Mars	Avril
P_1		**05M13**	05M04	**05M06**
P_2	**05M13**	05M09	**05M06**	
P_3	**05M13**	05M13		**05M06**
P_4	05M16	**05M13**	**05M06**	05M16
P_5	05M04	**05M13**	**05M06**	05M04
P_6		05M06		05M13
P_7		**05M13**	**05M06**	05M13
P_8	05M04	**05M13**	05M16	**05M06**
P_9		**05M13**	05M13	**05M06**
P_{10}		05M06	05M16	05M04
P_{11}		05M06	05M04	05M13
P_{12}	05M09	05M06	05M04	05M13
P_{13}	05M06	05M04	05M09	
P_{14}	05M06		05M13	05M09

5.1.1.3 Motif séquentiel contextuel

[Rabatel *et al.*, 2010b] ont étendu la notion de motif pour prendre en compte l'existence d'informations supplémentaires permettant de décrire les données.

Définition 6 (Contexte)
Un contexte c est une catégorie ou une modalité d'une variable (e.g. Homme ou Femme pour le sexe).

Définition 7 (Hiérarchie des contextes)
L'ensemble de tous les contextes muni d'une relation d'ordre partiel, $\leq$, constitue la hiérarchie des contextes H.

Définition 8 (Contexte général, contexte minimal)
Les contextes feuilles de H sont appelés les contextes minimaux. A contrario, plus on remonte dans l'arborescence de H plus le contexte est dit général.

Définition 9 (Motif séquentiel contextuel fréquent)
Soit c un contexte, H la hiérarchie des contextes et s une séquence. Une séquence s est fréquente dans un contexte c si son support dans c est supérieur ou égal à un support minimum $k_{\sigma_c} > 0$ fixé : $Freq_c(s) \geq k_{\sigma_c}$.

Définition 10 (Motif séquentiel c-général)
Une séquence s est générale dans c, dite c-générale, si s est fréquente dans tous les contextes descendants de c dans H.

Définition 11 (Motif séquentiel c-spécifique)
Une séquence s est spécifique à c (c-spécifique) ssi :

1) *s est fréquente dans c ;*
2) *s est c-générale ;*
3) *il n'existe pas de contexte c' tel que $c' > c$ et s soit c'-générale.*

Exemple : Jusqu'à présent, nous avons considéré la base comme un ensemble indivisible pour la recherche des motifs. Maintenant, nous allons prendre en compte les circonstances liées aux données : les contextes. Ces derniers impliquent l'existence de sous-ensembles de données rassemblant des propriétés similaires. Par exemple, nous pouvons intégrer des informations supplémentaires comme dans le tableau 5.3 qui associent à chaque patient son âge (*jeune* ou *âgé*) et son sexe (*homme* ou *femme*). Ces informations complémentaires peuvent être ordonnées suivant la hiérarchie des contextes représentée dans la figure 5.1.

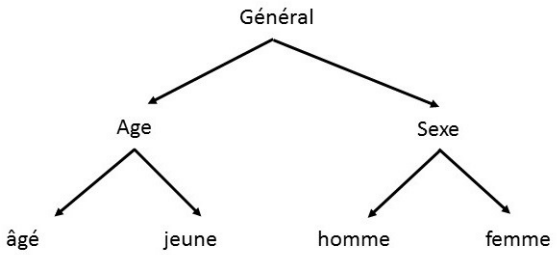


Figure 5.1 – Hiérarchie des contextes de l'exemple détaillé dans le tableau 5.3 .

Ces informations contextuelles peuvent avoir une influence non négligeable sur la séquence d'évènements. Ainsi, l'extraction de motifs doit rendre cette influence perceptible pour l'utilisateur afin de lui offrir une vue contextualisée des données. Considérons maintenant la séquence $\langle(05M13)(05M06)\rangle$ dans le tableau 5.3, nous constatons que :

- cette séquence de GHM est fréquente dans la population âgée (7 personnes âgées sur 8) mais pas dans la population jeune (seulement 1 personne sur 6) ;
- cette séquence de GHM est fréquente chez les personnes âgées quel que soit leur sexe (5 hommes âgés sur 5 et 2 femmes âgées sur 3).

Si, maintenant nous fixons un support minimum à 80%, nous extrairons le motif contextuel $\langle(05M13)(05M06)\rangle$ pour le contexte des personnes âgées. Il est donc c-spécifique pour le contexte des personnes âgées.

Tableau 5.3 – Mise en valeur du motif $\langle(05M13)(05M06)\rangle$ (en gras) avec les informations contextuelles sur l'âge et le sexe. Ce motif est spécifique aux personnes âgées. Une seule personne jeune est concernée.

Patients	**Age**	**Sexe**	**Janvier**	**Février**	**Mars**	**Avril**
P_1	âgé	homme		**05M13**	05M04	**05M06**
P_2	âgé	homme	**05M13**	05M09	**05M06**	
P_3	âgé	homme	**05M13**	05M13		**05M06**
P_4	âgé	homme	05M16	**05M13**	**05M06**	05M16
P_5	âgé	homme	05M04	**05M13**	**05M06**	05M04
P_6	âgé	femme		05M06		05M13
P_7	âgé	femme		**05M13**	**05M06**	05M13
P_8	âgé	femme	**05M13**	05M16	**05M06**	
P_9	jeune	homme		**05M13**	05M13	**05M06**
P_{10}	jeune	homme		05M06	05M16	05M04
P_{11}	jeune	homme		05M06	05M04	05M13
P_{12}	jeune	femme	05M09	05M06	05M04	05M13
P_{13}	jeune	femme	05M06	05M04	05M09	
P_{14}	jeune	femme	05M06		05M13	05M09

5.1.1.4 Motif séquentiel discriminant

Définition 12 (Taux de croissance)
Soit s, un motif dont les fréquences dans les contextes c et c' sont respectivement $Freq_c(s)$, $Freq_{c'}(s)$. Le taux de croissance de s dans c par rapport à c' est défini par le rapport des fréquences : $\frac{Freq_c(s)}{Freq_{c'}(s)}$.

Exemple : Considérons à nouveau le motif $\langle(05M13)(05M06)\rangle$. Déterminons le taux de croissance de ce motif dans le contexte des personnes âgées par rapport à celui des personnes jeunes. Nous obtenons : $\frac{7/8}{1/6} = 5,25$. Autrement dit, par rapport aux personnes jeunes, ce motif est 5,3 fois plus fréquent dans le contexte des personnes âgées.

Définition 13 (Motif séquentiel contextuel discriminant)
Un motif séquentiel est dit discriminant s'il est fréquent dans un contexte c (et ses sous-contextes) et non fréquent dans les autres contextes. Autrement dit, un motif M sera dit discriminant pour un contexte c, par rapport à un autre contexte c', si son taux de croissance est supérieur à un seuil $k_{\sigma_c^d} > 0$ fixé.

Exemple : En supposant que le professionnel de santé précise qu'il est intéressé par des GHM qui apparaissent dans au moins 50% des cas (support minimum) présents dans la base et par ceux qui sont deux fois plus présents dans un contexte par rapport aux autres, alors le motif $\langle(05M13)(05M06)\rangle$ est discriminant pour les personnes âgées.

Savoir qu'un comportement est général, spécifique ou discriminant est utile pour l'interprétation médicale. Dans cette approche, la propriété « d'être fréquent » dépend d'un contexte donné. La propriété « d'être caractéristique » ou « discriminant », se singularise par le fait d'être plus fréquent pour un contexte, par rapport à tous les autres contextes.

De manière générale, l'extraction de motifs est un problème difficile qui nécessite de naviguer dans un très grand espace de recherche. La prise en compte des contextes étend encore cet espace de recherche. Via des propriétés théoriques intéressantes basées sur la fréquence et sur les propriétés formelles associées au treillis formé par les contextes, [Rabatel *et al.*, 2010a] ont développé un algorithme très efficace pour l'extraction de ces motifs contextuels. Cette approche a été généralisée à d'autres mesures d'intérêt que la fréquence, utiles pour la sélection des motifs tels que le gain d'information, le taux d'émergence, la confiance dont l'objectif est de s'appuyer sur les caractéristiques statistiques des motifs pour isoler les plus intéressants au sens de critères experts. Dans la suite, nous allons mettre en œuvre cette approche sur nos données.

5.1.2 Description du processus de fouille

Le processus de fouille, schématisé dans la figure 5.2, s'effectue en plusieurs étapes.

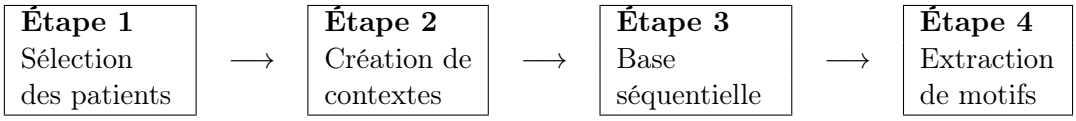


Figure 5.2 – Étapes du processus de fouille.

Étape 1 : Sélection des patients dans la base PMSI-MCO 2009-2014. Elle consiste à filtrer les données à partir de la base de données constituée dans le chapitre 3. Cette étape est effectuée à l'aide d'une requête SQL sur les bases de données de l'ATIH. Pour chaque patient, nous récupérons l'ensemble de ses séjours sur la période 2009 à 2014, excepté les séjours pour séances (radiothérapie, dialyses, chi-

miothérapie, *etc.*) et les prestations inter-établissements : lorsque l'on transfère un patient dans un autre établissement pour lui faire faire un acte (*e.g.* une coronarographie) car le premier n'a pas l'équipement pour le réaliser. De part les règles de codage du PMSI, ces séjours ont le même motif d'admission.

Étape 2 : Création de contextes. La population de patients a été découpée pour créer des sous-populations appelées contextes, à l'aide de covariables associées aux patients. Pour ce faire, nous avons pris en compte le genre (Homme/Femme), la classe d'âge du patient au moment de sa première apparition dans la période d'observation. Pour des raisons de cohérence de suivi médical, après discussion avec l'expert médical, Pr. Paul Landais (Chef de Service du BESPIM [2] à Nîmes et directeur de l'unité de recherche Inserm EA 2415), nous avons retenu un découpage des classes en trois catégories de sorte à prendre en compte trois aspects de la survenue d'un infarctus du myocarde : les patients ayant moins de 45 ans (l'occurrence de cet évènement est rare à exceptionnelle, elle concerne moins de 5% de la population), ceux ayant entre 45 et 65 ans (il s'agit de la population atteignant un âge où commence à se manifester l'insuffisance coronarienne qui peut se traduire par un infarctus. C'est une tranche d'âge à risque pour cette pathologie, elle représente environ 50% de la population) et les patients ayant plus de 65 ans (il s'agit également d'une population à risque mais qui présente généralement plusieurs facteurs de risque cumulés avec le vieillissement. Elle concerne 45% de la population). Enfin, nous avons retenu le nombre de séjours selon trois catégories pour qualifier un contexte médical : les 3-5 séjours (représentant en moyenne une hospitalisation par an sur 6 ans. Ces patients présentent surtout des évènements associés à leur pathologie cardiaque tels que angor, coronarographie, pose de stent... Cela concerne la majorité de la population, 54% des patients avec un nombre médian d'hospitalisations par patient de 4 et une moyenne égale à 4,46), les 5-60 séjours (représentant en moyenne une hospitalisation par mois sur 6 ans. Ces patients sont hospitalisés plus fréquemment que les patients du groupe précédent. Ils ont des pathologies associées nécessitant des prises en charge hospitalières réitérées comme par exemple pour le suivi d'un diabète, d'une rétinopathie ou d'une autre complication oculaire... Cela concerne environ 45% des patients avec un nombre médian d'hospitalisations par patient de 8 et une moyenne égale à 15,5) et enfin les plus de 60 séjours (il s'agit de patients très spécifiques, hospitalisés plus d'une fois par mois. Ils ont d'autres pathologies associées nécessitant des hospitalisations fréquentes, comme une prise en charge en dialyse par exemple. Cela concerne 0,04% des patients avec une médiane à 85 jours et une moyenne à 117 jours). Les combinaisons de ces modalités pour chaque covariable constituent des contextes minimaux ou feuilles : il y en a 18 dans notre cas. La hiérarchie de nos contextes est représentée dans la figure 5.3. À noter que les patients ayant moins de 3 séjours pourraient faire l'objet d'une étude particulière à l'aide d'outils comme ceux du package TraMineR [Gabadinho *et al.*, 2011] du logiciel R, permettant de réaliser une typologie de ces trajectoires courtes non dénuées d'intérêt.

2. Biostastique, Épidémiologie, Santé Publique et Information Médicale

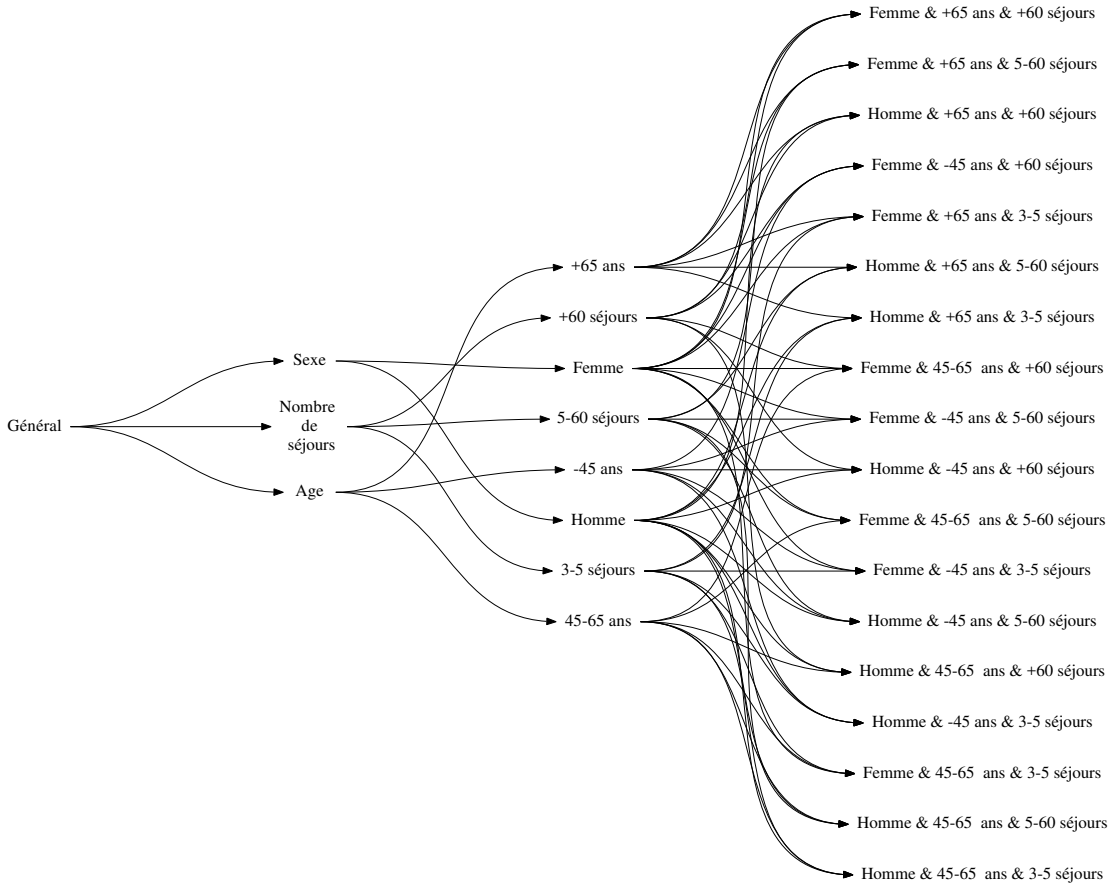


Figure 5.3 – Hiérarchie des contextes.

Étape 3 : Construction de la base séquentielle. Pour l'étape de construction de la base séquentielle, nous nous sommes intéressés aux GHM et aux DP. À chaque patient est associé une séquence de GHM et une séquence de DP. La longueur de ces séquences est égale au nombre de séjours effectués pendant ces six années. Pour chaque patient, nous avons nettoyé les données, en conservant la première hospitalisation du patient puis tous ses séjours d'hospitalisation liés à la cardiopathologie. Concrètement, nous avons retenu les séjours avec GHM commençant par '05' et/ou un motif d'hospitalisation pour un facteur de risque identifié comme tel par la médecine (voir Annexe C.1). En effet, suite à de premières expérimentations, nous avons pu constater que la fouille faisait essentiellement ressortir des évènements liés à la cardiologie (facteurs de risque et comorbidités). Ainsi, un certain nombre de motifs séquentiels ont mis en évidence des faits populationnels comme le développement de problèmes de cataracte pour les personnes âgées, par exemple, ou encore des pathologies gynécologiques pour les femmes. De plus, nous avons ajouté une information supplémentaire en intégrant le code *Décès* lorsque ce dernier avait été constaté.
Un exemple de construction de la trajectoire du patient est illustré dans la figure 5.4. Nous reconstituons la séquence d'évènements du patient en étape I. Le patient 1 est hospitalisé pour IM, pour diabète, pour obésité, pour fracture du poignet et

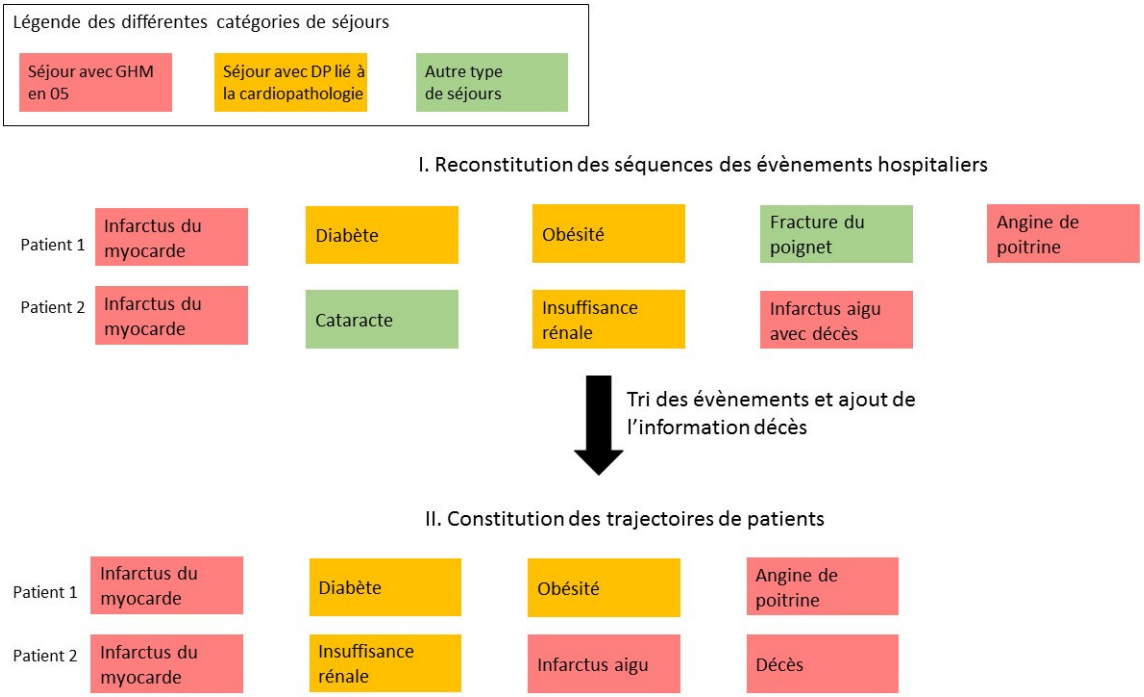


Figure 5.4 – Construction de la trajectoire du patient.

pour angine de poitrine. Dans cette séquence d'évènements, nous identifions un séjour non lié à la cardiologie : fracture du poignet. Cet évènement est retiré de la séquence du patient 1 dans l'étape II. Le patient 2 est hospitalisé pour IM, cataracte, insuffisance rénale, et décède suite à un IM aigu. Dans cette séquence d'évènements, nous identifions un séjour non lié à la cardiologie : cataracte. De plus nous intégrons une information supplémentaire : le décès. L'évènement cataracte est supprimé de la séquence du patient et l'information *Décès* est ajoutée dans l'étape II.
Ainsi, ces séquences de GHM et de DP triées constituent la trajectoire du patient. Dans la suite, nous étudierons donc deux types de trajectoires (voir section 2.2 pour la définition des termes GHM et DP).

Étape 4 : Extraction de motifs. Nous effectuons la recherche de motifs séquentiels contextuels. L'extraction de motifs est faite à l'aide de deux algorithmes [Rabatel, 2011] : CFPM (Contextual Frequent Pattern Mining) et CoPaM (Contextual discriminant Pattern Mining), permettant de détecter des motifs séquentiels contextuels fréquents et discriminants avec un support respectivement de 1% et 5%, et un seuil pour le taux de croissance fixé à 2 dans le cas des motifs discriminants.

5.2 Expérimentations

Dans les résultats présentés ici, nous nous sommes focalisés sur les contextes décrits dans le tableau 5.4. Quelques éléments importants seront évoqués pour les autres.

Tableau 5.4 – Détail des contextes étudiés.

Contexte	**Effectifs**
Homme & 45-65 ans & 5-60 séjours	50 663
Femme & 45-65 ans & 5-60 séjours	11 340
Homme & -45 ans & 5-60 séjours	6 108
Femme & -45 ans & 5-60 séjours	1 954

Dans la suite, nous faisons référence à des codes issus du PMSI pour les GHM et les codes CIM-10. L'ensemble de ces codes est explicité dans les tableaux C.2 et C.3 de l'annexe. Les résultats présentés dans cette section sont discutés dans la section 5.3.

5.2.1 Motifs fréquents

5.2.1.1 Trajectoires de GHM

L'ensemble des résultats est résumé dans le tableau 5.5. La fouille donne 24 motifs. Parmi ces motifs, 3 contiennent au moins trois items, 12 n'en contiennent que deux. Les autres motifs sont réduits à un seul item. Nous remarquons que les hommes de 45-65 ans ont plus de motifs par opposition aux femmes de -45 ans qui en ont le moins.

Nous allons nous intéresser plus particulièrement à ces deux tranches d'âge : les -45 ans et les 45-65 ans. Dans un premier temps, nous allons regarder les motifs communs aux contextes présentés dans le tableau 5.4. Dans un second temps, nous allons examiner ce qui distingue un contexte d'un autre en déterminant les motifs présents dans un contexte mais pas dans un autre et inversement.

Les motifs communs contenant un seul itemset [3] sont les poses d'endoprothèses avec ($\langle(05K05)\rangle$) et sans ($\langle(05K06)\rangle$) IM, puis IM aigu ($\langle(05M04)\rangle$), les actes diagnostiques ($\langle(05K10)\rangle$) et thérapeutiques ($\langle(05K13)\rangle$) par voie vasculaire, les dilatations coronaires ($\langle(05K24)\rangle$), l'athérosclérose coronarienne ($\langle(05M16)\rangle$). Les motifs contenant au moins deux items communs à tous les contextes sont les hospitalisations multiples pour pose d'endoprothèse avec ou sans IM, jusqu'à trois fois ($\langle(05K05)(05K06)\rangle$, $\langle(05K06)(05K06)\rangle$, $\langle(05K05)(05K06)(05K06)\rangle$, $\langle(05K06)(05K06)(05K06)\rangle$) et IM aigu suivi de pose de stent sans IM ($\langle(05M04\text{-}05K06)\rangle$). Nous pouvons faire deux remarques, pour tous les motifs évoqués :

1) les fréquences sont plus élevées chez les 45-65 ans que chez les -45 ans ;
2) les fréquences sont plus élevées chez les hommes que chez les femmes.

Dans le cas de la première remarque, nous constatons toutefois des exceptions. Nous listons, dans la suite, celles pour lesquelles la différence est significative avec test statistique approprié [4] au seuil de 5%. $\langle(05K05)\rangle$ est plus fréquent pour les -45 ans (par exemple pour les hommes, il est à 35% *vs* 20%), tout comme $\langle(05M04)\rangle$, avec

3. Ici, un itemset est aussi un item.
4. Test de proportion ou test exacte de Fisher pour des effectifs inférieurs à 5.

une fréquence de 14% pour les hommes de -45 ans *vs* 8% pour les hommes de 45-65 ans, ou encore ⟨(05M16)⟩ avec un taux égal à 1,4% pour les hommes de -45 ans *vs* 1% pour ceux de 45-65 ans. ⟨(05K05)(05K06)⟩ est également un contre-exemple (9% des hommes de -45 ans *vs* 6% des hommes de 45-65 ans), mais aussi ⟨(05K06)(05K05)⟩ (2,3% *vs* 1,6%) et ⟨(05M04)(05K06)⟩ (2% *vs* 1,4%) ou encore ⟨(05K05)(05K06)(05K06)⟩ (2,3% *vs* 1,3%).

Tableau 5.5 – Extrait de motifs séquentiels contextuels fréquents pour les trajectoires de GHM au seuil de 1%.

	Contexte (5-60 séjours)							
	45-65 ans				**-45 ans**			
	Homme (50 663)		**Femme** (11 340)		**Homme** (6 108)		**Femme** (1 954)	
Motif	**Eff**	**%**	**Eff**	**%**	**Eff**	**%**	**Eff**	**%**
⟨(05K05)⟩	10 072	19,88	2 217	19,55	2 126	34,81	598	30,6
⟨(05K06)⟩	36 721	72,48	7 632	67,3	3 594	58,84	957	48,98
⟨(05K10)⟩	2 391	4,72	528	4,66	221	3,62	94	4,81
⟨(05K13)⟩	2 381	4,7	440	3,88	238	3,9	67	3,43
⟨(05K24)⟩	1 388	2,74	303	2,67	174	2,85	43	2,2
⟨(05M04)⟩	4 033	7,96	1 342	11,83	833	13,64	344	17,6
⟨(05M06)⟩			138	1,22	63	1,03	25	1,28
⟨(05M16)⟩	532	1,05	150	1,32	105	1,72	42	2,15
⟨(Décès)⟩	578	1,14	169	1,49			24	1,23
⟨(05K05)(05K05)⟩			119	1,05	144	2,36	27	1,38
⟨(05K05)(05K06)⟩	2 792	5,51	560	4,94	565	9,25	115	5,89
⟨(05K05)(05M04)⟩					111	1,82	23	1,18
⟨(05K06)(05K05)⟩	811	1,6	145	1,28	138	2,26		
⟨(05K06)(05K06)⟩	9 717	19,18	1 765	15,56	979	16,03	192	9,83
⟨(05K06)(05K13)⟩	897	1,77	172	1,52	80	1,31		
⟨(05K06)(05K24)⟩	735	1,45	158	1,39	92	1,51		
⟨(05K06)(05M04)⟩	583	1,15	126	1,11	87	1,42		
⟨(05K10)(05K06)⟩	714	1,41						
⟨(05K13)(05K06)⟩	679	1,34						
⟨(05M04)(05K06)⟩	694	1,37	127	1,12	126	2,06	22	1,13
⟨(05M04)(05M04)⟩							22	1,13
⟨(05K05)(05K06)(05K06)⟩	643	1,27	129	1,14	138	2,26	20	1,02
⟨(05K06)(05K06)(05K06)⟩	2 330	4,6	350	3,09	227	3,72	51	2,61
⟨(05K06)(05K06)(05K06)(05K06)⟩	547	1,08						

Eff : Effectif.

Dans le cas de la deuxième remarque, nous constatons également des exceptions. L'IM aigu est plus fréquent chez les femmes que chez les hommes. Sa fréquence est de 12% contre 8% pour les 45-65 ans, elle est de 17,6% contre 13,6% pour les -45 ans. L'athérosclérose est également plus fréquente chez les femmes avec 1,3% *vs* 1% pour les 45-65 ans.

Nous allons maintenant nous intéresser à ce qui distingue un contexte d'un autre. Il s'agit d'examiner les motifs qui sont présents dans un contexte mais pas dans un autre et inversement. Comparons tout d'abord les contextes -45 ans et 45-65 ans. Nous remarquons que les motifs suivants : ⟨(05K10)(05K06)⟩, ⟨(05K13)(05K06)⟩, ⟨(05K06)(05K06)(05K06)(05K06)⟩, ne sont pas présents chez

les -45 ans. En revanche, ils le sont chez les hommes de 45-65 ans. *A contrario*, les motifs ⟨(05K05)(05M04)⟩ et ⟨(05M04)(05M04)⟩ ne sont pas présents chez les 45-65 ans alors qu'ils le sont chez les -45 ans. De plus, ⟨(05K05)(05M04)⟩ est fréquent dans ces contextes et seulement dans ces contextes.

Comparons maintenant au sein de chaque tranche d'âge ce qui distingue les hommes des femmes et réciproquement. Chez les 45-65 ans, les motifs ⟨(05M06)⟩ et ⟨(05K05)(05K05)⟩ ne concernent que les femmes. Par ailleurs les motifs : ⟨(05K10)(05K06)⟩, ⟨(05K13)(05K06)⟩ et ⟨(05K06)(05K06)(05K06)(05K06)⟩ ne concernent que les hommes. De même, chez les -45 ans, les motifs ⟨(Décès)⟩ et ⟨(05M04-05M04)⟩ ne concernent que les femmes. Notons que le double IM aigu est fréquent seulement pour les femmes de cette tranche d'âge [5] mais jamais chez les hommes. Par ailleurs, ⟨(05K06)(05K05)⟩, ⟨(05K06)(05K13)⟩, ⟨(05K06)(05K24)⟩ et ⟨(05K06)(05M04)⟩ ne concernent que les hommes.

5.2.1.2 Trajectoires de DP

L'extraction de motifs met en évidence 25 motifs (voir tableau 5.6). Parmi ces motifs, 17 contiennent au moins deux itemsets. Les autres motifs ne contiennent qu'un seul item. Notons qu'il y a plus de motifs que pour les trajectoires de GHM, mais il n'y en a pas de longueur supérieure à trois.
Analysons plus en détails par population les résultats obtenus. Nous allons exposer les résultats de la même manière que pour les trajectoires de GHM avec en premier les motifs communs à tous les contextes et ensuite, les différences entre ces contextes.

Après examen du tableau 5.6, nous détaillons les motifs communs. La fouille met en évidence des motifs liés à la sélection même des patients c'est-à-dire les codes CIM-10 du SCA de I20 à I25, puis les séjours pour angine de poitrine suivis de cardiopathie ischémique chronique (⟨(I20)(I25)⟩) ou angine de poitrine (⟨(I20)(I20)⟩), ou encore IM (⟨(I20)(I21)⟩), les séjours pour IM suivis d'angine de poitrine (⟨(I21)(I20)⟩) ou IM (⟨(I21)(I21)⟩) ou cardiopathie ischémique chronique (⟨(I21)(I25)⟩), les séjours pour cardiopathie ischémique chronique suivis d'angine de poitrine (⟨(I25)(I20)⟩) ou cardiopathie ischémique chronique (⟨(I25)(I25)⟩). Nous avons noté les mêmes exceptions que dans la section 5.2.1.1. De la même manière que précédemment seules les différences significatives à un seuil de 5% sont répertoriées.
Dans le cas de la première remarque, les motifs qui concernent l'infarctus aigu sont plus fréquents chez les -45 ans que chez les 45-65 ans. Si l'on compare pour les hommes, ⟨(I21)⟩ concerne 50% des -45 ans contre 30% des 45-65 ans. Le double IM (⟨(I21)(I21)⟩), concerne 6% des hommes de -45 ans contre 3% des hommes de 45-65 ans. Enfin, l'IM suivi d'angine de poitrine (⟨(I21)(I20)⟩) est présent pour 6% des patients de -45 ans contre 3% de ceux de 45-65 ans.
Dans le cas de la deuxième remarque, pour les 45-65 ans, l'angine de poitrine est plus fréquente chez les femmes avec 34% *vs* 33% chez les hommes. L'IM aigu a une fréquence de 33% chez les femmes *vs* 29% chez les hommes.

5. Plus précisément pour les femmes de +65 ans.

Tableau 5.6 – Extrait de motifs séquentiels contextuels fréquents pour les trajectoires de DP au seuil de 1%.

	Contexte (5-60 séjours)							
	45-65 ans				**-45 ans**			
	Homme (50 663)		**Femme** (11 340)		**Homme** (6 108)		**Femme** (1 954)	
Motif	**Effectif**	**%**	**Effectif**	**%**	**Effectif**	**%**	**Effectif**	**%**
⟨(I20)⟩	16 516	32,6	3 852	33,97	1 861	30,47	521	26,66
⟨(I21)⟩	14 930	29,47	3 794	33,46	3 035	49,69	1 018	52,1
⟨(I24)⟩	3 987	7,87	875	7,72	408	6,68	133	6,81
⟨(I25)⟩	21 481	42,4	3 958	34,9	1 876	30,71	450	23,03
⟨(I50)⟩	1 165	2,3	301	2,65	64	1,05		
⟨(R07)⟩			115	1,01	71	1,16	20	1,02
⟨(Décès)⟩	578	1,14	169	1,49			24	1,23
⟨(I20)(I20)⟩	3 101	6,12	694	6,12	367	6,01	84	4,3
⟨(I20)(I21)⟩	730	1,44	159	1,4	120	1,96	20	1,02
⟨(I20)(I25)⟩	2 249	4,44	423	3,73	247	4,04	40	2,05
⟨(I21)(I20)⟩	1 576	3,11	357	3,15	334	5,47	82	4,2
⟨(I21)(I21)⟩	1 388	2,74	307	2,71	362	5,93	79	4,04
⟨(I21)(I24)⟩					74	1,21		
⟨(I21)(I25)⟩	2 179	4,3	411	3,62	411	6,73	89	4,55
⟨(I24)(I24)⟩	699	1,38						
⟨(I24)(I25)⟩	603	1,19						
⟨(I25)(I20)⟩	1 591	3,14	293	2,58	155	2,54	38	1,94
⟨(I25)(I21)⟩	699	1,38			89	1,46		
⟨(I25)(I25)⟩	4 509	8,9	648	5,71	372	6,09	59	3,02
⟨(I20)(I20)(I20)⟩	664	1,31	137	1,21	67	1,1		
⟨(I21)(I21)(I21)⟩					62	1,02	20	1,02
⟨(I21)(I20)(I20)⟩					74	1,21		
⟨(I21)(I25)(I25)⟩					68	1,11		
⟨(I25)(I25)(I25)⟩	887	1,75			67	1,1		

Comparons les contextes -45 ans et 45-65 ans. Les motifs de cardiopathies ⟨(I24)(I24)⟩ et ⟨(I24)(I25)⟩ ne sont pas présents chez les -45 ans alors qu'ils le sont pour les hommes de 45-65 ans. En revanche, les motifs suivants : ⟨(I21)(I24)⟩, ⟨(I21)(I21)(I21)⟩, ⟨(I21)(I20)(I20)⟩ et ⟨(I21)(I25)(I25)⟩, ne sont pas présents chez les 45-65 ans alors qu'ils le sont chez les -45 ans. Notons que le triple IM est fréquent pour les -45 ans, et seulement pour ce contexte. Il est général pour ce contexte.
Détaillons ensuite au sein de chaque tranche d'âge les différences entre les hommes et les femmes. Pour les 45-65 ans, le motif douleur au niveau de la gorge et du thorax (⟨(R07)⟩) ne concerne que les femmes. Les motifs ⟨(I24)(I24)⟩ et ⟨(I24)(I25)⟩, ⟨(I25)(I21)⟩ et ⟨(I25)(I25)(I25)⟩ ne concernent que les hommes. Notons que les motifs suivants ⟨(I24)(I24)⟩, ⟨(I24)(I25)⟩ ne sont fréquents que chez les hommes de 45-65 ans (on les retrouve également dans le contexte Homme & +65 ans & 5-60 séjours). Chez les -45 ans, les motifs ⟨(I50)⟩ (insuffisance cardiaque), ⟨(I25)(I21)⟩, ⟨(I21)(I20)(I20)⟩, ⟨(I21)(I25)(I25)⟩ et ⟨(I25)(I25)(I25)⟩ ne concernent que les hommes. Le décès n'est pas fréquent au seuil de 1% pour les hommes de -45 ans. Pour finir, nous remarquons que les motifs ⟨(I25)(I21)⟩ et ⟨(I25)(I25)(I25)⟩ ne sont fréquents que pour les hommes. On les retrouve dans d'autres contextes chez les hommes mais jamais chez les femmes.

5.2.1.3 Autres résultats

Pour le contexte général, autrement dit la population de patients dans son intégralité, les GHM fréquents sont ⟨(05K05)⟩ et ⟨(05K06)⟩. De même, pour les trajectoires de DP, nous mettons en évidence les DP : ⟨(I20)⟩, ⟨(I21)⟩, ⟨(I25)⟩, par ordre de fréquence croissante. Les motifs fréquents sont ceux liés aux critères de sélection de la population.

Le décès est le plus fréquent [6] pour le contexte Femme & +65 ans & 3-5 séjours, mais également pour les hommes. C'est même le motif ⟨(I21)(Décès)⟩ qui est fréquent [6] pour les femmes avec une fréquence égalant pratiquement 7%. Dans cette même population la fréquence de décès est la plus élevée atteignant 8%. L'infarctus aigu entraînant le décès est également fréquent, autour de 3%, pour les femmes de +65 ans ayant entre 5-60 séjours mais aussi les hommes de +65 ans avec 3-5 séjours.

De manière générale, il est également fréquent d'avoir des séjours à répétition pour pose d'endoprothèse. Notons que l'IM aigu a de forts taux de fréquence pour chacune des trois classes d'âge, tout comme les cardiopathies ischémiques chroniques.

5.2.2 Motifs discriminants

La recherche est faite au seuil de 5%, avec un seuil minimal du taux de croissance fixé à 2. La liste des motifs obtenue étant beaucoup moins importante que dans le premier cas, nous répertorions l'intégralité des résultats pour les trajectoires de DP dans le tableau 5.7. Ici nous avons une information supplémentaire : le taux de croissance. Dans le tableau, la colonne Txc correspond au minimum des taux de croissance déterminés par rapport aux autres contextes.

Examinons les résultats obtenus dans le cas des trajectoires de DP. Par exemple, pour le contexte Femme ayant -45 ans avec +60 séjours, le motif ⟨(I20)⟩ (Angine de poitrine) est discriminant au seuil de 5%, il est 2,5 fois plus fréquent dans ce contexte que dans les autres contextes.
Intéressons nous maintenant aux autres contextes. Par exemple, un motif discriminant pour le contexte Femme ayant +65 ans avec +60 séjours est ⟨(I50)⟩ c'est-à-dire : Insuffisance cardiaque. Toutefois, étant donné les effectifs concernés par les motifs mis en évidence, nous ne pouvons pas affirmer qu'il y ait réellement un motif qui soit caractéristique du contexte. Nous pouvons également faire la même remarque pour le reste des contextes qui apparaissent dans le tableau 5.7. Les autres contextes que nous avions sélectionnés n'apparaissent pas, car il n'y a pas de motifs discriminants les concernant. Il n'y a donc pas de motif qui les caractérise.

Pour les trajectoires de GHM, nous pouvons faire le même constat que dans les trajectoires de DP : les effectifs concernés par les motifs mis en évidence ne nous permettent pas de découvrir des motifs caractéristiques de ces contextes.

6. Significatif au seuil de 5%.

Tableau 5.7 – Extrait de motifs séquentiels contextuels discriminants pour les trajectoires de DP au seuil de 5% et taux de croissance minimal de 2.

	Motif séquentiel discriminant			
Contexte (Effectif)	**Motif**	**%**	**Effectif**	**Txc**
Femme & +65 ans & +60 séjours (19)	⟨(I25)(I50)(I21)⟩	15,79	3	675,88
	⟨(Z75)⟩	5,26	1	596,84
	⟨(I21)(I25)(I50)⟩	5,26	1	321,47
	⟨(I22)(I23)⟩	5,26	1	321,47
	⟨(I22)(I23)(I21)⟩	5,26	1	321,47
	⟨(I21)(I50)(I21)⟩	5,26	1	107,16
	⟨(I23)(I21)⟩	5,26	1	80,37
	⟨(I22)(I21)⟩	5,26	1	80,37
	⟨(I50)(I21)⟩	5,26	1	21,57
	⟨(I25)(I50)⟩	5,26	1	21,52
	⟨(I70)⟩	5,26	1	19,25
	⟨(I21)(I50)⟩	5,26	1	13,52
	⟨(I22)⟩	5,26	1	10,50
	⟨(I21)(I25)(I21)⟩	5,26	1	10,05
	⟨(I50)⟩	0,16	3	3,32
	⟨(I21)(I25)(I50)(I21)⟩	0,05	1	4217,32
Homme & 45-65 ans & +60 séjours (64)	⟨(I20)(I24)⟩	2,44	2	2,30
Femme & 45-65 ans & +60 séjours (18)	⟨(I24)(I20)⟩	11,11	2	15,37
	⟨(I24)(I20)(I24)⟩	5,56	1	87,96
	⟨(I25)(I21)(I25)⟩	5,56	1	28,43
	⟨(I21)(I21)(I21)(I21)⟩	5,56	1	21,71
	⟨(I20)(Décès)⟩	5,56	1	14,78
	⟨(I21)(I21)(I21)⟩	5,56	1	5,43
	⟨(I20)(I24)⟩	5,56	1	3,56
Femme & -45 ans & +60 séjours (12)	⟨(I20)⟩	83,33	10	2,45
	⟨(I20)(I25)⟩	16,67	2	3,76
	⟨(I21)(I20)⟩	16,67	2	3,05
45-65 ans & + 60 séjours (32)	⟨(I20)(I24)⟩	2,44	1	2,30

5.3 Discussion

La recherche de motifs fréquents nous apporte des éléments intéressants dans le cadre de notre étude. Dans un premier temps, ce sont les évènements liés à la sélection de notre population qui ressortent comme motifs fréquents, tels que l'IM lui-même mais aussi la pose d'endoprothèse. Elle met en évidence des évènements déjà connus du domaine médical comme les hospitalisations à plusieurs reprises pour pose de stent. En effet, si un patient a un infarctus, c'est qu'il est atteint de maladie coronarienne. Il est donc à fort risque d'avoir d'autres artères atteintes par des plaques d'athérome [ANAES, 2004]. De plus, ces patients présentent probablement des facteurs de risques cardiovasculaires autres, tel que le diabète, contribuant à l'obstruction d'autres artères. Ainsi, la surveillance sur le plan cardiologique suite à un infarctus, des autres artères, peut entraîner la programmation d'angioplastie pour ces patients [Delahaye *et al.*, 2001]. Ceci se confirme par les résultats observés dans les trajectoires de DP : la motivation d'une hospitalisation n'est pas nécessairement pour un infarctus mais pour angine de poitrine, cardiopathie, *etc.* De plus, l'antécédent d'infarctus étant également un facteur de risque de récidive [ANAES, 2004], un patient peut donc avoir de nouveau un infarctus. Par ailleurs, les stents

sont à risque de resténose[7], ils sont donc également surveillés. D'après l'évaluation de la [HAS, 2009], ce risque de resténose n'est pas lié au type de stent utilisé : « Les stents actifs, par rapport aux stents nus, ont un bénéfice confirmé mais limité en termes de diminution du taux de resténose et de geste de revascularisation ». Ce risque de resténose demeure patient-dépendant : il dépend de l'observance du patient dans son traitement (*e.g.* antiagrégants plaquettaires), de sa lutte contre ses différents facteurs de risques (arrêt du tabac, pratique de sport, suivi d'un régime alimentaire, *etc.*), mais aussi des pathologies associées [Benamer *et al.*, 2007, Le Feuvre, 2009].

Dans un deuxième temps, l'analyse des résultats sur les contextes des 45-65 ans et des -45 ans, en distinguant les hommes des femmes, a permis de mettre en lumière trois éléments. Tout d'abord, nous avons pu remarquer que l'IM était plus fréquent chez les femmes que chez les hommes. Ceci vient confirmer un constat établi dans d'autres études [Dreyer *et al.*, 2014, Dreyer *et al.*, 2015, Gabizon et Lonn, 2015, Gabet *et al.*, 2016], à savoir que cette pathologie n'est plus à spécifique aux hommes. De plus, les taux bruts de mortalité sont plus élevés pour les femmes [Canto *et al.*, 2012, Singh *et al.*, 2014, Journath *et al.*, 2015]. Toutefois, une grande partie de la mortalité précoce accrue après IM chez les femmes peut s'expliquer par l'âge plus avancé et les caractéristiques de risque plus défavorables des femmes [Vaccarino *et al.*, 1995, Nohria *et al.*, 1998, Marrugat *et al.*, 1998, Ishihara *et al.*, 2008]. Enfin, cette pathologie qui est une maladie de la vieillesse, touche de plus en plus de personnes jeunes [Vaccarino *et al.*, 2001, Simon *et al.*, 2006]. La proportion des -45 ans avec IM est presque deux fois plus importante que celle des 45-65 ans et elle est assez proche de celle des +65 ans. L'IM prématuré est favorisé par un changement des habitudes et des comportements [Pasternack *et al.*, 1985, Wilhelmsson *et al.*, 1975, Rosenberg *et al.*, 1985]. En effet, d'après l'étude *Interheart* mesurant les facteurs de risques de l'IM dans 52 pays [Yusuf *et al.*, 2004] : les lipides anormaux, le tabagisme, l'hypertension artérielle, le diabète, l'obésité abdominale, les facteurs psychosociaux, la faible consommation de fruits, de légumes, l'augmentation de la consommation d'alcool et la sédentarité représentent la plus grande partie du risque d'IM dans le monde.

En revanche, la recherche de motifs discriminants n'a pas permis de mettre en exergue de façon probante, des motifs qui seraient propres à une population donnée, mis à part l'angine de poitrine pour les femmes de -45 ans avec +60 séjours. Ceci peut s'expliquer par notre choix de sélection des séjours. En effet, ce choix ne sélectionne pas des pathologies propres à une classe d'âge non nécessairement liées à la pathologie cardiaque, comme la cataracte pour les personnes âgées. À la liste des arguments pour étayer le nombre limité, voire l'absence dans certains contextes, de motifs discriminants, nous pouvons ajouter le temps d'observation. Le manque de recul suffisant ne nous permet sans doute pas de mettre en évidence des phénomènes particuliers. De plus, cela peut également s'expliquer par le fait que la pathologie en elle-même n'aurait pas de particularité populationnelle au sens où nous avons défini les populations. Une piste à explorer serait de créer

7. Reformation du rétrécissement précédemment supprimé d'un conduit ou d'un orifice de l'organisme.

des contextes différents, avec d'autres covariables comme le décès ou encore les comorbidités (dyslipidémie, tabac, hypertension, surpoids, *etc.*) pour déterminer s'il existe des parcours qui soient caractéristiques d'une sous-population. Des approches non supervisées peuvent également être envisagées pour identifier des contextes *a posteriori* à partir des motifs extraits [Perera *et al.*, 2009, Martinez-Maldonado *et al.*, 2013, Mei *et al.*, 2006].

Nous notons que bien souvent le DP apporte des informations plus précises sur l'événement, c'est-à-dire le contexte lié à l'hospitalisation. Cependant, de façon intrinsèque à l'usage même de ce code, il renseigne le motif d'admission mais ne donne pas toujours tous les détails du séjour. Par exemple une admission pour arrêt cardiaque, reste une information peu précise, alors que les GHM transplantation cardiaque ou pose d'un défibrillateur, sont en revanche plus riches sur le détail des évènements au cours de l'hospitalisation. Cet exemple, nous montre qu'il est délicat d'exploiter les bases de données PMSI. Le choix du type de trajectoire peut s'avérer judicieux suivant la problématique posée mais peut parfois être source de difficultés dans l'exploitation des résultats. Ici, les deux sources de résultats se recoupent et se complètent.

5.4 Conclusion

À partir des bases de données PMSI-MCO, nous avons examiné les trajectoires de patients ayant eu un IM au cours d'une période d'observation de 6 ans dans le but d'identifier des profils de parcours de soins types et/ou spécifiques d'une sous-population donnée. Nous avons établi d'une part que les parcours types pour les sous-populations sont relativement identiques d'un contexte à un autre, et que d'autre part, ces derniers concernent l'IM, l'angine de poitrine et les cardiopathies. Ce qui diffère essentiellement d'une sous-population à une autre, en particulier dans la comparaison entre genre, est la fréquence de ces parcours types.

Au vu des résultats, nous ne sommes pas parvenus à identifier des parcours qui soient caractéristiques d'une sous-population. Le contexte particulier de la cardiologie interventionnelle ou encore le choix stratégique des entités étudiées (GHM ou DP) pourraient amener des éléments d'explications.

Dans cette partie des travaux, nous avons pu constater que les trajectoires de DP nous renseignent sur les pathologies tandis que les trajectoires de GHM, nous renseignent sur les soins pratiqués avec les poses d'endoprothèse, les actes diagnostiques ou thérapeutiques ou encore les dilatations coronaires.

Pour aller plus loin dans l'exploration de ces données, il serait intéressant de mener des analyses similaires sur des contextes différents et plus particuliers, comme par exemple de distinguer les patients suivant les comorbidités. Il faudrait alors constituer des contextes en tenant compte du nombre de pathologies associées ou

encore du type de pathologie. Ces analyses supplémentaires pourraient permettre de caractériser ou de déterminer des parcours spécifiques des comorbidités. Une détermination des contextes sans *a priori* est également envisageable [Perera *et al.*, 2009].

Ce type d'approche pourrait également être généralisée à d'autres pathologies afin d'établir des profils de parcours de soins dans le cas de maladies chroniques, par exemple. Ces parcours types ou caractéristiques pourraient être la base de la construction de parcours de soins intégrés [Tang *et al.*, 2015] comme dans le cas du diabète [Goderis *et al.*, 2015] pour à la fois mieux prendre en charge le patient, mais aussi réduire les coûts.

Dans cette section, l'analyse poussée des motifs obtenus nous a permis de démontrer leur intérêt descriptif. Nous avons d'ailleurs retrouvé des résultats, discutés en section 5.3, en adéquation avec la littérature traitant de l'IM. Dans la suite de ces travaux, nous proposons d'intégrer ces résultats dans des modèles prédictifs afin de prévoir la mortalité à l'hôpital. Notre objectif est de pouvoir déterminer les trajectoires les plus à risque et de prévenir la survenue d'un événement fatal (voir chapitre 6 à suivre).

Chapitre

6

Prédire le décès

Avec 17,5 [1] millions de morts par an, les maladies cardiovasculaires représentent la première cause de mortalité dans le monde [World Health Organization, 2016]. Elles concernent 30% de l'ensemble des décès. Dans 7,4 millions des cas, une cardiopathie coronarienne est à l'origine du décès et dans 6,7 millions des cas, il s'agit d'un AVC. L'OMS estime que d'ici 2 030 près de 24 millions de personnes mourront de maladies cardiovasculaires et ces affections demeureront la première cause de mortalité. Selon la Direction de la santé publique et de l'évaluation des risques de la Commission européenne, les maladies cardiaques et vasculaires sont responsables d'environ 2 millions de décès par an en Europe. En outre, ces maladies sont à l'origine du plus grand nombre de décès prématurés, avant l'âge de 75 ans [2]. Le risque majeur associé à l'IM est le décès. En France, environ 120 000 personnes sont atteintes d'IM par an. 12 000 en décèdent lors de la crise, et 18 000 personnes décéderont dans l'année qui suit. De plus, un patient sur cinq traité pour un IM meurt dans les cinq ans.

Les maladies cardiovasculaires constituent donc une part importante de la consommation des soins. Elles représentent le poste de dépenses le plus important de la consommation de soins et de biens médicaux. On estime que les maladies cardiovasculaires coûtent à l'économie de l'Union Européenne 210 milliards d'euros par an [Wilkins *et al.*, 2017]. En France, les dépenses pour l'année 2002 concernant les maladies cardiovasculaires ont représenté 13,6% des dépenses publiques de santé soit 15,3 milliards d'euros [Heijink *et al.*, 2008]. À mesure que la population vieillit, les dépenses nationales de santé devraient augmenter considérablement [Goss, 2008, Heidenreich *et al.*, 2011].

1. Chiffres OMS 2012.
2. http://ec.europa.eu/health/major_chronic_diseases/diseases/cardiovascular_en

Compte tenu de ces enjeux, de nombreux chercheurs universitaires s'intéressent à la consolidation et à l'enrichissement des connaissances médicales, mais aussi à la prévision des risques de mortalité associés aux maladies cardiovasculaires. Par exemple, [Yan *et al.*, 2005] ont évalué la valeur prédictive des scores de comorbidités de Romano, Deyo et Elixhauser à partir des données médico-économiques pour le risque de mortalité. Ils ont conclu que ces scores avaient la même valeur prédictive qu'un score établi à partir du dossier patient. [Gott *et al.*, 2007] ont étudié les trajectoires de décès dans le cas de l'insuffisance cardiaque. Ils ont comparé les courbes des scores de qualité de vie et de limitation de l'activité physique afin de constituer des groupes de patients et d'identifier des trajectoires caractéristiques du décès. [Fox *et al.*, 2006] ont mené des investigations à travers 14 pays pour construire un modèle du risque de mortalité et d'IM à 6 mois. Ce type de modèle peut s'avérer très utile dans le triage et la prise en charge des patients atteints d'un syndrome coronarien aigu. [Freemantle *et al.*, 2013] ont développé un modèle évaluant l'excès de mortalité hospitalière à partir des données médico-administratives et du score de Charlson (voir section 3.1.4 du chapitre 3). Leur modèle s'est révélé plus performant que l'indice de mortalité hospitalière.
Dans [Weintraub *et al.*, 2012], les auteurs ont validé un modèle de survie à trois ans suite à une intervention coronaire percutanée (ICP) à partir des données de registres pour des patients ayant des antécédents cardiaques. [Aylin *et al.*, 2007] ont comparé les modèles pronostiques du décès intra-hospitalier constitués soit à partir des bases administratives, soit à partir des bases cliniques (registres nationaux vasculaires et cancers). Ils ont étudié des patients ayant eu un pontage coronarien, un anévrisme de l'aorte abdominale ou une excision de polypes dans le cas d'un cancer colorectal. Ils ont obtenu des résultats similaires avec les deux types de modèles et ont démontré l'intérêt d'utiliser les bases administratives dans le cas de modèles pronostiques. À l'inverse, [Geraci *et al.*, 2005] ont élaboré des modèles plus performants dans le cas d'utilisation des données cliniques que dans le cas d'utilisation des données administratives pour prédire la mortalité après un pontage coronarien. Toutefois, ils ont démontré qu'en ajoutant quelques variables cliniques il était possible d'améliorer la performance des modèles conçus avec les données administratives. [Siregar *et al.*, 2014] ont également comparé l'utilisation des données administratives et des données cliniques pour prédire la mortalité chez les patients ayant subi une chirurgie cardiaque. Ils ont établi que la qualité du codage des données administratives avait un impact non négligeable sur la performance des modèles. Enfin, [McNamara *et al.*, 2016] ont développé un outil de surveillance du risque de décès intra-hospitalier pour des patients atteints d'IM aigu, à partir de données issues de registre.

Étant donné le nombre de patients impliqués et la quantité de données à exploiter, les chercheurs ont également utilisé des méthodes de fouille de données [Rajalakshmi et Dhenakaran, 2015], parfois combinées à des méthodes statistiques plus classiques pour étudier des motifs ou évaluer le risque de mortalité. [Austin *et al.*, 2012] ont prédit la mortalité à 30 jours pour des patients atteints d'IM aigu ou d'insuffisance cardiaque à l'aide de données issues de registres de cardiologie et de bases hospitalières. Ils ont comparé les performances des modèles : arbre de régression, RL, bagging, random forests, boosted regression trees. [Kim *et al.*, 2011]

ont utilisé dix années de données des services de soins intensifs pour prédire la mortalité. Les auteurs ont évalué des modèles basés sur les réseaux de neurones, les séparateurs à vaste marge (SVM), les arbres de décision et la RL. [Le Duff *et al.*, 2004] se sont servis des données d'interventions des secours concernant les arrêts cardiaques pour déterminer les critères favorisant la survie. À l'aide des réseaux bayésiens, ils ont démontré que la survie était liée à cinq variables : l'âge, le sexe, le rythme cardiaque d'origine, l'origine de l'insuffisance cardiaque et les techniques de réanimations employées.

D'autres auteurs se sont intéressés aux données séquentielles pour construire des modèles prédictifs. Par exemple, [Dart *et al.*, 2003] ont analysé les parcours intra-hospitaliers des patients sur un an à partir des données du PMSI. Ils ont construit un modèle de prédiction, fondé sur des règles d'association, des cheminements du patient entre les différentes unités médicales. C'est un premier pas vers la création d'un outil de gestion d'allocation de ressources au sein d'une unité médicale ou de l'ensemble des unités d'un hôpital. [Fabregue *et al.*, 2011] ont développé une technique de classification basée sur des modèles séquentiels à partir de données de puces à ADN pour prédire le grade de la tumeur dans le cas du cancer du sein. [Wright *et al.*, 2015] ont élaboré un modèle prédictif des prescriptions médicamenteuses dans le cas du diabète. Ils ont extrait les motifs fréquents d'une base séquentielle constituée de l'historique des prescriptions de médicaments des patients. Puis, ils ont établi une base de règles afin de déterminer la prochaine étape du traitement médicamenteux de ces patients.

Dans cette partie des travaux, nous souhaitons mettre en évidence les parcours de soins les plus pronostiques du décès intra-hospitalier. La première phase, décrite dans le chapitre 5, a consisté à mettre en exergue des motifs fréquents dans des sous-populations (ou contextes), définies à l'aide de covariables. Les résultats de cette fouille sont utilisés à des fins de prédiction du décès intra-hospitalier. Nous intégrons ces motifs dans les modèles prédictifs à l'aide d'un score. Ce score mesure la similarité entre la trajectoire du patient et le(s) motif(s) : plus ce score est élevé, plus le motif est présent dans le parcours du patient. Nous comparons entre-elles les méthodes prédictives les plus utilisées dans la littérature, couplées à des mesures de similarités entre chaînes de caractères, afin de déterminer le meilleur couple *(modèle, score)* suivant le contexte étudié. Nous nous sommes appuyés sur la méthode TRIPOD [3] [Collins *et al.*, 2015] pour établir notre protocole de prédiction.

Dans ce chapitre, nous présentons le processus de prédiction dans la section 6.1 qui s'articule en quatre étapes. Ensuite, nous l'appliquons à nos données. Dans la section 6.2, nous présentons les résultats pour certains contextes. Pour chaque type de trajectoires, nous identifions les parcours les plus à risque. Enfin, nous commentons les résultats obtenus dans la section 6.3 et nous concluons par des perspectives dans la section 6.4.

3. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

6.1 Protocole de prédiction

À l'aide des motifs identifiés dans le chapitre 5, nous construisons des modèles de prédiction du décès intra-hospitalier. À partir des recommandations établies pour élaborer des modèles prédictifs à des fins de pronostic ou de diagnostic (la méthode TRIPOD [Collins *et al.*, 2015]), nous avons construit les différentes étapes de notre protocole de prédiction. Ce dernier est constitué de quatre étapes résumées dans la figure 6.1. Dans l'étape 1 (section 6.1.1), nous commençons par trier les motifs de sorte à obtenir une information condensée. Ensuite, dans l'étape 2 (section 6.1.2), nous constituons la base de données. Puis, dans l'étape 3 (section 6.1.3), nous procédons à la modélisation par contexte. Cette étape se décompose elle-même en plusieurs sous-étapes, dont l'une d'elle consiste à déterminer le « meilleur » modèle correspondant à un couple *(modèle, score)* à l'aide d'indicateurs classiques de performance. Enfin, dans l'étape 4 (section 6.1.4), nous validons le modèle sélectionné dans l'étape précédente. Ces étapes sont détaillées dans la suite de cette section. Les expérimentations ont été réalisées à l'aide du logiciel R version 3.3.1 [R Core Team, 2016].

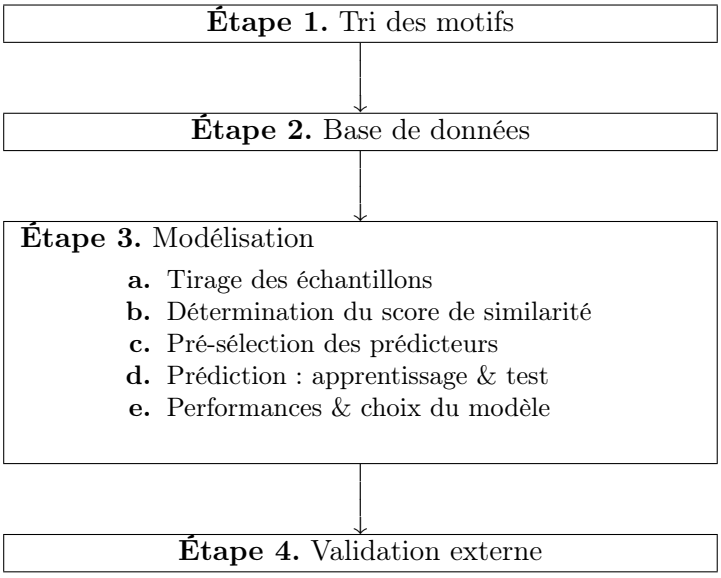


Figure 6.1 – Étapes du processus de modélisation.

6.1.1 Étape 1. Tri des motifs

Tout d'abord, nous nettoyons les données pour retirer l'information que nous souhaitons prédire, en l'occurrence, le décès. Tous les items contenant l'information Décès sont supprimés. Il s'agit des codes « Décès » mais aussi des codes GHM ou DP contenant cette information de façon intrinsèque comme le GHM 05M21 signifiant Infarctus aigu du myocarde avec décès : séjours de moins de 2 jours.

Ensuite, nous condensons l'information afin de réduire le risque d'introduire de la colinéarité dans le modèle. Nous conserverons tous les motifs maximaux [Gouda et Zaki, 2005], c'est-à-dire, qui ne sont pas inclus dans les autres, et ceci pour chaque contexte. Par exemple, si dans un contexte l'ensemble des motifs fréquents

est $M = \{\langle(A)\rangle, \langle(A)(B)(C)\rangle, \langle(A)(B)\rangle, \langle(B)(C)\rangle, \langle(B)(A)\rangle, \langle(A)(B)(D)\rangle\}$, il sera réduit à $M' = \{\langle(A)(B)(C)\rangle, \langle(B)(A)\rangle, \langle(A)(B)(D)\rangle\}$ pour être intégré dans un modèle prédictif.

À la fin de cette étape, nous produisons, par contexte, une liste de motifs maximaux à laquelle nous allons comparer les séquences constituées selon le protocole décrit dans l'étape 2.

6.1.2 Étape 2. Constitution de la base de données

À partir de la base de données décrite dans le chapitre 3, nous ajoutons deux critères de sélection afin de constituer la base de données qui sera utilisée en entrée du protocole de prédiction.

Tout d'abord, nous ne retenons que les patients ayant une trajectoire au moins de longueur quatre. Autrement dit, nous conservons les patients ayant eu au moins quatre évènements durant la période d'observation. Ceci nous permet de constituer une base de données avec des patients ayant un historique suffisant pour améliorer la capacité prédictive d'un modèle. Toutefois, ce critère de sélection réduit la base de données initiale (contenant 412 486 patients) à 5 199 patients. En outre, certains contextes, avec des effectifs trop faibles, ne sont pas retenus.

Ensuite, nous effectuons une sélection supplémentaire sur la région d'origine du patient pour effectuer, en étape 4 (section 6.1.4), une validation externe géographique [Moons *et al.*, 2012a]. La région d'origine du patient est déterminée selon les territoires médicaux de la carte des inter-régions du CeNGEPS [4] (Centre National de Gestion des Essais de Produits de Santé). Nous avons préalablement établi dans le chapitre 3 qu'il y avait un gradient Nord-Sud dans la mortalité. Afin d'éviter d'introduire ce biais de sélection dans nos modèles, nous conservons, pour la phase de modélisation, les patients originaires de toutes les régions exceptées Sud Méditerranée et Nord-Ouest. Ces dernières seront donc utilisées pour la validation externe : elles constitueront l'échantillon de validation. Nous obtenons pour la phase de modélisation un total de 3 576 patients. Nous construirons nos échantillons d'apprentissage et de test à partir de cet ensemble de patients. La séparation des données est illustrée dans la figure 6.2.

6.1.3 Étape 3. Modélisation par contexte

Nous souhaitons prédire la mortalité hospitalière suivant le parcours du patient, ainsi la variable binaire à expliquer est l'état final du patient : présumé vivant ou décédé dans un établissement de soins.
Cette étape se décompose en plusieurs sous-étapes listées dans la figure 6.1. Dans la suite de cette section, nous allons expliciter ces différentes sous-étapes. Tout d'abord, dans la section 6.1.3.1, nous expliquons de quelle façon nous avons formé les différents échantillons d'apprentissage et de test. Ensuite, dans la section 6.1.3.2, nous décrivons le calcul du score de similarité entre la trajectoire d'un patient et les motifs du

4. http://www.cengeps.fr

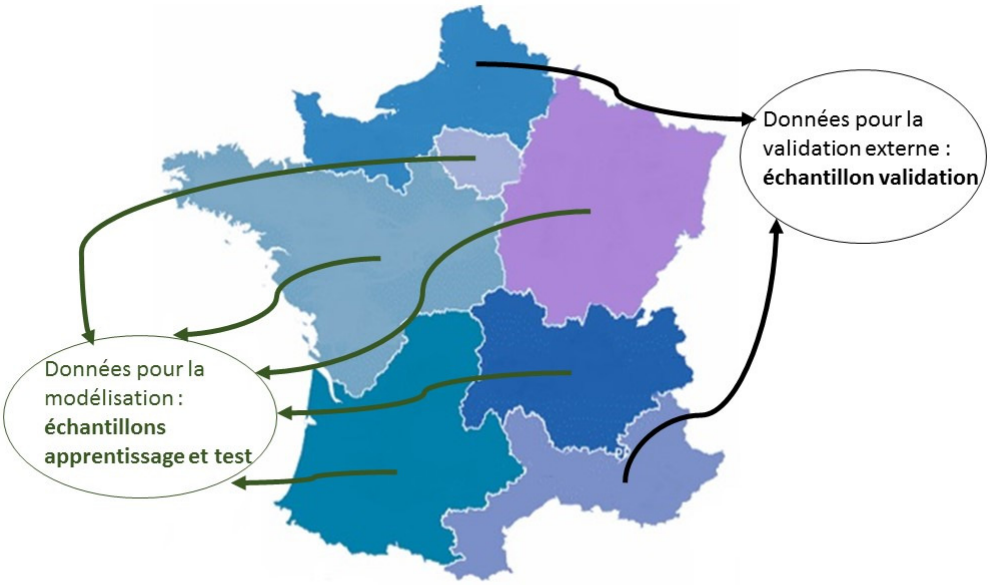


Figure 6.2 – Découpage de la base de données suivant trois échantillons : apprentissage, test et validation.

contexte concerné. Puis, nous détaillons, dans la section 6.1.3.3, la pré-sélection des prédicteurs inclus dans les modèles. Dans la section 6.1.3.4, nous présentons le principe de prédiction selon une modélisation en validation croisée avec les échantillons apprentissage et test. Enfin, dans la section 6.1.3.5, nous donnons les indicateurs de performances mesurés qui seront déterminants dans le choix du modèle.

6.1.3.1 Étape 3.a. Tirage d'échantillons

Chaque jeu de données est construit pour créer des échantillons équilibrés suivant la variable d'intérêt [Tillé, 2011]. Par cette technique, nous déterminerons des estimateurs biaisés mais robustes [Diallo, 2006]. Par ailleurs, [Bellman, 1954] a montré que des estimateurs avec biais ont dans plusieurs cas une convergence plus rapide, et donc une efficacité pratique bien plus grande.
Toutefois, les effectifs ne permettant pas toujours d'obtenir une répartition homogène, nous aurons dans certains cas des échantillons quasi-équilibrés. À noter que, la contrainte sur la longueur des séquences supprime des contextes pour lesquels il y a trop peu d'individus, ou encore pour lesquels il n'y a pas de cas de décès observé. Ainsi, parmi les contextes minimaux, nous pouvons conserver les suivants :

— Homme & +65 ans & 5-60 séjours ;

— Femme & +65 ans & 5-60 séjours ;

— Homme & +65 ans & 3-5 séjours ;

— Femme & +65 ans & 3-5 séjours ;

— Homme & 45-65 ans & 5-60 séjours ;
— Homme & 45-65 ans & 3-5 séjours.

6.1.3.2 Étape 3.b Détermination du score de similarité

Pour intégrer les motifs dans les modèles prédictifs, nous mesurons un score de similarité, entre les différents motifs et la trajectoire du patient comme défini ci-dessous. Nous rappelons (voir aussi chapitre 5) que la trajectoire du patient est une séquence de caractères (les codes GHM ou DP).

Définition 14 (Similarité entre trajectoire de patient et motif)
Pour un contexte donné c, soit $s_1^c, ..., s_k^c$, les k motifs du contexte. Soit P un patient du contexte c, avec une trajectoire T_P. Alors, le vecteur sim_P^c, de longueur k, donné par :

$$sim_P^c = (sim(T_P, s_1^c), ..., sim(T_P, s_k^c))$$

où sim est un score de similarité entre chaînes de caractères. sim_P^c mesure la similarité entre chacun des motifs du contexte c et la trajectoire T_P du patient.

Il existe un grand nombre de mesures permettant d'évaluer la similarité entre deux chaînes de caractères. Elles se divisent en trois catégories : les mesures basées sur les opérations d'édition, les mesures basées sur les q-grammes et les mesures heuristiques. Nous nous sommes interrogés sur la métrique à choisir. Généralement, le choix d'une mesure dépend de la nature de la séquence et de sa longueur. Les mesures basées sur les q-grammes sont adaptées pour des chaînes très longues contrairement aux autres [Navarro *et al.*, 2001]. Le choix d'une distance d'édition dépend du besoin de précision [Boytsov, 2011]. Par exemple, dans un cas où les différences entre les correspondances et les items d'un dictionnaire sont minces, une distance d'édition, avec plusieurs opérations d'édition, pourra donner de meilleurs résultats. En revanche, les mesures heuristiques comme celles de Jaro ont été conçues avec la dactylographie pour des chaînes relativement courtes [Bilenko *et al.*, 2003].

Au vu de ces considérations, il n'est pas aisé de déterminer quelle mesure nous devrions choisir, même si *a priori* les mesures de type q-grammes semblent plus appropriées pour les chaînes de caractères plutôt longues que nous traitons. Nous avons donc décidé d'intégrer la notion de mesure de similarité dans le choix du modèle. Nous avons retenu les mesures suivantes : la plus longue sous-chaîne commune (LCS), la distance de Levenshtein, la distance d'alignement optimal, la distance de Damerau-Levenshtein, les mesures q-gramme, Jaccard, cosinus, Jaro et Jaro-Winckler. Les définitions de l'ensemble de ces mesures sont détaillées dans l'annexe D.1. Nous avons utilisé le package R stringdist [Van der Loo, 2014] pour le calcul des scores de similarités.

6.1.3.3 Étape 3.c. Pré-sélection des prédicteurs

Dans la construction des modèles, nous prenons en compte le sexe, la classe d'âge et les scores entre la trajectoire et tous les motifs. Autrement dit, chaque motif est un prédicteur à part entière. Au sein du modèle prédictif, nous affectons un poids à chacun des motifs. De plus, l'intégration des scores est faite de deux façons : soit en variable continue, soit en variable discrétisée. Nous avons discrétisé le score selon trois classes :

— similarité faible pour un score inférieur à 0,4 ;

— similarité moyenne pour un score compris entre 0,4 et 0,6 ;

— similarité forte pour un score supérieur à 0,6.

Une étape préalable de sélection de covariables est effectuée pour le cas des variables qualitatives, en écartant celles qui n'ont qu'une seule modalité. Par exemple, dans le cas du contexte Homme & +65 ans, l'âge et le sexe ne seront pas pris en compte dans le modèle. Nous effectuons cette pré-sélection également dans le cas des scores discrétisés.

6.1.3.4 Étape 3.d. Prédiction : apprentissage & test

À l'issue de l'étape 3.c, nous avons choisi les prédicteurs. Il nous reste à implémenter différents modèles prédictifs, à partir des échantillons constitués dans l'étape 3.a pour prédire l'état final du patient : présumé vivant ou décédé dans un établissement de soins.

Modèles prédictifs. Il existe de nombreux modèles prédictifs. Nous en avons sélectionné quelques-uns parmi les plus populaires [Wu *et al.*, 2008, Howard et Bowles, 2012] afin de déterminer le meilleur couple *(modèle, score)* pour prédire la mortalité hospitalière par sous-population. Nous avons comparé les modèles suivants : la RL [Preux *et al.*, 2005], le modèle bayésien naïf (NB), le modèle des k plus proches voisins (KNN), le modèle d'arbre de régression et le modèle SVM.

Dans le cas de la RL, nous avons utilisé une méthode pas à pas (backward) suivant le critère AIC (Akaike Information Criterion) pour sélectionner les prédicteurs à l'aide du package R FWDselect [Sestelo *et al.*, 2015]. Étant donné le nombre de variables impliquées au départ dans le modèle, nous avons opté pour un modèle d'arbre à inférence conditionnelle (voir annexe D.2) afin de ne pas biaiser la sélection des prédicteurs. Ce modèle a été implémenté à l'aide des packages R party et partykit [Hothorn *et al.*, 2006]. Pour le cas des modèles SVM, nous entraînons des modèles à noyaux radial et polynomial. En effet, à l'aide d'une étude graphique des données, nous avons pu constater que le modèle à noyau linéaire n'était pas adapté. De plus, des expérimentations préalables ont montré que les modèles à noyau sigmoïde ne permettaient pas de détecter les cas de décès. Ainsi, nous avons choisi de ne pas retenir ce noyau dans cette deuxième phase d'expérimentations. Ces modèles ainsi que le modèle NB ont été implémentés à l'aide du package R e1071 [Meyer *et al.*, 2015]. Enfin, nous avons implémenté le modèle KNN à l'aide du package R FNN [Beygelzimer *et al.*, 2013].

Les modèles sont générés à partir des échantillons d'apprentissages, et ils sont validés à partir des échantillons de tests selon un principe de validation croisée.

6.1.3.5 Étape 3.e. Performances & choix du modèle

Dans cette dernière sous-étape, de la modélisation nous évaluons le pouvoir discriminant de ces modèles.

Mesures de discrimination. La discrimination est la capacité d'un modèle à séparer les individus positifs des autres, pour la variable d'intérêt (ici le décès) [Moons *et al.*, 2012b]. Il existe de nombreux indicateurs pour mesurer la discrimination d'un modèle. Nous rappelons leur définition dans l'annexe D.3. Pour comparer les modèles entre-eux, nous avons calculé l'accuracy, la précision, la sensibilité, la spécificité, la F-mesure, le taux d'erreur et l'aire sous la courbe ROC [5] (RAUC).

Pour déterminer le meilleur modèle, nous nous basons sur les mesures listées ci-dessus. Nous conservons le modèle représentant le meilleur compromis pour toutes ces mesures à l'aide du principe du vecteur maximum (voir annexe D.4) [Godfrey *et al.*, 2007].

6.1.4 Étape 4. Validation externe

C'est la dernière étape du processus. La validité d'un modèle se résume ainsi à sa capacité de simuler efficacement la réalité. Une fois le modèle sélectionné pour chaque contexte, nous procédons à sa validation externe à partir de l'échantillon validation (voir figure 6.2). Nous prenons donc la base de données constituée des patients ayant des trajectoires de longueur quatre et originaires des régions Sud Méditerranée et Nord-Ouest, soit au total 1 623 patients. Nous mesurons la discrimination à l'aide de la RAUC. Nous mesurons également la calibration [Altman *et al.*, 2009].

Mesures de calibration. La calibration est l'accord entre la probabilité de développer le résultat d'intérêt, dans un certain laps de temps, estimé par le modèle et les fréquences observées de ce résultat. Elle est idéalement évaluée graphiquement en traçant les fréquences de résultats observées par rapport aux probabilités prédites, au sein d'un sous-groupe de participants qui sont classés selon une probabilité estimée croissante. Elle est généralement accompagnée par un test de qualité d'ajustement comme le test d'Hosmer-Lemeshow, bien que ce type de test ait tendance à refléter une bonne qualité d'ajustement dû à un manque de puissance [Moons *et al.*, 2012a]. Par ailleurs, il existe également d'autres méthodes pour évaluer la calibration d'un modèle, comme par exemple, le score de Brier défini ci-dessous.

Définition 15 (Score de Brier)
Le score de Brier [Brier, 1950] est une fonction de score qui mesure l'exactitude des prédictions probabilistes. Il est applicable aux tâches dans lesquelles les prédictions doivent assigner des probabilités à un ensemble de résultats discrets mutuellement exclusifs.
L'ensemble des résultats possibles peut être de nature binaire ou catégoriel, et la

5. Receiver operating characteristic.

somme des probabilités assignées à cet ensemble de résultats doit être égale à 1 (chaque probabilité individuelle étant comprise entre 0 et 1). Il est calculé par :

$$Brier = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$$

où les p_i sont les probabilités prédites par le modèle, o_i les observations et n le nombre total d'observations.
Plus le score de Brier est bas pour un ensemble de prédictions, plus les prédictions sont calibrées. Notez que le score de Brier, dans sa formulation la plus courante, prend une valeur entre 0 et 1, puisque c'est la plus grande différence possible entre une probabilité prédite (qui doit être entre 0 et 1) et le résultat réel (dont les valeurs possibles sont uniquement 0 et 1).

Étant donné le nombre de modèles impliqués, nous utiliserons parmi ces méthodes la calibration à l'aide du score de Brier.

6.2 Expérimentations

Nous avons appliqué la procédure décrite précédemment, schématisée dans la figure 6.1. Ainsi, nous avons pris en compte 6 modèles avec 9 mesures de similarité (voir tableau 6.1) et 2 types de variables (scores continus ou discrétisés) : soit 108 modèles pour chaque contexte.

Tableau 6.1 – Récapitulatif des modèles et mesures de similarités utilisés pour la modélisation.

Modèles	**Mesures de similarités**
NB	LCS
KNN	Distance de Levenshtein (lv)
Arbre	Distance d'alignement optimal (osa)
RL	Distance de Damerau-Levenshtein (dl)
SVM à noyau radial	Mesure q-gramme (qgram)
SVM à noyau polynomial	Mesure de Jaccard (jaccard)
	Mesure cosinus (cosine)
	Mesure de Jaro (jaro)
	Mesure de Jaro-Winckler (jw)

Nous présentons, dans la section 6.2.1, les résultats concernant le choix du modèle (étape 3). Ensuite, dans la section 6.2.2, nous exposons les résultats de la validation externe (étape 4) de ces modèles. Enfin, dans la section 6.2.3, nous identifions les parcours les plus à risque de mortalité hospitalière à l'aide des modèles sélectionnés et validés précédemment.

6.2.1 Étape 3.e. Choix du modèle

Tout d'abord, nous détaillons les résultats obtenus pour deux contextes celui des patients ayant 3-5 séjours et celui des patients ayant 5-60 séjours, en nous intéressant aux types de couple *(modèle, score)* retenus dans cette sélection et à leurs performances. Dans un second temps, nous généralisons ces résultats à l'ensemble des contextes.

Le tableau 6.2 recense les modèles sélectionnés avec la méthode décrite dans la section 6.1.3.5. Pour les contextes 3-5 séjours et 5-60 séjours, les modèles retenus sont ceux de la RL, en majorité couplés à une distance d'édition et avec des scores non discrétisés. Les distances osa, lv et dl ont permis de construire des modèles avec des performances identiques pour les contextes 5-60 séjours et 3-5 séjours. Les taux d'erreurs sont beaucoup plus élevés pour les modèles du contexte 5-60 séjours avec des valeurs de l'ordre de 30% *vs* 9% pour le contexte 3-5 séjours. Globalement, les indicateurs de discrimination sont meilleurs pour les modèles du contexte 3-5 séjours.

Tableau 6.2 – Performances des meilleurs modèles retenus pour les contextes 3-5 séjours et 5-60 séjours.

Traj	Contexte	Eff	Nbd	Modèle	Score	Sim	Acc	TErr	Spec	Fmes	RAUC
GHM	5-60 séjours	2 980	362	RL	continu	osa	0,70	0,30	0,83	0,66	0,70
						lv	0,70	0,30	0,83	0,66	0,70
						dl	0,70	0,30	0,83	0,66	0,70
	3-5 séjours	593	95	**RL**	**continu**	**osa**	**0,92**	**0,08**	**0,95**	**0,95**	**0,92**
						lv	**0,92**	**0,08**	**0,95**	**0,95**	**0,92**
						dl	**0,92**	**0,08**	**0,95**	**0,95**	**0,92**
DP	5-60 séjours	2 980	362	RL	continu	qgram	0,71	0,29	0,80	0,69	0,71
	3-5 séjours	593	95	**RL**	**continu**	**lcs**	**0,91**	**0,09**	**0,95**	**0,91**	**0,91**

Traj : type de trajectoire étudiée GHM ou DP ; **Contexte** : libellé du contexte étudié ;
Eff : nombre de patients dans le contexte ; **Nbd** : nombre de décès observés dans le contexte ;
Modèle : Type de modèle ; **Score** : Type de score continu ou discrétisé ; **Sim** : mesure de similarité ;
Acc : Accuracy ; **TErr** : Taux d'erreur ; **Spec** : Spécificité ; **Fmes** : F-mesure ; **RAUC** : Area Under Roc Curve.
Pour chaque contexte, les modèles les plus performants apparaissent en gras.

Si l'on considère l'ensemble des contextes, comme pour les deux contextes particuliers étudiés, c'est en majorité la RL couplée à une distance d'édition qui donne les meilleures performances. De plus, les modèles avec les scores en variables continues ont également les meilleures performances. Toutefois, dans le cas des trajectoires de DP, les modèles avec des mesures basées sur les q-grammes pour des contextes correspondant aux classes d'âge 45-65 ans ou +65 ans donnent les meilleurs résultats. Nous trouvons 9 modèles avec des mesures heuristiques associées à la RL ou au modèle SVM avec des variables discrètes et ceci essentiellement pour des contextes avec la classe d'âge 45-65 ans.

Concernant la performance de ces modèles, la RAUC varie entre 0,6 et 0,98 pour les trajectoires de GHM et entre 0,64 et 0,93 pour les trajectoires de DP. Selon ce critère, les modèles les plus performants sont celui du contexte Homme & 45-65 ans & 3-5 séjours, pour les trajectoires de GHM, et celui du contexte 45-65 ans & 3-5 séjours, pour les trajectoires de DP. *A contrario*, les modèles les moins performants sont celui du contexte 45-65 ans & 3-5 séjours, pour les trajectoires de GHM et celui

du contexte Homme & 45-65 ans & 5-60 séjours, pour les trajectoires de DP.
De même, le taux d'erreur varie entre 3% (contexte Homme & 45-65 ans) et 32% (contexte général) pour les trajectoires de GHM, entre 7% (45-65 ans & 3-5 séjours) et 35% (+65 ans) pour les trajectoires de DP.

6.2.2 Étape 5. Validation externe

Dans cette section, nous examinons les résultats obtenus en termes de performance pour chacun des modèles. De la même manière que précédemment, nous considérons tout d'abord les contextes 3-5 séjours et 5-60 séjours. Ensuite, nous détaillons des résultats en examinant l'ensemble des contextes.

Pour le contexte 3-5 séjours, les modèles avec les trajectoires de GHM sont mieux calibrés que celui avec trajectoire de DP avec un score de Brier égal à 0,04 *vs* 0,05. En revanche, le modèle avec les trajectoires de DP a une meilleure discrimination (RAUC=0,95) que ceux des trajectoires de GHM (RAUC=0,88).
Pour le contexte 5-60 séjours c'est le contraire. Les modèles avec trajectoires de GHM sont plus discriminants et moins bien calibrés que ceux des trajectoires de DP. Nous remarquons, comme pour l'étape 4 (voir tableau 6.2), que dans le cas des trajectoires de GHM, quelle que soit la distance, la RL offre des performances équivalentes dans le cas de la validation externe.

Tableau 6.3 – Validation externe des modèles retenus dans l'étape 4.

Traj	Contexte	Eff	Nbd	Modèle	Score	Sim	RAUC	Brier
GHM	5-60 séjours	1 488	200	RL	continu	osa	**0,68**	0,19
						lv	**0,68**	0,19
						dl	**0,68**	0,19
	3-5 séjours	231	36	RL	continu	osa	0,88	**0,04**
						lv	0,88	**0,04**
						dl	0,88	**0,04**
DP	5-60 séjours	1 488	200	RL	continu	qgram	0,65	**0,18**
	3-5 séjours	231	36	RL	continu	lcs	**0,95**	0,05

Traj : type de trajectoire étudiée GHM ou DP ;
Contexte : libellé du contexte étudié ; **Eff** : nombre de patients dans le contexte ;
Nbd : nombre de décès observés dans le contexte ; **Modèle** : Type de modèle ;
Score : Type de score continu ou discrétisé ; **Sim** : mesure de similarité ;
RAUC : Area Under Roc Curve ; **Brier** : Score de Brier.
Pour chaque contexte, les meilleures performances apparaissent en gras.

De manière plus générale, la RAUC varie entre 0,35 et 0,99 pour les trajectoires de GHM et entre 0,5 et 0,9 pour les trajectoires de DP. La RAUC est la plus élevée pour le modèle du contexte Homme & 45-65 ans & 3-5 séjours pour les deux types de trajectoires. Le moins discriminant dans le cas des trajectoires de GHM est celui pour le contexte Homme & +65 ans & 3-5 séjours et dans le cas des trajectoires de DP, il s'agit du modèle pour le contexte Femme & 45-65 ans.

De même, le score de Brier varie entre 0,04 et 0,4 pour les trajectoires de GHM et entre 0,04 et 0,9 pour les trajectoires de DP. Les modèles les mieux calibrés, selon le score de Brier, sont ceux pour le contexte 3-5 séjours pour les deux types de trajectoires. En revanche, les modèles les moins bien calibrés sont celui pour le contexte Homme dans le cas des trajectoires de GHM et celui pour le contexte Femme & 45-65 ans pour le cas des trajectoires de DP.

6.2.3 Identification des parcours à risque

Notre deuxième objectif était d'identifier les parcours les plus à risque de la mortalité hospitalière. Nous allons y répondre en analysant plus en détails l'influence des variables impliquées tout d'abord dans les modèles pour les trajectoires de GHM, puis dans ceux pour les trajectoires de DP à la fois pour les contextes 5-60 séjours et 3-5 séjours.

Les résultats d'une RL sont interprétés en termes de facteurs de risque si l'odds-ratio (OR) est supérieur à 1 et que son intervalle de confiance ne contient pas la valeur 1, de facteurs protecteurs si l'OR est inférieur à 1 et que son intervalle de confiance ne contient pas la valeur 1, ou encore absence d'association entre l'évènement d'intérêt et la variable si l'OR est égal à 1 [Ulrike, 2010].

Dans le cas des trajectoires de GHM, nous avons noté plusieurs modèles de RL avec les distances osa, lv et dl ayant des performances identiques. Nous présentons donc un seul modèle par contexte dans le tableau 6.4.

Tableau 6.4 – Modèles logistiques pour les trajectoires de GHM.

	Modèle 5-60 séjours		
Variables	**Coeff**	**OR**	**IC 95%**
Constante	-3,13***		
Scores de :			
⟨**(05M04)(05K06)**⟩	9,63**	1,52e+04	38,83 à 9,37e+06
⟨**(05K10)**⟩	43,08***	5,12e+18	5,13e+09 à 2,56e+28
⟨**(05K06)(05K06)(05K06)**⟩	-8,34***	2,39e-04	5,33e-06 à 0,009
⟨**(05K13)**⟩	-20,87*	8,62e-10	4,82e-17 à 0,007

	Modèle 3-5 séjours		
Variables	**Coeff**	**OR**	**IC 95%**
Constante	-18,26***		
Classe d'âge			
45-65 ans	-2,26**	0,1	0,009 à 0,57
-45 ans	-2,96	0,05	3,89e-04 à 1,54
Scores de :			
⟨**(05K10)**⟩	137,38***	4,62e+59	7,37e+32 à 2,33e+101
⟨**(05K13)**⟩	-66,74**	1,03e-29	8,37e-56 à 6,41e-10
⟨**(05M04)**⟩	15,23	4,13e+06	4,24e-01 à 7,31e+14

Variables : variables retenues dans le modèle ;
Coeff : valeur des β_i du modèle
et test de nullité des β_i avec *p<0,05 ; **p<0,01 ; ***p<0,001 ;
OR : odds-ratio ; **IC 95%** : intervalle de confiance à 95% des OR.

Après examen de la première partie du tableau 6.4, nous identifions les motifs ⟨(05M04)(05K06)⟩ (IM aigu suivi de pose de stent) et ⟨(05K10)⟩ (Actes diagnostiques par voie vasculaire) comme étant des parcours augmentant le risque du décès intra-hospitalier pour le contexte 5-60 séjours. Les motifs ⟨(05K06)(05K06)(05K06)⟩ (3 séjours pour pose de stent) et ⟨(05K13)⟩ (Actes thérapeutiques par voie vasculaire) sont, au contraire, des facteurs protecteurs.
Pour le contexte 3-5 séjours, les résultats sont similaires. Dans la deuxième partie du tableau 6.4, les motifs ⟨(05M04)⟩ et ⟨(05K10)⟩ sont des facteurs de risque du décès alors que le motif ⟨(05K13)⟩ est identifié comme un facteur protecteur. Par ailleurs, l'examen des résultats concernant les classes d'âge indique une augmentation du risque pour la classe des +65 ans.

Tableau 6.5 – Modèles logistiques pour les trajectoires de DP.

		Modèle 5-60 séjours	
Variables	**Coeff**	**OR**	**IC 95%**
Constante	0,45		
Classe d'âge			
45-65 ans	-0,68*	0,5	0,29 à 0,87
-45 ans	-1,47*	0,26	0,05 à 0,78
Scores de :			
⟨**(I21)(I21)**⟩	15,5***	5,42e+06	3,36e+03 à 1,30e+10
⟨**(I20)(I21)**⟩	25,07***	7,7e+10	2,37e+06 à 9,26e+15
⟨**(I25)(I25)**⟩	-12,35***	4,32e-06	4,28e-09 à 3,24e-03
⟨**(I25)(I20)**⟩	19,02***	1,81e+08	1,92e+04 à 2,71e+12
⟨**(I20)(I25)**⟩	-14,26**	6,42e-07	1,13e-11 à 1,88e-02
⟨**(I21)(I20)**⟩	-23,83***	4,47e-11	1,09e-15 à 9,84e-01
⟨**(I20)(I20)**⟩	-6,97*	9,43e-04	1,63e-06 à 4,49e-01

		Modèle 3-5 séjours	
Variables	**Coeff**	**OR**	**IC 95%**
Constante	-30,7***		
Classe d'âge			
45-65 ans	-1,97**	0,14	0,009 à 1,1
-45 ans	-2,92*	0,05	0,0004 à 1,36
Scores de :			
⟨**(I21)(I21)**⟩	30,02***	1,09e+13	4,47e+06 à 3,1e+21
⟨**(I24)**⟩	28,78**	3,16e+12	5,74e+04 à 6,9e+21
⟨**(I25)(I25)**⟩	20,42*	7,4e+08	1,86e+02 à 1,94e+17
⟨**(I20)(I20)**⟩	12,89*	3,95e+05	6,77 à 5,79e+11

Variables : variables retenues dans le modèle ;
Coeff : valeur des β_i du modèle
et test de nullité des β_i avec *p<0,05 ; **p<0,01 ; ***p<0,001 ;
OR : odds-ratio ; **IC 95%** : intervalle de confiance à 95% des OR.

Nous examinons maintenant l'influence des variables dans le cas des trajectoires de DP. De la même manière que précédemment, nous nous focalisons tout d'abord sur le contexte 5-60 séjours. Dans la partie supérieure du tableau 6.5, les motifs ⟨(I21)(I21)⟩ (double IM aigu), ⟨(I20)(I21)⟩ (Angine de poitrine - IM aigu) et ⟨(I25)(I20)⟩ (Cardiopathie ischémique chronique - Angine de poitrine) sont identifiés comme favorisant une augmentation du risque, tandis que les autres motifs présents dans le tableau (⟨(I25)(I25)⟩, ⟨(I20)(I25)⟩, ⟨(I21)(I20)⟩, ⟨(I20)(I20)⟩) ont

un effet contraire. Par ailleurs, les valeurs des OR pour les classes d'âge comparées à la classe de référence (+65 ans), suggèrent une augmentation du risque pour la classe de référence.

Pour le contexte 3-5 séjours, nous examinons la partie inférieure du tableau 6.5. Le risque de décès est en augmentation pour les motifs suivants : $\langle$(I21)(I21)$\rangle$, $\langle$(I24)$\rangle$ (Autres cardiopathies), $\langle$(I25)(I25)$\rangle$ et $\langle$(I20)(I20)$\rangle$. Concernant l'âge, nous arrivons aux mêmes conclusions que pour le contexte 5-60 séjours.

6.3 Discussion

Dans cette section, nous revenons sur les objectifs de ce chapitre à savoir : 1) déterminer le meilleur couple *(modèle, score)*, dans la section 6.3.1 ; 2) identifier les parcours à risque du décès intra-hospitalier, dans la section 6.3.2. Enfin, dans la section 6.3.3, nous comparons nos résultats à d'autres du domaine en termes de performances.

6.3.1 Détermination du meilleur couple *(modèle, score)*

Tous les modèles appliqués dans nos expérimentations présentent des avantages et des inconvénients :

— la RL n'exige pas d'hypothèse sur la distribution des variables, en revanche, elle est sensible à la multicolinéarité ;

— le modèle NB, bien que s'appuyant sur une hypothèse simple qui selon [Domingos et Pazzani, 1997] est trop contraignante, se révèle robuste et efficace. Toutefois, tout comme la RL, il est sensible à la multicolinéarité ;

— KNN est un algorithme populaire de fait de sa simplicité et de sa vitesse de convergence. Cependant, KNN nécessite une grande capacité mémorielle. Ainsi, lorsque l'échantillon est de très grande taille, le temps de calcul est également important [Alpaydin, 1997] ;

— la modélisation par arbre présente l'avantage de la facilité d'interprétation, mais l'inconvénient majeur est le sur-apprentissage ;

— l'algorithme SVM est également très populaire, en particulier en classification de textes. Il présente une bonne précision et des garanties théoriques vis-à-vis du sur-ajustement. Avec le noyau approprié, il est possible d'obtenir de bons résultats même si les données ne sont pas linéairement séparables dans l'espace des prédicteurs. Cependant, déterminer le bon noyau peut être un vrai challenge. De plus, les résultats ou sorties sont difficiles à interpréter. Par ailleurs, ce type de modèle est très gourmand en mémoire en particulier pour l'ajustement et la détermination des bonnes constantes du modèle.

La comparaison des différents modèles montre que la RL couplée à une distance d'édition obtient les meilleures performances dans la plupart des contextes. De même [Austin, 2007] a comparé les modèles généralisés additifs (GAM), les modèles de régression par splines adaptatifs multivariés (MARS) et la RL dans la prédiction de la mortalité à 30 jours après un IM aigu. Austin en a conclu que la RL avait des

performances comparables aux GAM et MARS.
De plus, ce sont les modèles avec les scores en variables continues qui donnent les meilleures performances. En effet, la discrétisation implique une perte d'information. Sur ces points nous rejoignons les conclusions de [Steyerberg *et al.*, 2001] qui ont comparé diverses stratégies de modélisation, dans le cas de petits échantillons, pour prédire la mortalité à 30 jours des patients ayant eu un IM aigu.
En outre, nous avons remarqué que dans le cas des trajectoires de DP nous avons plus de modèles avec des mesures de type q-gramme et ceci plus particulièrement pour des contextes concernant les classes d'âge +65 ans et 45-65 ans. Si l'on compare par rapport aux trajectoires de GHM, ce résultat est alors lié à la longueur des séquences. En effet, de façon intrinsèque le codage est plus court dans celles des DP (code sur 3 caractères) que dans celles des GHM (code sur 5 caractères). Ainsi, avec un même nombre d'évènements, les chaînes de caractères sont plus courtes dans le cas des trajectoires de DP. Par ailleurs, le seul cas où nous obtenons un modèle avec une mesure q-gramme pour les trajectoires de GHM est le contexte des femmes avec 5-60 séjours. Or, de manière générale, les femmes ont des trajectoires plus courtes (dans notre sélection des données). En revanche, si l'on compare suivant le contexte de la classe d'âge, alors ce résultat n'est pas lié à la longueur des trajectoires car les trajectoires les plus courtes sont celles de la classe d'âge des -45 ans.

Finalement, une fois le couple *(modèle, score)* choisi, nous pouvons ensuite différencier les parcours à risque des parcours protecteurs. Nous synthétisons ces résultats dans la section suivante.

6.3.2 Identification des parcours à risque

La modélisation du décès à l'aide des motifs fréquents, extraits dans le chapitre 5, permet de distinguer les évènements hospitaliers favorisant une augmentation du risque de décès de ceux qui, au contraire, ont un effet protecteur.

Pour résumer, d'après les résultats de la section 6.2.3, les motifs préservant du décès, dans le cas des trajectoires de GHM, sont les actes thérapeutiques et le parcours IM aigu suivi d'une pose de stent. Ceci atteste de l'efficacité de la prise en charge [Cambou *et al.*, 2004, Falconnet *et al.*, 2009] avec un suivi des soins de la dilatation artérielle par divers moyens : endoprothèse, angioplastie... associés à un traitement médicamenteux [Danchin *et al.*, 2003]. En revanche, suivant l'état de gravité de la pathologie, un acte exploratoire, comme une artériographie ou une coronarographie, représentant déjà un risque pour le patient (comme tout acte invasif), favorisera d'autant plus une augmentation du risque [Mottier et Baba-Ahmed, 2006]. Ceci explique la présence de $\langle(05K10)\rangle$ dans les motifs à risque.

Si nous synthétisons les résultats établis dans le cas des trajectoires de DP, nous remarquons que le risque de décès est lié au profil d'évolution de la pathologie. Par exemple, un parcours comme angine de poitrine conduisant plus tard à un IM aigu induira une augmentation du risque. Tandis que le parcours IM aigu suivi plus tard par une angine de poitrine aura un effet contraire. Cela montre que le

suivi régulier d'un patient atteint d'IM est primordial. Toutefois, nous avons vu que certains parcours comme une double cardiopathie ischémique chronique ou une double angine de poitrine, n'ont pas le même effet suivant le contexte. Ils ont un effet délétère dans le contexte des 3-5 séjours, alors que dans le contexte des 5-60 séjours, ils ont un effet contraire. Diverses hypothèses peuvent être avancées pour expliquer ces résultats. Tout d'abord, suivant l'état de gravité de l'IM ou encore si l'état général du réseau coronarien du patient est fortement altéré, au dessus de tout recours, le pronostic est fatal [Nakache *et al.*, 1977]. D'autres éléments peuvent également être déterminants sur le pronostic, comme la réponse au traitement : certaines thérapies médicamenteuses s'avèrent plus efficaces que d'autres ou conviennent mieux à certains patients qu'à d'autres [Scheen, 2006]. Par ailleurs, de nombreux patients ne bénéficient pas de la réadaptation cardiaque. Il s'agit de la prise en charge globale du patient et de ses facteurs de risque. Or les bénéfices de l'entraînement physique sur les facteurs de risque sont démontrés par de nombreux travaux [Ghannem, 2010]. Enfin, les complications liées à l'IM (e.g l'insuffisance cardiaque, l'IM, la rupture myocardique...) sont responsables du décès du patient [Liozon *et al.*, 1992, Laissy *et al.*, 2004, Abitbol, 2005].

Ainsi, les motifs fréquents identifiés dans le chapitre 5 intégrés dans un modèle prédictif du décès intra-hospitalier, nous ont permis de mettre en évidence des parcours à risque dans la trajectoire du patient. Ils viennent souligner l'importance du suivi des patients atteints de cette pathologie sur une période d'un an voire au-delà, comme en atteste d'ailleurs la littérature sur ce sujet [DeBusk, 1994, Neff, 2004, Thygesen *et al.*, 2012, Oliver, 2014]. Il reste à savoir si cette approche est compétitive par rapport à d'autres. Nous allons aborder la question dans la section suivante.

6.3.3 Compétitivité de notre approche

Une étude comparative [Siontis *et al.*, 2012] des travaux de modélisation du risque de la mortalité, dans le cas de maladies cardiovasculaires, réalisés à partir de données cliniques, montre que les performances selon la RAUC varient de 0,71 à 0,88. Or nous obtenons dans le cas de la sélection des meilleurs modèles des performances variant de 0,6 à 0,98. Nos modèles ont donc des performances comparables à ceux évoqués plus haut. Par ailleurs, nous constatons que les résultats sont meilleurs pour des contextes à faibles effectifs. En effet, dans ces contextes, l'échantillonnage arrive à recouvrir plus de situations diverses. Ainsi, les échantillons sont plus représentatifs de la population et de fait les modèles ont de meilleures performances.

Toutefois, nous pourrions affiner nos résultats en intervenant sur différentes étapes de notre protocole. Pour commencer, dans l'étape 3 (voir 6.1.3), pour les modèles SVM, il existe diverses techniques pour la sélection des prédicteurs (features selection), telles que les méthodes enveloppantes [Guyon et Elisseeff, 2003] ou les méthodes filtrantes [Claeskens *et al.*, 2008]. Les deux approches ont des avantages et des inconvénients [Saeys *et al.*, 2007]. Les méthodes filtrantes sont habituellement plus efficaces d'un point de vue calculatoire que les méthodes enveloppantes, mais le critère de sélection n'est pas lié à l'efficience du modèle. Ainsi, la plupart des méthodes filtrantes évalue chaque prédicteur de manière

indépendante. Par conséquent, des prédicteurs fortement corrélés peuvent être sélectionnés et les interactions entre les variables ne sont pas quantifiables. Le désagrément des méthodes enveloppantes est que beaucoup de modèles sont évalués ce qui augmente le temps de calcul. Il y a aussi un risque accru de sur-ajustement avec les méthodes enveloppantes. Pour finir, à partir de l'étape 3.e (voir 6.1.3.5) du choix de modèle, il est également possible d'ajuster les paramètres du modèle avec des algorithmes d'optimisation [Marquardt, 1963, Foresee et Hagan, 1997] pour améliorer la qualité prédictive.

Néanmoins, ces résultats sont à nuancer du fait de l'incertitude liée au codage du PMSI (voir section 2.4.2 du chapitre 2). Dans d'autres domaines, des études [Chantry *et al.*, 2012, Grammatico, 2014] ont montré qu'il y avait des variations dans la façon de coder les séjours. Dans une étude de 2002, [Holstein *et al.*, 2002] ont comparé la qualité du codage, dans des services de cardiologie, sur l'impact de la valorisation des séjours et le détail des morbidités. Cette étude montre que les morbidités sont sous-renseignées lorsque les séjours sont codés par le clinicien par rapport à un médecin DIM. De plus, elle montre que dans 15% des cas il y a un changement de GHM dans la phase de recodage par le médecin DIM. Il serait intéressant de pouvoir compléter ces résultats par des études plus récentes dans le cas particulier de la cardiologie. Ceci nous permettrait d'actualiser ce pourcentage d'incertitude émis de 15%.

6.4 Conclusion

En utilisant les motifs séquentiels, extraits dans le chapitre 5, nous avons élaboré des modèles pour prédire le décès au sein d'un établissement de santé. Ces motifs ont été intégrés dans des scores en mesurant la similarité entre la trajectoire du patient et les motifs. Pour choisir le score nous avons mis en concurrence les mesures entre chaînes de caractères les plus connues. Nous avons construit un protocole de prédiction qui s'articule en plusieurs étapes en nous appuyant sur la méthode TRIPOD. Nous avons introduit les modèles prédictifs les plus couramment employés afin de les comparer. *In fine*, notre objectif était double : 1) déterminer le couple *(modèle, score)* ayant les meilleures performances pour chacun des contextes ; 2) identifier les motifs favorisant une augmentation du risque de décès.

Notre avons atteint notre premier objectif. Il résulte de la comparaison entre les différents modèles que la RL couplée à une distance d'édition est le modèle offrant les meilleures performances avec la conservation des scores de similarités en variables continues. D'autres perspectives de comparaisons et de modélisation sont envisageables à l'aide des modèles de survie tels que Cox [Timsit *et al.*, 2005]. Nous pourrions, par ailleurs, appréhender cette problématique à l'aide de modèles prédictifs basés sur les séquences [Rudin *et al.*, 2011].

Nous avons également atteint notre deuxième objectif en distinguant les motifs présentant un risque accru de décès. Ces motifs difficiles à interpréter par des experts médicaux, s'avèrent, en revanche utiles pour prédire le risque de mortalité hospitalière d'un patient. Nous retenons de nos expérimentations, présentées dans la section 6.2, que le risque de décès est fortement influencé par le profil d'évolution de la maladie et le suivi du patient après IM. Ceci témoigne de l'importance des recommandations de la société française de cardiologie [Delahaye *et al.*, 2001] sur la surveillance régulière des patients après un IM au moins durant une année. En effet, le risque de rechute et de décès est encore très élevé durant cette période et même encore au-delà.

De plus, notre approche est compétitive face à d'autres modèles prédictifs basés sur des données cliniques. Ainsi, bien que soumis à l'imprécision et l'incertitude intrinsèques à la codification du PMSI, ces résultats sont encourageants et nous incitent à proposer d'autres approches pour les améliorer. Pour affiner ces investigations, nous prévoyons d'employer d'autres types de motifs comme par exemple, les motifs obtenus avec la r-confiance [Mercadier *et al.*, 2016]. Nous souhaitons construire des modèles prédictifs en sélectionnant les parcours les plus représentatifs à la fois en fréquence, en taille et en confiance. Une autre approche pourrait également être envisagée, comme évoquée dans le chapitre 5, en tenant compte de l'état final du patient à la fin de la période d'observation dans la répartition des contextes. Cette approche consisterait alors à mettre en évidence des motifs qui soient spécifiques du décès. Ainsi, nous pourrions soit intégrer ces motifs dans notre protocole de prédiction, soit mettre en œuvre une méthode de classification des patients à partir de leur trajectoire comme dans [Fabregue *et al.*, 2011] afin de prédire le décès.

Partie IV

Des trajectoires de patients à la planification sanitaire

Deux prisonniers, l'un voit les barreaux de la prison,

et l'autre les étoiles.

Verlaine.

Table des matières

7 Extraction de motifs spatio-temporels **127**
7.1 Motifs spatio-temporels 129
7.1.1 Définitions préliminaires 130
7.1.2 Description du processus de fouille 132
7.2 Expérimentations 133
7.2.1 Extraction de motifs 134
7.2.2 Visualisation des flux de patients 137
7.3 Discussion 141
7.4 Conclusion 144

8 Profils de délais et de tarifs **147**
8.1 Classification de données longitudinales quantitatives 148
8.1.1 Définitions préliminaires 149
8.1.2 Description du processus de classification 152
8.2 Expérimentations 154
8.2.1 Trajectoires de délais inter-séjours 154
8.2.2 Trajectoires de tarifs de séjours 157
8.3 Discussion 159
8.3.1 Analyse des résultats 159
8.3.2 Limites de cette étude 161
8.4 Conclusion 162

Chapitre

7 Extraction de motifs spatio-temporels

Dans la partie III précédente, nous avons montré, qu'à partir des données issues du PMSI, nous pouvions extraire des motifs ayant une interprétation médicale puis, les utiliser pour prédire le décès à l'hôpital. Dans cette nouvelle partie, nous étudions les comportements de groupes de patients via la notion de flux que nous analyserons au travers des divers évènements hospitaliers. Les enjeux associés à la connaissance des flux de patients à divers niveaux sont décisifs [Jun *et al.*, 1999] : 1) pour prévoir les admissions de patients ; 2) pour adapter le parcours du patient et envisager les schémas de flux ; 3) pour gérer la disponibilité des ressources. Nous montrerons également que les données du PMSI sont pertinentes pour étudier ces flux comme elles l'ont été pour les aspects de prédiction.

La modélisation des flux de patients a pour objectif de mieux gérer l'allocation de ressources dans le cas d'attribution de lits [Chase, 2005], la planification d'opérations de chirurgie [Gallivan, 2005] ou encore la répartition des ressources entre établissements [Gunes et Yaman, 2005]. De manière générale, elle est utilisée pour gérer l'organisation des soins. Par exemple, [Biffl *et al.*, 2001] ont étudié les trajectoires de soins dans le cas d'une fracture du bassin. Ils ont identifié 5 étapes clés dans les parcours de soins. Dans un premier temps, ceci a permis d'organiser la gestion des flux de patients par l'équipe médicale. Le renfort de cette équipe par deux spécialistes en orthopédie a apporté de nouvelles techniques dans la pratique des soins. En conséquence, les flux de patients ont été modifiés. La prise en charge s'en est trouvée améliorée et la survie du patient également. [Broyles *et al.*, 2010] ont proposé un modèle de prédiction de la variation des flux de patients basé sur un modèle markovien. La connaissance des flux de patients entrant/sortant d'une unité de soins, à un instant donné, a permis de mieux planifier l'organisation des soins. [Pagnoni *et al.*, 2001] ont proposé un outil, nommé DoMiner, pour analyser les flux de patients. Basé sur la théorie des ensembles approximatifs (Rough set theory), DoMiner extrait des clusters d'informations transformés ensuite en règles d'association à partir desquelles sont bâties des règles de décision. DoMiner a été appliqué pour l'étude des flux de patients en provenance des urgences vers des

unités médicales. Ces travaux sont précurseurs dans la création d'un outil d'aide à la décision pour le triage des patients. Parallèlement, [Dart *et al.*, 2003] ont construit un modèle prédictif basé sur des règles de décision pour donner la tendance des flux de patients au travers des différentes unités médicales.
[Defossez *et al.*, 2014] ont représenté les différents parcours de soins, dans le cas du cancer du sein, dans le but d'identifier les différentes étapes de la prise en charge de cette pathologie. Cette représentation des flux de patients a ensuite servi de base dans l'élaboration de processus de soins intégrés. Par la suite, [Thompson *et al.*, 2015] ont également synthétisé les informations concernant les différentes étapes du traitement avec les acteurs médicaux rencontrés, dans le cas du cancer du sein. Cela leur a permis de mieux appréhender les flux de patients mais aussi, de mieux comprendre leurs décisions. Pour cette même pathologie, [Jay *et al.*, 2006] ont extrait et représenté les flux de patients entre les différents établissements de la région Lorraine par un treillis iceberg [1] [Stumme *et al.*, 2002] en s'appuyant sur des outils de l'analyse de concepts formels. Ainsi, ils ont distingué le réseau d'établissements travaillant en étroite collaboration des établissements plus isolés. Ces travaux ouvrent alors des perspectives dans la gestion de l'allocation de ressources inter-établissements.
L'étude des flux de patients permet également de comparer les différences de prises en charge entre établissements. Par exemple, [Suriadi *et al.*, 2014] ont comparé les flux de patients dans leur prise en charge aux urgences jusqu'à leur retour au domicile ou leur hospitalisation dans une unité médicale. Cette comparaison a été faite pour quatre hôpitaux différents afin de déterminer les variations dans les pratiques cliniques mais aussi leurs points communs. Ils ont mis en évidence des différences clés entre les divers parcours de prise en charge, mais surtout ils ont souligné l'intérêt de mieux comprendre l'origine des délais d'attente. Une autre façon d'analyser les prises en charge est proposée dans [Jay *et al.*, 2008]. À partir des données du PMSI, les auteurs ont mis en évidence la colonne vertébrale du réseau santé de la région Lorraine. De cette façon, il est plus aisé de distinguer les établissements qui sont le plus souvent en interaction de ceux qui, au contraire, sont plus isolés.
Enfin, modéliser les flux de patients dans le cas d'une pathologie permet d'en connaître les évolutions possibles [Harper, 2005]. Cette connaissance favorise ensuite l'adaptation des soins suivant les véritables besoins des patients, comme dans le cas des maladies chroniques [Wagner *et al.*, 2001].

Dans ce chapitre, nous explorons les divers parcours hospitaliers par contexte (voir chapitre 5) dans le but de mettre en évidence des phénomènes de groupes. Pour cela, nous nous intéressons aux motifs spatio-temporels [Compieta *et al.*, 2007, Benkert *et al.*, 2008]. Dans notre cas, la proximité spatiale, telle que définie dans les motifs initiaux de la littérature, est ici remplacée par une proximité liant les types d'évènements (ou hospitalisations) des patients. De plus, le temps considéré est lié à l'événement. Cette étude vise à caractériser les schémas de flux dans le cas de l'IM. Ainsi, nous pourrons comparer nos résultats avec les recommandations de

1. Il s'agit d'une méthode de classification hiérarchique.

l'HAS en termes de parcours de soins. Par ailleurs, l'analyse de ces schémas de flux permettra d'identifier les étapes clés dans les parcours hospitaliers de ces patients et alors de mettre en exergue tout ou partie des évolutions possibles de cette pathologie.

Dans la section 7.1, nous présentons les définitions générales du domaine puis, nous décrivons le processus d'extraction de connaissances mis en place, en particulier la transformation de la base de données utilisée en entrée de l'algorithme de fouille de données. Nous définissons un motif spatio-temporel sous la forme d'un couple de deux ensembles : les patients et les estampilles de temps pendant lesquelles ces patients sont réunis. Nous évoquons les techniques d'extraction de motifs spatio-temporels. Enfin, nous explorons les flux des patients, identifiés dans les motifs spatio-temporels, en reconstituant une partie de leur parcours de soins. Les résultats sont décrits dans la section 7.2. Pour finir, nous analysons les résultats obtenus dans la section 7.3.

7.1 Motifs spatio-temporels

Avec l'essor des nouvelles technologies (GPS, smartphones...), la génération et le stockage des données associées aux objets mobiles s'est accrue. Divers domaines d'applications nécessitent d'exploiter ce type de données. [Si *et al.*, 2009] ont associé les flux migratoires des oiseaux à la dissémination du virus H1N1. [Melnychuk *et al.*, 2010] ont étudié le flux migratoire des saumons. Connaître les mouvements migratoires des animaux peut contribuer à la préservation de la biodiversité [Tanaka *et al.*, 2015]. Les motifs spatio-temporels [2] trouvent des applications dans d'autres domaines tels que la régulation du trafic routier [Wilson, 2008] ou encore l'adaptation des infrastructures pour faciliter les déplacements des riverains, par l'étude des parcours des habitants d'une ville [Mao *et al.*, 2016]. Ainsi, l'extraction de motifs spatio-temporels s'est popularisée et de nombreux chercheurs se sont investis sur ce sujet.

Ces motifs servent à détecter des trajectoires identiques d'objets évoluant simultanément pendant un certain intervalle de temps. Il existe plusieurs définitions : les troupeaux (flocks) [Gudmundsson et van Kreveld, 2006], les convois (convoys) [Jeung *et al.*, 2008], les essaims (swarms) [Li *et al.*, 2010a], les essaims clos (closed swarms) [Li *et al.*, 2010b], les motifs de groupes (group pattern) [Wang *et al.*, 2006], les divergents (divergent objects) et les convergents (convergent objects) [Hai *et al.*, 2012]...
De nombreux algorithmes permettent d'extraire ces motifs, comme CuTS [Jeung *et al.*, 2008] pour les convois, ObjectGrowth [Li *et al.*, 2010a] pour les essaims clos, VG-growth [Wang *et al.*, 2006] pour les motifs de groupes... Ces algorithmes ne peuvent extraire qu'un seul type de motif. [Phan *et al.*, 2016] ont proposé

2. Dans la littérature, on évoque aussi bien les motifs spatio-temporels que les trajectoires pour désigner ces motifs. Ainsi, pour éviter de confondre avec le concept de trajectoire que nous avons introduit, nous utiliserons préférentiellement le terme de motifs spatio-temporels.

une approche unifiée pour extraire l'ensemble de ces trajectoires en utilisant notamment une extension de l'algorithme LCM [3]. Dans la suite, nous allons utiliser cet algorithme.

La recherche de phénomènes communs dans les parcours de soins est analogue à la détection de phénomènes de groupes pour les objets mobiles. En effet, un parcours de soins est une succession chronologique d'évènements à des estampilles de temps différentes. Ainsi, en redéfinissant la notion d'espace (ou de localisation géographique) par le type de séjour hospitalier et la notion de temps par l'occurrence d'un évènement, nous pouvons assimiler une trajectoire de patient à une trajectoire d'objet mobile. Nous avons, alors, appliqué une méthode de détection de motifs spatio-temporels aux données de santé décrites dans la section 7.1.2. Dans le cadre de notre étude, nous nous sommes intéressés aux essaims clos. Mettre en évidence des trajectoires similaires dans le suivi de cette pathologie à des moments particuliers du parcours de soins se révèle utile à la fois pour le patient mais aussi pour l'établissement de soins. Pour le patient, le suivi est amélioré en prévoyant des examens à des périodes programmées à l'avance, comme dans le cas du suivi des maladies chroniques [Wagner *et al.*, 2001]. Pour l'établissement de soins, les flux de patients sont anticipés afin d'améliorer la planification sanitaire.

Dans la suite de cette section, nous introduisons, dans la section 7.1.1, la terminologie propre au domaine de la détection de motifs spatio-temporels et l'illustrons par un exemple d'application. Puis, nous détaillons, dans la section 7.1.2, le processus de fouille appliquée aux données du PMSI.

7.1.1 Définitions préliminaires

La fouille de données spatio-temporelle consiste à discerner des ensembles d'objets restant groupés sur une même période. Il existe de nombreux motifs spatio-temporels. Nous présentons uniquement ceux utilisés dans la suite de ce chapitre. Commençons par quelques notations générales utilisées tout au long de cette section :

- $O = \{o_1, ..., o_n\}$ un groupe d'objets mobiles ;
- $T = \{t_1, ..., t_p\}$ un ensemble d'estampilles temporelles ;
- $x_{t_i}^{o_j}, y_{t_i}^{o_j}$ les informations spatiales de $o_j \in O$ au temps t_i ;
- min_o un support minimum, donné par l'utilisateur, correspondant au nombre d'objets minimum devant être ensemble ;
- min_t, le nombre minimum d'estampilles temporelles pendant lesquelles au moins min_o objets de O sont regroupés.

Définition 16 (Essaim)
Un couple (O, T) est un essaim (ou swarm en anglais) si :

1) *Il y a au moins un cluster contenant tout objet de O à chaque estampille temporelle de T : $\forall\ t_i \in T, \exists\ c$ tel que $O \subseteq c$ où c est un cluster ;*
2) *Il y a au moins min_o objets : $|O| \geq min_o$;*

3. Linear time Closed itemset Miner.

3) Il y a au moins min_t estampilles temporelles : $|T| \geq min_t$.

De manière intuitive, un essaim est un groupe d'objets contenant au moins ϵ éléments qui sont proches les uns des autres pour au moins min_t estampilles de temps.

Exemple : Pour illustrer ces définitions, nous allons considérer les évènements de 4 patients. Le temps est perçu comme une variable non continue. Il est divisé en estampilles temporelles correspondant à la survenue d'une hospitalisation. Le temps démarre à $t = 0$ pour la première hospitalisation des 4 patients pour un évènement lié à l'IM. Ces informations sont contenues dans la base de données du tableau 7.1. Comme dans l'exemple du chapitre 5, cette base décrit différents GHM (05M04 : infarctus aigu du myocarde ; 05M13 : Douleurs thoraciques ; 10M02 : Diabète ; 05M06 : Angine de poitrine ; 05M16 : Athérosclérose) associés au cours du temps par des professionnels de santé à des séjours hospitaliers. Dans notre exemple, un objet est un patient et la proximité spatiale est associée au type de séjour (GHM). Les trajectoires de ces patients sont représentées dans la figure 7.1.

Tableau 7.1 – Trajectoires de patients. Le temps est lié à la survenue d'une hospitalisation.

	Estampilles de temps			
Patient	**0**	**1**	**2**	**3**
P_1	05M04	10M02	05M06	05M16
P_2	05M04	05M13	05M06	
P_3	05M04	10M02	05M06	05M16
P_4	05M04	05M13	05M06	

Si $\epsilon = 2$ et $min_t = 2$, nous identifions les essaims suivants $\{(P_1, P_2, P_3, P_4), (0, 2)\}$ et $\{(P_1, P_3), (0, 3)\}$. En effet, les patients P_1, P_2, P_3 et P_4 ont le même GHM 05M04 au temps t_0 et au temps t_2, ils ont tous le GHM 05M06. De même, les patients P_1 et P_3 ont des parcours semblables avec les GHM 05M04 en t_0 et 05M16 en t_3.

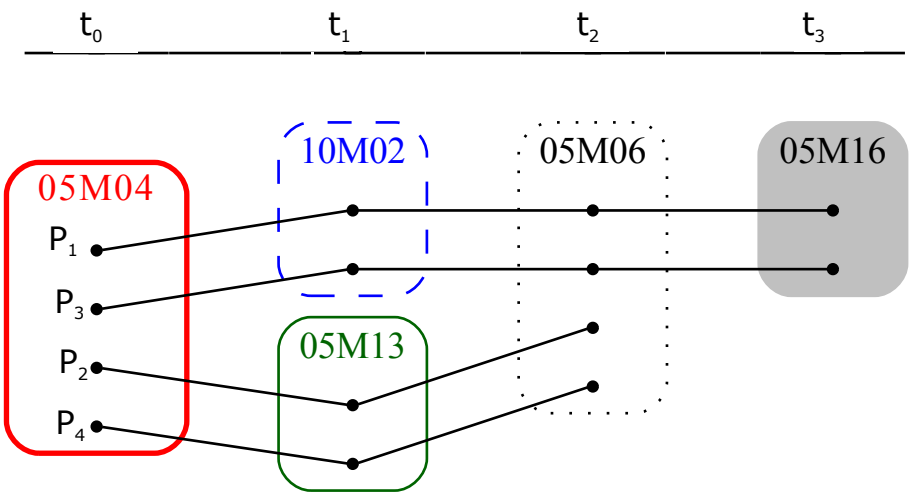


Figure 7.1 – Visualisation des trajectoires de patients. Le temps est lié à la survenue d'une hospitalisation.

Définition 17 (Essaim clos)
Un couple (O, T) est un essaim clos (closed swarm) s'il vérifie :

1) (O, T) est un essaim ;

2) (O, T) est objet-fermé : $\nexists\, O'$ tel que (O', T) est un essaim et $O \subset O'$;

3) (O, T) est temporel-fermé : $\nexists\, T'$ tel que (O, T') est un essaim et $T \subset T'$.

Exemple : Dans l'exemple précédent, nous pouvons également identifier les essaims suivants : $\{(P_1, P_3), (0, 1)\}$ et $\{(P_1, P_3), (2, 3)\}$. Ces groupes sont redondants et peuvent être réunis dans l'essaim $\{(P_1, P_3), (0, 2, 3)\}$ qui est un essaim clos.

Dans la suite, nous allons mettre en œuvre cette approche sur nos données.

7.1.2 Description du processus de fouille

Comme dans le cas des motifs fréquents, le processus de fouille s'effectue en plusieurs étapes dont certaines sont communes avec celles décrites dans le chapitre 5. Le schéma 7.2 ci-dessous résume les différentes étapes du processus.

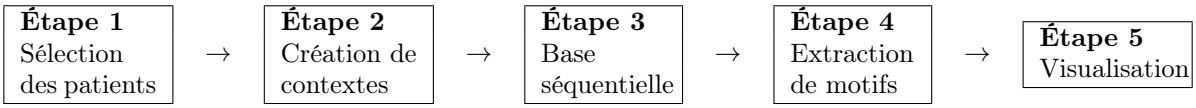


Figure 7.2 – Étapes du processus de fouille.

Étape 1 à Étape 2. La description de ces étapes est strictement identique à celle de la partie 5.1.2 du chapitre 5.

Étape 3 : Construction de la base séquentielle. La constitution de la base séquentielle est également identique à celle du chapitre 5. Nous ajoutons une action supplémentaire pour recaler les trajectoires selon un temps relatif correspondant à la survenue d'un évènement hospitalier. Un exemple de constitution des données séquentielles est schématisé dans la figure 7.3. Le patient 1 est hospitalisé la première fois en janvier pour un IM. Il est à nouveau hospitalisé au mois de juin pour angine de poitrine. Ces évènements sont rassemblés de la façon suivante : le patient 1 est hospitalisé au temps t_0 pour IM et au temps t_1 pour angine de poitrine. Le patient 2 est hospitalisé au mois de mars pour IM. Il est à nouveau hospitalisé au mois de mai pour des douleurs thoraciques et au mois de juin pour athérosclérose. Ces évènements sont rassemblés de la façon suivante : le patient 2 est hospitalisé au temps t_0 pour IM, au temps t_1 pour douleurs thoraciques et au temps t_2 pour athérosclérose.

Étape 4 : Extraction de motifs. L'extraction des essaims clos est faite à l'aide de l'algorithme Get_Move [Hai *et al.*, 2012]. La fouille génère de nombreux motifs qui sont filtrés en ne retenant que ceux qui concernent au moins 1% des patients du contexte étudié.

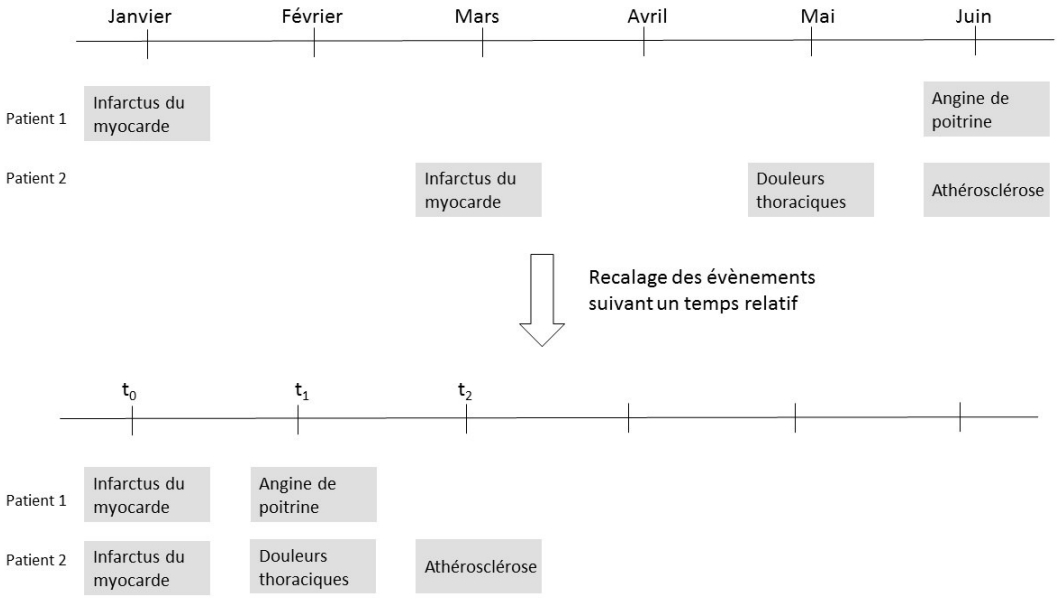


Figure 7.3 – Recalage des trajectoires de patients.

Étape 5 : Visualisation. À l'issue de l'étape 4, nous obtenons des fragments de parcours communs aux patients identifiés par l'algorithme. À partir des données originelles, nous récupérons les informations manquantes afin de reconstituer les parcours de ces patients depuis le départ (au temps t_0) jusqu'à l'évènement après le dernier évènement identifié par Get_Move. Cette reconstitution des parcours est ensuite représentée à l'aide d'un diagramme de Sankey. Il s'agit d'un diagramme de flux, dans lequel la largeur des bandes est proportionnelle au flux représenté, c'est-à-dire au nombre de patients considérés. Dans certains cas, nous obtenons de nombreux sommets, le graphique peut alors rapidement devenir illisible. Pour y remédier, nous recodons les sommets en réunissant les codes GHM (resp. DP) suivant une logique médicale. Pour cela, nous nous sommes appuyés sur des travaux en cardiologie [Laissy *et al.*, 2004, Abitbol, 2005, Dujardin et Fabre, 2008] portant sur les évolutions possibles de l'IM. Le détail des regroupements est répertorié dans les annexes E.1 et E.2.

7.2 Expérimentations

Dans cette section, nous nous concentrons sur les contextes suivants : Homme & +65 ans & 5-60 séjours et Femme & +65 ans & 5-60 séjours. Nous avons appliqué la méthode décrite dans la section 7.1.2 sur ces deux contextes. Dans la section 7.2.1, nous présentons les motifs mis en évidence par l'algorithme Get_Move. Dans la section 7.2.2, nous reconstituons les trajectoires des patients identifiés dans la section 7.2.2 afin de visualiser les flux de patients au travers des différents parcours de soins.

7.2.1 Extraction de motifs

Dans cette section, nous procédons à une simple description des motifs mis en évidence en y associant les nombres de patients concernés pour les trajectoires de GHM et de DP.

Dans les trajectoires de GHM, nous identifions 4 essaims clos à la fois chez les hommes et chez les femmes. Ces différents essaims sont représentés dans la figure 7.4 ci-dessous. Nous remarquons que dans les deux contextes les essaims sont similaires, seuls les effectifs changent. Nous les listons ci-après :

— essaim n°1 constitué de 10 105 hommes (resp. 3 387 femmes) ayant une pose de stent sans IM à t_0 puis à nouveau une pose de stent sans IM à t_1 ;

— essaim n°2 constitué de 2 448 hommes (resp. 704 femmes) ayant une pose de stent sans IM à t_0 puis à nouveau une pose de stent sans IM à t_2 ;

— essaim n°3 constitué de 2 029 hommes (resp. 828 femmes) ayant une pose de stent avec IM à t_0 puis une pose de stent sans IM à t_1 ;

— essaim n°4 constitué de 2 789 hommes (resp. 848 femmes) ayant une pose de stent sans IM à t_1 puis à nouveau une pose de stent sans IM à t_2.

Dans les trajectoires de DP, nous identifions 11 essaims clos chez les hommes et 7 chez les femmes. Ils sont représentés dans la figure 7.5 ci-dessous. Ici, nous notons des différences entre les hommes et les femmes. Tout d'abord, chez les femmes, il y a moins de motifs. Ensuite, nous n'observons pas de cluster au temps t_2 chez les femmes. Puisque les motifs sont plus nombreux chez les hommes, nous listons les essaims trouvés chez les hommes et nous mettrons entre parenthèses les résultats établis chez les femmes pour les essaims équivalents. Ainsi, nous identifions :

— essaim n°1 constitué de 1 821 hommes (essaim n°3 avec 615 femmes) ayant une angine de poitrine à t_0 puis une cardiopathie ischémique chronique à t_1 ;

— essaim n°2 constitué de 3 016 hommes (essaim n°2 avec 1 096 femmes) ayant une angine de poitrine à t_0 puis à nouveau une une angine de poitrine à t_1 ;

— essaim n°3 constitué de 4 052 hommes (essaim n°1 avec 1 201 femmes) ayant une cardiopathie ischémique chronique à t_0 puis à nouveau une cardiopathie ischémique chronique à t_1 ;

— essaim n°4 constitué de 1 085 hommes ayant une cardiopathie ischémique chronique à t_0 puis une angine de poitrine à t_1 ;

— essaim n°5 constitué de 887 hommes ayant une cardiopathie ischémique chronique à t_0 puis à nouveau une cardiopathie ischémique chronique à t_2 ;

— essaim n°6 constitué de 1 596 hommes (essaim n°5 avec 620 femmes) ayant un IM aigu à t_0 puis une cardiopathie ischémique chronique à t_1 ;

— essaim n°7 constitué de 1 396 hommes (essaim n°6 avec 789 femmes) ayant un IM aigu à t_0 puis à nouveau un IM aigu à t_1 ;

— essaim n°8 constitué de 921 hommes (essaim n°7 avec 775 femmes) ayant un IM aigu à t_0 et en décèdent à t_1 ;

— essaim n°9 constitué de 1 096 hommes (essaim n°4 avec 467 femmes) ayant un IM aigu à t_0 puis une angine de poitrine à t_1 ;

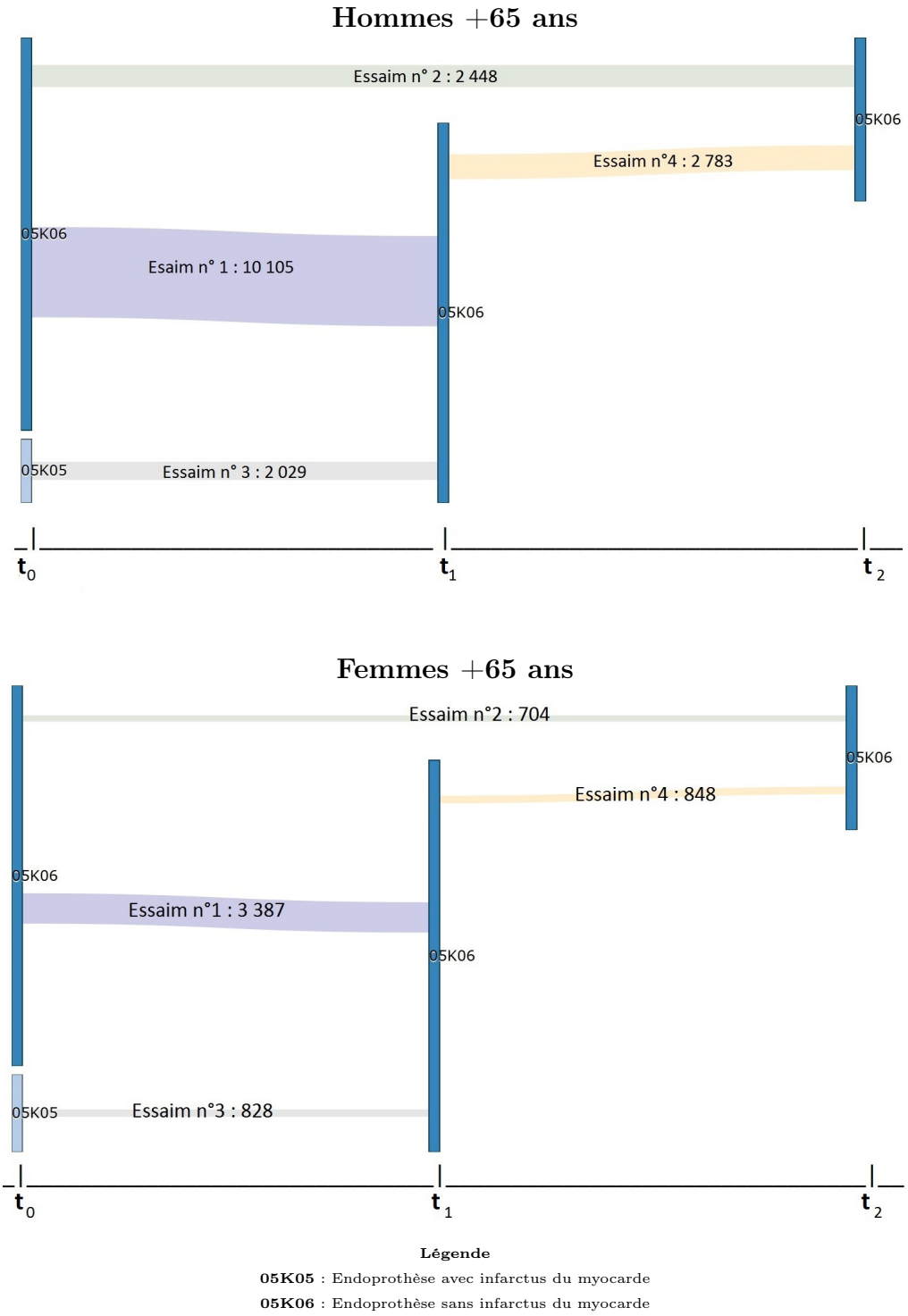
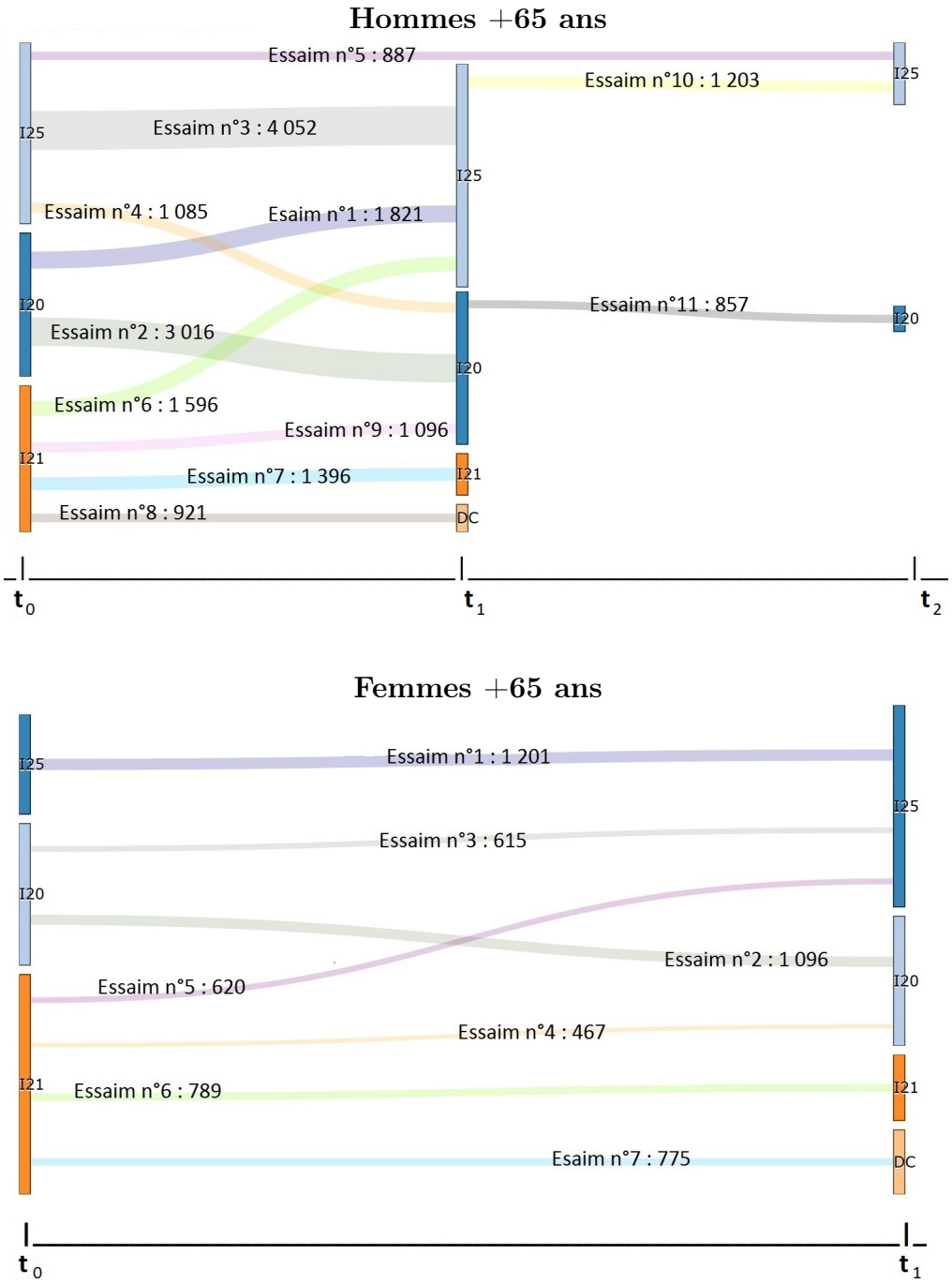


Figure 7.4 – Essaims clos dans les trajectoires de GHM.



Légende
I20 : Angine de poitrine
I21 : Infarctus aigu du myocarde
I25 : Cardiopathie ischémique chronique
DC : Décès

Figure 7.5 – Essaims clos dans les trajectoires de DP.

— essaim n°10 constitué de 1 203 hommes ayant une cardiopathie ischémique chronique à t_1 puis à nouveau une cardiopathie ischémique chronique à t_2 ;
— essaim n°11 constitué de 857 hommes ayant une angine de poitrine à t_1 puis à nouveau une angine de poitrine à t_2.

7.2.2 Visualisation des flux de patients

Dans cette section, nous nous intéressons aux flux de patients au travers des différentes étapes du parcours hospitalier. Lorsque les estampilles de temps ne sont pas consécutives comme pour l'essaim n°2, dans les trajectoires de GHM, nous souhaitons connaître les évènements hospitaliers survenus entre-temps. De plus, comme nous avons peu d'estampilles de temps (entre deux et trois) dans les essaims trouvés et que ces estampilles se situent généralement en début de parcours, nous cherchons également à connaître la suite de leur parcours et les schémas de flux de patients. Pour cela, nous avons donc reconstitué une partie des parcours de soins de l'ensemble de ces patients pour les deux types de trajectoires.

Dans la suite de cette section, nous décrivons les différents groupes constitués pour réaliser la visualisation des flux. Ensuite, nous décrivons les schémas de flux selon le point de départ.

Dans le cas des trajectoires de GHM, nous avons procédé à un rassemblement des patients identifiés précédemment dans les essaims afin d'éviter au maximum une redondance des patients d'un groupe à l'autre. Pour cela, nous affectons à un même groupe, les patients ayant le même sommet au temps t_0 de leur trajectoire. Nous pouvons visualiser les flux de patients pour les hommes et pour les femmes dans la figure 7.6. Pour les hommes comme pour les femmes, nous avons regroupé les essaims de la façon suivante :

— Groupe 1 (en vert) contient les patients des essaims n°1 et 2 ;
— Groupe 2 (en rouge) contient les patients des essaims n°3 ;
— Groupe 3 (en gris) contient les patients des essaims n°4.

Nous remarquons que les schémas de flux sont assez similaires d'un genre à l'autre. Toutefois, nous notons des différences au temps t_0 : les sommets CHIRO, ACOMP et TRYTH ne sont présents que chez les hommes et au temps t_1, le sommet CHIRC n'est présent que chez les hommes.
Le flux partant du sommet STEN1 (2 029 hommes et 828 femmes) passe ensuite par le sommet STEN2 pour 100% des effectifs. Ensuite, le flux se sépare en plusieurs branches vers les sommets STEN2 (92% des hommes et 75% des femmes), ACI (2,3% des hommes et 3% des femmes), ISCHE (0,1% des hommes), STEN1 (3% des patients), I.MYO (2% des hommes et 0,9% des femmes), CHIRC (0,2% des hommes), AUTRE (0,6% des hommes et 0,4% des femmes) et DECES (0,2% des hommes et 0,6% des femmes). Les branches s'affinent de plus en plus à mesure que l'on avance dans le temps. Au temps t_3, les sommets observés sont identiques aux précédents. Enfin, les derniers sommets observés au temps t_4, sont STEN2 (1% des hommes et 3% des femmes), STEN1 (0,1% des hommes et 0,2% des femmes), ACI (1,4% des hommes et 0,4% des femmes), I.MYO (0,1% des hommes et 0,4%

des femmes), TRYTH (0,1% des patients), CHIRC (0,1% des patients) et DECES (0,1% des hommes et 0,3% des femmes).
Le flux partant du sommet STEN2 (10 565 hommes et 3 504 femmes) se sépare vers les sommets STEN2 (95% des hommes et 97% des femmes) et ACI (2% des patients), STEN1 (1% des hommes et 0,9% des femmes), I.MYO (0,6% des hommes et 0,3% des femmes), AUTRE (0,6% des hommes et 0,2% des femmes), CHIRC (0,2% des hommes) et ISCHE (0,08% des patients). Ensuite, les différentes branches se rejoignent dans les sommets STEN2 (31% des hommes et 24% des femmes), ISCHE (0,1% des patients), AUTRE (4,6% des patients), STEN1 (1,4% des patients), I.MYO (1% des hommes et 0,8% des femmes), ACI (3% des patients), DECES (0,2% des hommes et 0,5% des femmes), CHIRC (0,4% des hommes et 0,3% des femmes) et TRYTH (0,02% des hommes). Ensuite, les schémas de flux sont identiques à ceux démarrant du sommet STEN1.
Les flux de patients démarrant des sommets restant se réunissent dans le sommet STEN2 (ce qui représente 149 hommes et 122 femmes). Au temps suivant, ils sont à nouveau dans STEN2 et suivent ensuite des schémas similaires à ceux décrits plus haut.

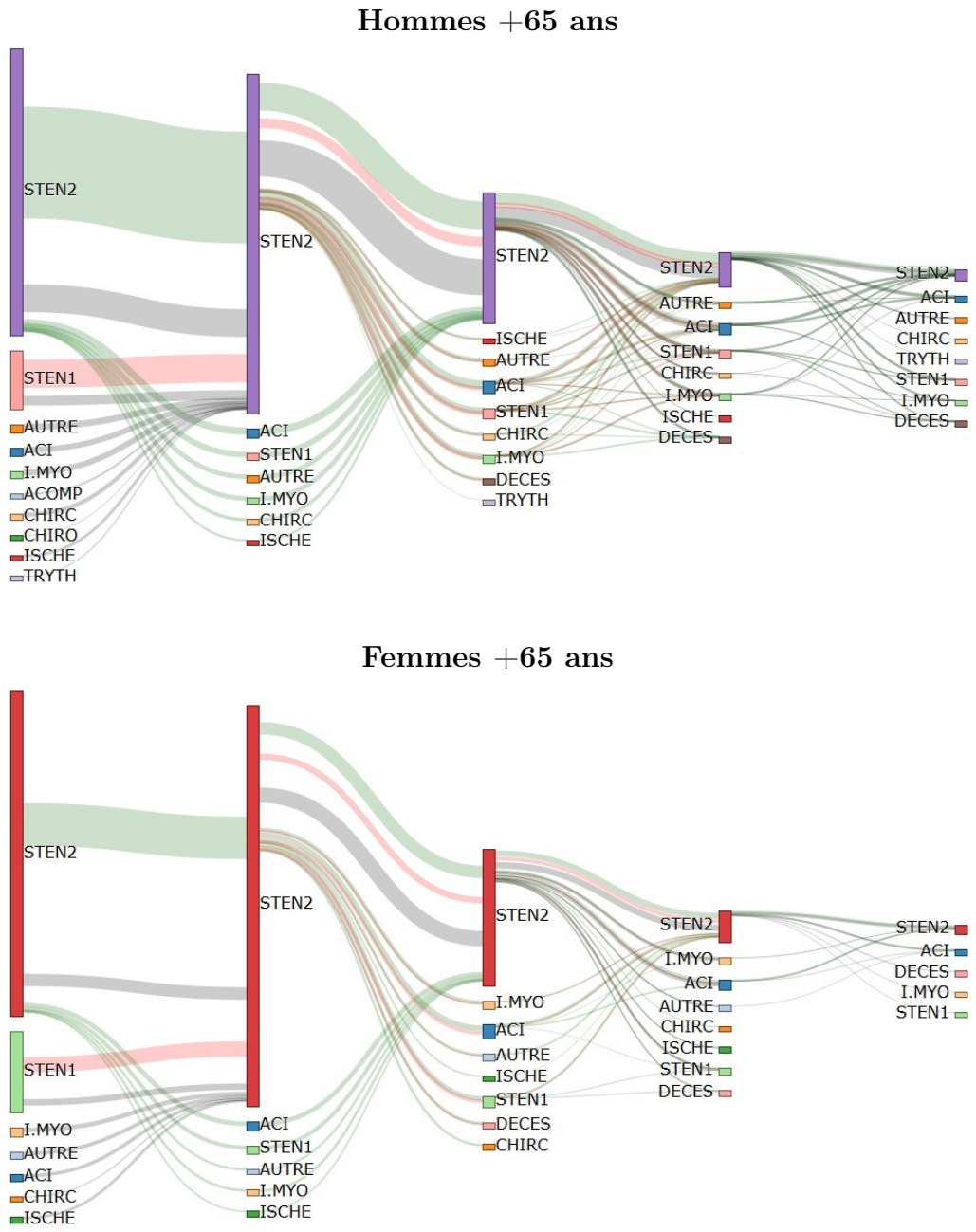
De même, dans le cas des trajectoires de DP, nous visualisons les flux de patients dans la figure 7.7. Nous avons regroupé les patients des différents essaims de la façon suivante :

— Groupe 1 (en vert) contient les patients des essaims n°1, 2 et 11 dans le cas des hommes, et les patientes des essaims n°2 et 3. Pour ce groupe le premier évènement est l'angine de poitrine ;

— Groupe 2 (en rouge) contient les patients des essaims n°3, 4, 5 et 10 dans le cas des hommes, et les patientes des essaims n°1. Pour ce groupe le premier évènement est la cardiopathie ischémique ;

— Groupe 3 (en gris) contient les patients des essaims n°6, 7, 8, 9 dans le cas des hommes, et les patientes des essaims n°4, 5, 6 et 7. Pour ce groupe le premier évènement est l'IM.

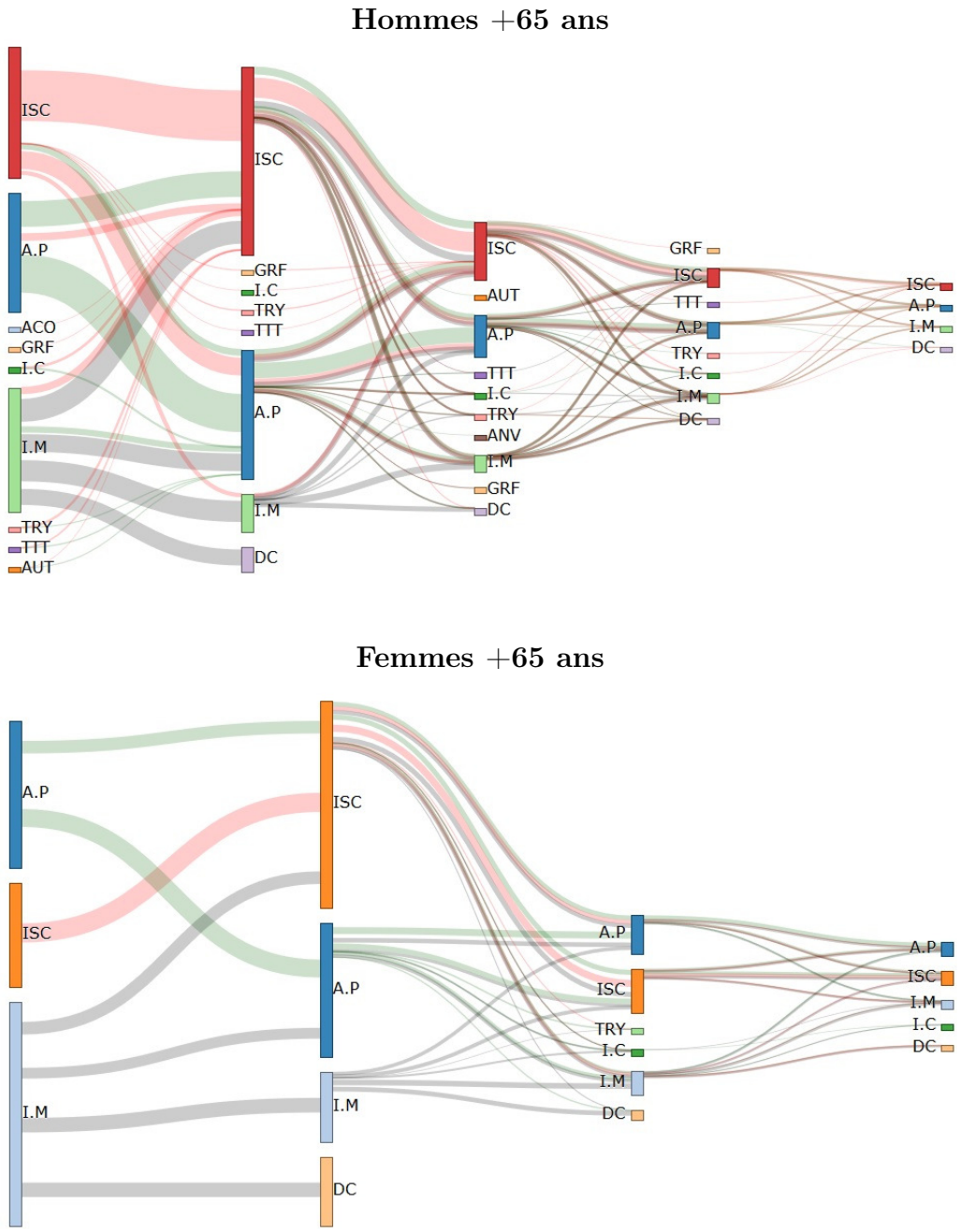
Nous remarquons des différences plus importantes entre les hommes et les femmes. Tout d'abord, il y a une estampille de temps supplémentaire chez les hommes et la diversité des sommets est plus grande dans le cas des hommes que dans celui des femmes à des temps identiques. Nous allons décrire les schémas de flux dans le cas des femmes et nous ajouterons au fur et à mesure les différences, le cas échéant, pour les hommes.
Le flux qui part du sommet I.M (4 943 hommes et 2 651 femmes) se sépare dans les sommets ISC (31% des hommes et 23% des femmes), A.P (22% des hommes et 18% des femmes), I.M (28% des hommes et 29% des femmes) et DC (19% des hommes et 29% des femmes). Au temps suivant, les branches sont plus fines et rejoignent les mêmes types de sommets avec en plus I.C, ISC, GRF, TRY et TTT. Au temps t_3, il reste les sommets A.P (2% des hommes et 0,08% des femmes), TTT (0,04% des hommes), ISC (2% des hommes et 0,08% des femmes), I.M (2% des hommes et 0,08% des femmes), I.C (0,1% des hommes et 0,01% des femmes) et DC (0,4% des patients).
Le flux qui part du sommet A.P (4 845 hommes et 1 711 femmes), se sépare en



Légende

ACI : Actes de cardiologie interventionnelle
ACOMP : Autres complications
AUTRE : Autres
CHIRC : Chirurgie cardiothoracique - chirurgie de revascularisation
CHIRO : Chirurgie orthopédique
DECES : Décès
I.MYO : Infarctus du myocarde
ISCHE : Ischémie
STEN1 : Pose d'endoprothèse avec IM
STEN2 : Pose d'endoprothèse sans IM
TRYTH : Troubles du rythme et de la conduction cardiaque

Figure 7.6 – Flux de patients dans les trajectoires de GHM.



Légende

A.P : Angine de poitrine
ACO : Autres complications
ANV : Anévrisme
AUT : Autres troubles
DC : Décès
GRF : Greffes cardiaques - complications de greffes - Suivi ajustement
I.C : Insuffisance cardiaque
I.M : Infarctus du myocarde
ISC : Ischémie
TRY : Troubles du rythme et de la conduction cardiaque
TTT : Observation, surveillance et suivi de traitement - Résultats anormaux d'examens

Figure 7.7 – Flux de patients dans les trajectoires de DP.

A.P (62% des hommes et 64% des femmes) et ISC (42% des hommes et 36% des femmes), puis, au temps suivant, les sommets sont I.M (2% des patients), ISC (5% des hommes et 9% des femmes), A.P (18% des hommes et 14% des femmes), DC (0,06% des hommes et 0,2% des femmes) et les sommets AUT, TTT, I.C, TRY, GRF, RYT et ANV pour 0,5% des hommes. Les branches sont de plus en plus fines, et rejoignent au temps t_4 les sommets A.P, ISC, I.M, I.C et DC.
Le flux qui part du sommet ISC (5 422 hommes et 1 201 femmes) se dirige vers le sommet ISC au temps suivant. Ensuite, se sépare dans les sommets A.P (4% des femmes), ISC (24% des femmes), TRY (0,2% des femmes) et I.M (2,4% des femmes). Les branches s'affinent au fur et à mesure, au temps t_3, il reste les sommets A. P (2% des femmes), ISC (2% des femmes), I.M (2% des femmes), I.C (0,02% des femmes) et DC (0,8% des femmes). Chez les hommes, le schéma des flux est plus complexe. Le flux partant de ISC se sépare en A.P (23% des hommes), ISC (78% des hommes), I.M (0,8% des hommes) et les sommets I.C, TRY, GRF et TTT pour 0,2% des hommes). Au temps suivant, on retrouve les mêmes sommets avec le DC en plus et la suite des schémas de flux est similaire à ceux déjà décrits pour les points de départ ISC et A.P.
Chez les hommes, les flux partant des sommets GRF, ACO, TRY, AUT, I.C et TTT se dirigent vers ISC (soit 34 hommes) et A.P (soit 18 hommes). Ensuite, les schémas de flux sont similaires à ceux décrits précédemment.

7.3 Discussion

La recherche de motifs spatio-temporels, comme la méthode décrite dans le chapitre 5, est basée sur la notion de fréquence. Les clusters identifiés dans ce chapitre correspondent aux motifs fréquents extraits précédemment. La différence principale entre les deux approches est la notion de temporalité. Avec les motifs séquentiels, nous pouvions par exemple conclure que fréquemment un patient qui a eu une pose de stent avec IM avait par la suite une pose de stent sans IM. Avec les motifs spatio-temporels, nous obtenons une information supplémentaire sur le moment de survenue de ces évènements. En effet, si l'on résume les résultats obtenus dans le cas des trajectoires de GHM, nous pouvons déterminer que ces hospitalisations répétées pour pose de stent sont successives dès le début du parcours du patient dans la majorité des cas. Certains sont concernés par des hospitalisations pour suivi de traitement et surveillance, pour de la chirurgie cardiologique (avec pose d'un défibrillateur). D'autres ont à nouveau un IM et/ou développent une cardiopathie ischémique chronique.
Si l'on résume maintenant les résultats concernant les trajectoires de DP, nous avons identifié trois étapes clés dans les schémas de flux de patients : l'IM, l'angine de poitrine et la cardiopathie ischémique. De plus, ces évènements sont consécutifs dans les trajectoires de patients. La majorité des patients manifestent des signes de récidives de l'IM par l'angine de poitrine. Beaucoup d'entre-eux font, par ailleurs, de nouveau un IM et/ou développent une cardiopathie ischémique. D'autres sont concernés par des affections telles que l'anévrisme, l'ischémie, les troubles

du rythme et de la conduction cardiaque ou encore par l'insuffisance cardiaque (I50). En outre, la part des décès représente environ 21% des hommes (resp. 32% des femmes). Ce dernier se produit le plus souvent lors de la première hospitalisation.

Nous avons cherché à lier ces résultats avec les parcours de soins recommandés par [HAS, 2016], dans le cas du suivi d'une personne atteinte de maladie coronarienne pendant un an. Selon ces recommandations, des visites régulières doivent être programmées pour garantir le contrôle des éventuelles modifications de l'état clinique et, le cas échéant, alerter en cas de suspicion de récidive. Ces visites visent également à contrôler les facteurs de risque cardiovasculaire. Ceci peut expliquer les hospitalisations pour suivi de traitement et surveillance et pour pose d'endoprothèse si le contrôle effectué par un examen non-invasif (ECG d'effort, coroscanner, coronarographie...) a confirmé le diagnostic d'une obstruction coronaire. Toutefois, ces hospitalisations à répétition pour pose de stent soulèvent des questions :

— S'agit-il de la particularité de ces patients ? Ont-ils un terrain propice à développer de l'athérome, du fait de leurs facteurs de risque et/ou de la non observance de leur traitement ? Parmi ces patients, avec les données du PMSI, nous identifions la répartition suivante : 75% des hommes (resp. 24% des femmes) ont de l'hypertension ; 59% des hommes (resp. 18% des femmes) ont des problèmes de dyslipidémies ; 37% des hommes (resp. 11% des femmes) sont diabétiques ; 18% des hommes (resp. 5% des femmes) souffrent d'obésité et 11% des hommes (resp. 1% des femmes) consomment des substances psychoactives (tabac, alcool...). Notons que des études soulignent l'insuffisance du contrôle des facteurs de risque cardiovasculaire ou de la réadaptation cardiaque [Al-Salameh *et al.*, 2016, HAS, 2012, Ghannem *et al.*, 2015].

— Le traitement est-il adapté à ces patients ? Le vif débat lancé sur l'usage des statines [Vallée, 2013, Halimi et Halimi, 2013], déclenche, en outre, des polémiques à propos de certaines pratiques de l'industrie pharmaceutique aussi bien dans la publication d'essais cliniques [Antes et Chalmers, 2003, Korn, 2000, Angell, 2000] que dans la diffusion des informations auprès des médecins [Moynihan, 2003, Abbasi et Smith, 2003, Campbell, 2008]. Le principe de précaution et le devoir d'informer les patients traités incitent les professionnels de santé à faire la part des choses entre science et marketing [Golomb *et al.*, 2007, Toussaint, 2010, Rédaction Prescrire, 2011, Golomb, 2015].

— Ces répétitions d'hospitalisation sont-elles liées au type de stent ? Ce dispositif médical intégré est sujet à controverse car il est susceptible de provoquer des cas de resténose [HAS, 2009]. Parmi les techniques permettant la revascularisation : angioplastie par ballonnet coronarien [Debbas *et al.*, 1995], stenting, revascularisation coronarienne avec une chirurgie cardiaque minimalement invasive [Blanc *et al.*, 1999], endartériectomie [Thompson, 1996], quelle est la plus adaptée selon le profil du patient ? Des études [Versaci *et al.*, 1997, Savage *et al.*, 1998] ont comparé l'angioplastie *versus* la pose de stent. Elles ont établi que les bénéfices étaient supérieurs dans le cas de la pose de stent. Néanmoins, les différences mises en évidence, dans les deux cas, sont significatives pour un seuil très bas. D'autres études [Mas *et al.*, 2006, Ederle *et al.*, 2010] ont été menées pour comparer les bénéfices d'une technique par rapport à une

autre notamment entre la pose d'endoprothèse et l'endartériectomie. Les deux études concluent à de meilleurs bénéfices dans la survie et le risque de récidive avec endartériectomie. Toutefois, elles préconisent un approfondissement avec un suivi sur le long terme. En revanche, dans [Massop *et al.*, 2009], le registre mondial SAPPHIRE[4] préconise le stenting à la place de l'endartériectomie chez les patients présentant un risque élevé de chirurgie en raison de facteurs de risque anatomiques. *In fine*, le domaine d'évaluation de l'influence sur le pronostic des différentes stratégies de suivi chez les patients ayant une maladie coronaire reste à approfondir. De plus, ces évaluations ne seront pas les mêmes suivant le type d'artère atteinte. En effet, les stratégies d'intervention sont différentes suivant le type d'artère à traiter. Dans le cas des coronaires il n'y a pas d'endartériectomie : c'est soit le stent, soit le pontage.

La visualisation des schémas de flux montre une différence des parcours entre genre. En effet, de manière générale, les trajectoires de patients sont plus longues chez les hommes[5]. De plus, la diversification des étapes dans les schémas de flux est plus importante chez les hommes également. Ces observations s'expliquent par le fait que : 1) la population des hommes est plus importante en termes d'effectifs, ainsi, les possibilités d'observer des trajectoires différentes sont accrues ; 2) le taux de mortalité est plus élevée dans la population des femmes (0,1% *vs* 0,08% pour les hommes). De plus, les femmes sont en moyenne significativement[5] plus âgées dans la même tranche d'âge (79 ans *vs* 75 ans pour les hommes). Ces deux derniers éléments peuvent expliquer des trajectoires plus courtes pour les femmes.

Par ailleurs, avec les motifs spatio-temporels, nous avons retrouvé une partie des évolutions possibles [Dujardin et Fabre, 2008] de cette pathologie, qui sont : les troubles du rythme et de la conduction cardiaque (les troubles du rythme du ventricule et les troubles de la conduction électrique du cœur), l'insuffisance cardiaque, les complications mécaniques, les autres complications (phlébite, embolie pulmonaire...), l'ischémie, l'anévrisme et la récidive. Nous mettons également en évidence les techniques de soins employées par la cardiologie interventionnelle (angioplastie, stenting...) ou celles de la chirurgie cardiologique avec la pose de défibrillateur, de pacemaker... ou encore la greffe cardiaque.

Cette méthode permet d'envisager une comparaison territoriale. En effet, dans le chapitre 3, nous avons mis en évidence une hétérogénéité géographique des hospitalisations, plus exactement de gradient Nord-Sud. L'exploration des flux en tenant compte de la région d'origine du patient en tant que paramètre contextuel offrirait la perspective comparative des flux soit en termes de soins, soit en termes d'évolution de la pathologie.

Les résultats obtenus par cette approche pourraient être intégrés dans un outil d'aide à la décision pour le médecin. En effet, ce dernier pourrait ainsi orienter ces recommandations envers son patient et le mettre en garde sur les risques de réhospitalisation en comparant son profil à celui de patients similaires. Ce

4. Stenting and Angioplasty with Protection of Patients with High Risk for Endarterectomy.
5. $p.value < 2,2.10^{-16}$.

type d'application existe d'ores et déjà. Par exemple, [Wang et Bajorek, 2016] proposent un outil à destination des professionnels de santé pour la prescription d'anti-thrombotiques chez des patients atteints de fibrillation atriale, afin d'éviter les risques d'interactions médicamenteuses, mais surtout les risques secondaires selon le profil du patient.

Dans ce chapitre, nous avons pris le parti de mettre en lumière les schémas de flux de patients les plus fréquents. Toutefois, nous pourrions explorer les cas rares et ainsi nous intéresser non plus à des phénomènes touchant une majorité de la population, induisant des généralités médicales, mais plutôt de revenir à l'individu dans sa variabilité biologique. Pour explorer les cas rares, nous proposons de nous focaliser sur les essaims ayant au moins 4 estampilles de temps. Par exemple, dans le cas des trajectoires de DP, nous identifions un essaim contenant 4 patients pour les estampilles de temps t_0, t_2, t_4, t_5 et t_6. Ces patients ont les parcours suivants : *Hypertension secondaire* (I15), puis *Insuffisance cardiaque* au temps t_2, *Maladies rhumatismales de la valvule mitrale* (I05) au temps t_4, *Autres atteintes des artères et artérioles* (I77) et enfin *Examen de contrôle après traitement d'affections autres que les tumeurs malignes* (Z09).

7.4 Conclusion

Dans ce chapitre, nous avons recherché des motifs spatio-temporels dans les trajectoires des patients ayant eu un IM au cours de la période 2009 à 2014. Dans notre cas, la notion de spatialité a été assimilée à celle de proximité de pathologies en termes de taxonomie. De plus, l'aspect temporel est lié à la survenue d'une hospitalisation. Ces analyses ont été menées sur 2 sous-populations de patients âgées de +65 ans, afin de mettre en lumière des comportements de groupes au sein de ces populations. Nous avons retrouvé les motifs fréquents du chapitre 5 et avons établi que les hospitalisations à répétition pour pose de stent étaient dans la majeure partie des cas, consécutives dès le début du parcours du patient. De plus, la visualisation des flux de patients nous renseigne sur les évolutions de la pathologie cardiaque. Nous avons retrouvé des éléments d'ores et déjà connus du domaine de la cardiologie.

Dans ce chapitre, nous avons détourné une méthode permettant de détecter des comportements similaires d'objets mobiles. Toutefois, nous pourrions également intégrer cette notion de localisation géographique de deux manières différentes : tout d'abord en considérant la localisation du patient au sein d'un établissement (*e.g.* unité médicale) ou en considérant la localisation géographique du patient sur le territoire français (*e.g.* la région de France, l'établissement de santé...). De la sorte, nous pourrions étendre les travaux de [Dart *et al.*, 2003] pour la répartition des flux de patients dans les unités médicales. De même, en associant la notion de spatialité à la géolocalisation du patient au sein d'une région, nous pourrions également étendre les travaux de [Jay *et al.*, 2008, Jay *et al.*, 2006] pour établir le réseau de santé d'une région. Ceci permettrait de faire émerger des comportements de groupes non pas en termes d'évolution de pathologies ou de type de soins mais en termes d'évolution de prise en charge ou de com-

paraison de prise en charge, ou encore de localisation d'épidémies par exemple dans le cas de maladies contagieuses. Par ailleurs, il existe d'autres recherches menées à l'aide des motifs spatio-temporels, comme ceux de [Newton *et al.*, 2015] pour comprendre et prédire la localisation des métastases dans le cas du cancer du sein. Ce type de recherche ouvre la voie à d'autres applications médicales à l'échelle du patient pour mieux comprendre et prévenir les évolutions d'une maladie.

L'étude des flux de patients a de surcroît une utilité prévisionnelle dans l'organisation sanitaire. En effet, [Jensen *et al.*, 2014] ont étudié les trajectoires de patients notamment dans le cas de maladies chroniques. Ils ont intégré leurs résultats dans un modèle prédictif afin de prévoir les flux de patients dans les années à venir et aussi adapter les infrastructures, le personnel et les dépenses à venir pour ces patients.

L'étude des flux de patients a diverses applications comme nous l'avons évoqué dans ce chapitre. Dans notre cas, il nous permet de mieux connaître les différentes étapes du parcours de soins des patients atteints d'un IM. Ceci peut avoir plusieurs applications : prévenir le patient de l'importance de l'observance de son traitement et de la mise en œuvre de la réduction des facteurs de risque ; améliorer la prise en charge par des visites de contrôle programmées ; mais aussi réduire les coûts. Pour mieux comprendre ces flux de patients, nous allons ajouter deux dimensions d'information à ces données qui sont le délai inter-séjours et le tarif du séjour.

Nous allons étudier les trajectoires de délais inter-séjours mais aussi celles des tarifs de ces séjours. L'objectif du chapitre 8 est de connecter ces résultats entre eux en établissant la répartition des patients dans des comportements types que ce soit en termes de délais de survenue d'un évènement hospitalier ou de tarifs du séjour hospitalier.

Chapitre

8 Caractériser des profils de délais et de tarifs

Dans le chapitre 7, nous avons caractérisé les flux de patients. Nous avons identifié les motifs des ré-hospitalisations (à travers les trajectoires de DP) et les types de soins pratiqués (à travers les trajectoires de GHM). Nous avons également évalué la proportion de patients concernée pour chaque flux. Cette étude des flux est intéressante pour anticiper les capacités d'accueil d'un établissement de santé ; l'évolution d'une pathologie ou d'une dégradation de l'état de santé d'un malade et permettra l'adaptation du traitement. Dans ce nouveau chapitre, nous allons utiliser les motifs caractérisant ces flux, tels que décrits dans le chapitre 7, pour évaluer les délais inter-hospitaliers et les tarifs de prise en charge. Dans le contexte difficile des réformes de santé [Safon, 2015], il s'agit d'une approche innovante visant l'amélioration de la planification sanitaire (avec une réduction des délais d'attente et la mise en place d'équipements adaptés) ainsi que la réduction des coûts qui sont deux leitmotivs d'un établissement de santé.

La planification sanitaire est notamment étudiée par l'analyse des délais d'attente de prise en charge. Cette problématique est abordée de diverses manières dans la littérature. La modélisation des files d'attente a été utilisée pour la gestion interne d'une unité de soins [El-Darzi *et al.*, 1998] ou plus particulièrement d'un service d'urgences [Laskowski *et al.*, 2009] ou encore pour la gestion des flux de patients au sein d'un hôpital entre les différentes unités de soins [Armony *et al.*, 2015]. [Rohleder *et al.*, 2005] ont employé des modèles de décision pour la gestion des salles d'opérations. [Potisek *et al.*, 2007] se sont servis de modèles de flux de patients basés sur des graphes pondérés afin de minimiser le délai d'attente et améliorer l'efficacité des visites pour les malades souffrant de douleurs chroniques.

La réduction des coûts peut se faire à plusieurs niveaux. Elle passe également par la gestion du temps. Par exemple, réduire les délais d'attente des patients afin d'optimiser l'utilisation des salles de chirurgie [Dexter *et al.*, 1999]. Elle est également liée à la gestion organisationnelle des soins [Armony *et al.*, 2015] : améliorer la coordination des tâches et des activités afin d'éviter une congestion

des processus hospitaliers. Elle peut, par ailleurs, être ajustée par la gestion d'allocations de ressources [Fassbender *et al.*, 2009, Gunes et Yaman, 2005]. Mais réduire les coûts peut aussi conduire à choisir une stratégie de soins plutôt qu'une autre [Buckley *et al.*, 2000, Ghosh *et al.*, 2001], sans compromettre la qualité de la prise en charge, ou alimenter une politique de vaccination, comme dans le cas du cancer du col de l'utérus [Ricciardi *et al.*, 2009].

Dans ce chapitre, nous nous intéressons à ces deux points d'intérêts : le délai entre deux hospitalisations et le coût de l'hospitalisation. Nous étudions le coût à travers la notion de tarif, puisque nous n'avons pas accès au coût réel de l'hospitalisation. Néanmoins, il convient de distinguer les coûts des tarifs. En effet, à un GHM est associé non pas un coût mais un tarif (voir chapitre 2 section 2.2.2). Ce tarif n'est pas forcément le reflet des coûts de prise en charge. Si, l'échelle nationale des coûts vise à assigner des coûts à des prises en charge, en revanche, l'autorité de régulation peut décider de diminuer unilatéralement des tarifs à des fins de politique de santé. Ainsi, dans cette situation les tarifs ne correspondent plus aux coûts [Milcent, 2017]. Notre objectif est d'établir à la fois des profils d'occurrence d'évènements et des profils d'évolution du tarif de la prise en charge. Nous envisageons le délai entre deux hospitalisations et le tarif de ces hospitalisations comme des séries chronologiques. Identifier des profils revient à regrouper dans une même catégorie des évolutions similaires. Autrement dit, identifier des profils revient à reconnaître des courbes qui évoluent de la même façon. Pour cela, nous avons utilisé un modèle, dérivé de l'algorithme de classification des k-moyennes, capable de classer des séries chronologiques.

Dans la section 8.1, nous présentons la méthode de classification des données longitudinales kmlShape. Puis, nous détaillons le processus de classification mis en place de l'extraction des données d'intérêt à la répartition des groupes de patients, créés dans le chapitre 7, dans les classes déterminées par kmlShape. Ensuite, dans la section 8.2, nous appliquons notre processus de classification afin de définir des profils de délais inter-séjours et de tarifs. Pour terminer, nous discutons les résultats dans la section 8.3 et nous suggérons des perspectives dans la section 8.4.

8.1 Classification de données longitudinales quantitatives

Les données longitudinales (ou séries chronologiques) sont des données dans lesquelles chaque variable a été mesurée plusieurs fois au cours du temps pour un même individu. Une manière d'analyser ces données est de les partitionner, c'est-à-dire de les diviser en des sous-groupes homogènes. Pour cela, plusieurs auteurs ont proposé des variantes de la méthode des k-moyennes [Dhillon *et al.*, 2004, Chandrasekhar *et al.*, 2011, Handhayani et Hiryanto, 2015]. D'autres méthodes reposent sur des modèles de mélange [Gaffney et Smyth, 1999, Chiou et Li, 2007, McNicholas et Murphy, 2010].
Dans la majorité des approches, deux individus sont rassemblés dans un même groupe lorsqu'ils ont des trajectoires proches à chaque temps de mesure. Cependant, ce type de méthodes ne prend pas en compte la forme générale des trajectoires.

Ainsi, la trajectoire moyenne d'un groupe ne renseigne pas sur la forme de la trajectoire, alors que, dans certains cas, l'évolution d'un phénomène peut avoir plus d'importance que son moment d'apparition. Pour remédier à cela, [Genolini *et al.*, 2016] proposent une méthode, nommée kmlShape, également basée sur les k-moyennes, décrite ci-dessous, afin de détecter des formes de courbes similaires. Cette méthode a déjà fait ses preuves dans le domaine de la santé [Genolini *et al.*, 2016]. Appliquée sur des données concernant la maladie d'Alzheimer, elle a permis de mettre en évidence quatre profils différents dans la perte des facultés de la personne malade. Appliquée, ensuite, sur des mesures de l'hormone LH chez des femmes fertiles, elle a permis de découvrir un nouveau profil de femmes ayant deux pics de LH. Alors qu'un seul pic avait initialement été établi comme référence dans ce domaine. Cette découverte pourrait, par ailleurs, apporter de nouveaux éléments dans la recherche en biologie de la reproduction.

Dans la suite de cette section, nous présentons, dans la section 8.1.1, le principe de fonctionnement de la méthode kmlShape et nous définissons les deux concepts clés de distance et de moyenne sur lesquels s'appuie cette méthode. Puis, dans la section 8.1.2, nous poursuivons avec le descriptif du processus de classification appliqué aux trajectoires de délais et de tarifs issues des données du PMSI.

8.1.1 Définitions préliminaires

La classification par les k-moyennes est un algorithme de partitionnement utilisé pour les données longitudinales. Elle alterne deux étapes : 1) le calcul des trajectoires moyennes de chaque groupe ; 2) le calcul des distances entre individu et la moyenne de son groupe. Cet algorithme affecte un individu dans le groupe où il est le plus proche. L'algorithme kmlShape [Genolini *et al.*, 2016] repose sur ces deux étapes. Il utilise la moyenne de Fréchet pour le calcul des trajectoires moyennes et la distance de Fréchet pour le calcul des distances entre individus et représentant de classe. Ces mesures confèrent à kmlShape la capacité de respecter la forme des courbes.

Définition 18 (Reparamétrisation)
Soit $t \in \mathbb{R}$ un réel et $[0, t]$ un intervalle de $\mathbb{R}$. Une reparamétrisation de $[0, t]$ est une fonction continue, croissante, surjective de $[0, t] \to [0, t]$.

Nous introduisons quelques notations pour les définitions de distances :

— d la distance euclidienne ;

— P et Q, deux courbes de $[0, t]$ dans $\mathbb{R}$;

— $\mathcal{A}$ l'ensemble de toutes les reparamétrisations de $[0, t]$;

— α et $\beta \in \mathcal{A}$ deux reparamétrisations ;

— $s \in [0, t]$ un réel.

Par analogie, une courbe P peut-être assimilée à la trajectoire d'un mobile voyageant à vitesse constante et $P \circ \alpha$ est la même trajectoire que P mais parcourue par un autre mobile avec une vitesse α.

Définition 19 (Distance de Fréchet)
La distance entre deux courbes P et Q reparamétrées par α et β au temps s est :
$d_{\alpha,\beta,s}(P,Q) = d(P \circ \alpha(s), Q \circ \beta(s)) = d\left(\begin{pmatrix} \alpha(s) \\ P(\alpha(s)) \end{pmatrix}, \begin{pmatrix} \beta(s) \\ Q(\beta(s)) \end{pmatrix}\right)$
La distance entre deux courbes P et Q reparamétrées par α et β est la distance maximale des distances $d_{\alpha,\beta,s}(P,Q)$ avec s variant dans $[0,t]$:

$$d_{\alpha,\beta}(P,Q) = Max_{s\in[0,t]}(d_{\alpha,\beta,s}(P,Q))$$

*Alors, la **distance de Fréchet** entre P et Q est le plus petit maximum possible entre P et Q après reparamétrisation de P et de Q :*

$$DistFrechet(P,Q) = Inf_{(\alpha,\beta)\in\mathcal{A}^2} d_{\alpha,\beta}(P,Q)$$

Si la variable d'intérêt est mesurée avec des échelles de temps différentes cela peut impacter le calcul de distance. Par voie de conséquence, cela peut affecter le résultat du partitionnement. Un exemple illustre ce propos dans la figure 8.1. Dans le schéma supérieur, selon la distance de Fréchet, la courbe Q est plus proche de la courbe R. Tandis que, dans le schéma inférieur, la courbe Q est plus proche de la courbe P. La définition qui suit permet de prendre en compte les changements d'échelle.

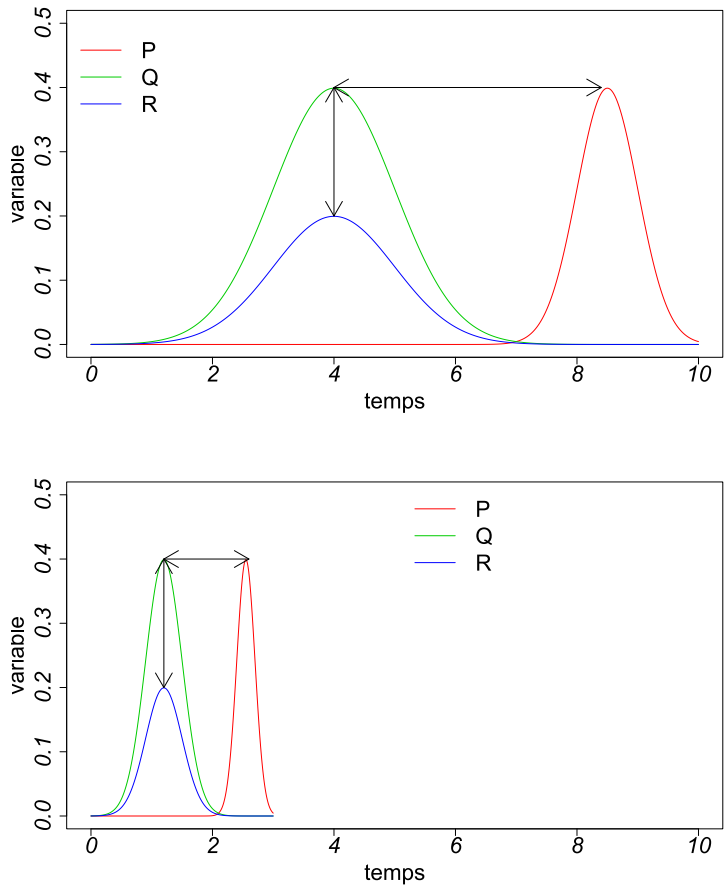


Figure 8.1 – Impact du changement d'échelle dans le calcul de la distance de Fréchet.

Définition 20 (Distance de Fréchet généralisée)
Soit λ le paramètre de l'échelle de temps. La distance de Fréchet généralisée au paramètre λ entre deux courbes P et Q est la distance de Fréchet obtenue après la transformation affine

$$\begin{array}{rccc} A: & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & (x,y) & \longmapsto & (\lambda x, y) \end{array}$$

c'est-à-dire $DistFrechet_\lambda(P,Q) = Inf_{(\alpha,\beta)\in\mathcal{A}^2} Max(P\circ\alpha\circ A, Q\circ\beta\circ A)$

Définition 21 (Moyenne de Fréchet)
La moyenne de Fréchet entre deux courbes P et Q est définie par :

$$MeanFrechet_\lambda(P,Q) = \left(\frac{P\circ\alpha\circ A + Q\circ\beta\circ A}{2}\right)$$

Cette définition peut se généraliser à n trajectoires en procédant par des regroupements deux par deux. Dans une population de n individus, il est possible de constituer des paires d'individus (avec un poids égal à 1) et de calculer les moyennes de Fréchet pour chaque paire. Les moyennes ainsi déterminées peuvent être à leur tour combinées par paires. Il est alors possible de calculer les moyennes de Fréchet pour chaque paire en les pondérant par le nombre d'individus utilisés pour le calcul de chacune. Ce principe est réitéré jusqu'à obtenir une moyenne unique. Cette méthode de calcul pas à pas réduit la complexité de l'algorithme de $O(t^n)$ à $O(nt^2)$.

Exemple : la figure 8.2 présente un exemple de calcul de distances moyennes, selon la méthode de Fréchet ou la méthode euclidienne, entre une série de courbes (dessinées en gris). Dans cet exemple, nous pouvons observer que la moyenne de Fréchet (courbe orange et trait épais) conserve la forme générale des courbes contrairement à la distance euclidienne (courbe noire et trait épais).

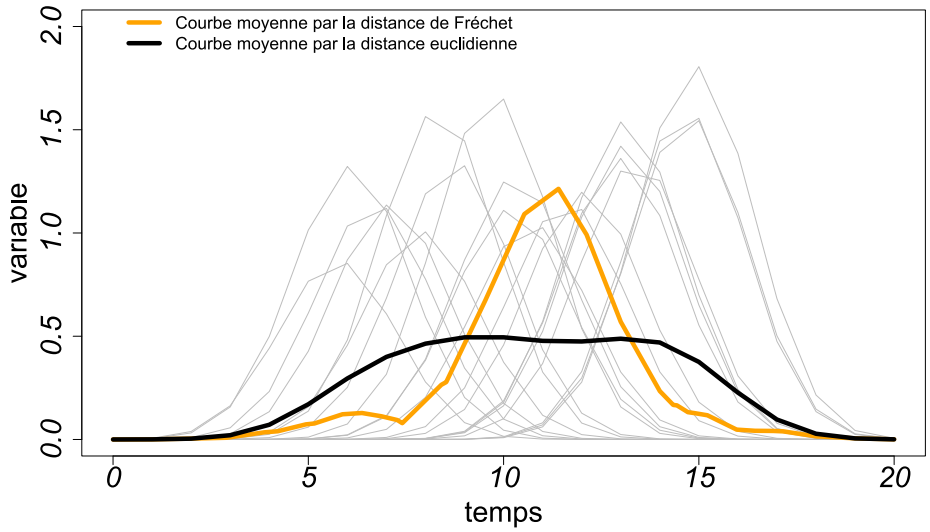


Figure 8.2 – Exemple de comparaison entre la moyenne de Fréchet et la distance euclidienne.

Nous avons intégré cette méthode de classification des données longitudinales dans un processus de classification de trajectoires que nous détaillons dans la section qui suit.

8.1.2 Description du processus de classification

Le processus de classification des données longitudinales, synthétisé dans la figure 8.3, se fait en plusieurs étapes.

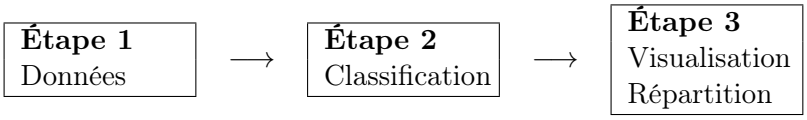


Figure 8.3 – Étapes du processus de classification.

Étape 1 : Acquisition et retraitement des données. Elle consiste à récupérer les données concernant les délais inter-séjours et la valorisation des tarifs des séjours. Les délais inter-séjours sont calculés par rapport au séjour suivant : il s'agit du nombre de jours écoulés entre le dernier jour d'une hospitalisation et le premier jour de l'hospitalisation suivante. Le tarif d'un séjour a été déterminé en tenant compte de la valorisation du séjour liée au GHS [1] et des suppléments [2]. Bien que cette valorisation ne soit pas un calcul exact du tarif de chacun des séjours, elle nous fournit une bonne estimation du tarif réel de l'hospitalisation.
Nous avons noté de fortes disparités entre les trajectoires de tarifs. Nous avons, alors, appliqué une transformation à l'aide du logarithme népérien afin de lisser les courbes et faciliter leur classification [3]. Dans la suite, ces deux types de trajectoires : délai et tarif sont traités de façon séparée. Pour chaque séjour, nous avons un délai au séjour suivant, si il y a un séjour ensuite, et un tarif de séjour. Ainsi, nous obtenons des mesures à chaque estampille temporelle. Ces données constituent les données longitudinales.

Étape 2 : Classification et choix du nombre de classes optimal. Nous classons, par contexte [4], les trajectoires de délais puis celles de tarifs à l'aide de la méthode kmlShape. Comme dans toute classification, se pose la question du choix du nombre de classes « optimal » [Dubes, 1987]. Il existe de nombreux indicateurs [Pal et Biswas, 1997, Maulik et Bandyopadhyay, 2002] plus ou moins spécifiques suivant les critères d'attente de l'analyse : hétérogénéité inter classes, homogénéité intra classe, ou un équilibre entre ces derniers critères. Ces critères qualité sont construits pour fonctionner avec des distances classiques comme par exemple la distance euclidienne. Par conséquent, dans notre cas, cela n'aurait pas de sens d'utiliser ces indicateurs pour mesurer la qualité de notre classification. De plus, n'ayant pas de « gold standard », nous optons, alors, pour une méthode plus analytique. Cette

1. Nous avons utilisé le tarif du GHS public.
2. Les suppléments pour la réanimation, les soins intensifs et les soins continus.
3. Pour éviter d'obtenir un trop grand nombre de classes ce qui deviendrait ardu à interpréter.
4. Tel que défini dans le chapitre 5.

méthode consiste à utiliser la répartition de classes, issue de la classification par kmlShape, dans un modèle prédictif du décès. La répartition offrant les meilleurs résultats est celle qui sera retenue. Plus concrètement, soit k le nombre de classes, nous classons les trajectoires en faisant varier k entre 2 et 6. À chaque itération, nous effectuons une régression logistique de l'état final du patient suivant le découpage obtenu par kmlShape. Nous définissons le k « optimal » comme étant celui pour lequel le modèle est le meilleur (critères de décision AIC et p.value). Le nombre k de classes est déterminé pour chaque contexte et pour chaque type de trajectoire. Il sera donc différent selon le contexte [5].

Étape 3 : Visualisation des représentants de classes et Répartition des patients identifiés par Get_Move selon la classe de kmlShape. À l'issue de l'étape 2, chaque patient du contexte concerné est affecté à une classe dont le représentant est la moyenne de Fréchet. Les courbes des représentants de classe sont lissées afin de faciliter la lecture graphique. Nous associons à ce graphique celui résumant les mesures de dispersion à chaque estampille de temps pour chacune des classes ainsi identifiés. L'objectif est de quantifier, au sein d'une même classe, les écarts autour de valeurs moyennes (au sens habituel). Puis, nous établissons la répartition des patients par classe selon les groupes constitués suite à l'étude des flux de patients par Get_Move (voir chapitre 7), comme illustré dans la figure 8.4.

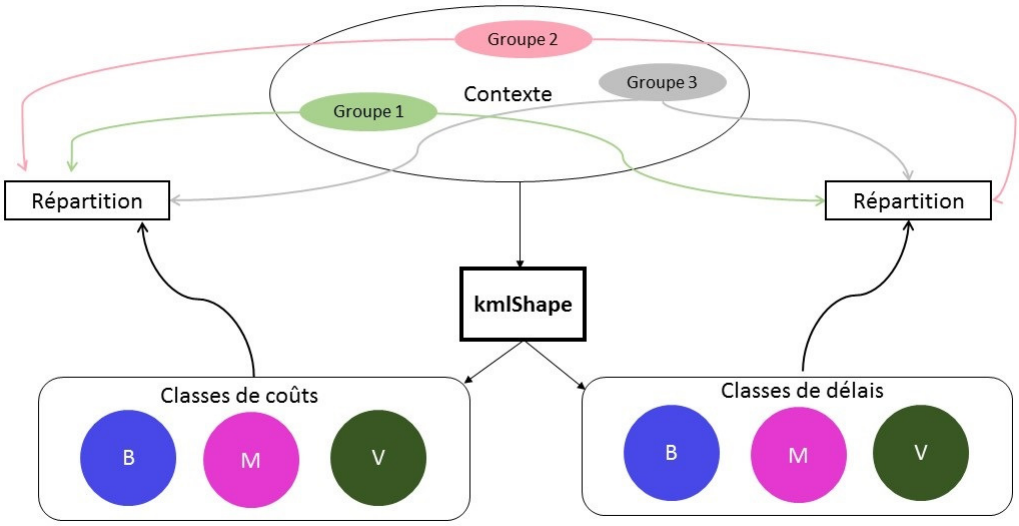


Figure 8.4 – Schéma de classification et de répartition des groupes dans les classes.

Dans la suite, nous avons appliqué ce processus de classification aux trajectoires de délais et de tarifs pour les mêmes contextes étudiés dans la section 7.2 à savoir : Homme & +65 ans & 5-60 séjours et Femme & +65 ans & 5-60 séjours.

5. On notera que dans les exemples présentés, dans la suite du manuscrit, k sera égal le plus souvent à 3, sauf pour le cas des hommes pour les trajectoires de tarifs.

8.2 Expérimentations

Dans le chapitre 7, nous avons établi des groupes à partir des essaims identifiés selon le premier évènement de la trajectoire. Pour chaque situation (contexte et type de trajectoire : GHM ou DP), nous avons rassemblé les patients en trois groupes différents. Nous rappelons ces différents groupes dans le cas des trajectoires de DP et entre parenthèses dans le cas des trajectoires de GHM :

— Groupe 1, le premier évènement est angine de poitrine (stenting sans IM).
— Groupe 2, le premier évènement est cardiopathie ischémique (stenting avec IM).
— Groupe 3, le premier évènement est IM (autres).

Il s'agit ici d'une synthèse des résultats. Cela ne signifie pas que le groupe 1 des trajectoires de DP, contient les mêmes patients que celui des trajectoires de GHM.

Nota Bene : Dans toute la suite du chapitre, lorsque le mot groupe sera employé, il fera, alors, référence à ceux construits dans le chapitre 7 et rappelés ci-dessus.

Dans la section 8.2.1, nous décrivons les résultats obtenus dans le cas des trajectoires de délais inter-séjours et dans la section 8.2.2, ceux concernant les trajectoires de tarifs de séjours. Pour chaque partie, nous commençons par décrire les résultats issus de la classification par kmlShape puis, nous détaillons les différents profils obtenus. Ensuite, nous examinons la dispersion de la classe autour de la moyenne pour chaque classe à chaque estampille temporelle. Enfin, nous examinons la répartition des groupes de patients présentés ci-dessus au sein de chaque classe, comme illustré dans la figure 8.4.

8.2.1 Trajectoires de délais inter-séjours

La classification des trajectoires de délais chez les +65 ans selon le sexe donne trois classes, représentées dans la partie droite de la figure 8.5, pour les deux sous-populations avec en abscisse les estampilles de temps et en ordonnée le délai entre deux séjours en jours. Ces classes sont des courbes différenciées par les couleurs : bleue (classe B), magenta (classe M) et verte (classe V). Elles caractérisent des profils différents de délais de ré-hospitalisations. En effet, nous pouvons voir dans la figure 8.5 que les courbes ont des formes très différentes. Comme il s'agit de délais inter-séjours, le graphique se lit de la façon suivante, à l'estampille de temps t_0 est affichée le temps qui sépare la première hospitalisation de la seconde. Par exemple, chez les femmes, la classe M, représentée par la courbe magenta (trait fin), a une ordonnée égale à 500 au temps t_0. Ainsi le délai entre la première hospitalisation et la deuxième hospitalisation, pour les femmes faisant partie de cette classe, est de 500 jours. Ensuite, à l'estampille de temps t_1, son ordonnée est égale à 515. Le délai entre la deuxième hospitalisation et la troisième est de 515 jours, ainsi de suite.

Dans le cas des hommes, la classe B contient des patients ayant des délais de plus en plus courts, donc des hospitalisations de plus en plus rapprochées. Elle concerne 65% des hommes.

La classe M regroupe 18% des patients. Elle représente des patients ayant des hospitalisations de plus en plus espacées allant de 4 à 5 mois jusqu'à plus de 2 ans.

La classe V regroupe 17% de la population. Elle caractérise des patients ayant des délais courts en début de parcours, un évènement très éloigné en milieu de parcours et des délais courts à nouveau.

Dans le cas des femmes, la classe B, regroupe 56% des patientes. Elle représente des patientes ayant des délais courts (inférieurs à 4 mois) qui augmentent puis diminuent.

La classe M regroupe 27% de la population. Elle concerne des patientes ayant des évènements de façon espacée en début de parcours puis plus rapprochés à la fin.

La classe V regroupe 17% des patientes. Elle décrit un profil similaire à la classe M chez les hommes.

Considérons les mesures de dispersion représentées dans la partie gauche de la figure 8.5. Nous avons représenté les dispersions au sein de chaque classe à chaque estampille de temps à l'aide de boîtes à moustaches, en conservant le même code couleur afin de différencier chaque classe. Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. On ajoute alors des segments aux extrémités menant jusqu'aux valeurs extrêmes. À chaque estampille de temps, il y a donc une boite à moustache par classe. Ce type de graphique nous apporte une information supplémentaire sur les variations, en termes de délais, au sein de chaque classe mais aussi à chaque instant de la trajectoire. En effet, nous avons à l'issue de la classification par kmlShape un représentant de classe selon une cohérence de trajectoire : par exemple si le représentant de classe est une courbe décroissante alors tous les patients de cette classe ont des délais inter-séjours de plus en plus courts mais nous ne savons pas si ces patients ont des délais qui sont proches de ceux du représentant de classe ou au contraire éloignés. Après examens des graphiques de dispersion, nous notons que, chez les hommes, la dispersion est importante pour la classe B au temps t_0 et pour la classe V aux temps t_1 et t_2. En revanche, chez les femmes, seule la classe M a une forte dispersion autour de sa moyenne quelle que soit l'estampille de temps.

Nous nous sommes, ensuite, intéressés aux groupes particuliers décrits en introduction de cette section. Nous avons regardé la répartition de ces patients dans les classes formées par kmlShape. Les résultats sont résumés dans le tableau 8.1. Ce dernier renseigne la répartition en pourcentage des groupes d'hommes dans les classes de délais. Les résultats concernant la population des femmes sont mis entre parenthèses. Dans ce tableau sont également répertoriés les effectifs totaux pour chaque groupe. Nous remarquerons que la répartition est identique à celle de la population de départ. Toutefois, nous observons quelques exceptions. Ainsi, chez les hommes comme chez les femmes, les différences sont importantes dans les groupes de trajectoires de DP et dans le groupe 3 des trajectoires de GHM.

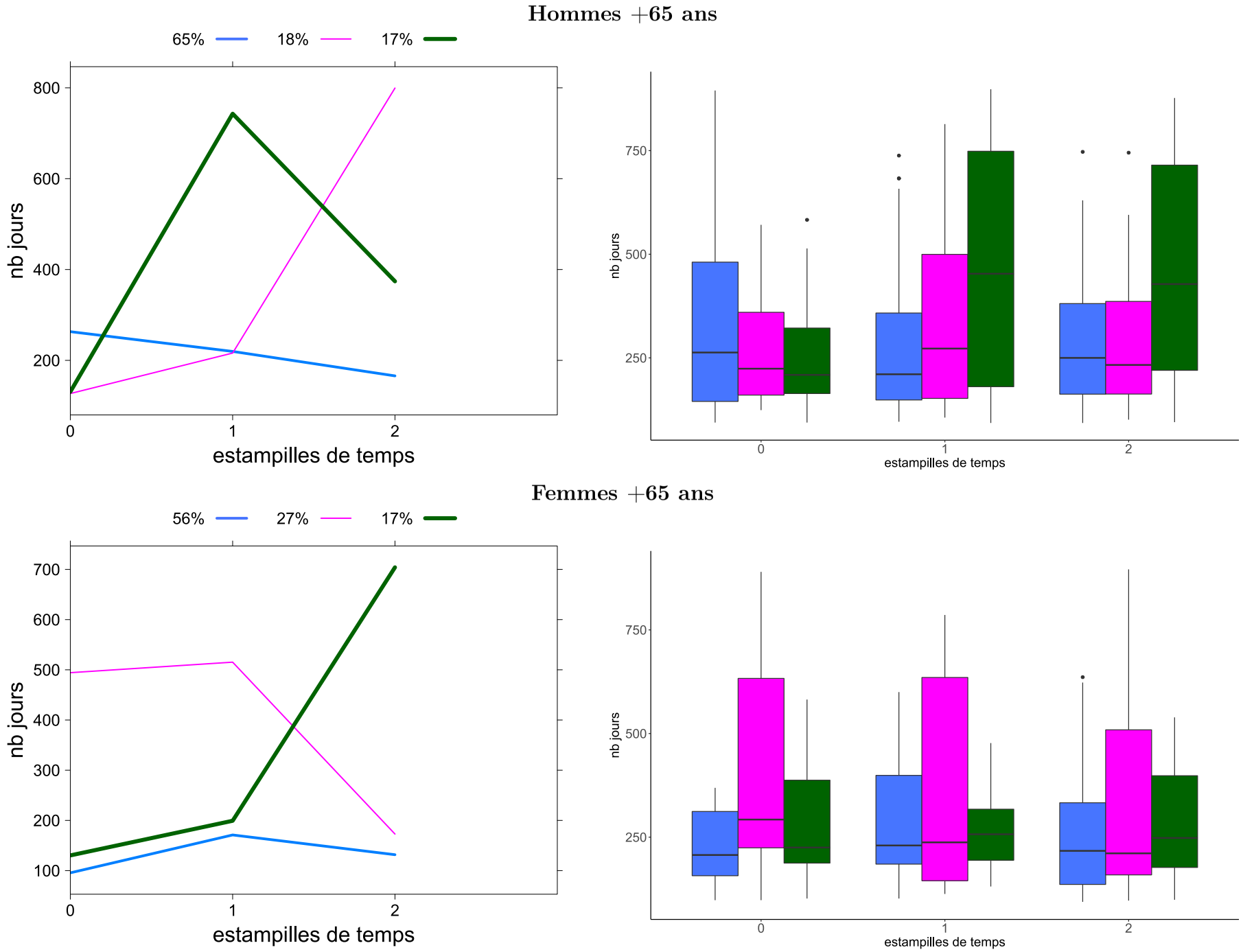


Figure 8.5 – Classes de délais à droite et mesures de dispersion à chaque estampille de temps par classe selon le sexe.

Tableau 8.1 – Répartition en pourcentage des hommes et des femmes (entre parenthèses) de chaque groupe dans les classes de délais.

Type de trajectoire	**Groupe** n°	Effectif	**Classes de délais** B	M	V
DP	1	4 845 (1 711)	51 (47)	30 (28)	19 (25)
	2	5 422 (1 201)	49 (52)	29 (32)	22 (16)
	3	4 943 (2 651)	50 (45)	32 (34)	18 (21)
GHM	1	10 565 (3 504)	61 (56)	20 (25)	19 (19)
	2	2 029 (828)	59 (60)	20 (23)	21 (17)
	3	149 (122)	64 (51)	21 (36)	15 (13)

8.2.2 Trajectoires de tarifs de séjours

Il résulte de la classification des trajectoires de tarifs de séjours chez les +65 ans, quatre classes chez les hommes et trois chez les femmes. Ces résultats sont représentés dans la partie droite de la figure 8.6 pour les deux sous-populations avec en abscisse les estampilles de temps et en ordonnée le tarif du séjour en euros. Comme dans la section 8.2.1, nous avons utilisé les mêmes labels pour les différentes classes. Nous ajoutons la classe R, correspondant à la courbe rouge chez les hommes. Contrairement au cas des trajectoires de délais, nous obtenons ici des profils vraiment très différents selon le genre. Ici, le graphique se lit de la façon suivante : par exemple, chez les hommes, pour la classe V, représentée par la courbe Verte (trait épais), son ordonnée est égale à 5 980 à l'estampille de temps t_0. Le tarif du premier séjour de la classe V est donc de 5 980€, ainsi de suite.

Chez les hommes, la classe B contient des patients ayant des séjours avec des tarifs relativement constants en début de parcours puis qui augmentent fortement au-delà de 4 000€ et diminuent ensuite. Elle concerne 31% des hommes. La classe M représente 26% des hommes. Elle caractérisent des patients ayant des séjours avec des tarifs oscillant autour de 2 500€. La classe V caractérise des patients ayant des séjours avec des tarifs diminuant fortement au fur à mesure du parcours puis augmentant légèrement en fin de parcours. Elle regroupe 24% des hommes. Enfin, la classe R concerne 19% des hommes. Elle caractérise des patients ayant des séjours avec des tarifs de plus en plus élevés démarrant à moins de 4 000€ pour finir à plus de 7 000€.

Chez les femmes, la classe B regroupe 35% des patientes. Elle représente des patientes ayant des séjours avec des tarifs augmentant au fur à mesure du parcours et diminuant légèrement. La classe M a un comportement opposé à la classe B. Elle caractérise, au contraire, des patientes ayant des séjours avec des tarifs en forte diminution, puis ré-augmentant fortement en fin de parcours. Elle regroupe 34% de la population. Pour finir, la classe V regroupe 32% de la population. Elle caractérise des patientes avec des tarifs de séjours qui augmentent en début de parcours puis diminuent jusqu'à la fin du parcours.

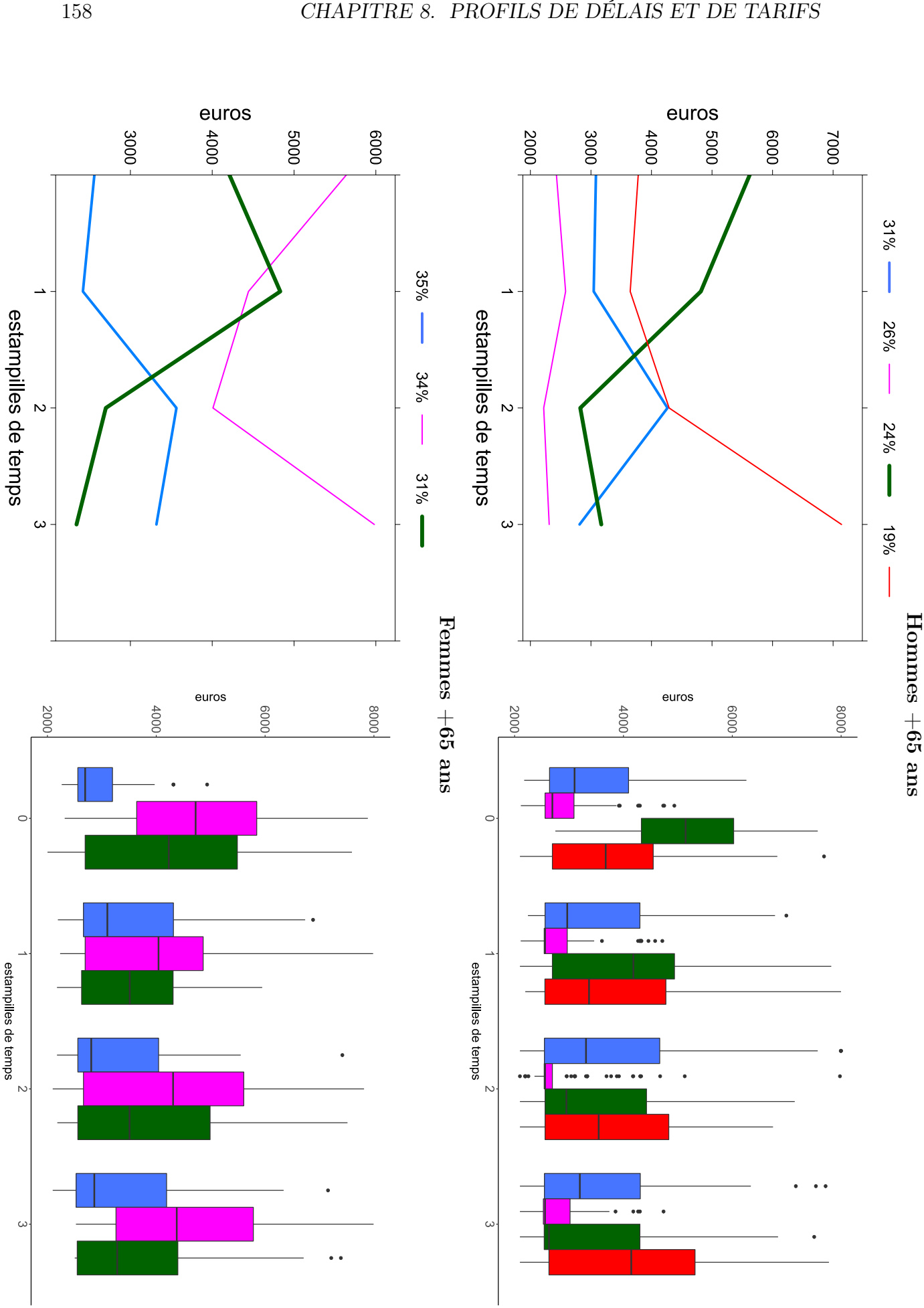


Figure 8.6 – Classes de tarifs à droite et mesures de dispersion à chaque estampille de temps par classe selon le sexe.

Considérons les mesures de dispersions représentées dans la partie gauche de la figure 8.6. Le principe de représentation est identique à celui décrit dans le cas des trajectoires de délais (voir section 8.2.1). De manière générale, les mesures de dispersions sont importantes pour toutes les classes de tarifs et à chaque estampille de temps. Néanmoins, ceci n'est pas observé ni pour la classe M, chez les hommes, ni pour la classe B au temps t_0, chez les femmes.

La répartition des groupes de patients est résumée dans le tableau 8.2. Le principe de présentation de ce tableau est identique à celui décrit dans la section précédente. À nouveau, nous constatons que la répartition est assez similaire à celle de la population cible avec toutefois des exceptions. Dans les deux cas (homme ou femme), le groupe 2 pour les trajectoires de DP et le groupe 3 pour les trajectoires de GHM se distinguent de la répartition dans le cas général.

Tableau 8.2 – Répartition en pourcentage des hommes et des femmes (entre parenthèses) de chaque groupe selon la classe de tarif.

Type de trajectoire	**Groupe** n°	Effectif	**Classes de tarifs** B	M	V	R
DP	1	4 845 (1 711)	31 (41)	29 (23)	24 (36)	17
	2	5 422 (1 201)	31 (52)	33 (19)	22 (29)	14
	3	4 943 (2 651)	27 (25)	15 (36)	43 (39)	15
GHM	1	10 565 (3 504)	32 (43)	31 (25)	20 (32)	17
	2	2 029 (828)	30 (17)	11 (35)	46 (48)	13
	3	149 (122)	29 (44)	33 (26)	24 (30)	14

8.3 Discussion

Cette section se décompose en deux sous-sections. Une première section 8.3.1, dans laquelle nous proposons une synthèse des résultats en les connectant avec les schémas de flux établis dans le chapitre 7. Puis, une deuxième section 8.3.2, dans laquelle nous évoquons les limites de nos analyses.

8.3.1 Analyse des résultats

Dans cette section, nous analysons les résultats obtenus suivant le même plan que leur présentation dans la section 8.2 : tout d'abord les trajectoires de délais et ensuite, les trajectoires de tarifs. Pour lier ces résultats à ceux de l'étude des flux de patients, nous évoquerons aussi bien les schémas de flux des trajectoires de DP, que ceux des trajectoires de GHM. En effet, nous avons pu constater que dans le cas de la répartition des groupes, il y avait un effet miroir entre les trajectoires de DP et celles de GHM.

Commençons donc, par la classification des trajectoires de délais inter-séjours. Chez les hommes, la répartition de la population cible du contexte dans les trois classes de délais nous renseigne sur la fréquence des occurrences des évènements hospitaliers. Pour la majorité d'entre-eux, ces occurrences sont de plus en plus rap-

prochées dans le temps avec une moyenne variant de 4 à 6 mois (classe B). De plus, lorsque l'on s'intéresse aux groupes, cette répartition est identique. Nous avions mis en évidence des hospitalisations à répétition pour des stentings. Ainsi, nous pouvons affirmer que dans la majorité des cas, quel que soit l'évènement de la première hospitalisation ces patients subissent des actes de cardiologie interventionnelle en moyenne tous les 4 à 6 mois. Néanmoins, une proportion non-négligeable (environ 30%) de ces patients a, quant à elle, des hospitalisations de plus en plus espacées dans le temps allant de 1 an à presque 3 ans (classe M).

Chez les femmes, comme chez les hommes, la répartition des groupes est assez proche de celle de la population cible du contexte. En revanche, les profils de délais sont assez différents de ceux des hommes. En effet, le profil majoritairement représenté est celui de patientes avec des hospitalisations rapprochées mais, avec en milieu de parcours deux séjours plus espacés (environ 7 mois au lieu de 3). Comme chez les hommes, une proportion non négligeable (environ 30%) a un profil tout à fait différent. Ce profil est celui des femmes qui ont des hospitalisations de moins en moins espacées dans le temps mais avec des occurrences en début de parcours séparées de plus d'un an (classe M). Autrement dit, si l'on s'intéresse par exemple aux femmes dont le premier évènement est un IM avec pose de stent, alors 60% de ces femmes ont à nouveau des ré-hospitalisations pour stenting au bout de 3 mois, puis 7 mois et à nouveau 3 mois. Presque 30% d'entre-elles ont des ré-hospitalisations après plus d'un an puis séparées de moins de 8 mois en fin de parcours.

Les rythmes d'occurrence dans les schémas de flux sont alors très différents mais que peut-on dire des tarifs ? À titre de référence, le tarif d'une hospitalisation pour IM varie entre 5 600€ pour un niveau de sévérité faible à 23 800€ pour un niveau de sévérité très élevé [Milcent, 2015].

Intéressons nous, maintenant, aux résultats concernant la classification des trajectoires de tarifs. L'évolution des tarifs de séjours est différente selon le sexe. Chez les hommes, la population est quasi équi-répartie dans les classes B, M et V. Une proportion moins importante se retrouve dans la classe R. La répartition des groupes est analogue à celle de la population cible. Ainsi, elle suggère que les flux de patients commençant par une angine de poitrine ou une cardiopathie ischémique ont soit une évolution à la baisse, soit une forte évolution à la baisse, soit encore une tendance à la baisse avec un pic à 4 200€ en milieu de parcours. Mais, pour le flux de patients dont le premier évènement est un IM, c'est un peu discordant avec la répartition générale. Pour la majorité d'entre-eux, les tarifs sont élevés en début de parcours (en moyenne 7 500€) puis, l'évolution est à la baisse avec une moyenne des tarifs avoisinant 3 000€. Peu de ces patients sont concernés par une évolution de tarifs de séjours avec des moyennes en-deçà de 2 000€.

Chez les femmes, la population est quasi équi-répartie dans les trois profils d'évolution de tarifs. En revanche, pour les groupes, nous observons des différences par rapport à cette répartition. Les flux de patientes ayant comme premier évènement une angine de poitrine ou une cardiopathie ischémique ont, pour la majeure partie d'entre-elles, une évolution à la hausse. Toutefois, les

tarifs moyens n'excèdent pas 4 000€. Au contraire, les flux de patientes démarrant avec un IM ont, pour la majeure partie d'entre-elles, une évolution à la baisse de leurs tarifs avec un pic à presque 5 000€ en moyenne en début de parcours.

Ainsi, si l'on compare ces résultats aux tarifs de référence cités plus haut, on s'aperçoit que quel que soit le profil de tarif, l'évolution moyenne des tarifs oscille dans une fourchette basse avec des tarifs inférieurs à 8 000€. Toutefois les graphes de dispersion nous montrent que dans une même classe la variabilité des tarifs autour de la moyenne est importante tout comme précisé dans [Milcent, 2015]. Par ailleurs, la synthèse de ces travaux soulèvent des interrogations au sujet des rythmes de ré-hospitalisation et notamment sur le lien entre le rythme des occurrences d'hospitalisation et évolution de tarif de séjour. Nous allons voir dans la section suivante qu'il n'est pas aisé d'y répondre.

8.3.2 Limites de cette étude

Nous trouvons plusieurs limites à notre étude. Tout d'abord, une première limite est comparative. En effet, la littérature regorge de travaux sur les délais entre l'apparition des premiers symptômes et l'accès aux soins [Albarqouni *et al.*, 2016, Miller *et al.*, 2017], entre l'arrivée aux urgences et la prise en charge de l'IM [Belle *et al.*, 2016] mais aussi, des délais de réadmissions pour récidive [Asche *et al.*, 2016]. Toutefois, le délai d'occurrence d'autres évènements liés à l'IM est assez peu étudié. Les chercheurs orientent leurs analyses sur des complications particulières, comme par exemple l'insuffisance cardiaque [Sulo *et al.*, 2016] et évaluent l'incidence et le délai d'apparition de cette pathologie suite à un IM. Les auteurs ont établi que, dans le cas de la population norvégienne, l'incidence de l'insuffisance cardiaque varie de 8% à 27% selon le sexe et l'âge. Sur un suivi de 3 ans, 12% des personnes ayant eu un premier épisode d'IM aigu ont été ré-hospitalisées pour une insuffisance cardiaque ou sont décédées. À notre connaissance, il n'existe pas d'étude plus générale sur la ré-hospitalisation, qu'elle soit liée de près ou de loin à l'IM et à la variation des délais d'apparition de ces évènements.
Par ailleurs, nous avons effectué l'analyse des trajectoires de tarifs uniquement d'un point de vue hospitalier, à l'aide du tarif des séjours selon le protocole défini par l'ATIH. Par défaut d'accès aux bases, nous n'avons pas intégré dans notre analyse ni les coûts de consultations de généralistes et/ou d'un cardiologue de ville, ni ceux de la consommation des traitements médicamenteux, ni ceux des interventions des services d'urgence avant l'arrivée dans un établissement de santé (comme par exemple une ambulance). Or, dans le cas d'études de coûts directs, [Blin *et al.*, 2016] se sont intéressés aux patients de +50 ans diabétiques ou avec une maladie rénale. Ils ont déterminés les coûts moyens de suivi de ces patients sur trois ans après un IM aigu. En tenant compte des frais d'hospitalisation et de la consommation de médicaments, ils ont établi que le coût moyen était de 20 000€. Dans une étude canadienne, [Cohen *et al.*, 2006] ont établi un coût moyen de l'ordre de 19 000€, mais le suivi est fait sur 6 ans sur une population de +65 ans. De plus, les auteurs ont pris en compte à la fois les soins de ville, les soins hospitaliers et la prescription de médicaments. En outre, les études comparatives de coûts de la littérature se basent, pour la plupart [Ades *et al.*, 1997, Kuntz *et al.*, 1996, Schneeweiss *et al.*,

2007, Mark *et al.*, 1995], sur un score mesurant un coût/efficacité, c'est-à-dire, un score mesurant un coût pondéré selon le nombre d'années de vie gagnées. Par exemple, [Fineberg *et al.*, 2010] ont comparé des stratégies de soins à l'admission de patients présentant les symptômes d'un IM. Ils suggèrent, pour ceux ayant un faible risque d'épisode aigu, qu'un accueil dans une unité de soins intermédiaire serait plus approprié. [Seidl *et al.*, 2017] ont également comparé des stratégies de soins, dans le suivi post-hospitalier, entre une prise en charge usuelle et la mise en place de soins infirmiers à domicile. Leur étude montre qu'en termes de coûts il n'y a pas de différence significative. En revanche, la deuxième stratégie apporte des changements significatifs en termes de survie.
De ce fait, il apparaît difficile de se comparer directement à d'autres travaux du domaine qui, soit prennent en compte tout ou partie de ces dépenses [Kucenic *et al.*, 2000, Philippe *et al.*, 2017, Lomas *et al.*, 2016], soit analysent un coût/efficacité [Kaul *et al.*, 2011, Fineberg *et al.*, 2010, Seidl *et al.*, 2017]. Toutefois, avec des données appropriées nous pourrions réaliser des calculs similaires avec l'approche proposée.

Une autre limite à ces expérimentations est la dissociation des analyses des trajectoires de délais. En effet, nous ne pouvons pas déterminer si une évolution des tarifs à la hausse est également liée à des occurrences d'hospitalisations éloignées dans le temps ou au contraire rapprochées. Ce qui, par ailleurs, n'a pas la même portée. Dans le premier cas, il s'agira plutôt d'une vision de la gestion de la trajectoire sur le long terme alors que dans l'autre cas il s'agira plutôt d'une vision à court terme. Cela pourrait éventuellement amener à reconsidérer la stratégie de soins en elle-même. Ainsi, établir des profils d'évolution conjointe permettrait d'adapter des politiques de pilotage de la trajectoire du patient selon la situation. Dans cette perspective, nous proposons dans la continuité de ces travaux d'utiliser une méthode de classification capable de traiter des trajectoires en trois dimensions. L'analyse des trajectoires jointe est d'ores et déjà utilisée. [Pingault *et al.*, 2013] ont clarifié les liens entre les symptômes de l'inattention/hyperactivité chez les enfants et le développement d'une dépendance/toxicomanie à l'âge adulte. Ils ont montré que l'association inattention/hyperactivité chez l'enfant avec la dépendance/toxicomanie est en grande partie attribuable à son association avec des problèmes de conflits pendant l'enfance. [Hall *et al.*, 2017] ont étudié les facultés de récupération des patients, suite à un épisode psychotique, en associant plusieurs critères (fonction cognitive, traitement...). Ils ont conclu que le traitement des patients ayant des trajectoires de faible récupération pouvait être plus agressif, de sorte à réduire les fonctions de détérioration et améliorer le rétablissement.

8.4 Conclusion

Dans ce chapitre, nous avons classé des trajectoires de délais et des trajectoires de tarifs dans les sous-populations des +65 ans différenciées selon le genre. Nous avons lié ces résultats aux groupes de patients créés dans l'étude des flux de patients afin de déterminer des profils types en termes d'occurrence d'hospitalisation mais

aussi de facturation de ces évènements. Pour cela, nous avons employé une méthode de classification de données longitudinales basée sur le modèle des k-moyennes et adaptée par [Genolini *et al.*, 2016] pour respecter les formes des trajectoires.

Nous retenons de nos expérimentations que dans les schémas de flux de patients débutant par un IM, la majorité des occurrences suivantes sont de plus en plus proches dans le temps avec une tendance de l'évolution des tarifs à la baisse de ces évènements. Néanmoins, un tiers de cette population est concerné par une augmentation des tarifs, ou des occurrences d'hospitalisations de plus en plus éloignées dans le temps.

Toutefois, le choix des variables analysées ainsi que notre stratégie d'étude s'imposent comme des limites à la comparaison à d'autres travaux du domaine. La stratégie d'analyse que nous avons choisie ne permet pas de faire le lien entre les différents profils de délais et de tarifs. Dans la suite de ces travaux, nous envisageons d'utiliser une autre méthode de classification [Genolini *et al.*, 2013], pour les données en trois dimensions, de sorte à associer à la fois les trajectoires de délais et de tarifs. Cette approche permettrait de prendre en compte la co-évolution des variables et ainsi, d'établir éventuellement un lien entre les délais inter-séjours et les tarifs de ces séjours. Mieux comprendre l'occurrence d'un évènement hospitalier en l'associant à son tarif pourrait permettre d'envisager des évolutions en termes de prise en charge des patients. De plus, il serait également intéressant de revenir aux coûts réels à l'aide d'une estimation des coûts de GHM [Landais *et al.*, 2013].

Partie V

Conclusion générale et perspectives

Science sans conscience n'est que ruine de l'âme.

Rabelais, Pantagruel.

Table des matières

9 Conclusions et perspectives **169**
9.1 Bilan 170
9.2 Perspectives 171
9.2.1 Méthodes 172
9.2.2 Applications au domaine de la santé 175

Chapitre

9

Conclusions et perspectives

Si la médecine a réduit la mortalité de l'IM par un facteur 3 en vingt ans, grâce notamment au développement des techniques non-invasives pour rétablir le flux sanguin de l'artère obstruée, l'IM demeure cependant un enjeu majeur de santé publique. C'est une maladie grave nécessitant une hospitalisation et une prise en charge dans une Unité de Soins Intensifs de Cardiologie (USIC).

La première motivation d'étude de cette pathologie est sa gravité : elle conduit au décès dans 25% des cas lors de la première année qui suit l'épisode. La deuxième motivation est l'amélioration de la planification sanitaire : l'impact de cette pathologie est important à prendre en compte dans le contexte de réduction des dépenses de santé. Les travaux présentés dans ce manuscrit s'inscrivent dans cette optique. Nous nous sommes intéressés aux données d'hospitalisation contenues dans le PMSI. Ces données recueillies à des fins de gestion et de tarification représentent une mine d'informations pour la recherche.

Contrairement à beaucoup d'études menées jusqu'à présent sur ce thème, nous avons souhaité étudier cette pathologie, non pas dans sa prise en charge immédiate, mais sur le suivi à long terme. Or, suivre le parcours d'un patient sur le long terme est une notion intrinsèque au concept de trajectoire. Ainsi, un nouvel enjeu a émergé : celui d'explorer les trajectoires de patients via les bases médico-administratives.

Dans ce but, nous avons proposé un canevas de méthodes issues de différentes disciplines comme l'informatique médicale, la fouille de données et la biostatistique. Par conséquent, chacune des parties qui composent ce manuscrit, est le fruit de la combinaison de techniques issues de ces trois disciplines. Nous présentons dans la section 9.1, un bilan des travaux réalisés et dans la section 9.2 des perspectives à plus long terme que celles formulées tout au long de ce mémoire.

9.1 Bilan

Dans cette thèse, nous nous sommes intéressés aux trajectoires de patients, présentant un IM, via les bases médico-administratives, à des fins à la fois de prédiction et de caractérisation des flux de patients pour la planification sanitaire. Dans cette section, nous dressons le bilan des différentes contributions.

Les données. Nous avons présenté le principe de fonctionnement du PMSI pour le cas du MCO (Médecine Chirurgie Obstétrique et odontologie). Nous avons expliqué comment retracer la trajectoire d'un patient à partir de ces bases de données par l'identifiant anonyme de patient en chaînant ses séjours. Bien que présentant un certain nombre de limites, nous avons pu constater, au travers d'exemples variés, que ces bases de données pouvaient être utilisées à des fins de recherche. Une première limite étant la qualité des données de chaînage, nous avons proposé un algorithme de sélection des identifiants patients basé sur des contrôles de cohérence des données. Ensuite, nous nous sommes appuyés sur un algorithme de repérage des pathologies, que nous avons enrichi avec un critère sur les actes de cardiologie interventionnelle afin de sélectionner tous les séjours hospitaliers concernant cette pathologie. Une fois notre base de données constituée, nous avons analysé ces données selon trois axes clés : les tendances de l'hospitalisation, la récidive et le décès intra-hospitalier. Nous avons établi des résultats en accord avec la littérature du domaine dont les points les plus marquants sont : 1) une augmentation du taux d'IM chaque année ; 2) une augmentation alarmante des cas chez les femmes, plus précisément les femmes jeunes dont le risque surpasse celui des hommes ; 3) une régionalisation de l'IM avec un contraste Nord-Sud ; 4) un risque de récidive notoire dans les 3 mois après le 1er épisode ; 5) 90% des décès se produisent lors du 1er séjour.

Les trajectoires dans la littérature. Avant d'explorer les trajectoires de patients à partir de la base de données constituée au préalable, il nous a fallu déterminer ce que le concept de trajectoire représentait dans le domaine biomédical. La simple recherche par mots clés ramène de nombreux documents qu'il faut explorer. Pour cela, nous avons proposé une méthode d'analyse semi-automatique originale de la littérature basée sur des outils de fouille de textes. Cette méthode nous a permis de délimiter les contours de notre étude en répondant à des questions sans *a priori* et de filtrer les documents. Ce travail s'est conclu par une revue systématique de la littérature qui a conjugué trois thématiques : les trajectoires de patients, l'IM et le PMSI. La méthode s'est avérée efficace pour explorer de manière transversale ces trois thématiques. À l'issue de cette revue, nous avons mis en évidence les différents concepts de la trajectoire du patient et ce qui motive son étude : pour en apprendre davantage sur l'évolution d'une pathologie grâce au suivi du patient, ou pour comparer les parcours de soins afin de mettre en place des stratégies qui utilisent des procédures de soins facilitant le travail des personnels de santé tout en fournissant un cadre rassurant au patient et en réduisant les coûts.

De la trajectoire du patient à la prédiction du décès. Une première partie d'analyse des trajectoires est destinée à la prédiction du décès et à l'identification des parcours les plus à risque. Tout d'abord, nous avons extrait les motifs fréquents dans des sous-populations à l'aide d'une méthode d'extraction de motifs séquentiels contextuels. L'objectif était de mettre en évidence des parcours types dans ces populations. Nous avons établi que ces parcours étaient relativement similaires d'un contexte à l'autre. Ensuite, nous avons modélisé le décès intra-hospitalier en intégrant ces motifs à l'aide d'un score de similarité. Ce dernier mesure alors la ressemblance entre la trajectoire du patient et le motif. Nous avons comparé plusieurs modèles prédictifs et plusieurs mesures de similarités afin de déterminer le meilleur modèle par contexte. À l'issue de nos expérimentations, nous avons établi que le plus souvent le modèle de régression logistique (RL) couplé avec une distance d'édition est le plus performant. Ce modèle s'est montré efficace avec une performance selon l'aire sous la courbe ROC allant jusqu'à 0,98. De plus, nous avons établi que le pronostic de décès était fortement lié à la fois au profil d'évolution de la maladie et au suivi du patient.

Des trajectoires de patients à la planification sanitaire. La deuxième partie d'analyse des trajectoires de patients consiste à caractériser les flux de patients. Pour cela, nous avons mis en évidence des phénomènes de groupes à l'aide des motifs spatio-temporels. Dans cette méthode, nous avons redéfini les notions d'espace et de temps pour pouvoir identifier les trajectoires dans nos données. Le diagramme de Sankey nous a permis, ensuite, d'obtenir une vision plus générale des schémas de flux de patients au travers des différents évènements hospitaliers. Enfin, nous avons poursuivi cette caractérisation des flux de patients par l'étude à la fois des délais inter-hospitaliers et des tarifs de séjours. Grâce à une méthode de classification des données longitudinales, nous avons déterminé des profils de patients en termes d'occurrences de ces évènements hospitaliers mais aussi de tendance évolutive de tarifs de ces évènements.

Chaque partie du mémoire peut mener à de nouvelles recherches, c'est ce que nous avons évoqué dans les conclusions de chacun des chapitres de ce manuscrit. Dans la section suivante, nous faisons le point sur des perspectives à plus long terme.

9.2 Perspectives

Le sujet de cette thèse est l'exploration des trajectoires de patients dans le cas de l'IM via les bases médico-économiques. Nous avons proposé des méthodes pour explorer les trajectoires de différentes façons. Nous avons également apporté une interprétation médicale des résultats obtenus. Les protocoles d'analyse proposés peuvent avoir d'autres applications que l'étude de l'IM. De plus, il existe d'autres moyens d'étudier les trajectoires de patients. De ce fait, les perspectives de ces travaux se décomposent en deux parties : les méthodes (section 9.2.1) pour l'analyse des trajectoires et les applications (section 9.2.2) de ces analyses de trajectoires dans le domaine de la santé. Dans la suite de cette section nous développons ces deux parties.

9.2.1 Méthodes

Dans cette section, nous détaillons les différentes approches possibles pour l'analyse des trajectoires dédiées à leur description, à leur utilisation dans un modèle à des fins explicatives, à leur visualisation. Une attention particulière sera portée sur l'étude de l'incertitude présente dans les données mais également dans les trajectoires issues de ces données.

Approche descriptive. Elle consiste à étudier la trajectoire dans son ensemble sans dissocier les événements les uns des autres mais en se focalisant sur leur enchaînement. Dans cette approche, la trajectoire est considérée comme le résultat d'une séquence d'évènements. L'objectif est alors de décrire, d'explorer les parcours et d'identifier les régularités ou les différences. Elle est d'une grande utilité lorsque la complexité des trajectoires ne permet pas de les classer manuellement selon des critères simples. Cette manière d'analyser les données permet de découvrir des structures cachées. Ces dernières pourront être utilisées pour réduire la complexité des données. C'est une approche non-paramétrique car elle ne fait pas d'hypothèse stochastique sur la genèse des parcours.
Pour l'analyse des séquences nous pourrions utiliser la méthode OMA (Optimal Matching Analysis) issue de la biostatistique [Abbott et Tsay, 2000]. Elle est beaucoup utilisée en biologie pour l'analyse de protéines et de séquences d'ADN. Son objectif est de rechercher dans un grand volume de données, des séquences ressemblant à une séquence particulière : une protéine donnée. Ces algorithmes s'appuient sur une mesure afin de calculer une distance entre les séquences. Ces mesures ont été introduites lors du chapitre 6. Toutefois, une des limites de cette méthode est qu'elle est sensible à la différence de longueur des séquences. Une alternative a été proposée par [Stovel et Bolan, 2004] en introduisant un coût variable selon la longueur des séquences. Ce coût est fixe lorsque les séquences sont de longueurs identiques et égal à environ un quart du coût fixe lorsque les séquences sont de longueurs différentes. Cette méthode pourrait être intégrée dans notre processus de prédiction du chapitre 6.
La fouille de données propose des méthodes pour l'analyse des séquences. Dans le chapitre 5, nous avons utilisé l'extraction de motifs et nous avons proposé des extensions en utilisant d'autres mesures ou d'autres types de motifs comme les motifs partiellement ordonnés clos présentés dans [Fabrègue *et al.*, 2013]. Il est possible d'envisager des approches par règles d'associations comme dans [Rudin *et al.*, 2011]. Nous avons également utilisé, dans le chapitre 7, la détection de motifs spatio-temporels et nous nous sommes concentrés sur un type de motif mais il en existe beaucoup d'autres que nous avons énumérés. Ce qui laisse d'autres pistes à explorer. Ces méthodes ont pour objectif de mettre en lumière des phénomènes communs au plus grand nombre mais on pourrait au contraire s'intéresser aux cas particuliers en détectant des évènements rares. Ceci permettrait d'étudier les trajectoires de patients dans le cas des maladies rares ou encore l'évolution de parcours de soins atypiques [Weiss et Hirsh, 1998].
À l'aide de la fouille de textes, il serait possible d'envisager la description des trajectoires d'une autre façon. Il faudrait redéfinir la trajectoire du patient en regroupant les libellés des Diagnostics Principaux (DP), des Groupes Homogènes de Malades

(GHM) et des actes par exemple. Ainsi, à chaque patient serait associé un paragraphe de mots listant tout ce qui concerne le séjour. L'étape suivante consisterait à procéder à une analyse textuelle comme celle décrite dans le chapitre 4 afin de classer les patients dans des groupes caractérisés par un ensemble de mots. Cela pourrait être une autre manière de comparer les parcours de soins selon l'évolution de la pathologie ou des actes pratiqués voire des comorbidités associées si l'on ajoute les Diagnostics reliés (DR) et les Diagnostics Associés (DAS) dans la trajectoire.

Approche causale. Elle consiste à étudier la trajectoire en se concentrant sur les évènements. Son objectif est d'apporter des explications permettant de répondre à des questions liant les conséquences aux observations. Un exemple est la prédiction du décès à l'aide des parcours fréquents ou encore la détermination des parcours à risque du décès, la mise en relation du parcours de soins avec l'évolution des coûts... C'est une approche paramétrique car elle part de l'hypothèse que la trajectoire est un processus stochastique complexe.
Elle s'intéresse à la modélisation des probabilités de durées ou de transition. Il existe de nombreux modèles pour répondre à ces questions notamment les modèles markoviens [Ycart, 2002] ou les modèles de survie [Timsit *et al.*, 2005, Alberti *et al.*, 2005]. Les modèles markoviens étudient les transitions d'un état à un autre et les modèles de survie, les délais d'occurrence d'un évènement particulier : une rechute de la maladie, le décès... Dans le chapitre 3, nous avons utilisé le modèle de survie pour modéliser le risque de récidive. En revanche, nous n'avons pas utilisé les modèles markoviens. Cela pourrait être une autre façon d'appréhender les données dans le cas des chapitres 7 et 8. Ainsi, un des premiers objectifs pourrait être la modélisation du temps de transition, ou la caractérisation d'un phénomène de stationnarité : les trajectoires de patients convergent-elles vers le même évènement ? On peut par ailleurs imaginer des extensions à ces investigations permettant de tenir compte d'informations contextuelles, comme les coûts de séjours, par pondération des probabilités de transition.
L'approche causale s'intéresse aux liens entre les données. Il existe de nombreux modèles [Benzécri, 1976] comme l'ACP (Analyse en Composantes Principales) pour des données quantitatives, l'ACM (Analyse des Composantes Multiples) généralisant la précédente, l'AFC (Analyse Factorielle des Correspondances) pour des données qualitatives, l'AFCM (Analyse Factorielle des Correspondances Multiples) généralisant la précédente...
D'autres modélisations spécifiques aux données longitudinales quantitatives sont issues de l'analyse des séries chronologiques. Ces méthodes étudient les tendances des séries, permettent de mettre en évidence des phénomènes de saisonnalité et sont utiles pour faire de la prévision. Par exemple, elles peuvent servir à déterminer quand aura lieu la prochaine hospitalisation. Il existe également de nombreux modèles [Box *et al.*, 2015] dans ce domaine : ARMA (AutoRegressive Moving Average), ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonnal AutoRegressive Integrated Moving Average)...
Une autre approche d'analyse des données pourrait également être envisagée à l'aide des outils issus de l'informatique médicale. L'idée serait d'utiliser la combinaison de l'analyse formelle, pour dégager des structures dans les données, avec des outils tels que le calcul d'évènements (Event Calculus) [Mueller, 2008] ou la logique d'actions

temporelles TAL (Temporal Action Logics) [Doherty et Kvarnström, 2008]. Ces méthodes reposent sur le principe que l'action et le changement sont liés avec le temps. Il serait alors intéressant de mettre en œuvre ces méthodes sur nos données afin de comparer les résultats obtenus dans le cas de la prédiction du décès.

La différence entre l'approche descriptive et l'approche causale est concrète dans la mesure où l'une est probabiliste alors que l'autre est exploratoire. Elles répondent à des questions différentes. La première s'interroge sur les principaux éléments qui différencient les trajectoires en les considérant dans leur ensemble. La seconde s'interroge sur l'impact du risque d'occurrence d'un évènement donné, sur ses caractéristiques ou encore sur les causes et les conséquences : c'est-à-dire les autres évènements reliés à celui-ci. Ces deux approches ne s'excluent pas l'une l'autre. Ce sont les questions de recherche et la nature des données qui orientent le choix vers la méthode la plus adaptée. Mais, les deux approches peuvent être complémentaires, comme nous l'avons montré à plusieurs reprises dans ce manuscrit. Nous avons utilisé des méthodes exploratoires pour synthétiser les données et ensuite nous avons exploité ces données à l'aide de méthodes explicatives telle que la prédiction.

Incertitude. Dans le chapitre 2, nous avons évoqué les limites dans l'utilisation du PMSI notamment dans les variations du codage. Par conséquent, nous pouvons associer à chaque code une certaine incertitude. Un champ à étudier est alors celui de la typologie de l'incertitude dans les bases médico-économiques. Une première partie de ces investigations pourrait consister à déterminer le degré d'incertitude du codage. Plusieurs auteurs [Kwakkel et Pruyt, 2013, Stirling, 2010, Taleb, 2008] ont proposé une classification des niveaux d'incertitude allant de la connaissance certaine à l'ignorance totale. Sur la base de ces conceptualisations, la suite de cette étude serait de proposer une modélisation à base de connaissances incertaines des trajectoires de patients. La théorie de la logique floue [Zadeh, 1996], la théorie des évidences [Shafer *et al.*, 1976] ou encore les réseaux bayésiens [Korb et Nicholson, 2010] sont des solutions à explorer. La logique floue, ou traitement des incertitudes, a pour objet d'étude la représentation des connaissances imprécises et le raisonnement approché. Elle s'appuie sur les évaluations de possibilité et de nécessité. La théorie des évidences ou théorie de Dempster-Shafer se base sur la notion de preuve. Elle utilise les fonctions de croyance et le raisonnement plausible. Son but est de combiner des preuves distinctes pour calculer la probabilité d'un évènement. Les réseaux bayésiens sont fondés sur une approche probabiliste subjective. Les probabilités sont des paramètres susceptibles d'être mis à jour (théorème de Bayes), caractérisant l'état de connaissances sur le système étudié plutôt que les caractéristiques objectives qui lui sont inhérentes. Quelles sont les avantages et les inconvénients de chacune de ces approches ? Quelle méthode est la plus adaptée dans ce cas de figure ? Ce nouveau travail permettrait de répondre à ces questions.

Visualisation. Enfin, il faut ajouter l'aspect visuel que nous n'avons pas évoqué jusqu'à présent. Dans le chapitre 7, nous avons proposé une visualisation des trajectoires de patients pour la schématisation des flux de patients. Ce type de visualisation pourrait être amélioré et combiné avec d'autres indicateurs comme les effectifs, le délai moyen entre les divers évènements, la durée moyenne passée

dans l'évènement, des informations sur le coût selon la trajectoire... Dans cette optique, on pourrait imaginer une plate-forme interactive pour la visualisation des motifs quels qu'ils soient : fréquents, rares, convois... Le principe serait de proposer à l'utilisateur une interface dans laquelle il pourrait définir ses critères de création de contextes, les faire varier, choisir également la pathologie qu'il souhaite étudier et ainsi visualiser les résultats directement avec des informations supplémentaires affichables par clic ou survol de souris telles que listées juste avant. Toutefois, pour réaliser cela, il faut pouvoir assurer à la fois la capacité de stockage de gros volumes de données et la rapidité du traitement de ces données. À l'ère du Big Data, des solutions sont avancées pour le stockage avec une architecture adaptée [Zhi *et al.*, 2011] et le traitement de grandes quantités de données avec par exemple le cloud computing [Agrawal *et al.*, 2011]. De plus, dans le domaine de la visualisation de données, il existe de nombreuses possibilités pour la représentation des trajectoires telles que les algorithmes de forces [Fruchterman et Reingold, 1991, Walshaw, 2000] pour la représentation des motifs fréquents par exemple. Suivant la quantité de données impliquées, des améliorations de ce type de représentation sont parfois nécessaires. Des solutions sont avancées pour le changement de l'affichage des arêtes [Selassie *et al.*, 2011], pour supprimer le chevauchement des labels [Sallaberry *et al.*, 2010, Fekete et Plaisant, 1999]. Un premier outil a été développé pour répondre à ces objectifs (voir annexe C.1).

In fine, quelle que soit la manière d'appréhender les trajectoires, il existe de nombreux outils issus de différentes disciplines pour y parvenir. Il reste des domaines à étudier pour comparer les méthodes entre-elles et/ou consolider des processus d'analyse des données du PMSI. Dans la section suivante, nous détaillons des applications possibles de l'étude des trajectoires concernant la santé et plus précisément en lien avec les travaux présentés dans ce mémoire.

9.2.2 Applications au domaine de la santé

Nous nous focalisons maintenant sur les applications dans le domaine de la santé des méthodes mentionnées précédemment. Le suivi des patients induit des applications diverses et variées. Nous en proposons quelques-unes dans la suite de cette section : de la prévention en passant par la surveillance de pathologies et/ou de dispositifs médicaux. Nous proposons également des applications dans la gestion des flux de patients aussi bien au sein d'un établissement que d'un territoire de santé.

Flux des patients au sein d'un établissement de santé. Au niveau de l'établissement de santé, les enjeux sont ceux mentionnés dans l'introduction du chapitre 7 de la partie IV à savoir la fluidité des flux de patients (réduire les délais d'attente, allouer les ressources de façon adéquate...) et la réduction des coûts tout en maintenant une prestation de soins de qualité. Pour cela, il est intéressant d'examiner la chronologie des hospitalisations afin de déterminer si certaines sont évitables [Cartier *et al.*, 2014]. Par conséquent, cela impacterait la prise en charge du patient en amont. Cette réflexion est par ailleurs à la fois économique (réduction des coûts si moins d'hospitalisations) et médicale, puisqu'une meilleure prise en charge en amont peut éviter des récidives, des rechutes ou des complications au patient. La modélisation

de la chronologie est alors essentielle pour le côté applicatif et pour se rapprocher des enjeux réels.
Si l'on s'intéresse à l'aspect économique pur, on peut s'apercevoir que la modélisation des coûts est complexe car dans le PMSI il s'agit d'un tarif, *i.e.* un coût moyen élaboré par l'ENC (Échelle Nationale des Coûts) [Perrier *et al.*, 2003]. Il est donc difficile d'estimer les véritables coûts et de les relier aux vraies dépenses de santé. Néanmoins, cela peut être à la base d'une véritable estimation du coût dans sa globalité en prenant en compte les coûts de ville et des médicaments à l'aide de la base SNIIRAM dans l'objectif de réduire les coûts selon des prises en charge rationalisées [Minvielle, 2000].

Flux de patients au sein d'un territoire de santé. Dans le cadre de la loi Santé 2016 [Quidu et Escaffre, 2017], chaque établissement de santé a adhéré au 1er juillet 2016 à un GHT (Groupe Hospitalier de Territoire). Il s'agit d'un dispositif régissant la coopération entre plusieurs établissements publics de santé d'un même territoire. Ces dispositifs sont décrits dans l'article 107 du projet de loi de modernisation de notre système de santé. Les objectifs sont : 1) La définition d'un projet médical commun avec une stratégie de prise en charge partagée ; 2) La mutualisation des fonctions support (systèmes d'informations, achats, plans de formation...). Les enjeux liés à l'apparition du GHT sont multiples : mettre en place une coopération territoriale, définir l'offre et la demande, mutualiser les activités et les ressources... Ces enjeux nécessitent des outils adaptés chargés d'apporter un éclairage aux décideurs de sorte à leur permettre de prendre des décisions stratégiques dans l'intérêt à la fois du GHT et du patient.
L'analyse des trajectoires de patients via les bases médico-économiques va alors s'avérer cruciale pour apporter des informations aux décideurs : 1) L'étude des trajectoires de patients d'un point de vue géographique permettrait d'envisager des modèles de décision dans l'allocation de ressources. Il s'agit alors de déterminer les établissements où le patient va se faire soigner selon son domicile mais aussi selon sa pathologie. En d'autres mots, mieux cerner la demande afin de déterminer les manques et les redondances en termes de prises en charges au niveau du territoire. Toutefois, il ne faut pas supprimer l'aspect humain dans cette décision économique. En effet, il faut également allouer les prestations de soins non seulement dans les lieux les plus « rentables » mais aussi dans des lieux stratégiques pour ne pas isoler le patient et lui faire prendre des risques [Ménard, 2002]. Néanmoins, cette prospection ne doit pas être restreinte aux bases PMSI. En effet, elles présentent l'inconvénient de ne pas recenser les consultations externes. Or, des patients atteints de pathologies graves comme la polyarthrite rhumatoïde bénéficient de ces consultations mais ne sont pas pour autant hospitalisés. Supprimer ce type d'accès aux soins peut représenter un réel danger. 2) L'étude des flux de patients mettrait en évidence les réseaux de soins afin de déterminer les établissements travaillant en étroite collaboration. Ceci permettrait, par exemple, de redéfinir les collaborations lorsqu'elles ne concernent pas des établissements d'un même territoire santé. Ceci permettrait également de renforcer des collaborations existantes par la mise en place de nou-

velles offres de soins. 3) Une autre application de l'analyse des trajectoires est la comparaison des prises en charge des établissements du GHT. Ceci permettrait de faciliter l'élaboration du projet médical commun dans une logique de prise en charge partagée [Rican et Vaillant, 2009].

Ces exemples soulignent l'intérêt des outils présentés dans ce manuscrit dans le cadre de la mise en place et du développement du GHT. Le cadre du GHT, regroupant des données de plusieurs établissements sur un même territoire, favorise alors la création de bases de connaissances, comme pour la surveillance des DMI, des IM, des insuffisances cardiaques, d'autres pathologies ou encore des pratiques de soins...

Surveillance des dispositifs médicaux implantables. Un dispositif médical implantable (DMI) est un produit de santé. Il est implanté en totalité ou en partie dans le corps humain ou placé dans un orifice naturel. Ce sont par exemple des endoprothèses, des stimulateurs cardiaques, prothèse de hanche... Bien que ces dispositifs aient pour but de soigner, ils ne sont pas dénués de risque pour la santé. Ils sont donc soumis à une surveillance de l'ANSM (Agence Nationale de la Sécurité du Médicament et des produits de santé). L'ANSM a lancé un plan d'action pour les DMI à long terme. Ces dispositifs sont difficiles à surveiller du fait des risques liés à la fois à l'exposition des patients à long terme et aux conséquences pour la santé en cas de nécessité d'une nouvelle intervention. L'objectif est alors de réaliser une étude bénéfice/risque. Ces études sont réalisées sur la base de données issues de la littérature et de celles fournies par le fabricant. D'autres sources de données sont envisageables par le biais d'essais cliniques. Toutefois, il n'existe pas à ce jour de base de données européennes référençant les incidents. Ce type de registre pourrait être élaboré à partir des bases PMSI. Cela consisterait à établir un algorithme de repérage des incidents liés aux DMI ayant nécessité une ré-hospitalisation. Ce type de dispositif peut être mis en place au sein d'un établissement ou d'un territoire de santé. Une autre façon de procéder à cette surveillance est de placer le patient au centre de l'action de son suivi [Basch *et al.*, 2014, Brédart *et al.*, 2014]. C'est lui qui transmet les données concernant son état de santé. Il faut ensuite analyser les trajectoires des impressions des patients [Patrick *et al.*, 2011a, Patrick *et al.*, 2011b]. Cependant, cette approche induit des difficultés de plusieurs ordres : 1) interpréter de façon médicale la perception du patient ; 2) trier les informations pertinentes de sorte à distinguer les manifestations de troubles consécutifs au DMI.

Prévention basée sur la surveillance. Au cours de ce mémoire, nous avons rapporté des exemples d'outils d'aide à la décision basés sur des données issues du suivi des patients comme l'aide au diagnostic [Séroussi *et al.*, 2013] ou encore la prescription de médicaments [Wang et Bajorek, 2016]. Beaucoup de ces dispositifs sont construits à partir de données cliniques. Certains utilisent des données issues du PMSI comme référence. Une idée serait de réaliser un dispositif équivalent pour prévenir la récidive et/ou le décès dans le cas de l'IM. Ce dispositif s'appuierait sur les données du PMSI, combinées éventuellement avec des données cliniques en appliquant la méthode proposée au chapitre 6 par exemple ou d'autres évoquées dans la section 9.2.1.

Les objets connectés. La surveillance des patients peut être imaginée avec d'autres dispositifs. La notion de trajectoire peut également être associée aux évolutions biologiques intra-corporelles. Dans ce contexte, des applications à l'analyse de la trajectoire s'orienteraient plus spécifiquement dans la détection de changements corporels annonciateurs du développement d'une maladie. C'est la révolution des objets connectés. Ces derniers deviennent alors de véritables outils de prévention santé. Ils sont utiles à la fois pour les malades et leurs médecins favorisant un diagnostic précoce [Kirtava *et al.*, 2013] et de fait une prise en charge adaptée. Il existe de nombreux exemples d'objets connectés. Récemment, l'entreprise Higia [1] a inventé une brassière pour le dépistage du cancer du sein chez la femme. Les mesures faites par les 200 capteurs sensoriels sont ensuite intégrées dans un modèle déterminant le risque de développement du cancer. Un autre exemple, E-Celcius [2] est un projet de capsule miniaturisée qui avalée, permet de mesurer et de transmettre toutes les 30 secondes la température du corps via une technologie de radiofréquence. Son utilité est de prévenir les infections post-opératoires. D'autres dispositifs similaires permettraient d'assurer des examens médicaux à des patients n'ayant pas un accès facile aux soins (délai de rendez-vous trop long, territoire de santé déserté, accès aux transports difficile...). Une idée similaire peut être introduite pour la gestion des flux de patients et la réduction des coûts. Le principe serait d'introduire des dispositifs médicaux [Martelli *et al.*, 2017] permettant la surveillance des patients à distance. Ainsi, le patient hospitalisé pourrait regagner son domicile plus facilement et l'hôpital récupérerait un lit. Le suivi des constantes du patient pourrait être effectué à distance par une infirmière, à l'aide du dispositif médical connecté, qui alerterait le médecin responsable et le patient en cas de suspicion de complications. Cependant, ces dispositifs médicaux de « consultation » doivent demeurer sous contrôle du médecin, seul garant du diagnostic et de la prescription. Les nouvelles technologies offrent des perspectives de pratique de la médecine différente sous forme de e-santé. Ces nouveaux outils n'en demeurent pas moins des outils d'aide à la décision soumis à des limites en termes de confiance dans les mesures retrouvées par ces appareils. Dans ce contexte, les méthodes prenant en compte l'incertitude, telles que celles évoquées dans la section 9.2.1, sont particulièrement intéressantes. De plus, ces nouvelles technologies posent des problèmes éthiques autour de la confidentialité des données [Béranger et Bouadi, 2014].

Le mot de la fin

Dans cette thèse, dont les deux applications principales sont le suivi et l'amélioration des soins des patients d'une part et la réduction des coûts au sein des établissements hospitaliers d'autre part, il peut apparaître une contradiction. Néanmoins, comme l'écrit le philosophe Hegel : « deux pensées opposées donnent lieu à l'élaboration d'une idée nouvelle qui aboutira à un progrès majeur ». Un exemple illustre ce propos dans le domaine médical s'agissant de la révolution des nanotechnologies et du numérique. Deux théories s'affrontent : la première se situant dans la tradition humaniste plaide pour une amélioration sociale et

1. http://higia.tech/
2. http://www.bodycap-medical.com/fr/produit/ecelsius

politique portée par les nouvelles technologies. Cette théorie est soutenue par deux penseurs américains [Sandel, 2009] et [Fukuyama, 2004]. La théorie opposée, portée par [Kurzweil, 2005] président de l'université de la singularité financée par Google, veut développer la « techno-fabrication » de la « posthumanité », afin de créer une nouvelle espèce hybridée avec des machines, dotées de capacités physiques et d'une intelligence artificielle sans limite. De ces deux versions antagonistes est née une série d'innovations appliquées à l'humain tant au niveau de la thérapie génique, par exemple pour le traitement des maladies orphelines [Cohen-Haguenauer, 2011] et neurodégénératives [Charlier *et al.*, 2014] mais aussi le développement des nanotechnologies utilisées pour la fabrication de nano-médicaments [Couvreur *et al.*, 1995] non détectés par les macrophages et permettant de traiter des cancers comme celui du foie. Ces progrès permettront l'allongement de la vie sur le plan qualitatif et réduiront les préjudices causés par le handicap ou les accidents grâce aux exosquelettes. Tous ces progrès bienfaiteurs s'accompagneront à moyen et long termes d'une réduction des coûts économiques de la santé et seront par conséquent rentables. Néanmoins, la notion éthique devra être prédominante pour que l'être humain conserve sa singularité en évitant ainsi l'aboutissement à un monde de robots et à l'uniformisation de la société [Ferry, 2016].

Partie VI

Annexes

Annexe

A

Données

Tableau A.1 – Codes CIM-10 et codes CCAM pour l'algorithme de repérage d'un IM dans les bases médico-administratives.

Codes CIM-10

Code	Libellé
I21	Infarctus aigu du myocarde
I22	Infarctus du myocarde à répétition
I23	Certaines complications récentes d'un infarctus aigu du myocarde
I24	Autres cardiopathies ischémiques aiguës

Codes CCAM

Code	Libellé
DDAF001	Dilatation intraluminale d'un vaisseau coronaire sans pose d'endoprothèse, par voie artérielle transcutanée
DDAF003	Dilatation intraluminale de 3 vaisseaux coronaires ou plus avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF004	Dilatation intraluminale de 2 vaisseaux coronaires avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF006	Dilatation intraluminale d'un vaisseau coronaire avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF007	Dilatation intraluminale de 2 vaisseaux coronaires avec artériographie coronaire, avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF008	Dilatation intraluminale d'un vaisseau coronaire avec artériographie coronaire, avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF009	Dilatation intraluminale de 3 vaisseaux coronaires ou plus avec artériographie coronaire, avec pose d'endoprothèse, par voie artérielle transcutanée
DDAF010	Dilatation intraluminale d'un vaisseau coronaire avec artériographie coronaire, sans pose d'endoprothèse, par voie artérielle transcutanée
DDFF001	Athérectomie intraluminale d'artère coronaire par méthode rotationnelle, par voie artérielle transcutanée
DDFF002	Athérectomie intraluminale d'artère coronaire par méthode directionnelle, par voie artérielle transcutanée

Tableau A.2 – Taux standardisés d'hospitalisation avec IM et pourcentage de variation annuelle des taux par région chez les hommes.

	Taux d'hospitalisation pour 10 000 hommes						**var an**
Région	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**%**
Alsace	42.4	50.38	52.32	56.46	53.94	59.73	7,37
Aquitaine	40.57	52.2	50.1	53.06	53.84	54.04	6,48
Auvergne	28.87	33.21	37.85	39.97	41.69	41.37	7,63
Basse-Normandie	35.99	44.57	48.6	49.38	47.6	46.79	5,84
Bourgogne	35.22	43.83	45.38	48.33	50.72	55.07	9,6
Bretagne	32.22	38.92	39.42	40.12	38.04	38.39	3,91
Centre	33.39	43.35	42.06	38.45	39.43	42.01	5,47
Champagne-Ardenne	34.89	43.26	43.8	42.74	42.86	43.31	4,83
Corse	46.46	70.44	62.3	61.65	60.06	69.72	10,51
Franche-Comté	35.5	40.37	39.25	40.48	38.15	41.82	3,59
Haute-Normandie	32.56	39.43	38.05	38.34	38.71	39.4	4,22
Île de France	39.42	46.9	46.34	48.13	47.47	49.56	4,94
Languedoc-Roussillon	39.4	47.76	45.27	46.13	47.99	48.93	4,78
Limousin	32.57	36.19	39.21	43.27	42.61	48.97	8,64
Lorraine	44.85	52.21	53.54	54.97	55.93	58.25	5,5
Midi-Pyrénées	41.29	50.08	51.15	53.49	54.64	58.03	7,27
Nord Pas de Calais	40.04	47.22	46.54	48.74	47.13	47.01	3,53
PACA	49.23	58.68	56.29	58.33	60.12	63.78	5,58
Pays de la Loire	27.6	32.47	31.61	34.55	35.3	37.76	6,69
Picardie	29.99	39.53	39.21	40.11	39.25	37.07	5,11
Poitou-Charentes	35.63	42.01	42.63	44.45	44.6	47.27	6
Rhône-Alpes	31.1	38.72	39.42	41.65	42.74	44.14	7,57

Tableau A.3 – Taux standardisés d'hospitalisation avec IM et pourcentage de variation annuelle des taux par région chez les femmes.

	Taux d'hospitalisation pour 10 000 femmes						**var an**
Région	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**%**
Alsace	12.99	15.98	15.88	17.52	16.45	17.68	6,81
Aquitaine	11.63	13.85	14.36	14.22	14.25	15.37	5,97
Auvergne	8.58	10.37	11.34	12.4	12.61	12.66	8,32
Basse-Normandie	11.5	13.81	14.33	15.19	14.67	14.04	4,43
Bourgogne	11.56	14.64	14.56	15.01	15.36	17.12	8,6
Bretagne	9.7	11.45	11.07	11.11	10.7	11.35	3,48
Centre	10.94	13.62	12.48	12.32	12.2	13.61	5,09
Champagne-Ardenne	11.01	13.33	14.32	14.06	13.03	14.13	5,56
Corse	12.48	17.69	15.61	12.92	15.31	16.02	7,17
Franche-Comté	12.94	14.52	12.95	13.78	11.6	12.34	60,31
Haute-Normandie	8.68	10.62	10.42	11.94	10.71	11.62	6,66
Île de France	10.27	12.22	12.35	12.55	12.48	12.98	5,02
Languedoc-Roussillon	10.85	13.11	13.62	14	13.57	13.52	4,83
Limousin	10.33	12.26	11.81	12.85	14.66	15.08	8,15
Lorraine	13.65	15.54	16.22	14.44	15.92	17.48	5,47
Midi-Pyrénées	12.17	14.42	14.87	15.36	15.41	17.11	7,26
Nord Pas de Calais	12.76	15.53	14.73	16.18	14.78	15.33	4,3
PACA	12.8	15.65	15.1	15.84	16.53	17.23	6,45
Pays de la Loire	7.76	8.97	8.94	9.81	9.84	10.76	6,93
Picardie	9.39	12.44	13.82	13.35	12.44	11.82	5,66
Poitou-Charentes	9.97	11.99	12.81	12.88	13.03	13.48	6,46
Rhône-Alpes	9.64	12.06	11.8	12.01	12.49	13.12	6,75

Annexe

B

Trajectoires dans la littérature

Tableau B.1: Description des items et catégories observés

Item	Catégories	Détails
Finalité	*Avancée Médicale*	Meilleure compréhension d'une pathologie et adaptation des soins en conséquence
	Recommandations santé	Consignes sanitaires pour l'amélioration de l'état de santé et/ou éviter sa dégradation
	Évaluation coût/qualité	Comparaison de traitements, de processus de soins, de nouveaux médicaments
	Planification sanitaire	Mise en place de procédure de soins pour améliorer la prise en charge
	Outil d'exploitation des données	Création d'un outil synthétique de visualisation de données ou création d'algorithmes pour regrouper et classer les données issues de plusieurs sources
	Autres	Tout le reste
Bases de données	*Registre*	Bases de registre

Item	Catégories	Détails
	Hosp & PMSI	Bases de données hospitalières ou PMSI
	Interviews	Tout type d'interview
	Questionnaires	Questionnaires et QCM
	Autres	Bases pharmacie, bases de la sécurité sociale, les données des médecins généralistes, les bases de la banque du sang, et les journaux de bord des patients.
Maladies	*Maladies cardiovasculaires*	AVC, IM, insuffisance cardiaque
	Diabète	
	Cancer	Colorectal, de la prostate, du sein, du poumon, de la vessie, du col de l'utérus, de l'endomètre
	Maladies pulmonaires	Bronchopneumopathie chronique obstructive, embolie pulmonaire
	Maladie rénale	Insuffisance rénale
	Maladie neurologiques	Sclérose en plaques, schizophrénie, dépression
	Autres	La goutte, l'arthrose, la scoliose idiopathique, craniotomie, blessure pénétrante du colon, douleurs, fractures pelviennes
Méthodes	*Modèle de survie*	Modèle de Cox, Kaplan-Meier
	Tests param et non-param	χ^2, test de Fisher, de Student, Kruskal Wallis, Mann-Withney...
	Classification	
	ANOVA-ANCOVA	
	Modèle linéaire/logistique	Modèles de régression linéaire ou logistique, ou modèle GLM, modèle logistique
	Autres	Modèles à variables latentes, le coefficient de Kappa, la méta-analyse...
Concept de trajectoire	*Coût*	Suivi des coûts dans le cas d'un traitement ou de procédure de soins
	Soins	Parcours de soins, avec l'historique des consultations, des motifs d'hospitalisation, soins prodigués
	Processus de soins	Série d'étapes par lesquelles passe le patient dans un parcours pré-programmé, ou une série d'étapes de fonctionnement d'une équipe de soins
	Évolution de l'état de santé	Symptômes, signes cliniques, évolutions de capacités cognitives
	Mesures biologiques	

Item	Catégories	Détails
	Risque	Évolution d'une mesure du risque
	Survie	Évolution de la survie
	Autres	Mesure du temps d'activité physique, les prises de décisions du patient

Tableau B.2: Références sources par item étudié

Item	Catégorie	T1 ∩ T2	T1 ∩ T3
Concept de trajectoire	*Coût*	[Couchoud *et al.*, 2015, Sundberg *et al.*, 2014, Ricciardi *et al.*, 2009, Popp *et al.*, 2002]	
	Soins	[Bettencourt-Silva *et al.*, 2015, Thompson *et al.*, 2015, Bossuyt *et al.*, 2015, Jensen *et al.*, 2015, Harlos *et al.*, 2015, Guldbrandt *et al.*, 2015, Defossez *et al.*, 2014, Palmer *et al.*, 2013a, Palmer *et al.*, 2013b, Ellis *et al.*, 2010, Myklebust *et al.*, 2010]	[Aeyels *et al.*, 2016, Kristoffersen *et al.*, 2015, Hagiwara *et al.*, 2014, Kesavan *et al.*, 2013, Kinsman *et al.*, 2012, Song *et al.*, 2010, O'Donnell *et al.*, 2006, Young *et al.*, 2004]
	Processus de soins	[Tang *et al.*, 2015, Goderis *et al.*, 2015, Yang *et al.*, 2014, Van Hoeve *et al.*, 2014, Krummenauer *et al.*, 2011, Naqvi *et al.*, 2009, Kinsman *et al.*, 2009, Baade *et al.*, 2007, Miller *et al.*, 2002, Ghosh *et al.*, 2001, Biffl *et al.*, 2001, Buckley *et al.*, 2000, Arko *et al.*, 2001]	[Laut et Foldspang, 2012, Mazzini *et al.*, 2008, Pelliccia *et al.*, 2004, Bestul *et al.*, 2004, Kucenic *et al.*, 2000]
	Évolution de l'état de santé	[Klinkhammer-Schalke *et al.*, 2015, Mastenbroek *et al.*, 2015, Jensen *et al.*, 2014, Strömberg *et al.*, 2014, Sieberg *et al.*, 2013, Cocchi *et al.*, 2013, Schwartz *et al.*, 2013, Veloso *et al.*, 2013, Ahmed *et al.*, 2011, Dely *et al.*, 2012, Jiwa *et al.*, 2010, Danielsson *et al.*, 2009, Diaz *et al.*, 2008]	[Lewis *et al.*, 2014, Myers *et al.*, 2014, Smith *et al.*, 2011, Martens *et al.*, 2008, Ginzburg *et al.*, 2003, Rankin *et al.*, 2002]
	Mesures biologiques	[Gebregziabher *et al.*, 2010]	
	Risque		[Dharmarajan *et al.*, 2015]
	Survie	[Noble *et al.*, 2015]	[Wang *et al.*, 2013, Gerber *et al.*, 2009]
	Autres	[Sverrisson *et al.*, 2014, Jayanti *et al.*, 2013]	[Myers *et al.*, 2011, Gerber *et al.*, 2011, Rosenfeld, 2004]
Finalité	*Avancée médicale*	[Harlos *et al.*, 2015, Mastenbroek *et al.*, 2015, Sverrisson *et al.*, 2014, Sieberg *et al.*, 2013, Schwartz *et al.*, 2013, Ahmed *et al.*, 2011, Gebregziabher *et al.*, 2010, Danielsson *et al.*, 2009]	[Myers *et al.*, 2014, Kesavan *et al.*, 2013, O'Donnell *et al.*, 2006]

Item	Catégorie	T1 ∩ T2	T1 ∩ T3
	Recommandations santé	[Guldbrandt *et al.*, 2015, Jensen *et al.*, 2014, Van Hoeve *et al.*, 2014, Veloso *et al.*, 2013, Biffl *et al.*, 2001]	[Kristoffersen *et al.*, 2015, Hagiwara *et al.*, 2014, Kinsman *et al.*, 2012, Myers *et al.*, 2011, Gerber *et al.*, 2011, Smith *et al.*, 2011, Mazzini *et al.*, 2008, Martens *et al.*, 2008, Rankin *et al.*, 2002]
	Évaluation coût/qualité	[Couchoud *et al.*, 2015, Goderis *et al.*, 2015, Yang *et al.*, 2014, Sundberg *et al.*, 2014, Strömberg *et al.*, 2014, Palmer *et al.*, 2013a, Krummenauer *et al.*, 2011, Dely *et al.*, 2012, Ricciardi *et al.*, 2009, Kinsman *et al.*, 2009, Diaz *et al.*, 2008, Popp *et al.*, 2002, Miller *et al.*, 2002, Ghosh *et al.*, 2001, Buckley *et al.*, 2000, Arko *et al.*, 2001]	[Aeyels *et al.*, 2016, Wang *et al.*, 2013]
	Planification sanitaire	[Klinkhammer-Schalke *et al.*, 2015, Cocchi *et al.*, 2013, Palmer *et al.*, 2013a, Ellis *et al.*, 2010, Myklebust *et al.*, 2010, Naqvi *et al.*, 2009]	[Laut et Foldspang, 2012, Pelliccia *et al.*, 2004, Young *et al.*, 2004, Bestul *et al.*, 2004, Kucenic *et al.*, 2000]
	Outil d'exploitation des données	[Bettencourt-Silva *et al.*, 2015, Thompson *et al.*, 2015, Bossuyt *et al.*, 2015, Defossez *et al.*, 2014]	
	Autres	[Noble *et al.*, 2015, Tang *et al.*, 2015, Jensen *et al.*, 2015, Jayanti *et al.*, 2013, Jiwa *et al.*, 2010, Baade *et al.*, 2007]	[Dharmarajan *et al.*, 2015, Lewis *et al.*, 2014, Song *et al.*, 2010, Gerber *et al.*, 2009, Rosenfeld, 2004, Ginzburg *et al.*, 2003]
Continent	*Amériques*	[Thompson *et al.*, 2015, Harlos *et al.*, 2015, Sverrisson *et al.*, 2014, Sieberg *et al.*, 2013, Schwartz *et al.*, 2013, Ahmed *et al.*, 2011, Gebregziabher *et al.*, 2010, Diaz *et al.*, 2008, Miller *et al.*, 2002, Ghosh *et al.*, 2001, Biffl *et al.*, 2001, Buckley *et al.*, 2000, Arko *et al.*, 2001]	[Dharmarajan *et al.*, 2015, Mazzini *et al.*, 2008, Young *et al.*, 2004, Rosenfeld, 2004, Bestul *et al.*, 2004, Rankin *et al.*, 2002, Kucenic *et al.*, 2000]
	Asie	[Tang *et al.*, 2015, Yang *et al.*, 2014, Naqvi *et al.*, 2009]	[Myers *et al.*, 2014, Myers *et al.*, 2011, Gerber *et al.*, 2011, Gerber *et al.*, 2009, Song *et al.*, 2010, Ginzburg *et al.*, 2003]
	Europe	[Noble *et al.*, 2015, Klinkhammer-Schalke *et al.*, 2015, Couchoud *et al.*, 2015, Bettencourt-Silva *et al.*, 2015, Bossuyt *et al.*, 2015, Jensen *et al.*, 2015, Goderis *et al.*, 2015, Mastenbroek *et al.*, 2015, Guldbrandt *et al.*, 2015, Jensen *et al.*, 2014, Sundberg *et al.*, 2014, Defossez *et al.*, 2014, Van Hoeve *et al.*, 2014, Strömberg *et al.*, 2014, Jayanti *et al.*, 2013, Cocchi *et al.*, 2013, Palmer *et al.*, 2013a, Palmer *et al.*, 2013b, Veloso *et al.*, 2013, Krummenauer *et al.*, 2011, Dely *et al.*, 2012, Myklebust *et al.*, 2010, Danielsson *et al.*, 2009, Ricciardi *et al.*, 2009, Popp *et al.*, 2002]	[Aeyels *et al.*, 2016, Kristoffersen *et al.*, 2015, Hagiwara *et al.*, 2014, Kesavan *et al.*, 2013, Laut et Foldspang, 2012, Smith *et al.*, 2011, Martens *et al.*, 2008, O'Donnell *et al.*, 2006, Pelliccia *et al.*, 2004]
	Intercontinental		[Lewis *et al.*, 2014, Wang *et al.*, 2013]
	Australasie	[Ellis *et al.*, 2010, Jiwa *et al.*, 2010, Kinsman *et al.*, 2009, Baade *et al.*, 2007]	[Kinsman *et al.*, 2012]

Item	Catégorie	T1 ∩ T2	T1 ∩ T3
Bases de données	*Questionnaires*	[Klinkhammer-Schalke *et al.*, 2015, Cocchi *et al.*, 2013, Krummenauer *et al.*, 2011, Jensen *et al.*, 2015, Mastenbroek *et al.*, 2015, Guldbrandt *et al.*, 2015, Sieberg *et al.*, 2013, Palmer *et al.*, 2013a, Veloso *et al.*, 2013, Ahmed *et al.*, 2011, Naqvi *et al.*, 2009, Strömberg *et al.*, 2014]	[Lewis *et al.*, 2014, Myers *et al.*, 2011, Smith *et al.*, 2011, Mazzini *et al.*, 2008]
	Interviews	[Noble *et al.*, 2015, Danielsson *et al.*, 2009, Baade *et al.*, 2007, Buckley *et al.*, 2000, Arko *et al.*, 2001]	[Gerber *et al.*, 2011, Song *et al.*, 2010, Young *et al.*, 2004, Rosenfeld, 2004]
	Hosp & PMSI	[Noble *et al.*, 2015, Bettencourt-Silva *et al.*, 2015, Palmer *et al.*, 2013b, Dely *et al.*, 2012, Ellis *et al.*, 2010, Naqvi *et al.*, 2009, Danielsson *et al.*, 2009, Diaz *et al.*, 2008, Popp *et al.*, 2002, Ghosh *et al.*, 2001, Tang *et al.*, 2015, Thompson *et al.*, 2015, Defossez *et al.*, 2014, Jiwa *et al.*, 2010, Ricciardi *et al.*, 2009]	[Kristoffersen *et al.*, 2015, Dharmarajan *et al.*, 2015, Myers *et al.*, 2014, Kesavan *et al.*, 2013, Kinsman *et al.*, 2012, Gerber *et al.*, 2009, Mazzini *et al.*, 2008, Martens *et al.*, 2008, O'Donnell *et al.*, 2006, Pelliccia *et al.*, 2004, Bestul *et al.*, 2004, Ginzburg *et al.*, 2003, Kucenic *et al.*, 2000, Myers *et al.*, 2011, Gerber *et al.*, 2011, Smith *et al.*, 2011]
	Registre	[Couchoud *et al.*, 2015, Thompson *et al.*, 2015, Bossuyt *et al.*, 2015, Jensen *et al.*, 2015, Harlos *et al.*, 2015, Goderis *et al.*, 2015, Mastenbroek *et al.*, 2015, Guldbrandt *et al.*, 2015, Jensen *et al.*, 2014, Yang *et al.*, 2014, Sverrisson *et al.*, 2014, Sundberg *et al.*, 2014, Defossez *et al.*, 2014, Van Hoeve *et al.*, 2014, Strömberg *et al.*, 2014, Sieberg *et al.*, 2013, Jayanti *et al.*, 2013, Schwartz *et al.*, 2013, Veloso *et al.*, 2013, Ahmed *et al.*, 2011, Gebregziabher *et al.*, 2010, Myklebust *et al.*, 2010, Jiwa *et al.*, 2010, Ricciardi *et al.*, 2009, Kinsman *et al.*, 2009, Baade *et al.*, 2007, Miller *et al.*, 2002, Biffl *et al.*, 2001, Buckley *et al.*, 2000, Arko *et al.*, 2001, Bettencourt-Silva *et al.*, 2015, Strömberg *et al.*, 2014]	[Aeyels *et al.*, 2016, Kristoffersen *et al.*, 2015, Dharmarajan *et al.*, 2015]
	Autres	[Tang *et al.*, 2015, Palmer *et al.*, 2013a, Klinkhammer-Schalke *et al.*, 2015, Couchoud *et al.*, 2015, Bossuyt *et al.*, 2015, Harlos *et al.*, 2015, Sundberg *et al.*, 2014, Schwartz *et al.*, 2013, Kinsman *et al.*, 2009, Ghosh *et al.*, 2001, Biffl *et al.*, 2001, Jiwa *et al.*, 2010]	[Hagiwara *et al.*, 2014, Rankin *et al.*, 2002, Hagiwara *et al.*, 2014]
Méthodes	*ANOVA - ANCOVA*	[Tang *et al.*, 2015, Klinkhammer-Schalke *et al.*, 2015, Biffl *et al.*, 2001, Klinkhammer-Schalke *et al.*, 2015]	[Myers *et al.*, 2011, Gerber *et al.*, 2011, Ginzburg *et al.*, 2003, Rankin *et al.*, 2002, Kucenic *et al.*, 2000, Wang *et al.*, 2013, Kinsman *et al.*, 2012]
	Classification	[Jensen *et al.*, 2014, Sundberg *et al.*, 2014, Thompson *et al.*, 2015, Schwartz *et al.*, 2013]	

Item	Catégorie	T1 ∩ T2	T1 ∩ T3
	Modèle linéaire/logistique	[Goderis *et al.*, 2015, Sieberg *et al.*, 2013, Jayanti *et al.*, 2013, Schwartz *et al.*, 2013, Palmer *et al.*, 2013b, Veloso *et al.*, 2013, Ahmed *et al.*, 2011, Myklebust *et al.*, 2010, Popp *et al.*, 2002, Miller *et al.*, 2002, Mastenbroek *et al.*, 2015, Strömberg *et al.*, 2014, Krummenauer *et al.*, 2011, Noble *et al.*, 2015, Guldbrandt *et al.*, 2015]	[Mazzini *et al.*, 2008, Lewis *et al.*, 2014]
	Tests Param et non param	[Klinkhammer-Schalke *et al.*, 2015, Guldbrandt *et al.*, 2015, Yang *et al.*, 2014, Van Hoeve *et al.*, 2014, Strömberg *et al.*, 2014, Cocchi *et al.*, 2013, Krummenauer *et al.*, 2011, Dely *et al.*, 2012, Naqvi *et al.*, 2009, Kinsman *et al.*, 2009, Biffl *et al.*, 2001, Buckley *et al.*, 2000, Arko *et al.*, 2001, Jensen *et al.*, 2014, Sverrisson *et al.*, 2014, Sundberg *et al.*, 2014, Jayanti *et al.*, 2013, Miller *et al.*, 2002, Mastenbroek *et al.*, 2015, Schwartz *et al.*, 2013]	[Lewis *et al.*, 2014, Wang *et al.*, 2013, Kinsman *et al.*, 2012, O'Donnell *et al.*, 2006, Bestul *et al.*, 2004, Kristoffersen *et al.*, 2015, Hagiwara *et al.*, 2014, Myers *et al.*, 2011, Gerber *et al.*, 2011, Gerber *et al.*, 2009, Pelliccia *et al.*, 2004, Kucenic *et al.*, 2000]
	Modèle de survie	[Harlos *et al.*, 2015, Sverrisson *et al.*, 2014, Jiwa *et al.*, 2010, Gebregziabher *et al.*, 2010, Guldbrandt *et al.*, 2015, Yang *et al.*, 2014]	[Kristoffersen *et al.*, 2015, Gerber *et al.*, 2009, Smith *et al.*, 2011, Wang *et al.*, 2013, Gerber *et al.*, 2011, O'Donnell *et al.*, 2006]
	Autres	[Noble *et al.*, 2015, Couchoud *et al.*, 2015, Thompson *et al.*, 2015, Bossuyt *et al.*, 2015, Jensen *et al.*, 2015, Mastenbroek *et al.*, 2015, Defossez *et al.*, 2014, Gebregziabher *et al.*, 2010, Ellis *et al.*, 2010, Danielsson *et al.*, 2009, Ricciardi *et al.*, 2009, Diaz *et al.*, 2008, Popp *et al.*, 2002, Jensen *et al.*, 2014, Dely *et al.*, 2012, Cocchi *et al.*, 2013]	[Dharmarajan *et al.*, 2015, Hagiwara *et al.*, 2014, Myers *et al.*, 2014, Kesavan *et al.*, 2013, Smith *et al.*, 2011, Song *et al.*, 2010, Martens *et al.*, 2008, Pelliccia *et al.*, 2004, Rosenfeld, 2004, Bestul *et al.*, 2004, Rankin *et al.*, 2002, Myers *et al.*, 2011, Gerber *et al.*, 2011, Gerber *et al.*, 2009, Kucenic *et al.*, 2000]
Maladies	*Maladies cardiovasculaires*	[Mastenbroek *et al.*, 2015, Yang *et al.*, 2014, Strömberg *et al.*, 2014, Palmer *et al.*, 2013b, Kinsman *et al.*, 2009, Buckley *et al.*, 2000, Arko *et al.*, 2001, Tang *et al.*, 2015]	tous
	Cancer	[Klinkhammer-Schalke *et al.*, 2015, Bettencourt-Silva *et al.*, 2015, Thompson *et al.*, 2015, Jensen *et al.*, 2015, Harlos *et al.*, 2015, Guldbrandt *et al.*, 2015, Sverrisson *et al.*, 2014, Defossez *et al.*, 2014, Van Hoeve *et al.*, 2014, Palmer *et al.*, 2013a, Veloso *et al.*, 2013, Jiwa *et al.*, 2010, Ricciardi *et al.*, 2009, Ghosh *et al.*, 2001, Noble *et al.*, 2015, Diaz *et al.*, 2008, Baade *et al.*, 2007]	
	Diabète	[Tang *et al.*, 2015, Bossuyt *et al.*, 2015, Goderis *et al.*, 2015, Gebregziabher *et al.*, 2010, Ellis *et al.*, 2010]	
	Maladies neurologiques	[Cocchi *et al.*, 2013, Schwartz *et al.*, 2013, Ahmed *et al.*, 2011, Myklebust *et al.*, 2010, Naqvi *et al.*, 2009, Danielsson *et al.*, 2009]	[Dharmarajan *et al.*, 2015]
	Maladies pulmonaires	[Dely *et al.*, 2012, Jensen *et al.*, 2014, Palmer *et al.*, 2013a]	

Item	Catégorie	T1 ∩ T2	T1 ∩ T3
	Maladies rénales	[Couchoud *et al.*, 2015, Jayanti *et al.*, 2013, Bossuyt *et al.*, 2015]	
	Autres	[Jensen *et al.*, 2014, Sundberg *et al.*, 2014, Krummenauer *et al.*, 2011, Popp *et al.*, 2002, Miller *et al.*, 2002, Biffl *et al.*, 2001]	[Kristoffersen *et al.*, 2015]

Annexe

C Extraction de motifs séquentiels contextuels

Tableau C.1 – Codes CIM-10 retenus pour la sélection des séjours dans les trajectoires de patients.

Codes	**Libellé**
A41	Sepsis
E10 à E14	Diabète sucré
E66	Obésité
E78	Dyslipidémie
F10 à F19	Consommation de substances psychoactives
I10 à I13	Hypertension
I50	Insuffisance cardiaque
I64 ; I67 ; I69	Accident vasculaire cérébral
J95 à J99	Insuffisance respiratoire
J44	Bronchopneumopathie chronique obstructive
K92	Saignement gastro-intestinaux
N17	Insuffisance rénale
S72	Fracture de la hanche

Tableau C.2 – Codes GHM présents dans le manuscrit.

Code	**Libellé**
04M11	Signes et symptômes respiratoires
05C05	Pontages aortocoronariens sans cathétérisme cardiaque, ni coronarographie
05C11	Autres interventions de chirurgie vasculaire
05C19	Poses d'un défibrillateur cardiaque
05K05	Endoprothèse avec infarctus du myocarde
05K06	Endoprothèse sans infarctus du myocarde
05K10	Actes diagnostiques par voie vasculaire
05K13	Actes thérapeutiques par voie vasculaire
05K24	Dilatations coronaires et autres actes thérapeutiques sur le cœur par voie vasculaire
05M04	Infarctus du myocarde aigu
05M06	Angine de poitrine
05M11	Cardiopathies congénitales et valvulopathies
05M16	Athérosclérose coronarienne
05M17	Autres affections de l'appareil circulatoire
05M21	Infarctus du myocarde avec décès
23M06	Autres facteurs influant sur l'état de santé
27C05	Transplantation cardiaque

Tableau C.3 – Codes CIM-10 présents dans le manuscrit.

Code	**Libellé**
I15	Hypertension secondaire
I20	Angine de poitrine
I21	Infarctus aigu du myocarde
I22	Infarctus du myocarde à répétition
I23	Certaines complications récentes d'un infarctus aigu du myocarde
I24	Autres cardiopathies ischémiques aiguës
I25	Cardiopathie ischémique chronique
I35	Atteintes non rhumatismales de la valvule aortique
I39	Endocardite et atteintes valvulaires cardiaques au cours de maladies classées ailleurs
I46	Arrêt cardiaque
I47	Tachycardie paroxystique
I50	Insuffisance cardiaque
R07	Douleur au niveau de la gorge et du thorax
Z04	Examen et mise en observation pour d'autres raisons

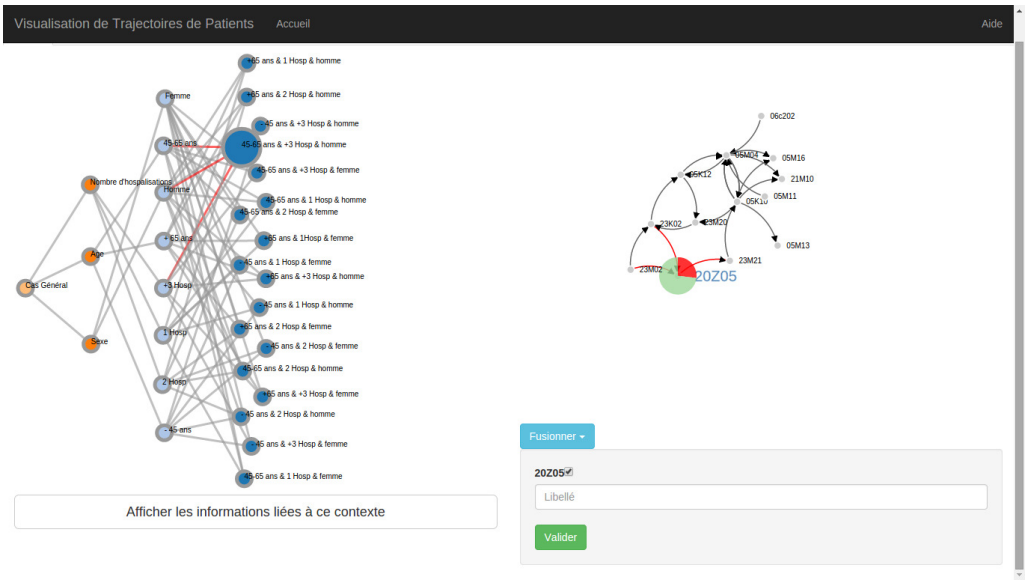


Figure C.1 – Capture d'écran de l'interface graphique créée pour analyser et explorer les données.

Annexe D

Prédire le décès

D.1 Distances de similarités entre chaînes de caractères

D.1.1 Distances basées sur des opérations d'édition

Le principe de ces distances est de compter le nombre d'opérations basiques nécessaires pour passer d'une chaîne à l'autre. Ces distances peuvent être catégorisées suivant les opérations autorisées. Ces dernières sont :

- la substitution d'un caractère : « houx » en « toux » ;
- l'effacement d'un caractère : « houx » en « hou » ;
- l'insertion d'un caractère : « main » en « matin » ;
- la transposition de caractères adjacents : « chien » en « chine ».

Dans la suite, nous notons a et b deux chaînes de caractères.

Définition 22 (Distance de Hamming)
La distance de [Hamming, 1950] est la distance d'édition la plus simple qui n'autorise que la substitution de caractères. Elle est donc définie seulement pour les chaînes de longueurs identiques. C'est une distance au sens mathématique du terme. À deux suites de symboles de même longueur, elle associe le nombre de positions où les deux suites diffèrent.

Définition 23 (Plus longue sous-chaîne commune)
LCS [Needleman et Wunsch, 1970] est définie comme étant la plus longue chaîne obtenue en appareillant les caractères de a et de b, tout en gardant l'ordre des caractères intact. La distance LCS est alors définie comme le nombre de caractères non appareillés. Elle est similaire à la distance d'édition avec seulement les opérations d'effacement et d'insertions avec un poids de 1.

Définition 24 (Distance de Levenshtein)
La distance de [Levenshtein, 1966] ou distance d'édition compte le nombre de suppressions, insertions et substitutions nécessaires pour changer a en b.

Définition 25 (Distance d'alignement optimal)
La distance d'alignement optimal est une variante de la distance de Levenshtein. Elle autorise les transpositions de caractères adjacents, mais cette opération ne peut être réalisée qu'une seule fois.

Définition 26 (Distance de Damerau-Levenshtein)
La distance de Damerau-Levenshtein [Damerau, 1964] calcule le nombre minimum d'opérations nécessaires pour transformer une chaîne de caractères en une autre, où une opération est définie comme l'insertion, la suppression, la substitution d'un simple caractère, ou comme une transposition de deux caractères adjacents.

D.1.2 Mesures basées sur les q-grammes

Définition 27 (q-gramme)
Un q-gramme est une sous-séquence de q caractères consécutifs d'une chaîne. Pour obtenir les q-grammes associés à une chaîne de caractères il faut faire glisser une fenêtre de q-caractères et enregistrer l'ensemble des q-grammes trouvés.

Définition 28 (Mesure q-gramme)
Soit v_a (resp. v_b) le vecteur de comptage du nombre de q-grammes dans a (resp. b), la mesure q-gramme est donnée par la différence absolue $|v_a - v_b|$. Elle compte le nombre de q-grammes non commun aux deux chaînes de caractères.

Définition 29 (Mesure de Jaccard)
La mesure de [Jaccard, 1901] est définie par :

$$1 - \frac{|Q_a \cap Q_b|}{|Q_a \cup Q_b|}$$

où Q_a est l'ensemble des q-grammes uniques dans a et Q_b celui de b.

Définition 30 (Mesure cosinus)
La mesure cosinus est calculée de la façon suivante :

$$1 - \frac{v_a v_b}{\|v_a\| \|v_b\|}$$

où v_a et v_b sont définis plus haut.

Remarques : Aucune de ces mesures basées sur les q-grammes ne satisfait la condition d'identité (*i.e* la mesure, entre deux chaînes de caractères a et b, basée sur les q-grammes valant 0, ne garantie pas que $a = b$). La mesure cosinus ne satisfait pas l'hypothèse de l'inégalité triangulaire.

D.1.3 Mesures heuristiques

Définition 31 (Mesure de Jaro)
La mesure de [Jaro, 1989] est un nombre entre 0 (correspondance exacte) et 1 (totale discordance) mesurant la dissimilarité entre deux chaînes. Elle est définie par :

$$1 - \frac{1}{3}\left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{(m-t)}{|m|}\right)$$

où

- *$|a|$ (resp. $|b|$) est la longueur de la chaîne a (resp. b) ;*
- *m représente le nombre de caractères correspondants entre les 2 chaînes ;*
- *t est le nombre de transpositions entre les caractères correspondants.*

Définition 32 (Mesure de Jaro-Winkler)
La mesure de Jaro-Winkler [Winkler, 1990] ajoute un terme de correction par rapport à la distance de Jaro. Elle utilise un coefficient de préfixe p qui favorise les chaînes de longueur l (avec $l \leq 4$). Elle est définie par

$$d - l \times p \times d$$

où

- *d est la distance de Jaro ;*
- *l est la longueur du préfixe commun ;*
- *p est un facteur de pénalité. La plupart du temps il est fixé à 0,1 dans les travaux de Winkler.*

Notons que dans les deux cas l'inégalité triangulaire n'est pas satisfaite.

D.2 Modèle par arbre à inférence conditionnelle

L'apprentissage par arbre de décision consiste à construire un arbre depuis un ensemble d'apprentissage constitué de n-uplets étiquetés. Un arbre de décision peut être décrit comme un diagramme de flux de données (ou flowchart) où chaque nœud interne décrit un test sur une variable d'apprentissage, chaque branche représente un résultat du test et chaque feuille contient la valeur de la variable cible (une étiquette de classe pour les arbres de classification, une valeur numérique pour les arbres de régression).

Usuellement, les algorithmes pour construire les arbres de décision sont construits en divisant l'arbre du sommet vers les feuilles en choisissant à chaque étape une variable d'entrée qui réalise le meilleur partage de l'ensemble d'objets, comme décrit précédemment. Pour choisir la variable de séparation sur un nœud, les algorithmes testent les différentes variables d'entrée possibles et sélectionnent celle qui maximise ou minimise un critère donné.

Dans le cas des arbres de régression, le même schéma de séparation peut être appliqué, mais au lieu de minimiser le taux d'erreur de classification, on cherche à maximiser la variance inter-classes (avoir des sous-ensembles dont les valeurs de la variable-cible soient les plus dispersées possibles). En général, le critère utilise le test du χ^2.

L'apprentissage par arbre de décision présente deux faiblesses. Tout d'abord, il peut amener des arbres de décision très complexes, qui généralisent mal l'ensemble d'apprentissage (il s'agit du problème de sur-apprentissage). On utilise des procédures d'élagage pour contourner ce problème, certaines approches comme l'inférence conditionnelle permettent de s'en affranchir [Strobl *et al.*, 2009].
Ensuite, dans les méthodes classiques, l'utilisation d'indicateurs dans la sélection de variables tels que l'indice de Gini, le gain d'information sont sources de biais [Strobl, 2005]. En effet, lorsque les données incluent des attributs ayant plusieurs catégories, le gain d'information dans l'arbre est biaisé en faveur de ces attributs. Dans le cas de variables continues, ce type de modèle a tendance à sélectionner les variables ayant le plus de valeurs manquantes. Cependant, le problème de la sélection de prédicteurs biaisés peut être contourné par des méthodes telles que l'inférence conditionnelle.

Le modèle par arbre à inférence conditionnelle est une méthode statistique basée sur l'utilisation de tests non-paramétriques comme critère de séparation. Elle utilise un test de significativité pour sélectionner les prédicteurs. Ce dernier est basé sur des tests de permutations de significativité. À chaque étape, les valeurs de la variable à prédire sont permutées pour chaque test de lien avec le prédicteur. Un score est ensuite calculé à partir de l'ensemble des p-values obtenues. Le prédicteur retenu est celui pour lequel le score est le plus faible.

Ce modèle procède en 2 étapes principales [Hothorn *et al.*, 2006]. Les arbres à inférence conditionnelle estiment une relation de régression par une partition récursive binaire dans un cadre d'inférence conditionnelle. L'algorithme fonctionne de la façon suivante :

1. Teste l'hypothèse d'indépendance globale entre les variables d'entrée et la réponse (qui peut être multivariée aussi). Lorsque cette hypothèse ne peut être rejetée, l'algorithme mesure chaque association entre Y et X_j $(j = 1, .., m)$ et sélectionne la variable d'entrée ayant l'association la plus forte avec la réponse. Cette association est mesurée par une p-value correspondant à un test pour une hypothèse partielle d'une variable d'entrée seule et la réponse.
2. Implémente une séparation binaire dans la variable d'entrée sélectionnée.
3. Répète les étapes 1) et 2) récursivement.

D.3 Mesures de performances d'un modèle

Mesures de discrimination

Nous nous plaçons dans le cas classique décrit dans le tableau D.1 où l'évènement positif est le décès. Ainsi, nous définissons les vrais positifs comme étant le cas dans lequel le modèle prédit un décès alors que ce dernier a été observé et les faux positifs dans le cas contraire. De même, nous définissons les vrais négatifs lorsque le modèle prédit que le patient est toujours en vie et que le décès n'est pas observé à l'hôpital et les faux négatifs dans le cas contraire.

Tableau D.1 – Matrice de confusion

		Prédit	
		Décès	Vivant
Observé	Décès	VP	FN
	Vivant	FP	VN

Les grandeurs calculées pour l'ensemble des modèles sont détaillées ci-après.

Définition 33 (La sensibilité / Le Rappel)
La sensibilité consiste à mesurer la capacité du modèle de détecter un cas positif lorsque celui-ci s'est produit. Elle résume la compétence du modèle dans la détection de tous les décès (et avoir le moins de faux négatifs). La sensibilité est donnée par la formule suivante :

$$Sensibilité = Rappel = \frac{VP}{VP + FN}$$

Définition 34 (La spécificité)
La spécificité, qui s'oppose à la sensibilité, consiste à mesurer la capacité d'un modèle de détecter un résultat négatif lorsque l'hypothèse n'est pas vérifiée. Ainsi la spécificité d'un modèle mesure son aptitude à ne prédire que les décès (et avoir le moins de faux positifs). La spécificité est donnée par la formule suivante :

$$Spécificité = \frac{VN}{VN + FP}$$

Définition 35 (La précision)
La précision mesure la capacité du modèle de repérer les cas pertinents parmi les cas positifs. Plus la précision est élevée plus le modèle est « précis ». Elle se calcule à l'aide de la formule :

$$Précision = \frac{VP}{VP + FP}$$

Définition 36 (La F-mesure)
La F-mesure est la moyenne harmonique de la précision et du rappel, elle est définie par :

$$F_{mes} = 2\frac{Précision.Rappel}{Précision + Rappel}$$

Définition 37 (Le taux d'erreur)
Le taux d'erreur correspond au nombre de mauvaises réponses. Il mesure la qualité du modèle dans la détection des cas positifs mais aussi des cas négatifs. Il se calcule par :

$$TErr = \frac{FP + FN}{VP + VN + FP + FN}$$

Définition 38 (Accuracy)
L'accuracy mesure l'exactitude des prédictions par rapport à la réalité observée. Elle est donnée par :

$$Acc = \frac{VP + VN}{VP + VN + FN + FP} = 1 - TErr$$

Définition 39 (L'Aire sous la courbe ROC)
L'Aire sous la courbe ROC, en anglais Area Under Curve (RAUC). Lorsque les unités sont normalisées, la RAUC est la probabilité qu'un classifieur choisisse une instance positive donnée, plutôt qu'une négative (sachant que les rangs des positives sont inférieurs à ceux des négatives). Graphiquement (voir figure D.1), il s'agit de représenter le taux de vrais positifs en fonction du taux de faux positifs ou encore la sensibilité en fonction du complément de la spécificité (1 − Spécificité).

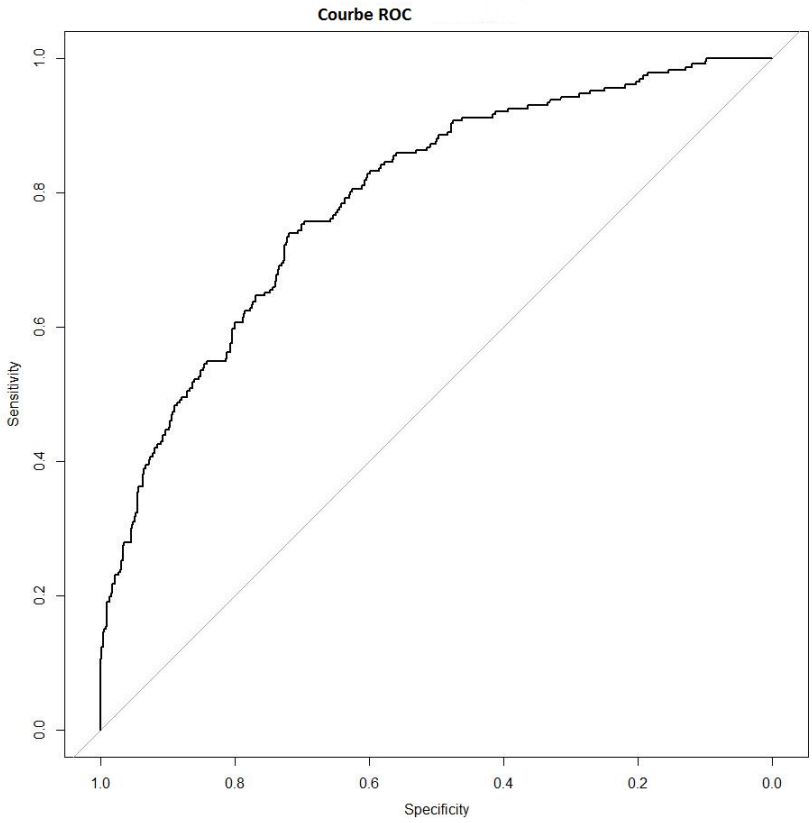


Figure D.1 – Exemple de courbe ROC.

D.4 Vecteur maximum

Nous allons expliquer le principe du vecteur maximum sur la base d'un exemple. Nous partons en vacances en bord de mer et nous souhaitons réserver un hôtel. Seulement, nous avons quelques exigences : tout d'abord nous voulons que l'hôtel soit le plus proche de la plage, qu'il soit de bon standing et que bien évidemment le prix de chambre soit le moins cher possible.

Après quelques recherches nous obtenons la liste d'hôtels présentée dans le tableau D.2. Dans chaque colonne est présentée successivement le nom de l'hôtel, le standing de l'hôtel (étoile), la distance à la plage en kilomètres (dist) et le prix total du séjour. En se basant sur ces trois derniers attributs, nous pouvons envisager les données comme un ensemble de triplets.

Le principe du vecteur maximum est de trouver les maximums parmi tous les uplets qui sont dominés par les autres. Par exemple, BestHotel peut être éliminé en comparaison avec Kyriad. Ibis peut être éliminé en comparaison avec BestHotel, Novotel ou Kyriad. BelHotel est éliminé en comparaison avec Mercure, Novotel ou Kyriad.

Tableau D.2 – Liste des hôtels.

nom	**étoile**	**dist**	**prix**
BestHotel	**	0,7	1 175
Ibis	*	1,2	1 237
Novotel	*	1,2	750
Mercure	***	0,6	2 250
BelHotel	***	0,5	2 550
Kyriad	**	0,5	980

Il peut y avoir plusieurs maximums, car un maximum n'a pas besoin d'être le meilleur sur tous les critères. Par exemple, Kyriad n'a pas le plus haut standing, il n'est pas le plus proche de la plage et n'est pas non plus le moins cher. Néanmoins, il présente un bon équilibre entre les différents critères.
Une manière de résoudre ce problème est de classer les hôtels selon leur rang pour chaque critère. Par exemple, nous souhaitons un hôtel de haut standing, donc Mercure et BelHotel auront un rang égal à 1 sur le critère étoile, Kyriad et BestHotel auront un rang égal à 2 et les autres un rang égal à 3. Nous procédons ainsi pour tous les critères et nous obtenons le tableau D.3. Dans la dernière colonne nous faisons la somme des rangs. Ainsi, il devient plus aisé de choisir l'hôtel présentant le meilleur compromis selon nos critères en prenant celui dont la somme des rangs est minimale.

Tableau D.3 – Classement des hôtels selon leurs critères.

	Rangs			
nom	**étoile**	**dist**	**prix**	**Somme**
BestHotel	3	4	3	10
Ibis	5	5	4	14
Novotel	5	5	1	11
Mercure	1	3	5	9
BelHotel	1	1	6	8
Kyriad	3	1	2	**6**

Ainsi, l'hôtel présentant le meilleur compromis selon nos critères est le Kyriad avec la plus faible somme des rangs égale à 6.

Annexe

E

Extraction de motifs spatio-temporels

Les tableaux sont colorés de sorte à identifier les catégories ayant les labels identiques et regroupant les mêmes notions d'un tableau à l'autre.

Tableau E.1: Regroupements des codes GHM pour la visualisation des flux de patients

Code	**Libellé**	**Nouveau code**	**Libellé nouveau code**
01C06	Interventions sur le système vasculaire précérébral	ISCHE	Ischémie
01M16	Accidents ischémiques transitoires et occlusions des artères précérébrales	ISCHE	Ischémie
01M32	Explorations et surveillance pour affections du système nerveux	AUTRE	Autres

Code	Libellé	Nouveau code	Libellé nouveau code
04M10	Embolies pulmonaires	ACOMP	Autres complications
04M11	Signes et symptômes respiratoires	ACOMP	Autres complications
04M13	Œdème pulmonaire et détresse respiratoire	ACOMP	Autres complications
04M17	Épanchements pleuraux	ACOMP	Autres complications
05C04	Pontages aortocoronariens avec cathétérisme cardiaque ou coronarographie	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C05	Pontages aortocoronariens sans cathétérisme cardiaque, ni coronarographie	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C08	Autres interventions cardiothoraciques ou vasculaires sans circulation extra-corporelle	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C10	Chirurgie majeure de revascularisation	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C11	Autres interventions de chirurgie vasculaire	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C12	Amputations du membre inférieur, sauf des orteils, pour troubles circulatoires	ISCHE	Ischémie
05C13	Amputations pour troubles circulatoires portant sur le membre supérieur ou les orteils	ISCHE	Ischémie
05C14	Poses d'un stimulateur cardiaque permanent avec IM aigu ou insuffisance cardiaque congestive ou état de choc	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C15	Poses d'un stimulateur cardiaque permanent sans IM aigu, ni insuffisance cardiaque congestive, ni état de choc	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C18	Autres interventions sur le système circulatoire	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C19	Poses d'un défibrillateur cardiaque	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05C21	Créations et réfections de fistules artérioveineuses	CHIRC	Chirurgie cardiothoracique -

Code	Libellé	Nouveau code	Libellé nouveau code
	pour affections de la CMD 05		Chirurgie de revascularisation
05C22	Remplacements de stimulateurs cardiaques permanents	CHIRC	Chirurgie cardiothoracique - Chirurgie de revascularisation
05K05	Endoprothèse avec IM	STEN1	Endoprothèse avec IM
05K06	Endoprothèse sans IM	STEN2	Endoprothèse sans IM
05K10	Actes diagnostiques par voie vasculaire	AUTRE	Autres
05K11	Traitements des troubles du rythme par voie vasculaire	TRYTH	Troubles du rythme et de la conduction cardiaque
05K13	Actes thérapeutiques par voie vasculaire sauf endoprothèses	ACI	Actes de cardiologie interventionnelle
05K14	Mise en place de certains accès vasculaires pour des affections de la CMD 05	ACI	Actes de cardiologie interventionnelle
05K19	Traitements majeurs de troubles du rythme par voie vasculaire	TRYTH	Troubles du rythme et de la conduction cardiaque
05K20	Autres traitements de troubles du rythme par voie vasculaire	TRYTH	Troubles du rythme et de la conduction cardiaque
05K21	Poses de bioprothèses de valves cardiaques par voie vasculaire	ACI	Actes de cardiologie interventionnelle
05K23	Ablations, repositionnements et poses de sondes cardiaques supplémentaires par voie vasculaire	ACI	Actes de cardiologie interventionnelle
05K24	Dilatations coronaires et autres actes thérapeutiques sur le cœur par voie vasculaire	ACI	Actes de cardiologie interventionnelle
05K25	Actes thérapeutiques sur les artères par voie vasculaire	ACI	Actes de cardiologie interventionnelle
05M04	Infarctus aigu du myocarde	I.MYO	Infarctus aigu du myocarde
05M06	Angine de poitrine	ISCHE	Ischémie
05M11	Cardiopathies congénitales et valvulopathies	ISCHE	Ischémie
05M16	Athérosclérose coronarienne	ISCHE	Ischémie
05M17	Autres affections de l'appareil circulatoire	AUTRE	Autres
05M21	Infarctus aigu du myocarde avec décès :	DECES	Décès

Code	Libellé	Nouveau code	Libellé nouveau code
	séjours de moins de 2 jours		
06M09	Autres affections digestives	AFDIG	Affections digestives
07C12	Autres interventions sur les voies biliaires sauf cholécystectomies isolées	AUTRE	Autres
08C22	Interventions pour reprise de prothèses articulaires	CHIRO	Chirurgie orthopédique
08C27	Autres interventions sur le rachis	CHIRO	Chirurgie orthopédique
08C32	Interventions sur la jambe	CHIRO	Chirurgie orthopédique
08C47	Prothèses de hanche pour traumatismes récents	CHIRO	Chirurgie orthopédique
10M02	Diabète, âge supérieur à 35 ans	DIABE	Diabète
10M11	Autres maladies métaboliques congénitales	CHOLE	Cholestérol
11C02	Interventions sur les reins et les uretères et chirurgie majeure de la vessie pour une affection tumorale	CHIRU	Chirurgie urologique
11C08	Autres interventions sur les reins et les voies urinaires	CHIRU	Chirurgie urologique
11M02	Lithiases urinaires	IRENA	Infections rénales
11M15	Autres affections des reins	DIABE	Diabète
	et des voies urinaires d'origine diabétique		
12C11	Interventions pelviennes majeures chez l'homme pour tumeurs malignes	AUTRE	Autres
23M06	Autres facteurs influant sur l'état de santé	AUTRE	Autres
23M10	Soins de contrôle chirurgicaux	AUTRE	Autres
23M20	Autres symptômes et motifs de recours aux soins de la CMD 23	AUTRE	Autres
Deces	Décès	DECES	Décès

Tableau E.2: Regroupements des codes CIM-10 pour la visualisation des flux de patients

Code	**Libellé**	**Nouveau code**	**Libellé nouveau code**
C61	Tumeur maligne de la prostate	AUT	Autres troubles
D62	Anémie posthémorragique aiguë	AUT	Autres troubles
Deces	Décès	DC	Décès
E11	Diabète sucré de type 2	DIA	Diabète
E13	Autres diabètes sucrés précisés	DIA	Diabète
E78	Anomalies du métabolisme des lipoprotéines et autres lipidémies	COL	Cholestérol
I10	Hypertension essentielle	TENS	Hypo/Hypertension
I11	Cardiopathie hypertensive	AUC	Autres cardiopathies
I20	Angine de poitrine	A.P	Angine de poitrine
I21	Infarctus aigu du myocarde	I.M	Infarctus du myocarde
I22	Infarctus à répétitions	I.M	Infarctus du myocarde
I23	Certaines complications récentes d'un IM aigu	COM	Complications mécaniques
I24	Autres cardiopathies ischémiques aiguës	ISC	Ischémie
I25	Cardiopathie ischémique chronique	ISC	Ischémie
I26	Embolie pulmonaire	ACO	Autres complications
I28	Autres maladies des vaisseaux pulmonaires	ACO	Autres complications
I34	Atteintes non rhumatismales de la valvule mitrale	AUC	Autres cardiopathies
I35	Atteintes non rhumatismales de la valvule aortique	ISC	Ischémie
I38	Endocardite, valvule non précisée	AUC	Autres cardiopathies
I42	Myocardiopathie	AUT	Autres troubles
I44	Bloc de branche gauche et auriculoventriculaire	TRY	Troubles du rythme et de la conduction cardiaque
I45	Autres troubles de la conduction	TRY	Troubles du rythme et de la conduction cardiaque
I46	Arrêt cardiaque	COM	Complications mécaniques

Code	Libellé	Nouveau code	Libellé nouveau code
I47	Tachycardie paroxystique	TRY	Troubles du rythme et de la conduction cardiaque
I48	Fibrillation et flutter auriculaires	TRY	Troubles du rythme et de la conduction cardiaque
I49	Autres arythmies cardiaques	TRY	Troubles du rythme et de la conduction cardiaque
I50	Insuffisance cardiaque	I.C	Insuffisance cardiaque
I65	Occlusion et sténose des artères précérébrales, n'entraînant pas un infarctus cérébral	ISC	Ischémie
I70	Athérosclérose	ISC	Ischémie
I71	Anévrismes aortiques et dissections	ANV	Anévrisme
I72	Autres anévrismes et dissections	ANV	Anévrisme
I74	Embolie et thrombose artérielles	ISC	Ischémie
I77	Autres atteintes des artères et artérioles	ISC	Ischémie
I82	Autres embolies et thromboses veineuses	ISC	Ischémie
I83	Varices des membres inférieurs	AUT	Autre troubles
I95	Hypotension	TENS	Hypo/Hypertension
J96	Insuffisance respiratoire, non classée ailleurs	ACO	Autres complications
N17	Insuffisance rénale aiguë	AUT	Autres troubles
Q21	Malformations congénitales des cloisons cardiaques	COM	Complications mécaniques
Q24	Autres malformations congénitales cardiaques	COM	Complications mécaniques
R00	Anomalies du rythme cardiaque	TRY	Troubles du rythme et de la conduction cardiaque
R06	Anomalies de la respiration	ACO	Autres complications
R07	Douleur au niveau de la gorge et du thorax	A.P	Angine de poitrine
R25	Mouvements involontaires anormaux	AUT	Autres troubles
R55	Syncope et collapsus	TRY	Troubles du rythme et de la conduction cardiaque
R57	Choc, non classé ailleurs	TRY	Troubles du rythme et de la conduction cardiaque
R93	Résultats anormaux d'imagerie diagnostique d'autres parties du corps	TTT	Observation, surveillance et suivi de traitement - Résultats anormaux d'examens
R94	Résultats anormaux d'explorations fonctionnelles	TTT	Observation, surveillance et suivi de traitement -

Code	Libellé	Nouveau code	Libellé nouveau code
			Résultats anormaux d'examens
T82	Complications de prothèses, implants et greffes cardiaques et vasculaires	GRF	Greffes cardiaques - complications de greffes - Suivi ajustement
Z03	Mise en observation et examen médical pour suspicion de maladies	TTT	Observation, surveillance et suivi de traitement - Résultats anormaux d'examens
Z04	Examen et mise en observation pour d'autres raisons	TTT	Observation, surveillance et suivi de traitement - Résultats anormaux d'examens
Z09	Examen de contrôle après traitement d'affections autres que les tumeurs malignes	TTT	Observation, surveillance et suivi de traitement - Résultats anormaux d'examens
Z45	Ajustement et entretien d'une prothèse interne	GRF	Greffes cardiaques - complications de greffes - Suivi ajustement
Z51	Autres soins médicaux	TTT	Observation, surveillance et suivi de traitement - Résultats anormaux d'examens
Z94	Greffe d'organe et de tissu	GRF	Greffes cardiaques - complications de greffes - Suivi ajustement
Z95	Implants et de greffes cardiaques et vasculaires	GRF	Greffes cardiaques - complications de greffes - Suivi ajustement

Bibliographie

[Abbasi et Smith, 2003] Abbasi, K. et Smith, R. (2003). No more free lunches. *British Medical Journal*, 326(7400):1155–1156.

[Abbott et Tsay, 2000] Abbott, A. et Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology : Review and prospect. *Sociological Methods & Research*, 29(1):3–33.

[Abitbol, 2005] Abitbol, E. (2005). Complications mécaniques de l'infarctus du myocarde récent. *Médecine Thérapeutique Cardio*, 1(6):559–569.

[Aboa-Eboulé *et al.*, 2013] Aboa-Eboulé, C., Mengue, D., Benzenine, E., Hommel, M., Giroud, M., Béjot, Y. et Quantin, C. (2013). How accurate is the reporting of stroke in hospital discharge data ? a pilot validation study using a population-based stroke registry as control. *Journal of Neurology*, 260(2):605–613.

[Ades *et al.*, 1997] Ades, P. A., Pashkow, F. J. et Nestor, J. R. (1997). Cost-Effectiveness of Cardiac Rehabilitation After Myocardia. *Journal of Cardiopulmonary Rehabilitation*, 17(4):222–231.

[Aeyels *et al.*, 2016] Aeyels, D., Van Vugt, S., Sinnaeve, P. R., Panella, M., Van Zelm, R., Sermeus, W. et Vanhaecht, K. (2016). Lack of evidence and standardization in care pathway documents for patients with ST-elevated myocardial infarction. *European Journal of Cardiovascular Nursing : Journal of the Working Group on Cardiovascular Nursing of the European Society of Cardiology*, 15(3):e45–e51.

[Agrawal *et al.*, 2011] Agrawal, D., Das, S. et El Abbadi, A. (2011). Big data and cloud computing : current state and future opportunities. *In Proceedings of the 14th International Conference on Extending Database Technology*, pages 530–533.

[Agrawal et Srikant, 1995] Agrawal, R. et Srikant, R. (1995). Mining sequential patterns. *In Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14.

[Ahmed *et al.*, 2011] Ahmed, S., Mayo, N., Scott, S., Kuspinar, A. et Schwartz, C. (2011). Using latent trajectory analysis of residuals to detect response shift in general health among patients with multiple sclerosis article. *Quality of Life Research*, 20(10):1555–1560.

[Al-Salameh *et al.*, 2016] Al-Salameh, A., Bucher, S., Becquemont, L. et Ringa, V. (2016). Contrôle insuffisant des facteurs de risque cardiovasculaire chez les femmes diabétiques âgées en soins primaires. *Revue d'Épidémiologie et de Santé Publique*, 64, Supplement 6:S302.

[Albarqouni *et al.*, 2016] Albarqouni, L., Smenes, K., Meinertz, T., Schunkert, H., Fang, X., Ronel, J. et Ladwig, K.-H. (2016). Patients' knowledge about symptoms and adequate behaviour during acute myocardial infarction and its impact on delay time. *Patient Education and Counseling*, 99(11):1845–1851.

[Alberti *et al.*, 2005] Alberti, C., Timsit, J. F. et Chevret, S. (2005). Analyse de survie : comment gérer les données censurées ? *Revue des Maladies Respiratoires*, 22(2):333–337.

[Alpaydin, 1997] Alpaydin, E. (1997). Voting over Multiple Condensed Nearest Neighbors. *Artificial Intelligence Review*, 11(5):115–132.

[Altman *et al.*, 2009] Altman, D. G., Vergouwe, Y., Royston, P. et Moons, K. G. M. (2009). Prognosis and prognostic research : validating a prognostic model. *British Medical Journal*, 338:b605.

[ANAES, 2004] ANAES (2004). Méthodes d'évaluation du risque cardio-vasculaire global. Rapport technique, Agence nationale d'accréditation et d'évaluation en santé. http://www.has-sante.fr/portail/upload/docs/application/pdf/Risque_cardio_vasculaire_rap.pdf.

[Andrés *et al.*, 2012] Andrés, E., Cordero, A., Magán, P., Alegría, E., León, M., Luengo, E., Botaya, R. M., Ortiz, L. G. et Casasnovas, J. A. (2012). Long-term mortality and hospital readmission after acute myocardial infarction : an eight-year follow-up study. *Revista Española de Cardiología (English Edition)*, 65(5):414–420.

[Angell, 2000] Angell, M. (2000). Is academic medicine for sale ? *The New England Journal of Medecine*, 342:1516–1518.

[Antes et Chalmers, 2003] Antes, G. et Chalmers, I. (2003). Under-reporting of clinical trials is unethical. *The Lancet*, 361(9362):978–979.

[Arko *et al.*, 2001] Arko, F. R., Bohannon, W. T., Mettauer, M., Lee, S. D., Patterson, D. E., Manning, L. G. et Buckley, C. J. (2001). Retroperitoneal Approach for Aortic Surgery : Is it Worth it ? *Cardiovascular Surgery*, 9(1):20–26.

[Armony *et al.*, 2015] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., Yom-Tov, G. B. *et al.* (2015). On patient flow in hospitals : A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194.

[Asaria *et al.*, 2017] Asaria, P., Elliott, P., Douglass, M., Obermeyer, Z., Soljak, M., Majeed, A. et Ezzati, M. (2017). Acute myocardial infarction hospital admissions and deaths in England : a national follow-back and follow-forward record-linkage study. *The Lancet Public Health*, 2(4):e191–e201.

[Asche *et al.*, 2016] Asche, C. V., Ren, J., Kirkness, C. S., Kim, M., Dong, Y. et Hippler, S. (2016). A Prediction Model to Identify Acute Myocardial Infarction (AMI) Patients at Risk for 30-day Readmission. *In Proceedings of the Summer Computer Simulation Conference*, pages 1–8.

[ATIH, 2014] ATIH (2014). Aide à l'utilisation des informations de chaînage. Rapport technique, Agence Technique de l'Information sur l'Hospitalisation. http://www.atih.sante.fr/aide-lutilisation-des-informations-de-chainage.

[ATIH, 2015] ATIH (2015). Manuel des GHM - Version définitive 11g. Guide et note technique, Texte officiel, Ministère de la Santé et des Sports. http://www.atih.sante.fr/manuel-des-ghm-version-definitive-11g.

[Austin, 2007] Austin, P. C. (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*, 26(15):2937–2957.

[Austin *et al.*, 2012] Austin, P. C., Lee, D. S., Steyerberg, E. W. et Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease : what improvement is achieved by using ensemble-based methods ? *Biometrical Journal*, 54(5):657–673.

[Aylin *et al.*, 2007] Aylin, P., Bottle, A. et Majeed, A. (2007). Use of administrative data or clinical databases as predictors of risk of death in hospital : comparison of models. *British Medical Journal*, 334(7602):1044–1052.

[Ayres *et al.*, 2002] Ayres, J., Flannick, J., Gehrke, J. et Yiu, T. (2002). Sequential pattern mining using a bitmap representation. *In Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining*, pages 429–435.

[Baade *et al.*, 2007] Baade, P., Youl, P., English, D., Mark Elwood, J. et Aitken, J. (2007). Clinical pathways to diagnose melanoma : a population-based study : Melanoma Research. *Melanoma Research*, 17(4):243–249.

[Bada, 2014] Bada, M. (2014). Mapping of biomedical text to concepts of lexicons, terminologies, and ontologies. *Methods in Molecular Biology*, 1159:33–45.

[Basch *et al.*, 2014] Basch, E., Snyder, C., McNiff, K., Brown, R., Maddux, S., Smith, M. L., Atkinson, T. M., Howell, D., Chiang, A., Wood, W. *et al.* (2014). Patient-reported outcome performance measures in oncology. *Journal of Oncology Practice*, 10(3):209–211.

[Batal *et al.*, 2011] Batal, I., Valizadegan, H., Cooper, G. F. et Hauskrecht, M. (2011). A Pattern Mining Approach for Classifying Multivariate Temporal Data. *In Proceedings of International Conference on Bioinformatics and Biomedicine*, pages 358–365.

[Bayat *et al.*, 2001] Bayat, S., Cuggia, M., Duff, F. L. et Mauduit, N. (2001). Les bases de données épidémiologiques. *Les Cahiers du Numérique*, 2(2):155–176.

[Beck *et al.*, 2010] Beck, F., Guignard, R., Richard, J.-B., Wilquin, J.-L. et Peretti-Watel, P. (2010). Premiers résultats du baromètre santé 2010 évolutions récentes du tabagisme en france. *Institut National de Prévention et d'Education pour la Santé*, 10:1–13.

[Belle *et al.*, 2016] Belle, L., Motreff, P., Mangin, L., Rangé, G., Marcaggi, X., Marie, A., Ferrier, N., Dubreuil, O., Zemour, G., Souteyrand, G., Caussin, C., Amabile, N., Isaaz, K., Dauphin, R., Koning, R., Robin, C., Faurie, B., Bonello, L., Champin, S., Delhaye, C., Cuilleret, F., Mewton, N., Genty, C., Viallon, M., Bosson, J. L., Croisille, P. et Investigators, o. b. o. t. M. (2016). Comparison of Immediate With Delayed Stenting Using the Minimalist Immediate Mechanical Intervention Approach in Acute ST-Segment–Elevation Myocardial Infarction. *Circulation : Cardiovascular Interventions*, 9(3):e003388.

[Bellman, 1954] Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515.

[Benamer *et al.*, 2007] Benamer, H., Lefèvre, J. J., Debure, A. et Gaultier, C. (2007). Coronaropathie etangioplastie coronaire dans l'insuffisance rénale dialysée. *Annales de Cardiologie et d'Angéiologie*, 56(1):10–15.

[Benkert *et al.*, 2008] Benkert, M., Gudmundsson, J., Hübner, F. et Wolle, T. (2008). Reporting flock patterns. *Computational Geometry*, 41(3):111–125.

[Benzécri, 1976] Benzécri, J.-P. (1976). Histoire et préhistoire de l'analyse des données. *Les Cahiers de l'Analyse des Données*, 1(1):9–32.

[Béranger et Bouadi, 2014] Béranger, J. et Bouadi, R. (2014). Approche éthico-juridique de l'usage des données médicales à caractère personnel. *Les Cahiers du Numérique*, 10(2):93–123.

[Bernier *et al.*, 2012] Bernier, M. O., Mezzarobba, M., Maupu, E., Caër-Lorho, S., Brisse, H. J., Laurier, D., Brunelle, F. et Chatellier, G. (2012). Utilisation des données du programme de médicalisation des systèmes d'information (PMSI) dans les études épidémiologiques : application à la Cohorte Enfant Scanner. *Revue d'Épidémiologie et de Santé Publique*, 60(5):363–370.

[Bestul *et al.*, 2004] Bestul, M. B., McCollum, M., Stringer, K. A. et Burchenal, J. (2004). Impact of a Critical Pathway on Acute Myocardial Infarction Quality Indicators. *Pharmacotherapy : The Journal of Human Pharmacology and Drug Therapy*, 24(2):173–178.

[Bettencourt-Silva *et al.*, 2015] Bettencourt-Silva, J. H., Clark, J., Cooper, C. S., Mills, R., Rayward-Smith, V. J. et de la Iglesia, B. (2015). Building Data-Driven Pathways From Routinely Collected Hospital Data : A Case Study on Prostate Cancer. *Journal of Medical Internet Research Medical Informatics*, 3(3):e26.

[Beyeme-Ondoua, 2007] Beyeme-Ondoua, J.-P. (2007). Evaluation of the quality of surveillance data for colorectal cancer from the national PMSI database in 2003. *Santé Publique*, 19(6):471–480.

[Beygelzimer *et al.*, 2013] Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D. et Li, S. (2013). *FNN : Fast Nearest Neighbor Search Algorithms and Applications*. https://CRAN.R-project.org/package=FNN.

[Biffl *et al.*, 2001] Biffl, W. L., Smith, W. R., Moore, E. E., Gonzalez, R. J., Morgan, S. J., Hennessey, T., Offner, P. J., Ray, C. E., Franciose, R. J. et Burch, J. M. (2001). Evolution of a Multidisciplinary Clinical Pathway for the Management of Unstable Patients With Pelvic Fractures. *Annals of Surgery*, 233(6):843–850.

[Bilenko *et al.*, 2003] Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P. et Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23.

[Blanc *et al.*, 1999] Blanc, P., Aouifi, A., Chiari, P., Bouvier, H., Jegaden, O. et Lehot, J. J. (1999). Chirurgie cardiaque mini-invasive : techniques chirurgicales et particularités anesthésiques. *Annales Françaises d'Anesthésie et de Réanimation*, 18(7):748–771.

[Blei *et al.*, 2003] Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

[Blin *et al.*, 2016] Blin, P., Dureau-Pournin, C., Lassalle, R., Jové, J., Thomas-Delecourt, F., Droz-Perroteau, C., Danchin, N. et Moore, N. (2016). Outcomes, health care resources use, and costs in patients with post-myocardial infarction : The horus cohort study in the egb french claims and hospital database. *Value in Health*, 19(3):A16.

[Bocquier *et al.*, 2011] Bocquier, A., Thomas, N., Zitouni, J., Lewandowski, E., Cortaredona, S., Jardin, M., Favier, O., Finkel, S., Champion, F., Bernardy, A., Trugeon, A. et Verger, P. (2011). Évaluation de la qualité du chaînage des séjours hospitaliers pour l'étude des variations spatiales de santé à partir des données du PMSI. Étude de faisabilité dans trois régions françaises. *Revue d'Épidémiologie et de Santé Publique*, 59(4):243–249.

[Boddaert *et al.*, 2015] Boddaert, J., Barondeau, M.-L., Khiami, F., Nion, N., Frandji, D. et Riou, B. (2015). Cost accounting of a geriatric perioperative unit. *Santé Publique*, 27(4):529–537.

[Bollegala *et al.*, 2010] Bollegala, D., Okazaki, N. et Ishizuka, M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management*, 46(1):89–109.

[Bossuyt *et al.*, 2015] Bossuyt, N., Van Casteren, V., Goderis, G., Wens, J., Moreels, S., Vanthomme, K. et De Clercq, E. (2015). Public Health Triangulation to inform decision-making in Belgium. *Studies in Health Technology and Informatics*, 210:855–859.

[Boudemaghe, 2016] Boudemaghe, T. (2016). *Élaboration d'un cadre méthodologique pour l'analyse de l'information médicale de la tarification à l'activité*. Thèse de doctorat, Université Montpellier.

[Boutault *et al.*, 1999] Boutault, F., Dodart, L., Gas, C., Paoli, J. R., Lauwers, F. et Chale, J. J. (1999). [Two or three things about the PMSI in stomatology and maxillofacial surgery]. *Revue de Stomatologie et de Chirurgie Maxillo-Faciale*, 100(6):279–287.

[Box *et al.*, 2015] Box, G. E., Jenkins, G. M., Reinsel, G. C. et Ljung, G. M. (2015). *Time series analysis : forecasting and control*. John Wiley & Sons.

[Boytsov, 2011] Boytsov, L. (2011). Indexing Methods for Approximate Dictionary Searching : Comparative Analysis. *Journal of Experimental Algorithmics*, 16:1–91.

[Brandes, 2001] Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

[Brédart *et al.*, 2014] Brédart, A., Marrel, A., Abetz-Webb, L., Lasch, K. et Acquadro, C. (2014). Interviewing to develop patient-reported outcome (pro) measures for clinical research : eliciting patients' experience. *Health and Quality of Life Outcomes*, 12(1):15.

[Brier, 1950] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

[Brossette *et al.*, 1998] Brossette, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T. et Moser, S. A. (1998). Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. *Journal of the American Medical Informatics Association*, 5(4):373–381.

[Broyles *et al.*, 2010] Broyles, J. R., Cochran, J. K. et Montgomery, D. C. (2010). A statistical Markov chain approximation of transient hospital inpatient inventory. *European Journal of Operational Research*, 207(3):1645–1657.

[Buckley *et al.*, 2000] Buckley, C., Lee, S., Arko, F., Bohannon, W., Mettauer, M., Patterson, D. et Manning, L. (2000). Economic considerations for aortic surgery : retroperitoneal approach–is it worth it ? *Acta Chirurgica Belgica*, 100(6):247–250.

[Bureau International du Travail, 2003] Bureau International du Travail (2003). Le stress dans le monde du travail. *In Le travail dans le monde*, chapitre 5, page 17.

[Burke *et al.*, 2015] Burke, T., Manglani, Y., Altawil, Z., Dickson, A., Clark, R., Okelo, S. et Ahn, R. (2015). A Safe-Anesthesia Innovation for Emergency and Life-Improving Surgeries When no Anesthetist is Available : A Descriptive Review of 193 Consecutive Surgeries. *World Journal of Surgery*, 39(9):2147–2152.

[Béjot *et al.*, 2011] Béjot, Y., Benzenine, E., Lorgis, L., Zeller, M., Aubé, H., Giroud, M., Cottin, Y. et Quantin, C. (2011). Comparative Analysis of Patients with Acute Coronary and Cerebrovascular Syndromes from the National French Hospitalization Health Care System Database. *Neuroepidemiology*, 37(3-4):143–152.

[Cambou *et al.*, 2004] Cambou, J. P., Danchin, N., Boutalbi, Y., Hanania, G., Humbert, R., Clerson, P., Vaur, L., Guéret, P., Blanchard, D., Genès, N. et Lablanche, J. M. (2004). Évolution de la prise en charge et du pronostic de l'infarctus du myocarde en France entre 1995 et 2000 : résultats des études USIK 1995 et USIC 2000. *Annales de Cardiologie et d'Angéiologie*, 53(1):12–17.

[Cambou *et al.*, 1997] Cambou, J.-P., Genes, N., Vaur, L., Renault, M., Etienne, S., Ferrieres, J. et Danchin, N. (1997). Epidémiologie de l'infarctus du myocarde en france : Spécificités régionales. *Archives des Maladies du Coeur et des Vaisseaux*, 90(11):1511–1519.

[Campbell, 2008] Campbell, D. J. (2008). Why do men and women differ in their risk of myocardial infarction ? *European Heart Journal*, 29(7):835–836.

[Canto *et al.*, 2012] Canto, J. G., Rogers, W. J., Goldberg, R. J., Peterson, E. D., Wenger, N. K., Vaccarino, V., Kiefe, C. I., Frederick, P. D., Sopko, G. et Zheng, Z.-J. (2012). Association of Age and Sex With Myocardial Infarction Symptom Presentation and In-Hospital Mortality. *Journal of the American Medical Association*, 307(8):813–822.

[Cartier *et al.*, 2014] Cartier, T., Naiditch, M. et Lombrail, P. (2014). Hospitalisations potentiellement évitables : une responsabilité des seuls soins de premier recours ? *Revue d'Épidémiologie et de Santé Publique*, 62(4):225–236.

[Célant *et al.*, 2014] Célant, N., Dourgnon, P., Guillaume, S., Pierre, A., Rochereau, T. et Sermet, C. (2014). L'enquête santé et protection sociale (esps) 2012. premiers résultats. *Questions d'Économie de la Santé*, 198:1–6.

[Chandrasekhar *et al.*, 2011] Chandrasekhar, T., Thangavel, K. et Elayaraja, E. (2011). Effective clustering algorithms for gene expression data. *International Journal of Computer Applications*, 32(4):25–29.

[Chantry *et al.*, 2012] Chantry, A. A., Deneux-Tharaux, C., Bal, G., Zeitlin, J., Quantin, C., Bouvier-Colle, M.-H. et 1, p. l. g. G. (2012). Le programme de médicalisation du système d'information (PMSI) – processus de production des données, validité et sources d'erreurs dans le domaine de la morbidité maternelle sévère. *Revue d'épidémiologie et de Santé Publique*, 60(3):177–188.

[Charlier *et al.*, 2014] Charlier, C. M., Gonzalez-Dunia, D. et Malnou, C. E. (2014). Bornavirus et cellules cibles : une amitié presque sincère. *Virologie*, 18(4):187–200.

[Charlson *et al.*, 1987] Charlson, M. E., Pompei, P., Ales, K. L. et MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies : development and validation. *Journal of Chronic Diseases*, 40:373–383.

[Chase, 2005] Chase, M. (2005). Beginning patient flow modeling in Vancouver Coastal Health. *Clinical and Investigative Medicine*, 28(6):323–325.

[Chevreul *et al.*, 2012] Chevreul, K., Prigent, A., Durand-Zaleski, I. et Steg, P. G. (2012). Does lay media ranking of hospitals reflect lower mortality in treating acute myocardial infarction ? *Archives of Cardiovascular Diseases*, 105(10):489–498.

[Chiou et Li, 2007] Chiou, J.-M. et Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 69(4):679–699.

[Chung *et al.*, 2015] Chung, S.-C., Sundström, J., Gale, C. P., James, S., Deanfield, J., Wallentin, L., Timmis, A., Jernberg, T. et Hemingway, H. (2015). Comparison of hospital variation in acute myocardial infarction care and outcome between sweden and united kingdom : population based cohort study using nationwide clinical registries. *British Medical Journal*, 351:h3913.

[Claeskens *et al.*, 2008] Claeskens, G., Croux, C. et Kerckhoven, J. V. (2008). An Information Criterion for Variable Selection in Support Vector Machines. *Journal of Machine Learning Research*, 9(Mar):541–558.

[Cocchi *et al.*, 2013] Cocchi, A., Meneghelli, A., Erlicher, A., Pisano, A., Cascio, M. T. et Preti, A. (2013). Patterns of referral in first-episode schizophrenia and ultra high-risk individuals : results from an early intervention program in Italy. *Social Psychiatry and Psychiatric Epidemiology*, 48(12):1905–1916.

[Cohen *et al.*, 2006] Cohen, A. M., Hersh, W. R., Peterson, K. et Yen, P.-Y. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

[Cohen-Haguenauer, 2011] Cohen-Haguenauer, O. (2011). Thérapie génique des maladies rares. *La Revue de Médecine Interne*, 32:S210–S212.

[Cohn, 1985] Cohn, P. F. (1985). Silent myocardial ischemia : classification, prevalence, and prognosis. *The American Journal of Medicine*, 79(3):2–6.

[Colin *et al.*, 2007] Colin, X., Lafuma, A. et Gueron, B. (2007). Costs of cardiovascular events of diabetic patients in the French hospitals. *Diabetes & Metabolism*, 33(4):310–313.

[Collins *et al.*, 2015] Collins, G. S., Reitsma, J. B., Altman, D. G. et Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) : the TRIPOD statement. *BioMed Central Medicine*, 13(1):g7594.

[Coloma *et al.*, 2013] Coloma, P. M., Valkhoff, V. E., Mazzaglia, G., Nielsson, M. S., Pedersen, L., Molokhia, M., Mosseveld, M., Morabito, P., Schuemie, M. J., van der Lei, J. *et al.* (2013). Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems : a validation study in three european countries. *British Medical Journal Open*, 3(6):e002862.

[Colonna *et al.*, 2012] Colonna, M., Mitton, N., Schott, A.-M., Remontet, L., Olive, F., Gomez, F., Iwaz, J., Polazzi, S., Bossard, N. et Trombert, B. (2012). Joint use of epidemiological and hospital medico-administrative data to estimate prevalence. Application to French data on breast cancer. *Cancer Epidemiology*, 36(2):116–121.

[Compieta *et al.*, 2007] Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F. et Kechadi, T. (2007). Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages and Computing*, 18(3):255–279.

[Conte *et al.*, 2016] Conte, C., Rueter, M., Laurent, G., Bourrel, R., Lapeyre-Mestre, M. et Despas, F. (2016). Psychotropic drug initiation during the first diagnosis and the active treatment phase of B cell non-Hodgkin's lymphoma : a cohort study of the French national health insurance database. *Supportive Care in Cancer*, 24(11):4791–4799.

[Couchoud *et al.*, 2015] Couchoud, C., Couillerot, A.-L., Dantony, E., Elsensohn, M.-H., Labeeuw, M., Villar, E., Ecochard, R. et Bongiovanni, I. (2015). Economic impact of a modification of the treatment trajectories of patients with end-stage renal disease. *Nephrology Dialysis Transplantation*, 30(12):2054–2068.

[Couvreur *et al.*, 1995] Couvreur, P., Dubernet, C. et Puisieux, F. (1995). Controlled drug delivery with nanoparticles : current possibilities and future trends. *European Journal of Pharmaceutics and Biopharmaceutics*, 41(1):2–13.

[Cresci *et al.*, 2011] Cresci, S., Wu, J., Province, M. A., Spertus, J. A., Steffes, M., McGill, J. B., Alderman, E. L., Brooks, M. M., Kelsey, S. F., Frye, R. L. et Bach, R. G. (2011). Peroxisome Proliferator-Activated Receptor Pathway Gene Polymorphism Associated With Extent of Coronary Artery Disease in Patients With Type 2 Diabetes in the Bypass Angioplasty Revascularization Investigation 2 Diabetes TrialClinical Perspective. *Circulation*, 124(13):1426–1434.

[Dalichampt *et al.*, 2014] Dalichampt, M., Weill, A., Raguideau, F., Ricordeau, P., Alla, F. et Zureik, M. (2014). Risque d'embolie pulmonaire, d'accident vasculaire cérébral ischémique et d'infarctus du myocarde chez les femmes sous contraceptif oral combiné en france : une étude de cohorte sur 5 millions de femmes de 15 à 49 ans à partir des données actualisées du SNIIRAM et du programme de médicalisation des systèmes d'information. *Revue d'Épidémiologie et de Santé Publique*, 62:S75.

[Damerau, 1964] Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications Association for Computing Machinery*, 7(3):171–176.

[Danchin *et al.*, 2003] Danchin, N., Hanania, G., Grenier, O., Vaur, L., Amelineau, E., Guéret, P., Blanchard, D., Ferrières, J., Genès, N., Lablanche, J. M., Cantet, C. et Cambou, J. P. (2003). Évolution du traitement de sortie après hospitalisation pour syndrome coronaire aigu en France entre 1995 et 2000 : données des études Usik 1995, Prévenir 1 et 2 et Usic 2000. *Annales de Cardiologie et d'Angéiologie*, 52(1):1–6.

[Danielsson *et al.*, 2009] Danielsson, U., Bengs, C., Lehti, A., Hammarström, A. et Johansson, E. E. (2009). Struck by lightning or slowly suffocating – gendered trajectories into depression. *BioMed Central Family Practice*, 10(1):56.

[Dart *et al.*, 2003] Dart, T., Cui, Y., Chatellier, G. et Degoulet, P. (2003). Analysis of hospitalised patient flows using data-mining. *Studies in Health Technology and Informatics*, 95:263–268.

[Davis *et al.*, 2014] Davis, L. A., Polk, B., Mann, A., Wolff, R. K., Kerr, G. S., Reimold, A. M., Cannon, G. W., Mikuls, T. R. et Caplan, L. (2014). Folic acid pathway single nucleotide polymorphisms associated with methotrexate significant adverse events in United States veterans with rheumatoid arthritis. *Clinical and Experimental Rheumatology*, 32(3):324–332.

[De Peretti et Bonaldi, 2010] De Peretti, C. et Bonaldi, C. (2010). Étalonnage du PMSI MCO pour la surveillance des infarctus du myocarde – année 2003. Rapport technique, Invs. http://opac.invs.sante.fr/doc_num.php?explnum_id=250.

[de Peretti *et al.*, 2012] de Peretti, C., Chin, F., Tuppin, P., Béjot, Y., Giroud, M., Schniztler, A. et Woimant (2012). Personnes hospitalisées pour accident vasculaire cérébral en France : tendances 2002-2008. *Bulletin Épidémiologique Hebdomadaire*, (10-11):125–130.

[De Peretti *et al.*, 2012] De Peretti, C., Chin, F., Tuppin, P. et Danchin, N. (2012). Personnes hospitalisées pour infarctus du myocarde en France : tendances 2002–2008. *Bulletin Épidémiologique Hebdomadaire*, 41:459–465.

[Debbas *et al.*, 1995] Debbas, N., Eeckhout, E., Stauffer, J. C., Kaufman, U., Vogt, P., Sigwart, U., Kappenberger, L. et Goy, J. J. (1995). Traitement par angioplastie au ballonnet de resténoses survenues sur prothèses endocoronaires. *Archives des Maladies du Coeur et des Vaisseaux*, 88(7):987–991.

[DeBusk, 1994] DeBusk, R. F. (1994). A Case-Management System for Coronary Risk Factor Modification after Acute Myocardial Infarction. *Annals of Internal Medicine*, 120(9):721–729.

[Defossez *et al.*, 2014] Defossez, G., Rollet, A., Dameron, O. et Ingrand, P. (2014). Temporal representation of care trajectories of cancer patients using data from a regional information system : an application in breast cancer. *BioMed Central Medical Informatics and Decision Making*, 14(1):24.

[Dégano *et al.*, 2015] Dégano, I. R., Salomaa, V., Veronesi, G., Ferriéres, J., Kirchberger, I., Laks, T., Havulinna, A. S., Ruidavets, J.-B., Ferrario, M. M., Meisinger, C. *et al.* (2015). Twenty-five-year trends in myocardial infarction attack and mortality rates, and case-fatality, in six european populations. *Heart*, 101(17):1413–1421.

[Delahaye *et al.*, 2001] Delahaye, F., Bory, M., Cohen, A., Danchin, N., De Gevigney, G., Dellinger, A., Fraboulet, J.-Y., Gayet, J.-L., Guize, L., Iung, P., Mabo, C., Monpère, P.-G., Steg, D. et Thomas (2001). Recommandations de la société française de cardiologie concernant la prise en charge de l'infarctus du myocarde après la phase aiguë. *Archives des Maladies du Cœur et des Vaisseaux*, 94(7):697–738.

[Delmas *et al.*, 2011] Delmas, M. C., Marguet, C., Raherison, C., Nicolau, J. et Fuhrman, C. (2011). Readmissions for asthma in France in 2002–2005. *Revue des Maladies Respiratoires*, 28(9):e115–e122.

[Dely *et al.*, 2012] Dely, C., Sellier, P., Dozol, A., Segouin, C., Moret, L. et Lombrail, P. (2012). Preventable readmissions of "community-acquired pneumonia" : Usefulness and reliability of an indicator of the quality of care of patients' care pathways. *Presse Médicale*, 41(1):e1–e9.

[Dent et Tutt, 2014] Dent, M. et Tutt, D. (2014). Electronic patient information systems and care pathways : The organisational challenges of implementation and integration. *Health Informatics Journal*, 20(3):176–188.

[Dexter *et al.*, 1999] Dexter, F., Macario, A., Traub, R., Hopwood, M. et Lubarsky, D. (1999). An Operating Room Scheduling Strategy to Maximize the Use of operating room block time : computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. *Anesthesia & Analgesia*, 89(1):7–20.

[Dharmarajan *et al.*, 2015] Dharmarajan, K., Hsieh, A. F., Kulkarni, V. T., Lin, Z., Ross, J. S., Horwitz, L. I., Kim, N., Suter, L. G., Lin, H., Normand, S.-L. T. et Krumholz, H. M. (2015). Trajectories of risk after hospitalization for heart failure, acute myocardial infarction, or pneumonia : retrospective cohort study. *British Medical Journal*, 350:h411.

[Dhillon *et al.*, 2004] Dhillon, I. S., Guan, Y. et Kulis, B. (2004). Kernel k-means : spectral clustering and normalized cuts. *In Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*, pages 551–556.

[Diallo, 2006] Diallo, M. S. (2006). L'approche prédictive dans la théorie de l'échantillonnage. Rapport technique, Université Laval Québec. http://archimede.mat.ulaval.ca/theses/MS-Diallo_06.pdf.

[Diaz *et al.*, 2008] Diaz, R. J., Laughlin, S., Nicolin, G., Buncic, J. R., Bouffet, E. et Bartels, U. (2008). Assessment of chemotherapeutic response in children with proptosis due to optic nerve glioma. *Child's Nervous System*, 24(6):707–712.

[Doherty et Kvarnström, 2008] Doherty, P. et Kvarnström, J. (2008). Temporal action logics. *Foundations of Artificial Intelligence*, 3:709–757.

[Domingos et Pazzani, 1997] Domingos, P. et Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130.

[DREES, 2015] DREES (2015). *L'État de santé de la population en France - Édition 2015*. Ministère des affaires sociales, de la santé et des droits des femmes.

[Dreyer *et al.*, 2014] Dreyer, R. P., Ranasinghe, I., Dharmarajan, K., Wang, Y., Hsieh, A., Bernheim, S. et Krumholz, H. (2014). Sex differences in 30-day readmission risk in young women and men with acute myocardial infarction. *Journal of the American College of Cardiology*, 12(63):A260.

[Dreyer *et al.*, 2015] Dreyer, R. P., Ranasinghe, I., Wang, Y., Dharmarajan, K., Murugiah, K., Nuti, S. V., Hsieh, A. F., Spertus, J. A. et Krumholz, H. M. (2015). Sex differences in the rate, timing and principal diagnoses of 30-day readmissions in younger patients with acute myocardial infarction. *Circulation*, 132(3):158–166.

[Dubes, 1987] Dubes, R. C. (1987). How many clusters are best ? - An experiment. *Pattern Recognition*, 20(6):645–663.

[Dujardin et Cambou, 2005] Dujardin, J.-J. et Cambou, J.-P. (2005). Épidémiologie de l'infarctus du myocarde. *Encyclopédie Médico-Chirurgicale - Cardiologie-Angéiologie*, 2(4):375–387.

[Dujardin et Fabre, 2008] Dujardin, J.-J. et Fabre, O. (2008). Complications de l'infarctus du myocarde. Évolution et pronostic. *Cardiologie*, 3(1):1–13.

[Ederle *et al.*, 2010] Ederle, J., Dobson, J., Featherstone, R. L., Bonati, L. H., van der Worp, H. B., de Borst, G. J., Lo, T. H., Gaines, P., Dorman, P. J., Macdonald, S., Lyrer, P. A., Hendriks, J. M., McCollum, C., Nederkoorn, P. J., Brown, M. M. et International Carotid Stenting Study investigators (2010). Carotid artery stenting compared with endarterectomy in patients with symptomatic carotid stenosis (International Carotid Stenting Study) : an interim analysis of a randomised controlled trial. *Lancet*, 375(9719):985–997.

[Egho *et al.*, 2013] Egho, E., Jay, N., Raïssi, C., Nuemi, G., Quantin, C. et Napoli, A. (2013). An approach for mining care trajectories for chronic diseases. *In Proceedings of the 14th Conference on Artificial Intelligence in Medicine.*, pages 258–267.

[El-Darzi *et al.*, 1998] El-Darzi, E., Vasilakis, C., Chaussalet, T. et Millard, P. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, 1(2):143–149.

[El Kalam *et al.*, 2004] El Kalam, A. A., Deswarte, Y., Trouessin, G. et Cordonnier, E. (2004). Gestion des données médicales anonymisées : problèmes et solutions. *In Proceedings of the 2ème Conférence Francophone en Gestion et Ingenierie des Systèmes Hospitaliers*, pages 9–11.

[Elixhauser *et al.*, 1998] Elixhauser, A., Claudia Steiner, Harris, R. et Coffey, R. (1998). Comorbidity Measures for Use with Administrative Data : Medical Care. *Medical care*, 36(1):8–27.

[Ellis *et al.*, 2010] Ellis, E., Ballance, K., Lunt, H. et Lewis, D. (2010). Diabetes outpatient care before and after admission for diabetic foot complications. *Journal of Wound Care*, 19(4):150–152.

[Fabrègue *et al.*, 2013] Fabrègue, M., Braud, A., Bringay, S., Le Ber, F. et Teisseire, M. (2013). Orderspan : Mining closed partially ordered patterns. *In Inproceedings of the International Symposium on Intelligent Data Analysis*, pages 186–197.

[Fabregue *et al.*, 2011] Fabregue, M., Bringay, S., Poncelet, P., Teisseire, M. et Orsetti, B. (2011). Mining microarray data to predict the histological grade of a breast cancer. *Journal of Biomedical Informatics*, 44(Suppl. 1):S12–S16.

[Falconnet *et al.*, 2009] Falconnet, C., Perrenoud, J.-J., Carballo, S., Roffi, M. et Keller, P.-F. (2009). Syndrome coronarien aigu : guidelines et spécificité gériatrique. *Revue Médicale Suisse*, 5(204):1137–1147.

[Fassbender *et al.*, 2009] Fassbender, K., Fainsinger, R. L., Carson, M. et Finegan, B. A. (2009). Cost trajectories at the end of life : the Canadian experience. *Journal of Pain and Symptom Management*, 38(1):75–80.

[Fekete et Plaisant, 1999] Fekete, J.-D. et Plaisant, C. (1999). Excentric labeling : dynamic neighborhood labeling for data visualization. *In Proceedings of the Special Interest Group on Computer-Human Interaction conference on Human Factors in Computing Systems*, pages 512–519.

[Ferry, 2016] Ferry, L. (2016). *La révolution transhumaniste.* Plon.

[Fetter et Freeman, 1986] Fetter, R. B. et Freeman, J. L. (1986). Diagnosis Related Groups : Product Line Management within Hospitals. *Academy of Management Review*, 11(1):41–54.

[Fetter *et al.*, 1980] Fetter, R. B., Shin, Y., Freeman, J. L., Averill, R. F. et Thompson, J. D. (1980). Case Mix Definition by Diagnosis-Related Groups. *Medical Care*, 18(2):1–53.

[Fineberg *et al.*, 2010] Fineberg, H. V., Scadden, D. et Goldman, L. (2010). Care of Patients with a Low Probability of Acute Myocardial Infarction. *The New England Journal of Medecine*, 310:1301–1307.

[Flament, 1981] Flament, C. (1981). L'analyse de similitude : une technique pour les recherches sur les représentations sociales. [Similarity analysis : A technique for researches in social representations.]. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 1(4):375–395.

[Foresee et Hagan, 1997] Foresee, F. D. et Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian learning. *In Proceedings of the International Conference on Neural Networks*, volume 3, pages 1930–1935.

[Fox *et al.*, 2006] Fox, K. A. A., Dabbous, O. H., Goldberg, R. J., Pieper, K. S., Eagle, K. A., Werf, F. V. d., Avezum, A., Goodman, S. G., Flather, M. D., Anderson, F. A. et Granger, C. B. (2006). Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome : prospective multinational observational study (GRACE). *British Medical Journal*, 333(7578):1091–1094.

[Frantzi *et al.*, 2000] Frantzi, K., Ananiadou, S. et Mima, H. (2000). Automatic recognition of multi-word terms :. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

[Freemantle *et al.*, 2013] Freemantle, N., Richardson, M., Wood, J., Ray, D., Khosla, S., Sun, P. et Pagano, D. (2013). Can we update the Summary Hospital Mortality Index (SHMI) to make a useful measure of the quality of hospital care ? An observational study. *British Medical Journal Open*, 3(1):e002018.

[Freisinger *et al.*, 2014] Freisinger, E., Fuerstenberg, T., Malyar, N. M., Wellmann, J., Keil, U., Breithardt, G. et Reinecke, H. (2014). German nationwide data on current trends and management of acute myocardial infarction : discrepancies between trials and real-life. *European Heart Journal*, 35(15):979–988.

[Fruchterman et Reingold, 1991] Fruchterman, T. M. J. et Reingold, E. M. (1991). Graph drawing by force-directed placement. *Journal of Software : Practice and Experience*, 21(11):1129–1164.

[Frunza *et al.*, 2011] Frunza, O., Inkpen, D., Matwin, S., Klement, W. et O'Blenis, P. (2011). Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, 51(1):17–25.

[Fukuyama, 2004] Fukuyama, F. (2004). *La fin de l'homme : les conséquences de la révolution biotechnique.*

[Gabadinho *et al.*, 2011] Gabadinho, A., Ritschard, G., Müller, N. et Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(1):1–37.

[Gabet *et al.*, 2016] Gabet, A., Danchin, N. et Olié, V. (2016). Infartus du myocarde chez la femme : évolutions des taux d'hospitalisation et de mortalité, France, 2002-2013. *Bulletin Épidémiologique Hebdomadaire*, pages 100–108.

[Gabizon et Lonn, 2015] Gabizon, I. et Lonn, E. (2015). Young women with acute myocardial infarction and the post-hospital syndrome. *Circulation*, 132(3):149–151.

[Gaffney et Smyth, 1999] Gaffney, S. et Smyth, P. (1999). Trajectory clustering with mixtures of regression models. *In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 63–72.

[Gallivan, 2005] Gallivan, S. (2005). Mathematical methods to assist with hospital operation and planning. *Clinical and Investigative Medicine*, 28(6):326–330.

[Gebregziabher *et al.*, 2010] Gebregziabher, M., Egede, L. E., Lynch, C. P., Echols, C. et Zhao, Y. (2010). Effect of Trajectories of Glycemic Control on Mortality in Type 2 Diabetes : A Semiparametric Joint Modeling Approach. *American Journal of Epidemiology*, 171(10):1090–1098.

[Genolini *et al.*, 2016] Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S. et Subtil, F. (2016). kmlShape : An Efficient Method to Cluster Longitudinal Data (Time-Series) According to Their Shapes. *Plos one*, 11(6): e0150738.

[Genolini *et al.*, 2013] Genolini, C., Pingault, J.-B., Driss, T., Côté, S., Tremblay, R. E., Vitaro, F., Arnaud, C. et Falissard, B. (2013). Kml3d : a non-parametric algorithm for clustering joint trajectories. *Computer Methods and Programs in Biomedicine*, 109(1):104–111.

[Geraci *et al.*, 2005] Geraci, J. M., Johnson, M. L., Gordon, H. S., Petersen, N. J., Shroyer, A. L., Grover, F. L. et Wray, N. P. (2005). Mortality After Cardiac Bypass Surgery : Prediction From Administrative Versus Clinical Data. *Medical Care*, 43(2):149–158.

[Gerber *et al.*, 2009] Gerber, Y., Benyamini, Y., Goldbourt, U. et Drory, Y. (2009). Prognostic Importance and Long-Term Determinants of Self-Rated Health After Initial Acute Myocardial Infarction :. *Medical Care - Official Journal of the Medical Care Section, American Public Health Association*, 47(3):342–349.

[Gerber *et al.*, 2011] Gerber, Y., Myers, V., Goldbourt, U., Benyamini, Y., Scheinowitz, M. et Drory, Y. (2011). Long-term trajectory of leisure time physical activity and survival after first myocardial infarction : a population-based cohort study. *European Journal of Epidemiology*, 26(2):109–116.

[Ghannem, 2010] Ghannem, M. (2010). La réadaptation cardiaque en post-infarctus du myocarde. *In Proceedings of Annales de Cardiologie et d'Angéiologie*, volume 59, pages 367–379.

[Ghannem *et al.*, 2015] Ghannem, M., Ghannem, L. et Ghannem, L. (2015). La réadaptation cardiaque en postinfarctus du myocarde. *Annales de Cardiologie et d'Angéiologie*, 64(6):517–526.

[Ghosh *et al.*, 2001] Ghosh, K., Downs, L. S., Padilla, L. A., Murray, K. P., Twiggs, L. B., Letourneau, C. M. et Carson, L. F. (2001). The Implementation of Critical Pathways in Gynecologic Oncology in a Managed Care Setting : A Cost Analysis. *Gynecologic Oncology*, 83(2):378–382.

[Gillum, 1994] Gillum, R. F. (1994). Trends in acute myocardial infarction coronary heart disease death in the united states. *Journal of the American College of Cardiology*, 23(6):1273–1277.

[Ginzburg *et al.*, 2003] Ginzburg, K., Solomon, Z., Koifman, B., Keren, G., Roth, A., Kriwisky, M., Kutz, I., David, D. et Bleich, A. (2003). Trajectories of Posttraumatic Stress Disorder Following Myocardial Infarction : A Prospective Study. *The Journal of Clinical Psychiatry*, 64(10):1217–1223.

[Goderis *et al.*, 2015] Goderis, G., Van Casteren, V., Declercq, E., Bossuyt, N., Van Den Broeke, C., Vanthomme, K., Moreels, S., Nobels, F., Mathieu, C. et Buntinx, F. (2015). Care trajectories are associated with quality improvement in the treatment of patients with uncontrolled type 2 diabetes : A registry based cohort study. *Primary Care Diabetes*, 9(5):354–361.

[Godfrey *et al.*, 2007] Godfrey, P., Shipley, R. et Gryz, J. (2007). Algorithms and analyses for maximal vector computation. *The International Journal on Very Large Data Bases*, 16(1):5–28.

[Goldberg *et al.*, 2016] Goldberg, M., Carton, M., Gourmelen, J., Genreau, M., Montourcy, M., Le Got, S. et Zins, M. (2016). L'ouverture du Système national d'information inter-régimes de l'assurance maladie (SNIIRAM) : des opportunités et des difficultés. L'expérience des cohortes Gazel et Constances. *Revue d'Épidémiologie et de Santé Publique*, 64(4):313–320.

[Goldberg *et al.*, 2008] Goldberg, M., Quantin, C., Guéguen, A. et Zins, M. (2008). Bases de données médico-administratives et épidémiologie : intérêts et limites. *Courrier des Statistiques*, (124):59–70.

[Golomb, 2015] Golomb, B. A. (2015). Misinterpretation of trial evidence on statin adverse effects may harm patients. *European Journal of Preventive Cardiology*, 22(4):492–493.

[Golomb *et al.*, 2007] Golomb, B. A., McGraw, J. J., Evans, M. A. et Dimsdale, J. E. (2007). Physician response to patient reports of adverse drug effects : implications for patient-targeted adverse effect surveillance. *Drug Safety*, 30(8):669–675.

[Goss, 2008] Goss, J. R. (2008). *Projection of Australian health care expenditure by disease, 2003 to 2033*. Health and Welfare Expenditure. Australian Institute of Health and Welfare.

[Gott *et al.*, 2007] Gott, M., Barnes, S., Parker, C., Payne, S., Seamark, D., Gariballa, S. et Small, N. (2007). Dying trajectories in heart failure. *Palliative Medicine*, 21(2):95–99.

[Gotz *et al.*, 2014] Gotz, D., Wang, F. et Perer, A. (2014). A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, 48:148–159.

[Gouda et Zaki, 2005] Gouda, K. et Zaki, M. J. (2005). GenMax : An Efficient Algorithm for Mining Maximal Frequent Itemsets. *Data Mining and Knowledge Discovery*, 11(3):223–242.

[Grammatico, 2014] Grammatico, L. (2014). *Intérêt et Limites du Programme de Médicalisation des Systèmes d'information (PMSI) dans la Surveillance des Infections de Prothèses Orthopédiques*. Thèse de doctorat, Université Pierre et Marie Curie.

[Grant *et al.*, 1996] Grant, J. B., Hayes, R. P., Pates, R. D., Elward, K. S. et Ballard, D. J. (1996). HCFA's Health Care Quality Improvement Program : The Medical Informatics Challenge. *Journal of the American Medical Informatics Association*, 3(1):15–26.

[Greene *et al.*, 2014] Greene, D., O'Callaghan, D. et Cunningham, P. (2014). How Many Topics ? Stability Analysis for Topic Models. *In Proceedings of the Machine Learning and Knowledge Discovery in Databases*, volume 8724, pages 498–513.

[Gudmundsson et van Kreveld, 2006] Gudmundsson, J. et van Kreveld, M. J. (2006). Computing longest duration flocks in trajectory data. *In Proceedings of the 14th International Symposium on Geographic Information Systems*, pages 35–42.

[Guldbrandt *et al.*, 2015] Guldbrandt, L. M., Fenger-Grøn, M., Rasmussen, T. R., Jensen, H. et Vedsted, P. (2015). The role of general practice in routes to diagnosis of lung cancer in Denmark : a population-based study of general practice involvement, diagnostic activity and diagnostic intervals. *BioMed Central Health Services Research*, 15(21):21.

[Gunes et Yaman, 2005] Gunes, E. D. et Yaman, H. (2005). Modeling change in a health system : implications on patient flows and resource allocations. *Clinical and Investigative Medicine.*, 28(6):331–333.

[Gusmano *et al.*, 2015] Gusmano, M., Rodwin, V., Weisz, D., Cottenet, J. et Quantin, C. (2015). Comparison of rehospitalization rates in France and the United States. *Journal of Health Services Research & Policy*, 20(1):18–25.

[Guyon et Elisseeff, 2003] Guyon, I. et Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(Mar): 1157–1182.

[Haesebaert *et al.*, 2013] Haesebaert, J., Termoz, A., Polazzi, S., Mouchoux, C., Mechtouff, L., Derex, L., Nighoghossian, N. et Schott, A.-M. (2013). Can hospital discharge databases be used to follow ischemic stroke incidence ? *Stroke*, 44(7):1770–1774.

[Hagiwara *et al.*, 2014] Hagiwara, M. A., Bremer, A., Claesson, A., Axelsson, C., Norberg, G. et Herlitz, J. (2014). The impact of direct admission to a catheterisation lab/CCU in patients with ST-elevation myocardial infarction on the delay to reperfusion and early risk of death : results of a systematic review including meta-analysis. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 22(1):67.

[Hai *et al.*, 2012] Hai, P. N., Poncelet, P. et Teisseire, M. (2012). Get_move : an efficient and unifying spatio-temporal pattern mining algorithm for moving objects. *In Proceedings of the International Symposium on Intelligent Data Analysis*, pages 276–288.

[Halimi et Halimi, 2013] Halimi, S. et Halimi, S. (2013). Cholestérol-statines, la polémique crée le trouble ! *Médecine des Maladies Métaboliques*, 7(5):397–398.

[Hall *et al.*, 2017] Hall, M.-H., Holton, K., Chittenden, T., Ongur, D., Eklund, K., Montrose, D. et Keshavan, M. (2017). 486 - Longitudinal Recovery Trajectories of Patients with First Episode Psychosis. *Biological Psychiatry*, 81(10, Supplement):S198.

[Hamming, 1950] Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell Labs Technical Journal*, 29(2):147–160.

[Handhayani et Hiryanto, 2015] Handhayani, T. et Hiryanto, L. (2015). Intelligent kernel k-means for clustering gene expression. *Procedia Computer Science*, 59:171–177.

[Harlos *et al.*, 2015] Harlos, C., Musto, G., Lambert, P., Ahmed, R. et Pitz, M. W. (2015). Androgen Pathway Manipulation and Survival in Patients with Lung Cancer. *Hormones and Cancer*, 6(2):120–127.

[Harper, 2005] Harper, P. (2005). Combining data mining tools with health care models for improved understanding of health processes and resource utilisation. *Clinical and Investigative Medicine*, 28(6):338–341.

[HAS, 2007] HAS (2007). Prise en charge de l'infarctus du myocarde à la phase aiguë en dehors des services de cardiologie. *Presse Médicale*, 36:1029–1037.

[HAS, 2009] HAS (2009). Évaluation des endoprothèses coronaires à libération de principe actif. Rapport technique, Haute Autorité de santé. http://www.has-sante.fr/portail/upload/docs/application/pdf/2009-11/synthese_de_levaluation_des_endoprotheses_coronaires_a_liberation_de_principe_actif.pdf.

[HAS, 2011] HAS (2011). Surpoids et obésité de l'adulte : prise en charge médicale de premier recours. *Recommandation de Bonne Pratique*, 133.

[HAS, 2012] HAS (2012). Indicateurs de pratique clinique Infarctus du myocarde (IDM) "Des 1ers signes au suivi à 1 an". Rapport technique, Haute Autorité de santé. http://www.has-sante.fr/portail/upload/docs/application/pdf/2012-07/04_indicateurs_idm_actualisation_2012_vf.pdf.

[HAS, 2016] HAS (2016). Guide du parcours de soins - maladie coronarienne stable. Rapport technique, Haute Autorité de santé. http://www.has-sante.fr/portail/upload/docs/application/pdf/2014-09/guide_mcs_web_2014-09-09_21-25-19_719.pdf.

[Heidenreich *et al.*, 2011] Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., Finkelstein, E. A., Hong, Y., Johnston, S. C., Khera, A., Lloyd-Jones, D. M., Nelson, S. A., Nichol, G., Orenstein, D., Wilson, P. W. F. et Woo, Y. J. (2011). Forecasting the Future of Cardiovascular Disease in the United States. *Circulation*, 123(8):933–944.

[Heijink *et al.*, 2008] Heijink, R., Noethen, M., Renaud, T., Koopmanschap, M. et Polder, J. (2008). Cost of illness : An international comparison. *Health Policy*, 88(1):49–61.

[Hernández *et al.*, 2004] Hernández, E. G., O'Callaghan, A. C., Doménech, J. C., Merino, V. L., Mañez, R. S., Errazti, I. E., Martín, J. V., Martínez, V. B., study research team, P. *et al.* (2004). Seasonal variations in admissions for acute myocardial infarction. the primvac study. *Revista Española de Cardiología (English Edition)*, 57(1):12–19.

[Holstein *et al.*, 2002] Holstein, J., Taright, N., Lepage, E., Razafimamonjy, J., Duboc, D., Feldman, L., Hittinger, L., Lavergne, T. et Chatellier, G. (2002). Quality of medical database to valorize the DRG model by ISA cost indicators. *Revue d'Épidémiologie et de Santé Publique*, 50(6):593–603.

[Hothorn *et al.*, 2006] Hothorn, T., Hornik, K. et Zeileis, A. (2006). Unbiased Recursive Partitioning : A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

[Howard et Bowles, 2012] Howard, J. et Bowles, M. (2012). The two most important algorithms in predictive modeling today. Presentation in the Strata Data Conference.

[Investigators *et al.*, 1988] Investigators, W. M. P. P. *et al.* (1988). The world health organization monica project (monitoring trends and determinants in cardiovascular disease) : a major international collaboration. *Journal of Clinical Epidemiology*, 41(2):105–114.

[Ishihara *et al.*, 2008] Ishihara, M., Inoue, I., Kawagoe, T., Shimatani, Y., Kurisu, S., Nakama, Y., Maruhashi, T., Kagawa, E., Dai, K., Matsushita, J. et Ikenaga, H. (2008). Trends in gender difference in mortality after acute myocardial infarction. *Journal of Cardiology*, 52(3):232–238.

[Jaccard, 1901] Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272.

[Jaro, 1989] Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420.

[Jay *et al.*, 2008] Jay, N., Kohler, F. et Napoli, A. (2008). Using Formal Concept Analysis for Mining and Interpreting Patient Flows within a Healthcare Network. *In Proceeedings of the Concept Lattices and Their Applications*, pages 263–268.

[Jay *et al.*, 2006] Jay, N., Napoli, A. et Kohler, F. (2006). Cancer patient flows discovery in DRG databases. *In Proceeedings of the Medical Informatics Europe*, pages 725–730.

[Jay *et al.*, 2013] Jay, N., Nuemi, G., Gadreau, M. et Quantin, C. (2013). A data mining approach for grouping and analyzing trajectories of care using claim data : the example of breast cancer. *BioMed Central Medical Informatics and Decision Making*, 13(1):130.

[Jayanti *et al.*, 2013] Jayanti, A., Wearden, A. J., Morris, J., Brenchley, P., Abma, I., Bayer, S., Barlow, J. et Mitra, S. (2013). Barriers to successful implementation of care in home haemodialysis (BASIC-HHD) :1. Study design, methods and rationale. *BioMed Central Nephrology*, 14:197.

[Jensen *et al.*, 2014] Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J. et Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, 5:4022.

[Jensen *et al.*, 2015] Jensen, H., Sperling, C., Sandager, M. et Vedsted, P. (2015). Agreement between patients and general practitioners on quality deviations during the cancer diagnostic pathway and associations with time to diagnosis. *Family Practice*, 32(3):329–335.

[Jeung *et al.*, 2008] Jeung, H., Yiu, M. L., Zhou, X., Jensen, C. S. et Shen, H. T. (2008). Discovery of convoys in trajectory databases. volume 1, pages 1068–1080.

[Jin *et al.*, 2008] Jin, H., Chen, J., He, H., Williams, G., Kelman, C. et O'Keefe, C. (2008). Mining Unexpected Temporal Associations : Applications in Detecting Adverse Drug Reactions. *IEEE Transactions on Information Technology in Biomedicine*, 12(4):488–500.

[Jiwa *et al.*, 2010] Jiwa, M., Maujean, E., Spilsbury, K. et Threlfal, T. (2010). The trajectory of lung cancer patients in Western Australia—A data linkage study : Still a grim tale. *Lung Cancer*, 70(1):22–27.

[Joachims, 1998] Joachims, T. (1998). Text categorization with Support Vector Machines : Learning with many relevant features. *In Proceedings of the 10th European Conference on Machine Learning*, volume 1398, pages 137–142.

[Jonnalagadda *et al.*, 2015] Jonnalagadda, S. R., Goyal, P. et Huffman, M. D. (2015). Automating data extraction in systematic reviews : a systematic review. *Systematic Reviews*, 4:78.

[Journath *et al.*, 2015] Journath, G., Hammar, N., Elofsson, S., Linnersjö, A., Vikström, M., Walldius, G., Krakau, I., Lindgren, P., de Faire, U. et Hellenius, M.-L. (2015). Time trends in incidence and mortality of acute myocardial infarction, and all-cause mortality following a cardiovascular prevention program in sweden. *PloS one*, 10(11):e0140201.

[Jun *et al.*, 1999] Jun, J. B., Jacobson, S. H. et Swisher, J. R. (1999). Application of Discrete-Event Simulation in Health Care Clinics : A Survey. *The Journal of the Operational Research Society*, 50(2):109–123.

[Juven, 2013] Juven, P.-A. (2013). Produire l'information hospitalière. *Revue d'Anthropologie des Connaissances*, 7(4):815–835.

[Kaul *et al.*, 2011] Kaul, P., McAlister, F. A., Ezekowitz, J. A., Bakal, J. A., Curtis, L. H., Quan, H., Knudtson, M. L. et Armstrong, P. W. (2011). Resource Use in the Last 6 Months of Life Among Patients With Heart Failure in Canada. *Archives of Internal Medicine*, 171(3):211–217.

[Kesavan *et al.*, 2013] Kesavan, S., Kelay, T., Collins, R. E., Cox, B., Bello, F., Kneebone, R. L. et Sevdalis, N. (2013). Clinical information transfer and data capture in the acute myocardial infarction pathway : an observational study. *Journal of Evaluation in Clinical Practice*, 19(5):805–811.

[Kim *et al.*, 2011] Kim, S., Kim, W. et Park, R. W. (2011). A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthcare Informatics Research*, 17(4):232–243.

[Kinjo *et al.*, 2005] Kinjo, K., Sato, H., Sakata, Y., Nakatani, D., Mizuno, H., Shimizu, M., Nishino, M., Matsu-ura, Y., Koretsune, Y., Nanto, S. *et al.* (2005). Identification of uncomplicated patients with acute myocardial infarction undergoing percutaneous coronary intervention are these patients suitable for early discharge ? *Circulation Journal*, 69(10):1163–1169.

[Kinsman *et al.*, 2009] Kinsman, L. D., Buykx, P., Humphreys, J. S., Snow, P. C. et Willis, J. (2009). A cluster randomised trial to assess the impact of clinical pathways on AMI management in rural Australian emergency departments. *BioMed Central Health Services Research*, 9:83.

[Kinsman *et al.*, 2012] Kinsman, L. D., Rotter, T., Willis, J., Snow, P. C., Buykx, P. et Humphreys, J. S. (2012). Do clinical pathways enhance access to evidence-based acute myocardial infarction treatment in rural emergency departments ? *The Australian Journal of Rural Health*, 20(2):59–66.

[Kirtava *et al.*, 2013] Kirtava, Z., Gegenava, T. et Gegenava, M. (2013). mHealth for cardiac patients telemonitoring and integrated care. *In In Proceedings of the IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom*, pages 21–25.

[Kivimaki *et al.*, 2012] Kivimaki, M., Nyberg, S., Batty, G., Fransson, E., Heikkila, K., Alfredsson, L., Bjorner, J., Borritz, M., Steptoe, A. et Theorell, T. (2012). Job strain as a risk factor for future coronary heart disease : collaborative meta-analysis of 2358 events in 197,473 men and women. *The Lancet*, 380:1491–1497.

[Klinkhammer-Schalke *et al.*, 2015] Klinkhammer-Schalke, M., Lindberg, P., Koller, M., Wyatt, J. C., Hofstädter, F., Lorenz, W. et Steinger, B. (2015). Direct improvement of quality of life in colorectal cancer patients using a tailored pathway with quality of life diagnosis and therapy (DIQOL) : study protocol for a randomised controlled trial. *Trials*, 16:460.

[Kohl *et al.*, 2012] Kohl, H. W., Craig, C. L., Lambert, E. V., Inoue, S., Alkandari, J. R., Leetongin, G., Kahlmeier, S. et Lancet Physical Activity Series Working Group & others (2012). The pandemic of physical inactivity : global action for public health. *The Lancet*, 380(9838):294–305.

[Korb et Nicholson, 2010] Korb, K. B. et Nicholson, A. E. (2010). *Bayesian artificial intelligence*. Florida : Chapman & Hall/CRC.

[Korn, 2000] Korn, D. (2000). Conflicts of interest in biomedical research. *Journal of the American Medical Association*, 284(17):2234–2237.

[Kristoffersen *et al.*, 2015] Kristoffersen, D. T., Helgeland, J., Waage, H. P., Thalamus, J., Clemens, D., Lindman, A. S., Rygh, L. H. et Tjomsland, O. (2015). Survival curves to support quality improvement in hospitals with excess 30-day mortality after acute myocardial infarction, cerebral stroke and hip fracture : a before–after study. *British Medical Journal Open*, 5(3):e006741.

[Krummenauer *et al.*, 2011] Krummenauer, F., Guenther, K.-P. et Kirschner, S. (2011). Cost effectiveness of total knee arthroplasty from a health care providers' perspective before and after introduction of an interdisciplinary clinical pathway - is investment always improvement ? *BioMed Central Health Services Research*, 11:338.

[Kucenic *et al.*, 2000] Kucenic, M. J., Meyers, D. G. et Meyers, D. G. (2000). Impact of a Clinical Pathway on the Care and Costs of Myocardial Infarction. *Angiology*, 51(5):393–404.

[Kuntz *et al.*, 1996] Kuntz, K. M., Tsevat, J., Goldman, L. et Weinstein, M. C. (1996). Cost-effectiveness of Routine Coronary Angiography After Acute Myocardial Infarction. *Circulation*, 94(5):957–965.

[Kurzweil, 2005] Kurzweil, R. (2005). *The singularity is near : When humans transcend biology*. Penguin.

[Kwakkel et Pruyt, 2013] Kwakkel, J. H. et Pruyt, E. (2013). Exploratory Modeling and Analysis, an approach for model-based foresight under deep uncertainty. *Technological Forecasting and Social Change*, 80(3):419–431.

[Laaidi *et al.*, 2002] Laaidi, M., Laaidi, K. et Besancenot, J.-P. (2002). Synergie entre pollens et polluants chimiques de l'air : les risques croisés. *Environnement, Risques & Santé*, 1(1):42–9.

[Lainay *et al.*, 2015] Lainay, C., Benzenine, E., Durier, J., Daubail, B., Giroud, M., Quantin, C. et Béjot, Y. (2015). Hospitalization Within the First Year After Stroke. *Stroke*, 46(1):190–196.

[Laissy *et al.*, 2004] Laissy, J. P., Sablayrolles, J. L., Sénéchal, Q., Deux, J. F., Sebban, V. et Serfaty, J. M. (2004). Complications de l'infarctus du myocarde. *Journal de Radiologie*, 85(10):1687–1693.

[Landais *et al.*, 2013] Landais, P., Duclos, C. et Le Bihan, C. (2013). L'aide à la décision médico-économique. *In Informatique médicale, e-Santé*, pages 199–236. Springer.

[Landais *et al.*, 1998] Landais, P., Stengel, B., Fumeron, C. et Jacquelinet, C. (1998). L'insuffisance rénale terminale traitée en France : épidémiologie et système d'information. *Médecine Thérapeutique*, 4(7):533–42.

[Laskowski *et al.*, 2009] Laskowski, M., McLeod, R. D., Friesen, M. R., Podaima, B. W. et Alfa, A. S. (2009). Models of emergency departments for reducing patient waiting times. *PloS one*, 4(7):e6127.

[Laut et Foldspang, 2012] Laut, K. G. et Foldspang, A. (2012). The effects on length of stay of introducing a fast track patient pathway for myocardial infarction : a before and after evaluation. *Health Services Management Research*, 25(1):31–34.

[Le Bihan-Benjamin, 2011] Le Bihan-Benjamin, C. (2011). Données PMSI chaînées : attention ! Un patient peut en cacher un autre ! Presentation in the Évaluation, Management, Organisation, Information, Santé.

[Le Bihan-Benjamin *et al.*, 2013] Le Bihan-Benjamin, C., Landais, P. et Chatellier, G. (2013). L'amélioration du chaînage des séjours dans la base nationale du PMSI MCO entre 2006 et 2009 : analyse et conséquences. *Journal d'Économie Médicale*, 30(1):17–30.

[Le Duff *et al.*, 2004] Le Duff, F., Muntean, C., Cuggia, M. et Mabo, P. (2004). Predicting survival causes after out of hospital cardiac arrest using data mining method. *Studies in Health Technology and Informatics*, 107(Pt 2):1256–1259.

[Le Feuvre, 2009] Le Feuvre, C. (2009). Maladie coronaire chez les patients diabétiques. *La Presse Médicale*, 38(6):964–972.

[Le Manach *et al.*, 2015] Le Manach, Y., Collins, G., Bhandari, M., Bessissow, A., Boddaert, J., Khiami, F., Chaudhry, H., De Beer, J., Riou, B., Landais, P. *et al.* (2015). Outcomes after hip fracture surgery compared with elective total hip replacement. *Journal of the American Medical Association*, 314(11):1159–1166.

[Le Quellec-Nathan, 2002] Le Quellec-Nathan, M. (2002). Prévenir les maladies cardio-vasculaires. *Actual Dossier Sante Publique*, 41:6–9.

[Lebart *et al.*, 1998] Lebart, L., Salem, A. et Berry, L. (1998). *Exploring Textual Data*, volume 4 de *Text, Speech and Language Technology*. Springer Science & Business Media.

[Lefebvre *et al.*, 2013] Lefebvre, C., Glanville, J., Wieland, L. S., Coles, B. et Weightman, A. L. (2013). Methodological developments in searching for studies for systematic reviews : past, present and future ? *Systematic Reviews*, 2:78.

[Leitner et Valencia, 2008] Leitner, F. et Valencia, A. (2008). A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Letters*, 582(8):1178–1181.

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

[Lewis *et al.*, 2014] Lewis, E. F., Li, Y., Pfeffer, M. A., Solomon, S. D., Weinfurt, K. P., Velazquez, E. J., Califf, R. M., Rouleau, J.-L., Kober, L., White, H. D., Schulman, K. A. et Reed, S. D. (2014). Impact of Cardiovascular Events on Change in Quality of Life and Utilities in Patients After Myocardial Infarction. a VALIANT Study (Valsartan In Acute Myocardial Infarction). *Journal of the American College of Cardiology : Heart Failure*, 2(2):159–165.

[Li *et al.*, 2010a] Li, Z., Ding, B., Han, J. et Kays, R. (2010a). Swarm : Mining relaxed temporal moving object clusters. *In Proceedings of the Very Large Database Endowment*, volume 3, pages 723–734.

[Li *et al.*, 2010b] Li, Z., Ji, M., Lee, J.-G., Tang, L.-A., Yu, Y., Han, J. et Kays, R. (2010b). Movemine : mining moving object databases. *In Proceedings of the Association Computing Machinery of International Conference on Management of data*, pages 1203–1206.

[Lin *et al.*, 2008] Lin, J. M., Bohland, J. W., Andrews, P., Burns, G. A. P. C., Allen, C. B. et Mitra, P. P. (2008). An Analysis of the Abstracts Presented at the Annual Meetings of the Society for Neuroscience from 2001 to 2006. *Plos one*, 3(4):e2052.

[Liozon *et al.*, 1992] Liozon, F., Vidal, E., Gaches, F., Venot, J., Liozon, E., Cransac, M., Loustaud, V. et Berdah, J. F. (1992). Les décès dans la maladie de Horton, Facteurs de pronostic. *La Revue de Médecine Interne*, 13(3):187–191.

[Lloyd-Jones *et al.*, 2010] Lloyd-Jones, D., Adams, R. J., Brown, T. M., Carnethon, M., Dai, S., De Simone, G., Ferguson, T. B., Ford, E., Furie, K., Gillespie, C. *et al.* (2010). Heart disease and stroke statistics—2010 update a report from the american heart association. *Circulation*, 121(7):e46–e215.

[Lomas *et al.*, 2016] Lomas, J. R., Asaria, M., Bojke, L., Richardson, G. et Walker, S. (2016). Which costs matter ? Costs included in economic evaluation and their impact on decision uncertainty : the example of acute myocardial infarction. *Value in Health*, 19(7):A363.

[Lombrail *et al.*, 1994] Lombrail, P., Minvielle, E., Comar, L. et Gottot, S. (1994). Programme de médicalisation des systèmes d'information et épidémiologie : une liaison qui ne va pas de soi. *Revue d'Épidémiologie et de Santé Publique*, 42(4):334–344.

[Lorgis *et al.*, 2013] Lorgis, L., Cottenet, J., Molins, G., Benzenine, E., Zeller, M., Aube, H., Touzery, C., Hamblin, J., Gudjoncik, A., Cottin, Y. et Quantin, C. (2013). Outcomes after acute myocardial infarction in HIV-infected patients : analysis of data from a French nationwide hospital medical information database. *Circulation*, 127(17):1767–1774.

[Loughnan *et al.*, 2008] Loughnan, M. E., Nicholls, N. et Tapper, N. J. (2008). Demographic, seasonal, and spatial differences in acute myocardial infarction admissions to hospital in melbourne australia. *International Journal of Health Geographics*, 7(1):1.

[Mao *et al.*, 2016] Mao, F., Ji, M. et Liu, T. (2016). Mining spatiotemporal patterns of urban dwellers from taxi trajectory data. *Frontiers of Earth Science*, 10(2):205–221.

[Mark *et al.*, 1995] Mark, D. B., Hlatky, M. A., Califf, R. M., Naylor, C. D., Lee, K. L., Armstrong, P. W., Barbash, G., White, H., Simoons, M. L., Nelson, C. L., Clapp-Channing, N., Knight, J. D., Harrell, F. E. J., Simes, J. et Topol, E. J. (1995). Cost Effectiveness of Thrombolytic Therapy with Tissue Plasminogen Activator as Compared with Streptokinase for Acute Myocardial Infarction. *New England Journal of Medicine*, 332(21):1418–1424.

[Mark et Newby, 2003] Mark, D. B. et Newby, L. K. (2003). Early hospital discharge after uncomplicated myocardial infarction : are further improvements possible ? *European Heart Journal*, 24(18):1613–1615.

[Marquardt, 1963] Marquardt, D. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441.

[Marrugat *et al.*, 2004] Marrugat, J., Elosua, R., Aldasoro, E., Tormo, M. J., Vanaclocha, H., Segura, A., Fiol, M., Moreno-Iribas, C., Perez, G., Arteagoitia, J. M. *et al.* (2004). Regional variability in population acute myocardial infarction cumulative incidence and mortality rates in spain 1997 and 1998. *European Journal of Epidemiology*, 19(9):831–839.

[Marrugat *et al.*, 1998] Marrugat, J., Sala, J., Masiá, R., Pavesi, M., Sanz, G., Valle, V., Molina, L., Serés, L., Elosua, R., Investigators, R. *et al.* (1998). Mortality differences between men and women following first myocardial infarction. *Journal of the American Medical Association*, 280(16):1405–1409.

[Martelli *et al.*, 2017] Martelli, N., Puc, C., Szwarcensztein, K., Beuscart, R., Coulonjou, H., Degrassat-Théas, A., Dutot, C., de Fleurian, A.-A. E., Favrel-Feuillade, F., Hounliasso, I. *et al.* (2017). Rôle et place de l'évaluation des technologies de santé à l'hôpital : schéma cible appliqué aux dispositifs médicaux. *Thérapie*, 72(1):105–113.

[Martens *et al.*, 2008] Martens, E. J., Smith, O. R. F., Winter, J., Denollet, J. et Pedersen, S. S. (2008). Cardiac history, prior depression and personality predict course of depressive symptoms after myocardial infarction. *Psychological Medicine*, 38(2):257–264.

[Martinez-Maldonado *et al.*, 2013] Martinez-Maldonado, R., Kay, J. et Yacef, K. (2013). An Automatic Approach for Mining Patterns of Collaboration around an Interactive Tabletop. *In Proceedings of the Artificial Intelligence in Education*, pages 101–110.

[Mas *et al.*, 2006] Mas, J.-L., Chatellier, G., Beyssen, B., Branchereau, A., Moulin, T., Becquemin, J.-P., Larrue, V., Lièvre, M., Leys, D., Bonneville, J.-F. et others (2006). Endarterectomy versus stenting in patients with symptomatic severe carotid stenosis. *New England Journal of Medicine*, 355(16):1660–1671.

[Massop *et al.*, 2009] Massop, D., Dave, R., Metzger, C., Bachinsky, W., Solis, M., Shah, R., Schultz, G., Schreiber, T., Ashchi, M. et Hibbard, R. (2009). Stenting and Angioplasty with Protection in Patients at High-Risk for Endarterectomy : SAPPHIRE Worldwide Registry First 2,001 Patients. *Catheterization and Cardiovascular Interventions*, 73(2):129–136.

[Mastenbroek *et al.*, 2015] Mastenbroek, M. H., Denollet, J., Versteeg, H., van den Broek, K. C., Theuns, D. A. M. J., Meine, M., Zijlstra, W. P. et Pedersen, S. S. (2015). Trajectories of Patient-Reported Health Status in Patients With an Implantable Cardioverter Defibrillator. *The American Journal of Cardiology*, 115(6):771–777.

[Maulik et Bandyopadhyay, 2002] Maulik, U. et Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654.

[Mazzini *et al.*, 2008] Mazzini, M. J., Stevens, G. R., Whalen, D., Ozonoff, A. et Balady, G. J. (2008). Effect of an American Heart Association Get With the Guidelines Program-Based Clinical Pathway on Referral and Enrollment Into Cardiac Rehabilitation After Acute Myocardial Infarction. *The American Journal of Cardiology*, 101(8):1084–1087.

[McNamara *et al.*, 2016] McNamara, R. L., Kennedy, K. F., Cohen, D. J., Diercks, D. B., Moscucci, M., Ramee, S., Wang, T. Y., Connolly, T. et Spertus, J. A. (2016). Predicting In-Hospital Mortality in Patients With Acute Myocardial Infarction. *Journal of the American College of Cardiology*, 68(6):626–635.

[McNicholas et Murphy, 2010] McNicholas, P. D. et Murphy, T. B. (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, 38(1):153–168.

[Mei *et al.*, 2006] Mei, Q., Xin, D., Cheng, H., Han, J. et Zhai, C. (2006). Generating semantic annotations for frequent patterns with context analysis. *In Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining*, pages 337–346.

[Melnychuk *et al.*, 2010] Melnychuk, M. C., Welch, D. W. et Walters, C. J. (2010). Spatio-Temporal Migration Patterns of Pacific Salmon Smolts in Rivers and Coastal Marine Waters. *Plos one*, 5(9):e12916.

[Ménard, 2002] Ménard, B. (2002). Questions de géographie de la santé. *L'Espace Géographique*, 31(3):264–275.

[Mercadier *et al.*, 2016] Mercadier, Y., Pinaire, J., Azé, J., Bringay, S. et Teisseire, M. (2016). La r-confiance pour l'identification de trajectoires de patients. *In Proceedings of 16ème Journées Francophones Extraction et Gestion des Connaissances*, pages 535–536.

[Meyer *et al.*, 2015] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. et Leisch, F. (2015). *e1071 : Misc Functions of the Department of Statistics, Probability Theory Group (Formerly : E1071), TU Wien*. https://CRAN.R-project.org/package=e1071.

[Michel *et al.*, 2008] Michel, E., Bocquier, A. et Verger, P. (2008). La fiabilité des causes médicales de décès en Provence-Alpes-Côte d'Azur. *Santé Publique*, 20(1):29–38.

[Milcent, 2015] Milcent, C. (2015). Tarification et variabilité des coûts hospitaliers : Le cas de l'infarctus du myocarde. Rapport technique 28, École d'Économie de Paris. https://halshs.archives-ouvertes.fr/halshs-01202684.

[Milcent, 2017] Milcent, C. (2017). Premier bilan de la tarification à l'activité (t2a) sur la variabilité des coûts hospitaliers. *Economie & Prévision*, (1):45–67.

[Miller *et al.*, 2017] Miller, A. L., Simon, D., Roe, M. T., Kontos, M. C., Diercks, D., Amsterdam, E. et Bhatt, D. L. (2017). Comparison of Delay Times from Symptom Onset to Medical Contact in Blacks Versus Whites With Acute Myocardial Infarction. *The American Journal of Cardiology*, 119(8):1127–1134.

[Miller *et al.*, 2002] Miller, P. R., Fabian, T. C., Croce, M. A., Magnotti, L. J., Elizabeth Pritchard, F., Minard, G. et Stewart, R. M. (2002). Improving Outcomes Following Penetrating Colon Wounds. *Annals of Surgery*, 235(6):775–781.

[Minvielle, 2000] Minvielle, E. (2000). Réconcilier standardisation et singularité : les enjeux de l'organisation de la prise en charge des malades. *Ruptures, Revue transdisciplinaire en Santé*, 7(1):8–22.

[Mo *et al.*, 2015] Mo, Y., Kontonatsios, G. et Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. *Systematic Reviews*, 4:172.

[Moher, 2009] Moher, D. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses : The PRISMA Statement. *Annals of Internal Medicine*, 151(4):264–269.

[Moons *et al.*, 2012a] Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G. et Woodward, M. (2012a). Risk prediction models : II. External validation, model updating, and impact assessment. *Heart*, 98(9):691–698.

[Moons *et al.*, 2012b] Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G. et Grobbee, D. E. (2012b). Risk prediction models : I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*, 98(9):683–690.

[Mottier et Baba-Ahmed, 2006] Mottier, D. et Baba-Ahmed, M. (2006). Anticoagulants et gestes invasifs. *Médecine Thérapeutique*, 12(1):48–52.

[Moynihan, 2003] Moynihan, R. (2003). Who pays for the pizza ? Redefining the relationships between doctors and drug companies. 1 : Entanglement - ProQuest. *British Medical Journal*, 326(7400):1189.

[Mueller, 2008] Mueller, E. T. (2008). Event calculus. *Foundations of Artificial Intelligence*, 3:671–708.

[Myers *et al.*, 2011] Myers, V., Drory, Y. et Gerber, Y. (2011). Sense of coherence predicts post-myocardial infarction trajectory of leisure time physical activity : a prospective cohort study. *BioMed Central Public Health*, 11(1):708.

[Myers *et al.*, 2014] Myers, V., Drory, Y., Gerber, Y. et Israel Study Group on First Acute Myocardial Infarction (2014). Clinical relevance of frailty trajectory post myocardial infarction. *European Journal of Preventive Cardiology*, 21(6):758–766.

[Myklebust *et al.*, 2010] Myklebust, L. H., Sørgaard, K. W., Bjorbekkmo, S., Eisemann, M. R. et Olstad, R. (2010). Time-trends in the utilization of decentralized mental health services in Norway - A natural experiment : The VELO-project. *International Journal of Mental Health Systems*, 4(1):5.

[Nakache *et al.*, 1977] Nakache, J.-P., Lorente, P., Benzécri, J.-P. et Chastang, J.-F. (1977). Aspects pronostiques et thérapeutiques de l'infarctus myocardique aigu compliqué d'une défaillance sévère de la pompe cardiaque. Application des méthodes de discrimination. *Les Cahiers de l'Analyse des Données*, 2(4):415–434.

[Naqvi *et al.*, 2009] Naqvi, H. A., Hussain, S., Zaman, M. et Islam, M. (2009). Pathways to Care : Duration of Untreated Psychosis from Karachi, Pakistan. *Plos one*, 4(10):e7409.

[Navarro *et al.*, 2001] Navarro, G., Baeza-Yates, R. A., Sutinen, E. et Tarhio, J. (2001). Indexing methods for approximate string matching. *IEEE Data Enginering Bulletin*, 24(4):19–27.

[Needleman et Wunsch, 1970] Needleman, S. B. et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

[Neff, 2004] Neff, M. J. (2004). Practice Guidelines : ACC/AHA Release Guidelines on Management of Patients with STEMI : Hospital and Long-Term Management. *American Family Physician*, 70(10):2011–2021.

[Newton *et al.*, 2015] Newton, P. K., Mason, J., Venkatappa, N., Jochelson, M. S., Hurt, B., Nieva, J., Comen, E., Norton, L. et Kuhn, P. (2015). Spatiotemporal progression of metastatic breast cancer : a Markov chain model highlighting the role of early metastatic sites. *Nature Partner Journals Breast Cancer*, 1:15018.

[Nguyen *et al.*, 2017] Nguyen, T.-L., Boudemaghe, T., Leguelinel-Blache, G., Eiden, C., Kinowski, J.-M., Le Manach, Y., Peyrière, H. et Landais, P. (2017). Identifying life-threatening admissions for drug dependence or abuse (iliadda) : Derivation and validation of a model. *Scientific Reports*, 7:44428.

[Nikfarjam et Gonzalez, 2011] Nikfarjam, A. et Gonzalez, G. H. (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. *In Proceedings of the American Medical Informatics Association Annual Symposium*, pages 1019–1026.

[Noble *et al.*, 2015] Noble, S. I., Nelson, A., Fitzmaurice, D., Bekkers, M.-J., Baillie, J., Sivell, S., Canham, J., Smith, J. D., Casbard, A., Cohen, A., Cohen, D., Evans, J., Fletcher, K., Johnson, M., Maraveyas, A., Prout, H. et Hood, K. (2015). A feasibility study to inform the design of a randomised controlled trial to identify the most clinically effective and cost-effective length of Anticoagulation with Low-molecular-weight heparin In the treatment of Cancer-Associated Thrombosis (ALICAT). *Health Technology Assessment*, 19(83):vii–xxiii.

[Nohria *et al.*, 1998] Nohria, A., Vaccarino, V. et Krumholz, H. M. (1998). Gender Differences in Mortality after Myocardial Infarction : Why Women Fare Worse Than Men. *Cardiology Clinics*, 16(1):45–57.

[Norén *et al.*, 2008] Norén, G. N., Bate, A., Hopstadius, J., Star, K. et Edwards, I. R. (2008). Temporal Pattern Discovery for Trends and Transient Effects : Its Application to Patient Records. *In Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining*, pages 963–971.

[Nubukpo, 2014] Nubukpo, P. (2014). Système opioïde endogène et stratégies thérapeutiques dans la dépendance à l'alcool. *L'Encéphale*, 40(6):457–467.

[O'Donnell *et al.*, 2006] O'Donnell, S., Condell, S., Begley, C. et Fitzgerald, T. (2006). Prehospital care pathway delays : gender and myocardial infarction. *Journal of Advanced Nursing*, 53(3):268–276.

[Oliver, 2014] Oliver, D. (2014). Readmission rates reflect how well whole health and social care systems function. *British Medical Journal*, 348:g1150.

[Ornato *et al.*, 1996] Ornato, J. P., Peberdy, M. A., Chandra, N. C. et Bush, D. E. (1996). Seasonal pattern of acute myocardial infarction in the national registry of myocardial infarction. *Journal of the American College of Cardiology*, 28(7):1684–1688.

[O'Mara-Eves *et al.*, 2015] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. et Ananiadou, S. (2015). Using text mining for study identification in systematic reviews : a systematic review of current approaches. *Systematic Reviews*, 4(1):5.

[Pagnoni *et al.*, 2001] Pagnoni, A., Parisi, S. et Lombardo, S. (2001). Analysis of patient flows via data mining. *Studies in Health Technology and Informatics*, (2):1379–1383.

[Pal et Biswas, 1997] Pal, N. R. et Biswas, J. (1997). Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6):847–857.

[Palmer *et al.*, 2013a] Palmer, J., Bozas, G., Stephens, A., Johnson, M., Avery, G., O'Toole, L. et Maraveyas, A. (2013a). Developing a complex intervention for the outpatient management of incidentally diagnosed pulmonary embolism in cancer patients. *BioMed Central Health Services Research*, 13:235.

[Palmer *et al.*, 2013b] Palmer, W. L., Bottle, A., Davie, C., Vincent, C. A. et Aylin, P. (2013b). Meeting the ambition of measuring the quality of hospitals' stroke care using routinely collected administrative data : a feasibility study. *International Journal for Quality in Health Care*, 25(4):429–436.

[Parikh *et al.*, 2009] Parikh, N. I., Gona, P., Larson, M. G., Fox, C. S., Benjamin, E. J., Murabito, J. M., O'Donnell, C. J., Vasan, R. S. et Levy, D. (2009). Long-term trends in myocardial infarction incidence and case fatality in the national heart, lung, and blood institute's framingham heart study. *Circulation*, 119(9):1203–1210.

[Park *et al.*, 2015] Park, J. Y., Lee, S.-H., Shin, M.-J. et Hwang, G.-S. (2015). Alteration in Metabolic Signature and Lipid Metabolism in Patients with Angina Pectoris and Myocardial Infarction. *Plos one*, 10(8):e0135228.

[Pascal, 2009] Pascal, L. (2009). Effets à court terme de la pollution atmosphérique sur la mortalité. *Revue Française d'Allergologie*, 49(6):466–476.

[Pasternack *et al.*, 1985] Pasternack, P. F., Colvin, S. B. et Baumann, F. G. (1985). Cocaine-induced angina pectoris and acute myocardial infarction in patients younger than 40 years. *The American Journal of Cardiology*, 55(6):847.

[Patrick *et al.*, 2011a] Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E. et Ring, L. (2011a). Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (pro) instruments for medical product evaluation : Ispor pro good research practices task force report : part 1—eliciting concepts for a new pro instrument. *Value in Health*, 14(8):967–977.

[Patrick *et al.*, 2011b] Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E. et Ring, L. (2011b). Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (pro) instruments for medical product evaluation : Ispor pro good research practices task force report : part 2—assessing respondent understanding. *Value in Health*, 14(8):978–988.

[Paynter *et al.*, 2016] Paynter, R., Bañez, L. L., Berliner, E., Erinoff, E., Lege-Matsuura, J., Potter, S. et Uhl, S. (2016). *EPC Methods : An Exploration of the Use of Text-Mining Software in Systematic Reviews*. Agency for Healthcare Research and Quality (US).

[Pedersen *et al.*, 2015] Pedersen, E. R., Tuseth, N., Eussen, S. J. P. M., Ueland, P. M., Strand, E., Svingen, G. F. T., Midttun, ø., Meyer, K., Mellgren, G., Ulvik, A., Nordrehaug, J. E., Nilsen, D. W. et Nygård, O. (2015). Associations of Plasma Kynurenines With Risk of Acute Myocardial Infarction in Patients With Stable Angina PectorisSignificance. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 35(2):455–462.

[Pei *et al.*, 2004] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U. et Hsu, M.-C. (2004). Mining sequential patterns by pattern-growth : The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440.

[Pelliccia *et al.*, 2004] Pelliccia, F., Cartoni, D., Verde, M., Salvini, P., Mercuro, G. et Tanzi, P. (2004). Critical pathways in the emergency department improve treatment modalities for patients with ST-elevation myocardial infarction in a European hospital. *Clinical Cardiology*, 27(12):698–700.

[Perera *et al.*, 2009] Perera, D., Kay, J., Koprinska, I., Yacef, K. et Zaïane, O. R. (2009). Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6): 759–772.

[Perlbarg *et al.*, 2014] Perlbarg, J., Allonier, C., Boisnault, P., Daniel, F., Le Fur, P., Szidon, P. et Bourgueil, Y. (2014). Faisabilité et intérêt de l'appariement de données individuelles en médecine générale et de données de remboursement appliqué au diabète et à l'hypertension artérielle. *Santé Publique*, 26(3):355–363.

[Perrier *et al.*, 2003] Perrier, L., Borella, L. et Philip, T. (2003). Le coût du cancer en france : approches macro et micro-économiques, évolution vers la tarification à l 'activité. *Bulletin du Cancer*, 90(11):1005–1009.

[Peters *et al.*, 2011] Peters, B. J. M., Pett, H., Klungel, O. H., Stricker, B. H. C., Psaty, B. M., Glazer, N. L., Wiggins, K. L., Bis, J. C., de Boer, A. et Maitland-van der Zee, A.-H. (2011). Genetic variability within the cholesterol lowering pathway and the effectiveness of statins in reducing the risk of MI. *Atherosclerosis*, 217(2):458–464.

[Phan *et al.*, 2016] Phan, N., Poncelet, P. et Teisseire, M. (2016). All in one : mining multiple movement patterns. *International Journal of Information Technology & Decision Making*, 15(05):1115–1156.

[Philippe *et al.*, 2017] Philippe, F., Blin, P., Bouée, S., Laurendeau, C., Torreton, E., Gourmelin, J., Velkovski-Rouyer, M., Levy-Bachelot, L. et Steg, G. (2017). Costs of healthcare resource consumption after a myocardial infarction in France : An estimate from a medicoadministrative database (GSB). *Annales de Cardiologie et d'Angéiologie*, 66(2):74–80.

[Pingault *et al.*, 2013] Pingault, J.-B., Côté, S. M., Galéra, C., Genolini, C., Falissard, B., Vitaro, F. et Tremblay, R. E. (2013). Childhood trajectories of inattention, hyperactivity and oppositional behaviors and prediction of substance abuse/dependence : a 15-year longitudinal population-based study. *Molecular Psychiatry*, 18(7):806–812.

[Popp *et al.*, 2002] Popp, A. J., Scrime, T., Cohen, B. R., Feustel, P. J., Petronis, K., Habiniak, S., Waldman, J. B. et Vosburgh, M. M. (2002). Factors affecting profitability for craniotomy. *Journal of Neurosurgery*, 12(4):1–5.

[Potisek *et al.*, 2007] Potisek, N. M., Malone, R. M., Shilliday, B. B., Ives, T. J., Chelminski, P. R., DeWalt, D. A. et Pignone, M. P. (2007). Use of patient flow analysis to improve patient visit efficiency by decreasing wait time in a primary care-based disease management programs for anticoagulation and chronic pain : a quality improvement study. *BioMed Central Health Services Research*, 7:8.

[Preux *et al.*, 2005] Preux, P. M., Odermatt, P., Perna, A., Marin, B. et Vergnenègre, A. (2005). Qu'est-ce qu'une régression logistique ? *Revue des Maladies Respiratoires*, 22(1):159–162.

[Quan *et al.*, 2005] Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E. et Ghali, W. A. (2005). Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data. *Medical Care*, 43(11):1130–1139.

[Quantin et Cnamts, 2015] Quantin, C. et Cnamts (2015). Méthodologie de repérage des pathologies et de répartition des dépenses par pathologie. Rapport technique, Caisse Nationale d'Assurance Maladie des Travailleurs Salariés. https://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/etudes-en-sante-publique/cartographie-des-pathologies-et-des-depenses/methodologie.php.

[Quantin *et al.*, 2005] Quantin, C., Gouyol, B., Allaert, F.-A. et Cohen, O. (2005). Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales. *Journal de la Société Française de Statistique*, 146(3):19–38.

[Quidu et Escaffre, 2017] Quidu, F. et Escaffre, J.-P. (2017). Objectif : bouleversements des pouvoirs. technique : un brouhaha institutionnel permanent. *L'information Psychiatrique*, 93(1):13–19.

[R Core Team, 2016] R Core Team (2016). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[Rabatel, 2011] Rabatel, J. (2011). *Extraction de motifs contextuels : Enjeux et applications dans les données séquentielles*. Thèse de doctorat, Université Montpellier II.

[Rabatel *et al.*, 2010a] Rabatel, J., Bringay, S. et Poncelet, P. (2010a). Aide à la décision pour la maintenance ferroviaire préventive. *In Proceedings of Extraction et Gestion des Connaissances*, pages 363–368.

[Rabatel *et al.*, 2010b] Rabatel, J., Bringay, S. et Poncelet, P. (2010b). Contextual sequential pattern mining. *In Proceedings of the International Conference on Data Mining Workshops*, pages 981–988.

[Rajalakshmi et Dhenakaran, 2015] Rajalakshmi, K. et Dhenakaran, S. S. (2015). Analysis of Datamining Prediction Techniques in Healthcare Management System. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(4):1343–1347.

[Rankin *et al.*, 2002] Rankin, S. H., de Leon, J. F., Chen, J.-L., Butzlaff, A. et Carroll, D. L. (2002). Recovery Trajectory of Unpartnered Elders After Myocardial Infarction : An Analysis of Daily Diaries. *Rehabilitation Nursing*, 27(3):95–102.

[Ratinaud et Déjean, 2009] Ratinaud, P. et Déjean, S. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. Presentation in the Modèles et Apprentissage en Sciences Humaines et Sociales.

[Ratinaud et Marchand, 2012] Ratinaud, P. et Marchand, P. (2012). Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux" : analyse du "CableGate" avec IRaMuTeQ. *Actes des 11eme Journées Internationales d'Analyse Statistique des Données Textuelles*, pages 835–844.

[Reinert, 1983] Reinert, A. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse des Données*, VIII(2):187–198.

[Remontet *et al.*, 2008] Remontet, L., Mitton, N., Couris, C. M., Iwaz, J., Gomez, F., Olive, F., Polazzi, S., Schott, A. M., Trombert, B., Bossard, N. et Colonna, M. (2008). Is it possible to estimate the incidence of breast cancer from medico-administrative databases ? *European Journal of Epidemiology*, 23(10):681–688.

[Rican et Vaillant, 2009] Rican, S. et Vaillant, Z. (2009). Territoires et santé : enjeux sanitaires de la territorialisation et enjeux territoriaux des politiques de santé. *Sciences Sociales et Santé*, 27(1):33–42.

[Ricciardi *et al.*, 2009] Ricciardi, A., Largeron, N., Giorgi Rossi, P., Raffaele, M., Cohet, C., Federici, A. et Palazzo, F. (2009). Incidence of invasive cervical cancer and direct costs associated with its management in Italy. *Tumori*, 95(2):146–152.

[Rohleder *et al.*, 2005] Rohleder, T. R., Sabapathy, D. et Schorn, R. (2005). An operating room block allocation model to improve hospital patient flow. *Clinical and Investigative Medicine*, 28(6):353.

[Rosenberg *et al.*, 1985] Rosenberg, L., Kaufman, D. W., Helmrich, S. P., Miller, D. R., Stolley, P. D. et Shapiro, S. (1985). Myocardial Infarction and Cigarette Smoking in Women Younger Than 50 Years of Age. *Journal of the American Medical Association*, 253(20):2965–2969.

[Rosenfeld, 2004] Rosenfeld, A. G. (2004). Treatment-Seeking Delay Among Women With Acute Myocardial :Decision Trajectories and Their Predictors. *Nursing Research*, 53(4):225–236.

[Rudin *et al.*, 2011] Rudin, C., Letham, B., Salleb-Aouissi, A., Kogan, E. et Madigan, D. (2011). Sequential event prediction with association rules. *In Proceedings of the Computational Learning Theory*, pages 615–634.

[Ruidavets *et al.*, 2010] Ruidavets, J.-B., Ducimetière, P., Evans, A., Montaye, M., Haas, B., Bingham, A., Yarnell, J., Amouyel, P., Arveiler, D., Kee, F. *et al.* (2010). Patterns of alcohol consumption and ischaemic heart disease in culturally divergent countries : the prospective epidemiological study of myocardial infarction (prime). *British Medical Journal*, 341:c6077.

[Rusmevichientong et Williamson, 2006] Rusmevichientong, P. et Williamson, D. P. (2006). An adaptive algorithm for selecting profitable keywords for search-based advertising services. *In Proceedings of the 7th Association for Computing Machinery Conference on Electronic Commerce*, pages 260–269.

[Rédaction Prescrire, 2011] Rédaction Prescrire (2011). Profils d'effets indésirables de médicaments cardiovasculaires. *Revue Prescrire*, 31(338):60–76.

[Saeys *et al.*, 2007] Saeys, Y., Inza, I. et Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

[Safon, 2015] Safon, M.-O. (2015). Historique des réformes hospitalières en France. Rapport technique, IRDES. http://www.irdes.fr/documentation/syntheses/historique-des-reformes-hospitalieres-en-france.pdf.

[Sallaberry *et al.*, 2010] Sallaberry, A., Zaidi, F., Pich, C. et Melançon, G. (2010). Interactive visualization and navigation of web search results revealing community structures and bridges. *In Proceedings of the Graphics Interface conference of Canadian Information Processing Society*, pages 105–112.

[Salle *et al.*, 2009] Salle, P., Bringay, S. et Teisseire, M. (2009). Mining Discriminant Sequential Patterns for Aging Brain. *In Proceedings of the Conference on Artificial Intelligence in Medicine in Europe*, pages 365–369.

[Sandel, 2009] Sandel, M. J. (2009). *The case against perfection*. Harvard University Press.

[Savage *et al.*, 1998] Savage, M. P., Fischman, D. L., Rake, R., Leon, M. B., Schatz, R. A., Penn, I., Nobuyoshi, M., Moses, J., Hirshfeld, J., Heuser, R., Baim, D., Cleman, M., Brinker, J., Gebhardt, S. et Goldberg, S. (1998). Efficacy of Coronary Stenting Versus Balloon Angioplasty in Small Coronary Arteries fn1. *Journal of the American College of Cardiology*, 31(2):307–311.

[Scheen, 2006] Scheen, A. (2006). L'étude clinique du mois. L'étude IDEAL comparant simvastatine 20-40 mg versus atorvastatine 80 mg en prévention après un infarctus du myocarde : entre deux idées de l'idéal. *Revue Médicale de Liège*, 61(1):53–59.

[Schmidt *et al.*, 2015] Schmidt, A., Heroum, C., Caumette, D., Le Lay, K. et Bénard, S. (2015). Acute Ischemic Stroke (AIS) patient management in French stroke units and impact estimation of thrombolysis on care pathways and associated costs. *Cerebrovascular Diseases*, 39(2):94–101.

[Schmidt *et al.*, 2012] Schmidt, M., Jacobsen, J. B., Lash, T. L., Bøtker, H. E. et Sørensen, H. T. (2012). 25 year trends in first time hospitalisation for acute myocardial infarction, subsequent short and long term mortality, and the prognostic impact of sex and comorbidity : a danish nationwide cohort study. *British Medical Journal*, 344:e356.

[Schneeweiss *et al.*, 2007] Schneeweiss, S., Patrick, A. R., Maclure, M., Dormuth, C. R. et Glynn, R. J. (2007). Adherence to Statin Therapy Under Drug Cost Sharing in Patients With and Without Acute Myocardial Infarction. *Circulation*, 115(16):2128–2135.

[Schott *et al.*, 2002] Schott, A.-M., Hajri, T., Colin, C., Grateau, F., Gilly, F. N., Tissot, É., Couchoud, C., Morestin, C. et Trillet-Lenoir, V. (2002). Intérêt de l'utilisation du pmsi pour l'analyse d'activité cancérologique d'une structure de soins multidisciplinaire : l'expérience de la coordination de cancérologie des hospices civils de lyon. *Bulletin du Cancer*, 89(11):969–73.

[Schwartz *et al.*, 2013] Schwartz, C. E., Quaranto, B. R., Healy, B. C., Benedict, R. H. et Vollmer, T. L. (2013). Cognitive Reserve and Symptom Experience in Multiple Sclerosis : A Buffer to Disability Progression Over Time ? *Archives of Physical Medicine and Rehabilitation*, 94(10):1971–1981.e1.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *Association for Computing Machinery Computing Surveys*, 34(1): 1–47.

[Seidl *et al.*, 2017] Seidl, H., Hunger, M., Meisinger, C., Kirchberger, I., Kuch, B., Leidl, R. et Holle, R. (2017). The 3-Year Cost-Effectiveness of a Nurse-Based Case Management versus Usual Care for Elderly Patients with Myocardial Infarction : Results from the KORINNA Follow-Up Study. *Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 20(3):441–450.

[Selassie *et al.*, 2011] Selassie, D., Heller, B. et Heer, J. (2011). Divided edge bundling for directional network data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2354–2363.

[Sestelo *et al.*, 2015] Sestelo, M., M. Villanueva, N. et Roca-Pardinas, J. (2015). *FWDselect : Selecting Variables in Regression Models*. https://CRAN.R-project.org/package=FWDselect.

[Shafer *et al.*, 1976] Shafer, G. *et al.* (1976). *A mathematical theory of evidence*, volume 1. Princeton university press Princeton.

[Si *et al.*, 2009] Si, Y., Skidmore, A. K., Wang, T., Boer, W. F. d., Debba, P., Toxopeus, A. G., Li, L. et Prins, H. H. T. (2009). Spatio-temporal dynamics of global H5n1 outbreaks match bird migration patterns. *Geospatial Health*, 4(1):65–78.

[Sieberg *et al.*, 2013] Sieberg, C. B., Simons, L. E., Edelstein, M. R., DeAngelis, M. R., Pielech, M., Sethna, N. et Hresko, M. T. (2013). Pain Prevalence and Trajectories Following Pediatric Spinal Fusion Surgery. *The Journal of Pain*, 14(12):1694–1702.

[Simon *et al.*, 2006] Simon, T., Mary-Krause, M., Cambou, J.-P., Hanania, G., Guéret, P., Lablanche, J.-M., Blanchard, D., Genès, N. et Danchin, N. (2006). Impact of age and gender on in-hospital and late mortality after acute myocardial infarction : increased early risk in younger womenResults from the French nation-wide USIC registries. *European Heart Journal*, 27(11):1282–1288.

[Singh *et al.*, 2014] Singh, J. A., Lu, X., Ibrahim, S. et Cram, P. (2014). Trends in and disparities for acute myocardial infarction : an analysis of medicare claims data from 1992 to 2010. *BioMed Central medicine*, 12(1):190.

[Siontis *et al.*, 2012] Siontis, G. C. M., Tzoulaki, I., Siontis, K. C. et Ioannidis, J. P. A. (2012). Comparisons of established risk prediction models for cardiovascular disease : systematic review. *British Medical Journal (Clinical research edition)*, 344:e3318.

[Siregar *et al.*, 2014] Siregar, S., Pouw, M. E., Moons, K. G. M., Versteegh, M. I. M., Bots, M. L., Graaf, Y. v. d., Kalkman, C. J., Herwerden, L. A. v. et Groenwold, R. H. H. (2014). The Dutch Hospital Standardised Mortality Ratio (HSMR) method and cardiac surgery : benchmarking in a national cohort using hospital administration data versus a clinical database. *Heart*, 100(9):702–710.

[Skinner *et al.*, 2014] Skinner, I., Smith, C. et Jaffray, L. (2014). Realist Review to Inform Development of the Electronic Advance Care Plan for the Personally Controlled Electronic Health Record in Australia. *Telemedicine and e-Health*, 20(11):1042–1048.

[Smith *et al.*, 2011] Smith, O. R. F., Kupper, N., Denollet, J. et Jonge, P. d. (2011). Vital exhaustion and cardiovascular prognosis in myocardial infarction and heart failure : predictive power of different trajectories. *Psychological Medicine*, 41(4):731–738.

[Song *et al.*, 2010] Song, L., Yan, H., Hu, D., Yang, J. et Sun, Y. (2010). Prehospital care-seeking in patients with acute myocardial infarction and subsequent quality of care in Beijing. *Chinese Medical Journal*, 123(6):664–669.

[Spencer *et al.*, 1998] Spencer, F. A., Goldberg, R. J., Becker, R. C. et Gore, J. M. (1998). Seasonal distribution of acute myocardial infarction in the second national registry of myocardial infarction 1. *Journal of the American College of Cardiology*, 31(6):1226–1233.

[Srikant et Agrawal, 1996] Srikant, R. et Agrawal, R. (1996). Mining sequential patterns : Generalizations and performance improvements. *In Proceedings of the Fifth International Conference on Extending Database Technology*, pages 3–17.

[Steyerberg *et al.*, 2001] Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E. et Habbema, J. D. (2001). Prognostic modeling with logistic regression analysis : in search of a sensible strategy in small data sets. *Medical Decision Making : An International Journal of the Society for Medical Decision Making*, 21(1):45–56.

[Stirling, 2010] Stirling, A. (2010). Keep it complex. *Nature*, 468(7327):1029–1031.

[Stovel et Bolan, 2004] Stovel, K. et Bolan, M. (2004). Residential trajectories : Using optimal alignment to reveal the structure of residential mobility. *Sociological Methods & Research*, 32(4):559–598.

[Strobl, 2005] Strobl, C. (2005). Statistical Sources of Variable Selection Bias in Classification Tree Algorithms Based on the Gini Index. Rapport technique Discussion Paper 420, Collaborative Research Center 386. https://epub.ub.uni-muenchen.de/1789/1/paper_420.pdf.

[Strobl *et al.*, 2009] Strobl, C., Malley, J. et Tutz, G. (2009). An introduction to recursive partitioning : Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4):323–348.

[Strömberg *et al.*, 2014] Strömberg, A., Fluur, C., Miller, J., Chung, M. L., Moser, D. K. et Thylén, I. (2014). ICD Recipients' Understanding of Ethical Issues, ICD Function, and Practical Consequences of Withdrawing the ICD in the End-of-Life. *Pacing and Clinical Electrophysiology*, 37(7):834–842.

[Stumme *et al.*, 2002] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N. et Lakhal, L. (2002). Computing iceberg concept lattices with titanic. *Data & Knowledge Engineering*, 42(2):189–222.

[Sulo *et al.*, 2016] Sulo, G., Igland, J., Vollset, S. E., Nygård, O., Ebbing, M., Sulo, E., Egeland, G. M. et Tell, G. S. (2016). Heart failure complicating acute myocardial infarction - burden and timing of occurence : A nation-wide analysis including 86771 patients from the cardiovascular disease Norway (CVDNOR) project. *Journal of the American Heart Association*, 5(1):e002667.

[Sundberg *et al.*, 2014] Sundberg, T., Petzold, M., Kohls, N. et Falkenberg, T. (2014). Opposite Drug Prescription and Cost Trajectories following Integrative and Conventional Care for Pain – A Case-Control Study. *Plos one*, 9(5):e96717.

[Suresh *et al.*, 2014] Suresh, R., Li, X., Chiriac, A., Goel, K., Terzic, A., Perez-Terzic, C. et Nelson, T. J. (2014). Transcriptome from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction. *Journal of Molecular and Cellular Cardiology*, 74:13–21.

[Suriadi *et al.*, 2014] Suriadi, S., Mans, R. S., Wynn, M. T., Partington, A. et Karnon, J. (2014). Measuring Patient Flow Variations : A Cross-Organisational Process Mining Approach. *In Proceedings of the Asia Pacific Business Process Management*, pages 43–58.

[Sverrisson *et al.*, 2014] Sverrisson, E. F., Zens, M. S., Liang Fei, D., Andrews, A., Schned, A., Robbins, D., Kelsey, K. T., Li, H., DiRenzo, J., Karagas, M. R. et Seigne, J. D. (2014). Clinicopathological correlates of Gli1 expression in a population-based cohort of patients with newly diagnosed bladder cancer. *Urologic Oncology : Seminars and Original Investigations*, 32(5):539–545.

[Séroussi *et al.*, 2013] Séroussi, B., Soulet, A., Spano, J.-P., Lefranc, J.-P., Cojean-Zelek, I., Blaszka-Jaulerry, B., Zelek, L., Durieux, A., Tournigand, C., Messai, N., Rousseau, A. et Bouaud, J. (2013). Which patients may benefit from the use of a decision support system to improve compliance of physician decisions with clinical practice guidelines : a case study with breast cancer involving data mining. *Studies in Health Technology and Informatics*, 192:534–538.

[Tada *et al.*, 2015] Tada, Y., Hiroshima, K., Shimada, H., Morishita, N., Shirakawa, T., Matsumoto, K., Shingyoji, M., Sekine, I., Tatsumi, K. et Tagawa, M. (2015). A clinical protocol to inhibit the HGF/c-Met pathway for malignant mesothelioma with an intrapleural injection of adenoviruses expressing the NK4 gene. *SpringerPlus*, 4:358.

[Taleb, 2008] Taleb, N. N. (2008). The fourth quadrant : a map of the limits of statistics. *An Edge Original Essay.*

[Tanaka *et al.*, 2015] Tanaka, P. S., Vieira, M. R. et Kaster, D. S. (2015). Efficient Algorithms to Discover Flock Patterns in Trajectories. *In Proceedings of the GeoInfo*, pages 56–67.

[Tang *et al.*, 2015] Tang, W., Sun, X., Zhang, Y., Ye, T. et Zhang, L. (2015). How to build and evaluate an integrated health care system for chronic patients : study design of a clustered randomised controlled trial in rural China. *International Journal of Integrated Care*, 15(1):e007.

[Tardif, 2007] Tardif, L. (2007). Étude méthodologique du chaînage des séjours Base PMSI MCO 2004. Rapport technique 116, Direction de la Recherche, des Études, de l'Évaluation et des Statistiques. http://drees.social-sante.gouv.fr/IMG/pdf/seriestat116.pdf.

[Thomas *et al.*, 2011] Thomas, J., McNaught, J. et Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14.

[Thompson *et al.*, 2015] Thompson, C. A., Kurian, A. W. et Luft, H. S. (2015). Linking Electronic Health Records to Better Understand Breast Cancer Patient Pathways Within and Between Two Health Systems. *Generating Evidence & Methods to Improve Patients Outcomes*, 3(1):1127.

[Thompson, 1996] Thompson, J. E. (1996). The Evolution of Surgery for the Treatment and Prevention of Stroke. *Stroke*, 27(8):1427–1434.

[Thorvaldsen *et al.*, 1995] Thorvaldsen, P., Asplund, K., Kuulasmaa, K., Rajakangas, A.-M., Schroll, M. *et al.* (1995). Stroke incidence, case fatality, and mortality in the who monica project. *Stroke*, 26(3):361–367.

[Thygesen *et al.*, 2012] Thygesen, K., Alpert, J. S., Jaffe, A. S., Simoons, M. L., Chaitman, B. R. et White, H. D. (2012). Third universal definition of myocardial infarction. *Circulation*, 126(16):2020–2035.

[Tillé, 2011] Tillé, Y. (2011). Dix années d'échantillonnage équilibré par la méthode du cube : une évaluation. *Techniques d'Enquête*, 37(2):233–246.

[Timsit *et al.*, 2005] Timsit, J.-F., Alberti, C. et Chevret, S. (2005). Le modèle de Cox. *Revue des Maladies Respiratoires*, 22(6):1058–1064.

[Toussaint, 2010] Toussaint, B. (2010). Évitons une autre affaire Mediator ! *Le Monde.fr*.

[Trombert-Paviot *et al.*, 2008] Trombert-Paviot, B., Couris, C.-M., Couray-Targe, S., Rodrigues, J.-M., Colin, C. et Schott, A.-M. (2008). Qualité et utilité d'un identifiant patient anonyme et unique pour le chaînage des séjours hospitaliers dans les bases de données médicoéconomiques françaises. *Revue d'Épidémiologie et de Santé Publique*, 55(3):203–211.

[Tuppin *et al.*, 2010] Tuppin, P., Neumann, A., Danchin, N., de Peretti, C., Weill, A., Ricordeau, P. et Allemand, H. (2010). Evidence-based pharmacotherapy after myocardial infarction in france : adherence-associated factors and relationship with 30-month mortality and rehospitalization. *Archives of Cardiovascular Diseases*, 103(6):363–375.

[Tzoulaki *et al.*, 2009] Tzoulaki, I., Molokhia, M., Curcin, V., Little, M. P., Millett, C. J., Ng, A., Hughes, R. I., Khunti, K., Wilkins, M. R., Majeed, A. et Elliott, P. (2009). Risk of cardiovascular disease and all cause mortality among patients with type 2 diabetes prescribed oral antidiabetes drugs : retrospective cohort study using UK general practice research database. *British Medical Journal*, 339:b4731.

[Ulrike, 2010] Ulrike, H. (2010). Qu'est-ce que l'Odds ratio et à quoi sert-il ? *Forum Medical Suisse*, 10(37):634–635.

[Vaccarino *et al.*, 1995] Vaccarino, V., Krumholz, H. M., Berkman, L. F. et Horwitz, R. I. (1995). Sex differences in mortality after myocardial infarction is there evidence for an increased risk for women ? *Circulation*, 91(6):1861–1871.

[Vaccarino *et al.*, 2001] Vaccarino, V., Krumholz, H. M., Yarzebski, J., Gore, J. M. et Goldberg, R. J. (2001). Sex differences in 2-year mortality after hospital discharge for myocardial infarction. *Annals of Internal Medicine*, 134(3):173–181.

[Vacheron, 2010] Vacheron, A. (2010). La prévention des maladies cardiovasculaires, un enjeu majeur de santé publique. *Cahiers de l'Académie des Sciences Morales et Politiques*, pages 105–119.

[Valensi *et al.*, 2011] Valensi, P., Lorgis, L. et Cottin, Y. (2011). Prevalence, incidence, predictive factors and prognosis of silent myocardial infarction : a review of the literature. *Archives of Cardiovascular Diseases*, 104(3):178–188.

[Vallée, 2013] Vallée, J.-P. (2013). Statines : peut-on «bouger les lignes» ? *Médecine*, 9(10):436–437.

[Van der Loo, 2014] Van der Loo, M. P. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1):111–122.

[Van Eck et Waltman, 2011] Van Eck, N. J. et Waltman, L. (2011). Text mining and visualization using VOSviewer. *International Society for Scientometrics and Infometrics Newsletter*, 7(3):50–54.

[Van Hecke *et al.*, 2015] Van Hecke, A., Heinen, M., Fernandez-Ortega, P., Graue, M., Hendriks, J., Høy, B., Köpke, S., Lithner, M. et van Gaal, B. (2015). Access to effective healthcare : effective self-management support intervention for patients with a chronic condition and a low social economic status : a systematic review. *BioMed Central Nursing*, 14(1):S7.

[Van Hoeve *et al.*, 2014] Van Hoeve, J., de Munck, L., Otter, R., de Vries, J. et Siesling, S. (2014). Quality improvement by implementing an integrated oncological care pathway for breast cancer patients. *The Breast*, 23(4):364–370.

[Veloso *et al.*, 2013] Veloso, A. G., Sperling, C., Holm, L. V., Nicolaisen, A., Rottmann, N., Thayssen, S., Christensen, R. d., Knudsen, J. L. et Hansen, D. G. (2013). Unmet needs in cancer rehabilitation during the early cancer trajectory – a nationwide patient survey. *Acta Oncologica*, 52(2):372–381.

[Versaci *et al.*, 1997] Versaci, F., Gaspardone, A., Tomai, F., Crea, F., Chiariello, L. et Gioffrè, P. A. (1997). A Comparison of Coronary-Artery Stenting with Angioplasty for Isolated Stenosis of the Proximal Left Anterior Descending Coronary Artery. *New England Journal of Medicine*, 336(12):817–822.

[Wagner *et al.*, 2001] Wagner, E. H., Austin, B. T., Davis, C., Hindmarsh, M., Schaefer, J. et Bonomi, A. (2001). Improving chronic illness care : translating evidence into action. *Health Affairs*, 20(6):64–78.

[Walshaw, 2000] Walshaw, C. (2000). A multilevel algorithm for force-directed graph drawing. *In Proceedings of the International Symposium on Graph Drawing*, pages 171–182.

[Wang *et al.*, 2013] Wang, W., McKinnie, S. M. K., Patel, V. B., Haddad, G., Wang, Z., Zhabyeyev, P., Das, S. K., Basu, R., McLean, B., Kandalam, V., Penninger, J. M., Kassiri, Z., Vederas, J. C., Murray, A. G. et Oudit, G. Y. (2013). Loss of Apelin Exacerbates Myocardial Infarction Adverse Remodeling and Ischemia-reperfusion Injury : Therapeutic Potential of Synthetic Apelin Analogues. *Journal of the American Heart Association*, 2(4):e000249.

[Wang et Bajorek, 2016] Wang, Y. et Bajorek, B. (2016). Selecting antithrombotic therapy for stroke prevention in atrial fibrillation : Health professionals' feedback on a decision support tool. *Health Informatics Journal*, page 1460458216675498.

[Wang *et al.*, 2006] Wang, Y., Lim, E. et Hwang, S. (2006). Efficient mining of group patterns from user movement data. *Data Knowledge Engineering*, 57(3): 240–282.

[Waterson *et al.*, 2012] Waterson, P., Eason, K., Tutt, D. et Dent, M. (2012). Using HIT to deliver integrated care for the frail elderly in the UK : current barriers and future challenges. *Work*, 41(Suppl 1):4490–4493.

[Weintraub *et al.*, 2012] Weintraub, W. S., Grau-Sepulveda, M. V., Weiss, J. M., DeLong, E. R., Peterson, E. D., O'Brien, S. M., Kolm, P., Klein, L. W., Shaw, R. E., McKay, C., Ritzenthaler, L. L., Popma, J. J., Messenger, J. C., Shahian, D. M., Grover, F. L., Mayer, J. E., Garratt, K. N., Moussa, I. D., Edwards, F. H. et Dangas, G. D. (2012). Prediction of Long-Term Mortality After Percutaneous Coronary Intervention in Older Adults Clinical Perspective. *Circulation*, 125(12):1501–1510.

[Weiss et Hirsh, 1998] Weiss, G. M. et Hirsh, H. (1998). Learning to predict rare events in event sequences. *In Proceedings of the Knowledge Discovery and Data Mining*, pages 359–363.

[Wilhelmsson *et al.*, 1975] Wilhelmsson, C., Elmfeldt, D., Vedin, J. A., Tibblin, G. et Wilhelmsen, L. (1975). Smoking and Myocardial Infarction. *The Lancet*, 305(7904):415–420.

[Wilkins *et al.*, 2017] Wilkins, E., Wilson, L., Wickramasinghe, K., Bhatnagar, P., Leal, J., Luengo-Fernandez, R., Burns, R., Rayner, M. et Townsend, N. (2017). European Cardiovascular Disease Statistics 2017. Rapport technique, European Heart Network. http://www.ehnheart.org/cvd-statistics.html.

[Wilson, 2008] Wilson, R. E. (2008). Mechanisms for spatio-temporal pattern formation in highway traffic models. *Philosophical Transactions of the Royal Society of London : Mathematical, Physical and Engineering Sciences*, 366(1872):2017–2032.

[Winkler, 1990] Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*, pages 354–359.

[World Health Organization, 2016] World Health Organization (2016). New initiative launched to tackle cardiovascular disease, the world's number one killer. *WHO*. http://www.who.int/cardiovascular_diseases/global-hearts/Global_hearts_initiative/en/.

[Wright *et al.*, 2015] Wright, A. P., Wright, A. T., McCoy, A. B. et Sittig, D. F. (2015). The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, 53:73–80.

[Wu *et al.*, 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J. et Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

[Yan *et al.*, 2005] Yan, Y., Birman-Deych, E., Radford, M. J., Nilasena, D. S. et Gage, B. F. (2005). Comorbidity indices to predict mortality from Medicare data : results from the national registry of atrial fibrillation. *Medical Care*, 43(11): 1073–1077.

[Yang *et al.*, 2014] Yang, J. M., Park, Y. S., Chung, S. P., Chung, H. S., Lee, H. S., You, J. S., Lee, S. H. et Park, I. (2014). Implementation of a clinical pathway based on a computerized physician order entry system for ischemic stroke attenuates off-hour and weekend effects in the ED. *The American Journal of Emergency Medicine*, 32(8):884–889.

[Ycart, 2002] Ycart, B. (2002). *Modèles et algorithmes markoviens*, volume 39. Springer Science & Business Media.

[Yeh *et al.*, 2010] Yeh, R. W., Sidney, S., Chandra, M., Sorel, M., Selby, J. V. et Go, A. S. (2010). Population trends in the incidence and outcomes of acute myocardial infarction. *New England Journal of Medicine*, 362(23):2155–2165.

[Young *et al.*, 2004] Young, W., McShane, J., O'Connor, T., Rewa, G., Goodman, S., Jaglal, S. B., Cash, L. et Coyte, P. (2004). Registered nurses' experiences with an evidence-based home care pathway for myocardial infarction clients. *Canadian Journal of Cardiovascular Nursing*, 14(3):24–31.

[Yusuf *et al.*, 2004] Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J. *et al.* (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study) : case-control study. *The Lancet*, 364(9438):937–952.

[Yusuf *et al.*, 2001] Yusuf, S., Reddy, S., Ôunpuu, S. et Anand, S. (2001). Global burden of cardiovascular diseases part ii : variations in cardiovascular disease by specific ethnic groups and geographic regions and prevention strategies. *Circulation*, 104(23):2855–2864.

[Zadeh, 1996] Zadeh, L. A. (1996). Possibility theory and soft data analysis. *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems*, pages 481–541.

[Zaki, 2001] Zaki, M. J. (2001). Spade : An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1-2):31–60.

[Zhang *et al.*, 2014] Zhang, Y., Nester, C. M., Martin, B., Skjoedt, M.-O., Meyer, N. C., Shao, D., Borsa, N., Palarasah, Y. et Smith, R. J. (2014). Defining the complement biomarker profile of C3 glomerulopathy. *Clinical Journal of the American Society of Nephrology*, 9(11):1876–1882.

[Zhao *et al.*, 2015] Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y. et Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BioMed Central Bioinformatics - Proceedings of the 12th Annual MidSouth Computational Biology and Bioinformatics Society Conference*, 16(13):S8.

[Zhi *et al.*, 2011] Zhi, Q., LIN, Z.-w. et Yan, M. (2011). Research of hadoop-based data flow management system. *The Journal of China Universities of Posts and Telecommunications*, 18:164–168.