UNIVERSITE LILLE 2 DROIT ET SANTE

FACULTE DE MEDECINE HENRI WAREMBOURG

Année : 2014

THESE POUR LE DIPLOME D'ETAT
DE DOCTEUR EN MEDECINE

Construction et évaluation de règles de prédiction de diagnostics à partir des bases de données hospitalières : application au contrôle qualité des données médico-administratives

Présentée et soutenue publiquement le 21 novembre 2014 à 18h00
au Pôle Formation

Par Mehdi DJENNAOUI

_____

JURY
Président :
Monsieur le Professeur Jean-Louis SALOMEZ
Assesseurs :
Monsieur le Professeur Régis BEUSCART
Monsieur le Docteur Emmanuel CHAZARD
Directeur de Thèse :
Monsieur le Docteur Grégoire FICHEUR
Service de l'Information et des Archives Médicales – CHRU de Lille

_____

Avertissement

La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Liste des abréviations

AMO	Assurance Maladie Obligatoire
ANAES	Agence Nationale d'Accréditation et d'Evaluation en Santé
ANAP	Agence Nationale d'Appui à la Performance
ANSSI	Agence Nationale de la Sécurité des Systèmes d'Information
ARS	Agence Régionale de Santé
ATIH	Agence Technique de l'Information Hospitalière
CCAM	Classification Commune des Actes Médicaux
CCMSA	Caisse Centrale de la Mutualité Sociale Agricole
CERIM	Centre d'Etudes et de Recherche en Informatique Médicale
CIM-10	Classification Internationale des Maladies 10ème révision
CISS	Collectif Inter Associatif sur la Santé
CM	Catégories Majeures
CMD	Catégories Majeures de Diagnostic
CMA	Complications et Morbidités Associées
CNAMTS	Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés
CNIL	Commission Nationale de l'Informatique et des Libertés
CNSA	Caisse Nationale de Solidarité pour l'Autonomie
CSG	Contribution Sociale Généralisée
DAS	Diagnostics Associés Significatifs
DG	Dotation Globale
DGCIS	Direction Générale de la Compétitivité de l'Industrie et des Services
DGFIP	Direction Générale des Finances Publiques
DIM	Département d'Information Médicale
DMS	Durée Moyenne de Séjour
DP	Diagnostic Principal
DR	Diagnostic Relié
DREES	Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques
DRG	Diagnosis Related Groups

EHESP	Ecole des Hautes Etudes en Santé Publique
ENC	Etude Nationale des Coûts
ESPIC	Etablissements de Santé Privés d'Intérêt Collectif
EXB	Extrême Basse
EXH	Extrême Haute
FAG	Forfait Annuel Greffes
FAU	Forfait d'Accueil aux Urgences
FEHAP	Fédération des Etablissements Hospitaliers et d'Aide à la Personne
FHF	Fédération Hospitalière de France
FHP	Fédération de l'Hospitalisation Privée
FINESS	Fichier National des Établissements Sanitaires et Sociaux
FNCLCC	Fédération Nationale des Centres de Lutte Contre le Cancer
FNORS	Fédération Nationale des Observatoires de Santé
FOIN	Fonction d'Occultation des Informations Nominatives
GHM	Groupes Homogènes de Malades
HAD	Hospitalisation à Domicile
HAS	Haute Autorité de Santé
IGAS	Inspection Générale des Affaires Sociales
INSEE	Institut National de la Statistique et des Etudes Economiques
IRDES	Institut de Recherche et de Documentation en Economie de la Santé
LFSS	Loi de Financement de la Sécurité Sociale
MAGIC	Module d'Anonymisation et de Gestion des Informations de Chaînage
MCO	Médecine Chirurgie Obstétrique
MECCS	Mission d'Evaluation et de Contrôle de la Sécurité Sociale
MERRI	Missions d'Enseignement, de Recherche, de Référence et d'Innovation
MIG-AC	Missions d'Intérêt Général – Aide à la Contractualisation
OMS	Organisation Mondiale de la Santé
ONDAM	Objectifs Nationaux de Dépenses d'Assurance Maladie
OQN	Objectif Quantifié National
PMSI	Programme de Médicalisation du Système d'Information hospitalière
PSPH	Participant au Service Public Hospitalier
RSA	Résumé de Sortie Anonyme
RSI	Régime Social des Indépendants
RSS	Résumé de Sortie Standardisé
RUM	Résumé d'Unité Médicale

SAE	Statistique Annuelle des Etablissements
SIAM	Service de l'Information et des Archives Médicales
SIRET	Système d'Identification du Répertoire des Etablissements
SNATIH	Système National d'Information sur l'Hospitalisation
SNIIRAM	Système National d'Information Inter-Régimes de l'Assurance Maladie
SPADE	Sequential PAttern Discovery using Equivalence classes
SSR	Soins de Suite et de Réadaptation
T2A	Tarification à l'Activité
UM	Unité Médicale
UNOCAM	Union Nationale des Organismes d'Assurance Maladie Complémentaire
UNPS	Union Nationale des Professions de Santé
UNRS	Union Nationale des Régimes Spéciaux

TABLE DES MATIERES

RESUME

Contexte : Dans le cadre du Programme de Médicalisation du Système Informatique (PMSI), l'instauration de la Tarification à l'Activité (T2A) a incité les établissements de santé à établir des procédures de contrôle qualité optimisant la rémunération des séjours, le repérage et le recodage de Complications et Morbidités Associées (CMA) manquantes constituant une de ces procédures. Les méthodes de data mining suscitent beaucoup d'intérêt dans l'analyse des bases de données. Ces méthodes produisent des règles prédictives simples, intuitives et faciles à appliquer.

Nous avions pour objectif de produire des règles de prédiction de CMA applicables au contrôle qualité des séjours en soumettant les données de la base nationale PMSI aux méthodes de data mining.

Méthode : Notre échantillon de travail était constitué à partir de la base nationale PMSI pour les années 2007 à 2010. Les CMA à prédire devaient être fréquentes et relever d'une pathologie chronique, la prédiction se faisant à partir des codes diagnostiques et d'actes des séjours. Notre étude se poursuivait par l'évaluation des règles produites en calculant la confiance qualité des règles et en appréciant le gain en termes de recodage de CMA.

Résultats : Notre échantillon comportait 59170 séjours. Les CMA ciblées étaient les codes E11 « diabète sucré non insulinodépendant », I48 « fibrillation atriale » et I50 « insuffisance cardiaque ». Nous avons extrait trois règles d'association et six règles séquentielles, et validé à l'issue de la procédure de contrôle trois règles prédictives, deux pour le code E11 et une pour le code I48. Les

trois règles validées ont toutes une confiance supérieure à 0.60 et un gain de recodage de CMA supérieur à 28 %.

Conclusion : Notre étude nous a permis d'extraire à partir de la base nationale PMSI, par data mining, des règles de prédiction de CMA valides, fiables et simples d'application dans le cadre du contrôle qualité des séjours.

INTRODUCTION

.I. Le financement des établissements de santé en France

L'inflation des dépenses de santé occupe une place notable parmi les évolutions économiques et sociologiques ayant marqué la France (1), ce dès la fin du siècle précédent ; ce phénomène, observé dans tous les pays industrialisés (2,3), est secondaire à une demande de soins en expansion rapide, elle-même apparentée à plusieurs facteurs (4,5) : non seulement le développement de nouvelles technologies de santé, forcément plus onéreuses, mais également l'amélioration des conditions de vie et l'allongement de la durée de vie avec pour inévitables corollaires l'expansion démographique et surtout le vieillissement de la population, d'où la modification du paysage sanitaire inhérente, les maladies chroniques évoluant désormais au premier plan. A ceci s'ajoute l'émergence de la santé comme préoccupation essentielle des populations, avec une attention plus accrue portée sur les questions de santé, aiguillonnée par les nouvelles problématiques portant sur l'environnement, l'alimentation, la iatrogénie etc.

Ainsi, cette prise de conscience sanitaire, pourvoyeuse d'impacts positifs tant en termes de santé publique qu'individuelle, dans un esprit de droit à la santé pour tous, a cependant favorisé la consommation de biens de santé.

Dans ce contexte, le financement de la santé en général et des établissements hospitaliers en particulier s'est positionné comme une préoccupation essentielle des pouvoirs publics.

Les instances gouvernantes disposaient de peu de leviers pour parvenir à une meilleure maîtrise des dépenses de santé (6) ; certaines mesures, comme l'augmentation des taux de prélèvements afin d'accroître les recettes publiques (à titre d'exemple la Contribution Sociale Généralisée (CSG)), et le transfert d'une partie des dépenses sur le budget des ménages afin de réduire la part portée par les finances publiques (le ticket modérateur par exemple), étaient fondamentalement mal perçues par le grand public et trouvaient vite leurs limites du fait de cette impopularité.

Il était alors logique de se tourner vers les prestataires de soins et de les cibler par un ensemble de mesures visant à mieux contrôler les dépenses de santé, dans un esprit d'efficience et de qualité de ces prestations.

Plusieurs approches de modes de financement de l'hôpital se sont ainsi succédées et complétées au fil des gouvernements (6), aboutissant à l'instauration de la Tarification à l'Activité (T2A), principale mode de financement actuel des établissements de santé publics et privés.

L'analyse succincte de ces modes de financement (7) permet de mieux appréhender la genèse et les enjeux de la réforme de la T2A. En concordance avec l'objectif de ce travail, cet historique se focalise en particulier sur le financement des activités de Médecine, Chirurgie et Obstétrique (MCO) pour les secteurs public et privé à but non lucratif.

A. Les cycles de financement de l'hôpital

1. Le prix de journée (1946)

Le contexte d'après-guerre des années quarante se distingue par un changement profond dans le domaine des politiques sociales, dominé par la création de la Sécurité sociale en 1945, avec l'émergence du principe de « solidarité » qui gouverne le système de santé français actuel.

De 1946 à 1984, le financement des établissements hospitaliers publics repose sur le principe d'un tarif à la journée d'hospitalisation et d'hébergement qui s'apparente à un forfait « tout compris », garant de la solidarité sociale à l'égard des patients hospitalisés.

Ce mode de financement a été critiqué principalement du fait de ses effets inflationnistes. En effet, il suffisait à certains établissements d'allonger les durées de séjour et de saturer leurs lits pour augmenter les recettes. En outre, la saturation des lits permettait ensuite d'argumenter l'ouverture de nouveaux lits l'année suivante. Dans le même temps, l'activité chirurgicale était graduellement transférée vers le secteur privé lucratif qui, lui, était intéressé par un paiement à l'acte en sus du prix de journée.

Malgré ces limites, ce système a fonctionné pendant près de 40 ans, favorisé en cela par le contexte contemporain de croissance économique, qui permettait encore de dépenser sans réellement compter.

En faveur de ce système, il convient de souligner qu'il a favorisé la rénovation de l'hôpital, le contexte économique favorable des trente glorieuses permettant des avancées considérables en matière sanitaire et sociale.

A l'issue de cette période, l'entrée de la France dans une période économiquement moins favorable a incité les pouvoirs publics à prendre des mesures de restriction des dépenses de santé. Ainsi, en 1983, afin de mieux contrôler la croissance budgétaire des hôpitaux, un nouveau mode de financement se met en place : le « budget global » ou « dotation globale de fonctionnement ».

2. La dotation globale de fonctionnement (1984)

Ce système de financement se base sur le principe d'un budget annuel alloué à chaque établissement par reconduction du budget de l'année précédente selon un « taux directeur national d'augmentation des dépenses hospitalières ».

Bien que ne tenant pas directement compte de l'activité de l'hôpital, ce nouveau système de financement n'incitait plus les établissements à allonger les durées de séjour, cette pratique n'ayant plus d'impact sur les recettes. Inversement, dans ce système de financement un lit vide rapportait autant qu'un lit plein.

Cette réforme a été complétée par une série de mesures visant à rationaliser le système de santé, axées sur des instruments de maîtrise quantitative et qualitative.

La maîtrise quantitative des dépenses reposait sur la planification sanitaire. Elle concernait en premier lieu la rationalisation du nombre de lits d'hospitalisation, puis s'est élargie aux équipements et à l'activité de soins. Les hôpitaux ont également développé le concept de management de projet sur la base du projet d'établissement. En 1996, le plan Juppé a confirmé cette volonté de contrôle des finances au-delà du monde de l'hospitalisation, en faisant voter par le Parlement les objectifs de dépenses et les prévisions de recettes (définition d'Objectifs Nationaux de Dépenses d'Assurance Maladie (ONDAM)).

Les instruments de maîtrise qualitative des activités hospitalières consistaient en l'accréditation et l'évaluation. L'accréditation visait l'assurance de la qualité des prestations hospitalières et était réalisée en partenariat avec l'Agence Nationale d'Accréditation et d'Evaluation en Santé (ANAES), devenue Haute Autorité de Santé (HAS) en 2009. L'évaluation a quant à elle abouti en 1984 à la création du Programme de Médicalisation du Système d'Information hospitalière (PMSI), avec la constitution d'une base de données d'informations médicales sur les séjours des patients.

Ainsi, avec la création de ces outils de description, de mesure et de comparaison de l'activité médicale hospitalière et d'allocation des budgets, les pouvoirs publics ont introduit une démarche plus gestionnaire de l'hôpital.

Néanmoins, les impacts de la réforme de la dotation globale furent limités ; notamment, l'amélioration de la performance médico-économique attendue ne s'est pas concrétisée et la stimulation de l'innovation et du dynamisme de l'organisation a été très modeste. Elle a de plus favorisé les conflits avec les services, ces derniers étant peu concertés pour la répartition des dotations, et créé des inégalités de financement entre hôpitaux. Surtout, en l'absence de création d'outils de comptabilité analytique, ce mode de financement ne permettait pas d'estimer le coût réel des activités de soins (6).

Au vu de l'évolution décrite, la mise en œuvre d'une modalité de tarification à l'activité semblait inévitable. Cette option, évoquée dès les années quatre-vingt-dix, a été jugée nécessaire dès 2004, compte tenu de l'accroissement du déficit budgétaire des hôpitaux.

3. La Tarification à l'Activité (T2A) (2004)

La mise en place de la Tarification à l'Activité s'est faite dans le contexte préalablement décrit d'augmentation croissante de la demande et du coût de l'offre de soins.

Le système de financement par Dotation Globale (DG) s'est avéré insuffisant pour endiguer ces dépenses, d'autant plus que ce système ne concernait que les hôpitaux publics et les hôpitaux privés Participant au Service Public Hospitalier (PSPH), les cliniques privées étant financées par un système de tarification à la journée et de forfaits liés aux actes, encadrés par un Objectif Quantifié National (OQN). Cette situation a contribué à créer des rentes de financement pour certains établissements avec en contrepartie une pénurie de moyens pour d'autres structures et un sous-financement pour certaines activités.

Dès lors, l'instauration de la Tarification à l'Activité a été envisagée dans une perspective de plus grande transparence et d'équité de traitement entre les hôpitaux,

l'encadrement des dépenses hospitalières reposant sur le diptyque « équité-productivité ».

B. Les principes fondateurs du PMSI et de la Tarification à l'Activité

Le PMSI est inspiré d'un concept imaginé par R.B. Fetter, professeur d'économie à l'université de Yale (8) ; ce concept aborde le classement des séjours hospitaliers selon une double optique, médicale reposant sur la pathologie prise en charge et économique reposant sur la nature et l'importance des moyens alloués à cette prise en charge.

Ce système a été adopté par le Congrès des Etats-Unis d'Amérique au début des années quatre-vingt devant l'explosion des dépenses hospitalières. La solution à cette problématique a été la création d'un système de rémunération se basant sur une classification médico-économique des séjours, la classification Diagnosis Related Groups (DRG). Ce mode de paiement attribuait à chaque DRG un tarif fixe, indépendant de la durée du séjour, introduisant la notion de Prospective Payment System, inspirateur de la Tarification à l'Activité.

L'ébauche du PMSI en France sera mise en place dans le milieu des années quatre-vingt, avec la création du pendant français de la classification DRG, la classification des Groupes Homogènes de Malades (GHM) (9,10). Cette mesure semblait inévitable ; en effet, si l'on considère l'hôpital comme une structure de production de soins, la médicalisation de la description de cette production était logique et indispensable.

Initialement construit pour être un outil de suivi d'activité, l'implication du PMSI dans le domaine du financement hospitalier se développera progressivement.

Comme décrit précédemment, face à l'augmentation des dépenses de santé, les pouvoirs publics disposaient de peu de leviers pour endiguer ce phénomène ;

l'incitation à une plus grande maîtrise de l'activité dans une optique d'efficience cadrait avec cet objectif, d'où l'instauration de la Tarification à l'Activité.

La T2A, définie comme « un mode de financement, qui vise à fonder l'allocation des ressources aux établissements de santé publics et privés sur la nature et le volume de leur activité réalisée, mesurée, pour l'essentiel, sur la base des données issues du programme de médicalisation des systèmes d'information (PMSI) » (Ministère de la Santé, octobre 2003), constitue de ce fait un système dynamique et prospectif, ayant pour socle la mesure en temps réel de l'activité des établissements, le paiement des soins réalisés étant tributaire de cette activité, selon des procédures d'attribution définies par des règles économiques et statistiques complexes.

La T2A s'inscrivait dans un contexte plus large de restructuration du paysage hospitalier français (7). Ainsi, le « Plan Hôpital 2007 » prévoyait un projet de nouvelle gouvernance hospitalière, avec d'une part la création d'une instance de pilotage médico-administratif (le conseil exécutif), et d'autre part la création de pôles d'activités.

Par ailleurs, le PMSI s'avérait utile pour de nombreuses utilisations internes aux hôpitaux : production de rapports d'activité, définition des stratégies, gestion des services (6) etc.

Ceci reflétait bien la volonté de responsabilisation des différents acteurs et d'amélioration de la coordination entre administratifs et médicaux-soignants, le Département d'Information Médicale (DIM) occupant de facto une place stratégique dans l'organisation et le développement de cet environnement médical, administratif et financier.

La mise en place de cette réforme ne pouvait se faire que progressivement, considérant l'importance des changements envisagés et des moyens déployés pour réaliser ces changements.

Du fait de leurs spécificités respectives, il était d'emblée nécessaire d'individualiser le secteur public et les établissements PSPH (renommés depuis Etablissements de Santé Privés d'Intérêt Collectif (ESPIC)) agrégés sous la dénomination ex-DG et le secteur privé désormais connu comme ex-OQN. Cette distinction était forcément source de difficultés supplémentaires dans l'instauration de la T2A.

Une expérimentation des nouveaux modes de financement des établissements publics et privés a été lancée en 2000 sur un panel d'établissements volontaires ; l'instauration de la T2A a par la suite débuté en 2004 pour les établissements ex-DG et en 2005 pour les établissements ex-OQN, en limitant initialement le financement du budget des établissements ex-DG par T2A à 10 % des recettes perçues, et ce uniquement pour les activités de Médecine, Chirurgie et Obstétrique (MCO) (les établissements ex-OQN étant d'emblée à 100 %), avec pour objectif d'étendre progressivement la réforme tant en termes de volume de financement que de champs d'activités concernés : Hospitalisation à Domicile (HAD), Soins de Suite et de Réadaptation (SSR) et Psychiatrie.

A ce jour, la T2A concerne 100 % de la part Assurance Maladie Obligatoire (AMO) des séjours en hospitalisation pour les activités MCO et HAD, l'extension de systèmes comparables au SSR et à la Psychiatrie étant envisagée dans un avenir prochain.

Signalons que, malgré une liste commune de GHM, les tarifs de rémunération diffèrent entre les deux secteurs ex-DG et ex-OQN. Ceci est lié non seulement aux différences de statuts des praticiens (d'une part salariés et d'autre part libéraux donc rémunérés par des honoraires), mais également à des méthodes différentes de calcul des tarifs, les tarifs du secteur ex-DG étant estimés à partir d'une Etude Nationale des Coûts (ENC) (11) issue d'un échantillon d'établissements, tandis que

les tarifs du secteur ex-OQN proviennent des données antérieures de facturation à l'Assurance Maladie.

Cet état de fait a été critiqué, du fait de l'hétérogénéité des dotations induite (12,13), d'où un objectif affiché de réduction des écarts de financement public-privé et de convergence tarifaire entre ces deux secteurs ; initialement fixée à 2012, la réalisation de cet objectif d'adéquation a été reportée à 2018 (14) dans l'attente des résultats d'études complémentaires.

A noter qu'une partie du financement hospitalier demeure indépendante de la Tarification à l'Activité.

Ainsi, le forfait hospitalier journalier (15), qui consiste en une participation du patient aux frais d'hébergement, est par définition hors du champ de la T2A, puisqu'à la charge du patient.

De même, le financement de produits de santé (médicaments et dispositifs médicaux) particulièrement onéreux (16) obéit à un mode de financement de produits médicaux en sus dont les listes sont déterminées par Arrêté ministériel.

Deux grandes activités soumises à autorisation sont également financées par des forfaits annuels (17), le Forfait d'Accueil aux Urgences (FAU) et le Forfait Annuel Greffes (FAG).

Enfin, des dotations annuelles complètent le financement hospitalier dans le cadre de l'enveloppe MIG-AC-MERRI (Missions d'Intérêt Général – Aide à la Contractualisation – Missions d'Enseignement, de Recherche, de Référence et d'Innovation), couvrant notamment le rôle de l'hôpital dans les domaines de maintien de la permanence des soins, de prévention, de promotion et d'éducation à la santé, de recherche et d'enseignement (18).

Se basant sur les expériences étrangères, outre les impacts attendus sur la réduction des durées de séjours et la réaffectation d'une partie de l'activité

d'hospitalisation vers l'ambulatoire et le soin à domicile, des effets délétères potentiels de la T2A étaient à craindre : fragmentation des séjours, sortie trop précoce des patients, transformation artificielle d'actes externes en hospitalisations, navettes entre établissements, surcodage des séjours (6) etc. A ce titre, le législateur a aussi institué des procédures de contrôle externe (19) avec remboursement d'indus et sanctions financières le cas échéant, d'où un enjeu supplémentaire en matière de qualité du codage et de ses conséquences.

C. Outils du PMSI et de la T2A

Considérant les enjeux financiers évoqués, le recueil de l'activité des établissements de santé, mis en place par l'Etat et géré par les établissements, revêt une importance fondamentale, les critères de ce recueil devant être respectés tant au niveau quantitatif en matière d'exhaustivité qu'au niveau qualitatif en matière de respect des règles et consignes édictées par les instances tutélaires.

L'ensemble des outils et des procédures liées au PMSI et à la T2A sont élaborés et gérés par l'Agence Technique de l'Information Hospitalière (ATIH) (20).

L'agence définit la nature des informations à recueillir et les modalités de recueil et de codage de ces informations selon les classifications règlementaires. En s'appuyant sur les informations recueillies, elle construit une classification médico-économique des séjours.

Les données ainsi produites sont agrégées par l'ATIH et constituent la base nationale regroupant l'ensemble des informations des établissements. Le traitement de cette base nationale sert à l'analyse, à la restitution et à la diffusion de l'information hospitalière par l'agence.

L'ensemble des procédures et consignes de recueil de l'information médicale sont compilées dans le guide méthodologique édité par l'ATIH, dont la mise à jour

annuelle est publiée au Bulletin officiel et disponible sur le site de l'agence (21). La version actuelle de la classification de groupage est la version 11f.

Par souci de concision et en accord avec l'objectif principal de ce travail, nous décrirons les outils affectés au PMSI et à la T2A uniquement pour le champ MCO.

1. Champ du recueil et définitions

L'hospitalisation d'un patient est le déclencheur de la production de l'information médicale. L'admission de tout patient dans le cadre du champ MCO se fait dans une Unité Médicale (UM) ; par Unité Médicale, on entend un ensemble individualisé de moyens matériels et humains assurant des soins à des patients.

Tous les types d'hospitalisations sont concernés par le recueil, aussi bien l'hospitalisation complète que l'hospitalisation à temps partiel (hospitalisation de jour et de nuit, anesthésie et chirurgie ambulatoires, séances). Les consultations externes et les actes médicaux et paramédicaux réalisés à titre externe ne sont donc pas concernés.

La conséquence de l'enregistrement administratif d'une admission dans une Unité Médicale d'hospitalisation sera la production d'un Résumé d'Unité Médicale (RUM) à la fin du séjour dans l'unité.

L'ensemble des RUM produits au décours de l'hospitalisation constitue le Résumé de Sortie Standardisé (RSS), dont l'anonymisation est l'origine du Résumé de Sortie Anonyme (RSA) qui est transmis à l'Agence Régionale de Santé (ARS) dont dépend l'établissement de santé.

2. Contenu des recueils d'informations relatives à l'activité

a) Le Résumé d'Unité Médicale (RUM)

Le Résumé d'Unité Médicale est constitué de rubriques de natures administrative et médicale. Les informations du RUM doivent être conformes au contenu du dossier médical du patient.

i. Informations administratives et démographiques

Ces informations sont dévolues à l'identification du patient. Certains champs sont obligatoires quelle que soit l'activité de soin alors que d'autres ne sont renseignés que dans le cadre de soins particuliers.

ii. Informations médicales

C'est le praticien responsable d'une structure médicale ou médico-technique ou le praticien ayant dispensé les soins qui est garant de l'exhaustivité et de la qualité des informations qu'il transmet pour traitement au médecin DIM de l'établissement.

Ne peuvent être codés dans le RUM que des affections ou des problèmes de santé présents et actifs au moment de l'hospitalisation et des actes réalisés pendant l'hospitalisation.

De même que pour les informations administratives, certains champs sont obligatoires tandis que d'autres sont propres à certaines activités, les diagnostics et les actes représentant les variables essentielles du recueil.

- *Diagnostics*

La nomenclature utilisée pour le codage des diagnostics est la Classification Internationale des Maladies dans sa dixième révision (CIM-10) de l'Organisation Mondiale de la Santé (OMS). Sa table analytique (chapitres I à XXII) est divisée en catégories dont les codes alphanumériques sont constitués de trois caractères. La majorité des catégories sont subdivisées en sous-catégories codées avec quatre ou cinq caractères.

Pour le recueil d'informations du PMSI, la règle est de coder avec quatre caractères chaque fois qu'une catégorie est subdivisée ; un code à trois caractères n'est admis que lorsqu'il correspond à une catégorie non subdivisée. Le recueil standard d'informations du PMSI utilise aussi des codes étendus au-delà du quatrième caractère.

Les diagnostics doivent figurer dans le RUM sous forme codée selon la plus récente mise à jour de la CIM-10 et selon les extensions nationales données dans la plus récente version du Manuel des GHM édité par l'ATIH et publié au Bulletin officiel (22).

- *Diagnostic Principal (DP)*

Le diagnostic principal est le problème de santé qui a motivé l'admission du patient dans l'UM, pris en charge pendant le séjour et déterminé à la sortie de l'UM, en connaissance de l'ensemble des informations médicales le concernant, y compris les résultats d'examens qui parviendraient ultérieurement à la sortie.

- *Diagnostic Relié (DR)*

Le diagnostic relié, en association avec le DP, rend compte de la prise en charge du patient en termes médico-économiques lorsque le DP est insuffisant pour cela.

Le DR n'est mentionné que lorsque le DP appartient au chapitre XXI de la CIM–10 (Facteurs influant sur l'état de santé et motifs de recours aux services de santé - codes Z), certains de ces diagnostics étant imprécis.

- *Diagnostics Associés Significatifs (DAS)*

Il s'agit d'une affection, d'un symptôme ou de tout autre motif de recours aux soins coexistant avec le DP et constituant un problème de santé distinct supplémentaire ou une complication de la morbidité principale ou de son traitement.

Il doit obligatoirement être pris en charge à titre diagnostique ou thérapeutique ou majorer l'effort de prise en charge du patient.

Lorsqu'un patient atteint d'une maladie chronique est hospitalisé pour un autre motif, la maladie chronique est un DAS, à condition que son traitement n'ait pas été interrompu pendant le séjour ou qu'elle ait bénéficié d'une surveillance.

Les antécédents guéris, les maladies stabilisées ou les facteurs de risque n'ayant bénéficié d'aucune prise en charge ne sont donc pas des DAS.

- *Actes*

Les actes médicaux sont codés selon la plus récente version en vigueur de la Classification Commune des Actes Médicaux (CCAM), classification élaborée par l'Assurance Maladie. Ses règles d'utilisation sont indiquées dans un Guide de lecture et de codage publié au Bulletin Officiel (23).

Le codage d'un acte avec la CCAM comprend :

- Son code principal (sept caractères alphanumériques) ;
- Sa phase, presque toujours codée « 0 » (seul un petit nombre d'actes connaissent une réalisation en phases distinctes) ;
- La ou les activités autorisées réalisées ;
- Le nombre de réalisations de l'acte ;
- La date de réalisation de l'acte ;
- Une éventuelle extension documentaire.

b) Le Résumé de Sortie Standardisé (RSS)

Le RSS est constitué de l'ensemble des RUM relatifs au même séjour d'un patient dans le champ d'activité MCO. Sa production est sous le contrôle du médecin responsable de l'information médicale.

Si le patient n'a fréquenté qu'une seule unité médicale, on parle de séjour mono-unité et le RSS équivaut au RUM (« RSS mono-RUM »).

Si le patient a fréquenté plusieurs unités médicales, on parle de séjour multi-unité et le RSS est formé par la suite ordonnée chronologiquement des RUM issus des séjours dans les différentes unités (« RSS multi-RUM »).

Dans un établissement de santé donné, pour un séjour-patient donné, entre une date d'entrée et une date de sortie données, il ne peut être produit qu'un RSS et un seul.

Le groupage du RSS, effectué par le logiciel groupeur de l'établissement de santé, permet de classer le séjour selon la classification des GHM, donnant lieu au RUM-RSS groupé, enregistrement enrichi des résultats du groupage.

i. La classification des GHM et le groupage des RSS

Depuis la mise en place de la T2A, la classification des GHM ainsi que la fonction de groupage ont connu de nombreuses modifications et adaptations ; la version actuelle de classification est la version 11f. Elle comporte 668 racines GHM pour un total de 2591 GHM. Les spécificités de la version v11f ainsi que l'historique des différentes versions sont disponibles dans le Manuel des GHM (22) qui est édité annuellement par l'ATIH et fait l'objet d'une publication au Bulletin officiel.

Les GHM constituent un système de classification médico-économique ; chaque RSS est classé dans un GHM donné au terme d'un processus de classement effectué selon un algorithme de groupage ou arbre de décision. A chaque GHM est associé un ou plusieurs Groupes Homogènes de Séjour (GHS). Ceux-ci déterminent le tarif de prise en charge du séjour par les régimes de l'Assurance Maladie.

Les GHM sont classés par racines qui se subdivisent en différentes catégories. La racine d'un GHM est constituée par les cinq premiers caractères du numéro de GHM. Les deux premiers chiffres désignent la Catégorie Majeure de Diagnostic (CMD), la lettre qui suit correspond à la sous-CMD, chirurgicale (C), technique non opératoire (K), médicale (M) ou indifférenciée (Z). Les deux derniers chiffres identifient le numéro d'ordre de la racine dans la sous-CMD.

La segmentation des racines GHM s'opère selon le sixième caractère qui décrit la complexité ou la sévérité du GHM. Il peut être de types 1, 2, 3 ou 4, le niveau 1 étant assimilé au niveau sans Complications et Morbidités Associées (CMA). Le tarif alloué au séjour, porté par le GHS, est alors proportionnel au niveau de sévérité.

ii. Les éléments déterminants du classement dans un GHM

- *Orientation vers une racine de GHM*

Le premier niveau de classement des RSS correspond aux Catégories Majeures (CM). Elles correspondent le plus souvent à un système fonctionnel et sont alors dites Catégories Majeures de Diagnostic (CMD) car c'est en général le DP du RSS qui détermine le classement.

L'algorithme de la classification repose sur des listes de diagnostics, des listes d'actes et un arbre de décision qui consiste en un ensemble de tests faits sur les informations du RSS pour l'orienter dans un GHM.

Le déroulement de l'algorithme de classement d'un séjour se déroule selon un certain nombre d'étapes :

Le RSS est classé dans une CMD selon le DP du RSS (plus rarement selon le DR ou selon l'un des actes), le DP étant lui-même sélectionné parmi les DP des RUM par la fonction de groupage en fonction du rang du RUM, de sa durée, de la nature du DP du RUM et des actes classants. Tous les diagnostics non retenus dans ce processus comme DP unique sont considérés comme des DAS, après élimination des doublons.

La présence d'un acte classant est ensuite recherchée dans le RSS ; un acte classant est un acte susceptible de modifier le classement en GHM. S'il existe un acte opératoire classant dans la CMD choisie selon le DP, le séjour est classé dans un groupe chirurgical, défini par la nature de l'intervention effectuée ; en l'absence d'acte classant, le RSS est orienté selon le DP dans un groupe médical.

La dernière étape du classement en GHM consiste en un nombre variable de tests sur les autres informations du RSS : ce sont très souvent l'âge et les DAS du RSS, plus rarement le mode de sortie. Des diagnostics peuvent également intervenir pour l'orientation vers un GHM chirurgical. De même, des actes classants non opératoires peuvent aussi intervenir dans le classement du séjour.

Enfin, d'autres facteurs interviennent dans le classement du GHM, comme le sexe (GHM définissant les pathologies de l'appareil génital), le mode d'entrée (entrée par transfert) ou le mode de sortie du séjour (GHM définissant les séjours se terminant par un décès).

- *De la racine de GHM au GHM : importance des CMA*

La notion de Complications et Morbidités Associées (CMA) est importante à mentionner. L'équipe de Yale définissait initialement les CMA comme : « les diagnostics associés dont la présence, toutes choses étant égales par ailleurs, augmentait la durée de séjour d'au moins 24 heures, dans au moins 75 % des cas ». Cette définition a été ajustée et affinée selon les évolutions de la classification, et la dernière version du Manuel des GHM définit les CMA sur le critère d'augmentation de la Durée Moyenne de Séjour (DMS) d'au moins 2 jours et d'au moins 25 % de la DMS de référence, avec au moins 55 % des séjours avec ce DAS concernés par cette augmentation de la durée de séjour par rapport à leur médiane de référence.

Ces CMA sont répertoriées dans une liste valable pour toute la classification des GHM. Les diagnostics qui font partie de la liste des CMA sont repérés par une caractéristique de diagnostics dont l'existence est testée sur tous les DAS du RSS.

Il existe cependant des listes d'exclusion de CMA, qui sont définies comme des DAS dont le caractère de CMA est exclu lorsque le DP ou le DR du RSS prend une certaine valeur. Pour chaque DP ou DR concerné, est mentionnée une liste de

diagnostics qui sont des CMA mais qui perdent cette propriété lorsqu'ils sont codés en DAS avec le DP ou DR mentionné.

En plus des listes d'exclusion, une durée de séjour minimum est requise pour accéder aux différents niveaux de sévérité. Ce seuil est de trois jours pour le niveau 2, quatre jours pour le niveau 3 et cinq jours pour le niveau 4. Si cette condition n'est pas respectée, le RSS est groupé dans le niveau de sévérité immédiatement inférieur sous réserve du respect de la durée de séjour seuil requise pour ce niveau.

De plus, certains GHM particuliers ne se subdivisent pas selon le niveau de sévérité, l'information portée par le sixième caractère étant alors représentée par une lettre. Z indique qu'il n'y a pas de niveau de sévérité, E est réservé à certains GHM dont l'ensemble des séjours se terminent par un décès, J identifie certains GHM dont l'activité est strictement ambulatoire (zéro nuit) et T correspond à certains GHM regroupant les séjours de très courte durée.

Enfin, la dernière version de la fonction de groupage intègre des équivalences de CMA. Il s'agit d'une part du décès dont l'influence est limitée au niveau 1 et d'autre part de l'âge pour lequel trois bornes ont été retenues : < 2 ans, > 69 ans et > 79 ans.

- *Effet de la durée sur le tarif du séjour*

Bien que ne faisant pas strictement partie de la fonction de groupage, il convient en marge de celle-ci de décrire l'influence de la durée de séjour sur le financement du séjour, avec la notion de bornes basses et de bornes hautes pour certains GHM.

En effet, pour chaque GHM, une Durée Moyenne de Séjour (DMS) est déterminée à partir des données des séjours antérieurs, cette DMS étant bornée par des valeurs limites correspondant aux bornes ou extrêmes (borne basse et borne haute).

Lorsque la durée d'un séjour est inférieure à la borne basse admise pour le GHM de ce séjour (situation dite « extrême basse » ou EXB), la valorisation du séjour est minorée, selon le GHM, soit d'un tarif EXB (forfait EXB multiplié par le nombre de jours manquants pour atteindre la borne EXB), soit d'un forfait EXB fixe quel que soit le nombre de jours manquants.

A l'inverse, lorsque la durée d'un séjour est supérieure à la borne haute admise pour le GHM de ce séjour (situation dite « extrême haute » ou EXH), la valorisation du séjour est majorée d'un tarif EXH (forfait EXH multiplié par le nombre de jours supplémentaires).

Ainsi, le RSS groupé comprend suite à la procédure de groupage :

- Le numéro du RUM ayant fourni le DP du RSS ;
- Le GHM du RSS ;
- Le GHS du RSS ;
- La situation de l'hospitalisation pour la durée de séjour par rapport à la borne extrême basse ou à la borne extrême haute.

c) Le Résumé de Sortie Anonyme (RSA)

Il est issu de l'anonymisation du RSS groupé et produit sous le contrôle du médecin responsable de l'information médicale.

Sa création se fait par un module logiciel fourni par l'ATIH, GENRSA pour les établissements de santé publics et les ESPIC (ex-DG), AGRAF-MCO pour les établissements de santé privés (ex-OQN).

Le RSA est toujours un enregistrement unique, y compris dans le cas d'un RSS multi-unité.

L'anonymisation se fait selon un principe de chaînage anonyme qui permet de relier entre elles les hospitalisations d'un même patient, quel que soient le lieu d'hospitalisation (hospitalisations issues d'établissements différents), le secteur

d'hospitalisation (secteur public ou privé) ou le champ d'activité (MCO, SSR, HAD, psychiatrie).

Le chaînage anonyme aboutit à la création d'un numéro anonyme propre à chaque patient. Les hospitalisations d'un même patient peuvent ainsi être identifiées et chaînées mais il demeure impossible d'identifier le patient à partir de son numéro de chaînage.

Les RSA sont transmis mensuellement à l'Agence Régionale de Santé via la plateforme d'échange e-PMSI sous le contrôle du médecin responsable de l'information médicale.

Les données recueillies dans le cadre du PMSI sont protégées par le secret professionnel, sous la responsabilité du médecin DIM de l'établissement.

La création des fichiers et les traitements de données sont soumis à l'avis préalable de la Commission Nationale de l'Informatique et des Libertés (CNIL).

Il convient de noter que le RUM et le RSS sont considérés comme indirectement nominatifs au regard de la loi, leur contenu ne peut ainsi être porté qu'à la connaissance des acteurs légalement ou règlementairement autorisés dans le cadre prévu par leurs fonctions.

Le médecin responsable de l'information médicale sauvegarde le fichier de RSS ; la durée de conservation de tous les fichiers constitués au titre d'une année, non anonymes et anonymes, est de cinq ans.

On comprend dès lors le rôle central joué par le médecin responsable de l'information médicale, qui conseille les praticiens pour la production des informations, organise le recueil, la circulation et le traitement des données médicales et veille à la qualité des données qu'il confronte, si besoin, avec les dossiers médicaux et les fichiers administratifs. Dans le cadre des contrôles, il doit

pouvoir établir la correspondance entre le dossier médical du patient et le numéro du RSS correspondant.

L'ensemble des RSA ainsi transmis à l'ATIH par les établissements de santé sont agrégés pour constituer une base nationale pour chaque année civile, la base nationale PMSI MCO, clé de voûte de notre travail, dont nous détaillons la structure plus avant.

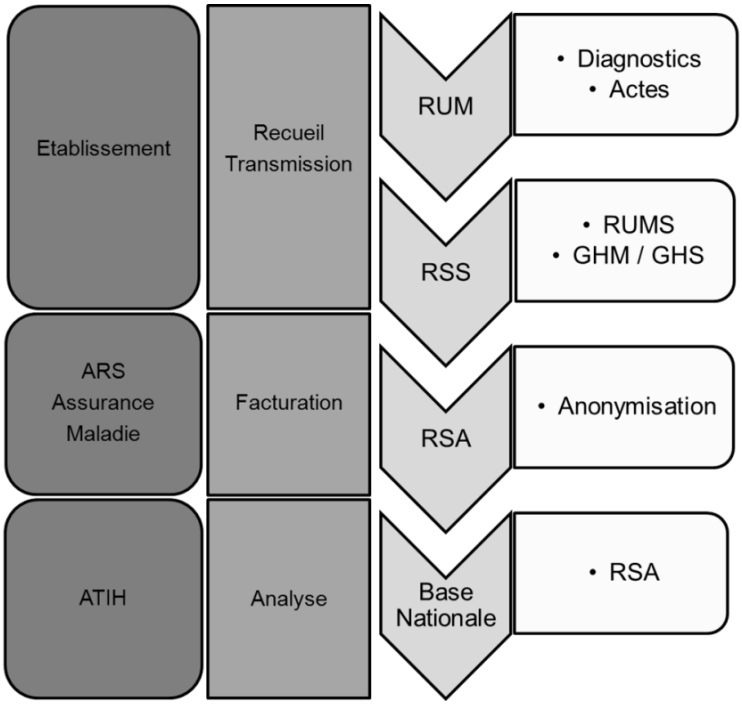


Figure 1 – Etapes du recueil PMSI

.II. Les bases de données de santé en France

A. Les bases nationales PMSI MCO

Les bases nationales PMSI MCO sont issues de la transmission des données d'activités des établissements ayant une activité d'hospitalisation en MCO (24). Suite à la généralisation du PMSI MCO, cette transmission de données est devenue obligatoire à compter du second semestre de 1994 pour les établissements ex-DG et à compter de janvier 1997 pour les établissements ex-OQN, les premières statistiques ayant été réalisées à partir de 1995.

Les bases de données nationales incluent l'ensemble des transmissions faites par les établissements de santé. Depuis 2004, cette procédure se fait par télétransmission internet sécurisée via la plateforme e-PMSI. Seule la dernière transmission en date et validée de chaque établissement pour une année civile est prise en considération pour la constitution de la base de données nationale, les fichiers se devant de cumuler l'activité réalisée à compter du début de l'année.

L'exploitation de ces bases de données par l'ATIH donne lieu à la production de plusieurs présentations statistiques, consultables en ligne sur le site internet de l'ATIH. Ces statistiques résument les différents champs de l'activité MCO : cartes de répartition géographique, statistiques par GHM, fréquence des séjours selon les codes diagnostiques ou d'actes, tableau des molécules onéreuses et dispositifs médicaux implantables de la liste en sus etc.

Surtout, et en plus de l'information agrégée apportée par ces statistiques, les bases nationales PMSI MCO sont, sous conditions légales définies, accessibles pour exploitation.

Les bases MCO disponibles sont celles des années 2005 à 2013. Avec plusieurs millions de séjours enregistrés, ces bases se situent sans conteste dans le champ du big data (25).

Elles constituent une source précieuse d'informations sur les données de santé des populations, leur utilisation pouvant se faire dans d'autres champs que celui de la tarification, tels l'épidémiologie (26,27), l'évaluation de la morbi-mortalité (28,29), la qualité et la sécurité des soins (30,31), voire dans des domaines innovants, comme l'analyse des filières de santé et du parcours de soins des patients (32,33).

La création de ces bases n'a pas été sans faire évoquer les problématiques de confidentialité et d'éthique (34) ; les données de santé sont par essence sensibles et leur exploitation ne peut se faire sans respecter les règles d'anonymat et de secret

professionnel. Face à cette situation propre à susciter les inquiétudes des professionnels et des usagers du système de santé, un cadre législatif strict a été mis en place (35) afin d'encadrer l'utilisation de ces données sans pour autant freiner et réduire les bénéfices attendus en matière de connaissances, de santé publique et de qualité des soins.

Seules les structures disposant d'un numéro SIRET (Système d'Identification du Répertoire des Etablissements), nécessaire pour effectuer une demande, peuvent accéder aux bases, les ARS et les Caisses Nationales d'Assurance Maladie bénéficiant de dispositions particulières.

Toute demande nécessite au préalable l'obtention d'une « autorisation évaluation des pratiques de soins » auprès de la Commission Nationale de l'Informatique et des Libertés (CNIL).

L'obtention du fichier de chaînage (ANO) doit être explicitement spécifiée lors de la déclaration à la CNIL ainsi que sur l'autorisation délivrée par cet organisme.

Afin de préserver l'anonymisation des bases, certaines contraintes peuvent être éventuellement imposées par la CNIL, avec la transformation de certaines variables de la base : discrétisation de l'âge en classes quinquennales, mode de sortie « décédé » remplacé par mode de sortie « à domicile », code géographique de résidence remplacé par le numéro de département.

Le tarif en vigueur pour l'obtention des bases est de 34 centimes d'euro pour mille enregistrements majoré d'un forfait de traitement de 250 €. A titre d'exemple, la base nationale MCO 2013 complète des établissements publics et privés (24.2 millions de RSA validés) est facturée environ 8 510 €.

Le demandeur doit disposer des moyens informatiques, statistiques et des compétences nécessaires pour traiter les bases de données commandées. Il est seul responsable de l'utilisation de ces données.

La structure des bases MCO est calquée sur celle du RSA ; en plus des variables provenant du RSA, il est possible de déterminer pour chaque séjour un identifiant patient anonyme, grâce la procédure de chaînage anonyme (36).

Ce procédé a été initié en 2001 pour les champs MCO et SSR, puis étendu en 2005 à l'HAD et en 2006 à la Psychiatrie.

La production et l'utilisation de cette information sont strictement encadrées par la loi, le dispositif ayant été accepté par la CNIL. A noter que lors d'une demande de bases PMSI, l'obtention des informations de chaînage doit être explicitement spécifiée lors de la déclaration à la CNIL ainsi que sur l'autorisation dispensée par cet organisme.

Les informations de chaînage sont contenues dans un fichier ANO, généré par l'établissement de santé via des outils informatiques fournis par l'ATIH. Une clé de chaînage est créée en appliquant le logiciel MAGIC (Module d'Anonymisation et de Gestion des Informations de Chaînage), intégrant le procédé sécurisé dénommé Fonction d'Occultation des Informations Nominatives (FOIN), fonction créée par la Caisse Nationale d'Assurance maladie des Travailleurs Salariés (CNAMTS) et validée par la CNIL et par l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI).

Le fichier ANO issu de cette procédure contient la clé de chaînage et un numéro administratif de séjour. Ce dernier est couplé aux fichiers PMSI médicaux par l'intermédiaire du numéro administratif avant leur anonymisation et leur télétransmission vers les services de tutelle. Un second hachage est appliqué au numéro anonyme au moment de l'intégration des fichiers dans la plateforme e-PMSI, afin de rendre impossible la mise en correspondance avec les fichiers de chaînage générés par les établissements.

La clé de chaînage ou numéro anonyme est généré à partir du numéro d'assuré social, de la date de naissance et du sexe du patient. Il est composé de 16 caractères, pouvant chacun prendre la valeur d'une lettre majuscule ou d'un chiffre soit potentiellement 36 numéros différents, et d'une lettre, variant en fonction du niveau de l'algorithme FOIN.

Le procédé de chaînage comporte des limites pouvant entraîner une rupture du chaînage (impossibilité de relier entre elles les informations PMSI d'un même patient) ou, plus rarement, un chaînage en excès (résumés PMSI de patients différents reliés entre eux par erreur).

L'utilisation du numéro d'assuré social pose problème dans le cas des éventuels ayant droits (enfants, conjoints, ascendants). Lors d'un changement de statut vis-à-vis de l'Assurance Maladie, le numéro anonyme d'un même patient sera différent.

Plus particulier est le cas des jumeaux de même sexe, car nés le même jour avec le même ayant droit et donc, par construction, avec le même numéro anonyme (0.7 % des naissances).

Les défauts de production de l'information source par les établissements de santé sont également une limite de cette procédure. Cependant, l'obligation d'enregistrer les informations de prise en charge par l'Assurance Maladie dans le fichier de chaînage pour obtenir un paiement, à la suite de l'instauration de la T2A en 2007, a considérablement amélioré l'exhaustivité de la production d'un numéro anonyme pour tout résumé PMSI.

Le procédé de chaînage anonyme est un élément essentiel des bases de données PMSI. En effet, le chaînage permet de relier entre eux les RSA correspondant à un même patient, quels que soient le lieu, le type ou le champ

d’activité de prise en charge hospitalière ; ceci permet entre autres les analyses statistiques par patient et l’étude des trajectoires hospitalières.

B. Les autres bases de l’ATIH

Les bases de données PMSI ne sont pas les seuls enregistrements de données de santé gérés par l’ATIH ; l’agence propose d’autres éléments statistiques sur son site, notamment via le Système National d'Information sur l'Hospitalisation (SNATIH), qui met à disposition des acteurs du système de santé des informations financières et d'activité des établissements hospitaliers publics et privés (37). Ces informations se présentent sous la forme de données médicales statistiques de synthèse issues du PMSI ; sont aussi disponibles des données centrées sur les établissements de santé, relatives à leurs performances et à l’analyse de leur activité. Le SNATIH comprend également l’application Hospi Diag (38), outil d’aide à la décision, qui permet de mesurer la performance d’un établissement de santé dans le champ MCO, en comparaison avec d’autres établissements. Les indicateurs d'Hospi Diag sont calculés par l'ATIH, l’Agence Nationale d'Appui à la Performance des établissements de santé et médico-sociaux (ANAP) assurant l’hébergement, le développement et la maintenance de l’outil.

Parmi les autres outils de restitution proposés par le SNATIH, nous pouvons citer les données de la Direction Générale des Finances Publiques (DGFIP) qui produit des tableaux de bords financiers sur les établissements publics de santé, les données de la Direction de la Recherche, des Etudes, de l’Evaluation et des Statistiques (DREES), via la Statistique Annuelle des Etablissements (SAE) et les données de population provenant de l’Institut National de la Statistique et des Etudes Economiques (INSEE).

C. Les autres bases de données de santé en France

De nombreuses autres bases de données, incluant la dimension médico-économique, sont désormais accessibles. Parmi celles-ci, les bases du Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM) (39) sont particulièrement intéressantes.

Créé en 1999 par la Loi de Financement de la Sécurité Sociale (LFSS), le SNIIRAM est une base de données nationale dont les objectifs sont de contribuer à une meilleure gestion de l'Assurance Maladie et des politiques de santé, d'améliorer la qualité des soins et de transmettre aux professionnels de santé les informations pertinentes sur leur activité.

La CNAMTS est chargée de la gestion de cette base. Elle est responsable du système d'information au regard de la CNIL, qui donne son avis sur le périmètre d'activité du SNIIRAM, ses finalités, son alimentation et l'accès aux données.

Parmi les objectifs dédiés du SNIIRAM, figure ainsi la transmission aux prestataires de soins des informations relatives à leur activité et à leurs recettes, conférant une dimension médico-économique à ces données.

La base de données gérée par l'Institut de Recherche et de Documentation en Economie de la Santé (IRDES) est également remarquable ; l'IRDES, producteur de données et d'analyses en économie de la santé, propose via les bases de données Eco-Santé (40), des séries statistiques dans le domaine sanitaire et social.

Accessibles en ligne, ces bases de données comprennent plus de 7 millions de données chiffrées sur l'économie de la santé : dépenses de santé, état de santé, professions de santé, hôpitaux, secteur pharmaceutique, protection sociale, démographie, indicateurs économiques. Ces données sont disponibles sur le plan national, régional et départemental. Régulièrement mises à jour, elles s'étendent sur des périodes de temps considérables.

Outre les problèmes évidents d'éthique et de confidentialité liés à ces enregistrements (41), la multiplication des bases données dédiées à la santé peut également devenir problématique quant au recensement de toutes ces bases et au recoupement des informations qu'elles contiennent pour une utilisation pertinente et optimale.

Ainsi, des organismes de recensement des bases de données se sont constitués ; ces organismes ont l'avantage de compiler les bases de données existantes, de vérifier la qualité des informations fournies et de permettre un recoupement des informations des différentes bases administrées. Ils fournissent de nombreuses statistiques de santé et gèrent également les demandes d'accès aux informations en se conformant au strict respect de la réglementation en termes de confidentialité et d'éthique.

A titre d'exemple, la plateforme "data.gouv.fr" (42) permet aux services publics de publier des données publiques et à la société civile de les enrichir, modifier, interpréter en vue de coproduire des informations d'intérêt général. Dans le domaine de la santé, cette plateforme propose des données relatives à l'offre et à la consommation de soins, à l'efficacité du système de santé et à la santé publique. Ces séries de données lui sont fournies par le Ministère des Affaires Sociales et de la Santé, l'Assurance Maladie, l'HAS, l'ATIH, les établissements publics sous tutelle et les collectivités territoriales.

Nous pouvons également citer l'Institut des Données de Santé (IDS) (43). Cet organisme est constitué de 14 membres :

- Le Ministère des Affaires Sociales et de la Santé ;
- La Direction Générale de la Compétitivité de l'Industrie et des Services (DGCIS) ;
- La CNAMTS ;

- La Caisse Centrale de la Mutualité Sociale Agricole (CCMSA) ;
- La Caisse du Régime Social des Indépendants (RSI) ;
- La Caisse Nationale de Solidarité pour l'Autonomie (CNSA) ;
- L'Union Nationale des Régimes Spéciaux (UNRS) ;
- L'Union Nationale des Professions de Santé (UNPS) ;
- L'Union Nationale des Organismes d'Assurance Maladie Complémentaire (UNOCAM) ;
- Le Collectif Inter Associatif sur la Santé (CISS) ;
- La Fédération Hospitalière de France (FHF) ;
- La Fédération des Etablissements Hospitaliers et d'Aide à la Personne (FEHAP) ;
- La Fédération de l'Hospitalisation Privée (FHP) ;
- La Fédération Nationale des Centres de Lutte Contre le Cancer (FNCLCC).

Il compte également des membres associés, parmi lesquels on peut citer l'HAS, l'IRDES, l'Ecole des Hautes Etudes en Santé Publique (EHESP) ou encore la Fédération Nationale des Observatoires de Santé (FNORS).

L'IDS s'est fixé pour mission d'améliorer la connaissance du fonctionnement et du financement du système de santé, afin d'optimiser la gouvernance de ce système. Cette mission passe par une utilisation optimale des bases de données de santé par les membres de l'institut et les organismes de recherche, en permettant notamment la mise en commun des données issues des différents organismes et le partage de ces données, dans le strict respect des obligations de secret médical et de respect des libertés individuelles, de l'éthique et de la déontologie.

L'IDS propose actuellement cinq bases de données en partage :

- Les données du SNIIRAM fournies par l'Assurance Maladie ;
- Les données du PMSI fournies par l'ATIH ;
- Les données du SNATIH ;
- Les données comptables des établissements de santé publics fournies par la DGFIP ;
- Les données relatives à la situation financière des établissements privés fournies par la DREES.

D'autres bases de données seront prochainement partagées (données sur le handicap, données issues des mutuelles et des organismes d'assurance maladie complémentaire, données issues des plans « Cancer »).

Le partage de ces données est encadré par une charte de déontologie qui définit les conditions d'utilisation, d'extraction, d'exploitation et de diffusion des données, dans les conditions de respect de l'anonymat des données ; l'IDS entretient dans ce cadre des liens étroits avec la CNIL, il est également en relation avec les usagers de santé via le Collectif Inter Associatif sur la Santé (CISS).

La richesse de l'offre proposée en matière de bases de données incite à réfléchir à de nouveaux outils pour l'exploitation de ces bases, les outils traditionnels atteignant dans ce contexte leurs limites, tant du fait de l'abondance des données que de leur complexité ; parmi ces outils, les méthodes de data mining, encore dénommées fouilles de données, semblent constituer une perspective intéressante. Ayant déjà fait leurs preuves dans des domaines autres que la santé, notamment dans le domaine commercial, ces méthodes suscitent un engouement croissant dans les applications liées à la gestion des bases de données de santé (32,33,44,45).

.III. L’exploitation des bases de données de santé

A. Besoins

Ces dernières années ont vu les quantités de données de santé collectées devenir de plus en plus conséquentes. Ceci a créé un besoin d’analyse et d’interprétation afin d’extraire de ces données de nouvelles connaissances utiles, non sans susciter des interrogations et des craintes, concernant notamment la confidentialité et la protection de ces données sensibles, ainsi que leur détournement pour des utilisations à des fins commerciales (46).

B. Data reuse, big data et data mining

L’exploitation de ces bases se situe en droite ligne du data reuse, c'est-à-dire la réutilisation d’informations à d’autres fins que celles pour lesquelles elles sont détenues ou élaborées (47,48), ces procédés de data reuse suscitant un intérêt croissant dans leurs applications aux données de santé (49–51).

Au vu du volume des données disponibles, ces données entrent également dans la définition du big data, ce qui là encore soulève des questions concernant les outils d’exploitation spécifiques et adaptés au big data (52).

Parmi les méthodes dévolues à cet objectif, les techniques de data mining se sont particulièrement distinguées.

Y. Kodratoff (53) définit le data mining comme un « processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par l'utilisateur-analyste qui y joue un rôle central ».

Autrement dit, le principe du data mining est d’extraire de façon automatique ou semi-automatique, à partir d’importantes quantités de données brutes, des informations inédites et pertinentes, en vue d’une utilisation industrielle ou

opérationnelle de ces informations (54). Il peut également révéler des associations et des tendances et donc servir d'outil prévisionnel.

Ces procédés, utilisés initialement dans le domaine du marketing ont vu leur terrain d'application s'étendre au web usage mining, à la bio-informatique (par exemple en génétique), aux diagnostics médicaux, etc.

Ces méthodes se placent dans le champ de l'extraction de connaissances à partir de données, l'extraction des nouvelles informations dans le cadre du data mining se faisant de telle sorte qu'aucun des attributs sur lesquels se fonde cette extraction ne se voit accorder une place plus importante.

C. Règles d'association

Parmi ces méthodes d'apprentissage, une d'entre elles présente un intérêt tout particulier dans la construction de règles de prédiction simples dans la formulation : les règles d'association.

Ces méthodes ont été développées à l'origine pour l'analyse de bases de données de transactions de ventes, bases constituées de listes d'articles achetés, afin d'identifier les groupes d'articles vendus le plus fréquemment ensemble.

Une forme particulière de règles d'association permet d'introduire une notion de temporalité (c'est-à-dire de conserver une contrainte d'ordre) dans les règles produites, on parle alors de règles d'association temporelles ou règles séquentielles (55).

L'élaboration de ces règles repose sur des algorithmes d'extraction. De nombreux algorithmes ont été développés dans le cadre de ces méthodes d'apprentissage, l'algorithme APriori étant considéré comme précurseur dans ce domaine.

Les règles produites par les méthodes de data mining sont intuitivement faciles à interpréter et à appliquer dans le domaine dont elles sont issues.

Cependant, il faut noter que ces méthodes peuvent aussi produire des règles triviales (déjà bien connues des intervenants du domaine) ou inutiles (car provenant de particularités de l'ensemble d'apprentissage) : on parle alors de « false knowledge discovery ».

La conséquence de cet état de fait est l'obligation de recourir à une étape de validation experte des résultats obtenus, pour ne retenir que les règles pertinentes.

L'apport des méthodes de data mining dans la gestion et l'analyse des bases issues du domaine de la santé pourrait s'avérer très intéressant ; en particulier, ces méthodes trouvent leur utilité dans l'analyse des parcours de soins des patients.

.IV. Le parcours de soins

La notion de parcours de soins constitue une préoccupation émergente de la gestion du système de santé français ; à titre d'exemple, parmi les objectifs du SNIIRAM, figure l'identification des parcours de soins des patients, dans le cadre d'une amélioration de la gestion des politiques de santé (39).

Les parcours de soins peuvent se concevoir comme l'organisation d'une prise en charge globale et continue des patients et usagers, au plus proche de leur lieu de vie (56). Cette notion s'est imposée essentiellement du fait de la progression des maladies chroniques. Elle implique une évolution du système de santé, historiquement axé sur le soin, vers une prise en charge plus complète centrée sur l'individu.

En effet, la prise en charge transversale des maladies chroniques implique de multiples intervenants, la spécialisation croissante des professionnels de santé

amplifiant ce phénomène. L'optimisation des parcours des patients et des usagers s'impose alors comme un axe transversal essentiel, structurant le système de santé.

Un parcours de soins est défini comme la trajectoire globale des patients et usagers dans leur territoire de santé, avec une attention toute particulière portée à l'individu et à ses choix (57).

Ce mode organisationnel nécessite une action coordonnée entre les différents acteurs de la prévention, du sanitaire, du médico-social et du social, intégrant les facteurs déterminants de la santé que sont l'hygiène, le mode de vie, l'éducation, le milieu professionnel et l'environnement.

On peut résumer les objectifs de cette approche par le paradigme suivant : faire en sorte qu'une population reçoive les bons soins par les bons professionnels dans les bonnes structures au bon moment, le tout au meilleur coût.

Cette approche nécessite donc de cartographier les étapes du trajet du patient tout au long de son parcours de santé, en prenant en compte la double dimension temporelle (organiser une prise en charge coordonnée tout au long de la maladie du patient) et spatiale (organiser une prise en charge sur un territoire congruent au bassin de vie du patient).

En plus des maladies chroniques, de nombreuses initiatives ont été engagées sur l'optimisation des parcours dans diverses populations prioritaires : personnes âgées, personnes handicapées, mères et enfants dans le cadre de la périnatalité, jeunes et addictions, personnes en situation de précarité et détenus etc.

Si le parcours d'une personne donnée est par essence unique, on peut, à l'échelle d'une population, repérer et organiser des typologies de parcours a priori en termes d'endroits, de manières, de coût et d'efficacité des interventions de santé, ce qui permet de calibrer et d'anticiper les ressources nécessaires en matière d'offre et

d'organisation des prises en charge, avec la création de réseaux de soins et de professionnels dédiés à la prise en charge de pathologies spécifiques.

Les enjeux de cette démarche s'appréhendent aisément : concevoir des parcours lisibles, accessibles, complets et de qualité, pour une prise en charge globale et coordonnée des patients, dans un esprit d'équité et d'égalité d'accès aux soins, la satisfaction des usagers et des professionnels se trouvant améliorée par l'adaptation de l'offre de soins aux besoins et la fluidification des prises en charge, en respectant les points cardinaux d'efficience et de rationalisation des dépenses de santé.

Ce dernier point mérite d'être approfondi, car il est en lien direct avec les perspectives de réforme des systèmes de financement, en particulier du système de la T2A et du PMSI.

Le rapport 2011 au Parlement sur la Tarification à l'Activité (58) conclut ainsi au caractère indissociable de la T2A et de l'organisation territoriale de l'offre de soins, alors que le rapport 2012 de l'Inspection Générale des Affaires Sociales (IGAS) sur l'évaluation de la tarification des soins hospitaliers et des actes médicaux (59) pointe le fait que les outils tarifaires actuellement en place ne favorisent pas les parcours de soins et recommande de renforcer cette logique de parcours pour réduire entre autres la segmentation des financements.

En réponse à ces constats, les pouvoirs publics ont élaboré des perspectives de financement des parcours de soins, dans un esprit de réforme et d'évolution des systèmes de financement actuels.

Dans cette démarche, la prise en charge des patients ne serait plus réduite à une addition de phases techniques segmentées et évoluerait vers des prises en charge coordonnées et personnalisées en fonction de la pathologie et de l'environnement du patient, en adéquation avec la vie de l'usager, introduisant la

notion de « parcours de vie », renforçant le rôle de l'usager dans sa prise en charge et augurant d'une médecine dite « 4P » : prédictive, préventive, personnalisée et participative (60).

Cette approche innovante a de plus le mérite de prendre en considération les enjeux en termes de création et d'accès aux bases de données de santé, notamment dans le cadre de la télémédecine, de la médecine génomique et personnalisée et des autres volets s'intégrant dans cette logique de financement.

Ces nouveaux modes de prise en charges en appellent ainsi à de nouveaux modes de financement.

Bien sûr, d'autres pistes pour l'amélioration de l'efficience des systèmes de financement sont en parallèle évoquées, telles que celles intégrant la qualité des soins aux modalités de financement (61). Cependant, nous avons particulièrement insisté sur la notion de parcours de soins, car elle se trouve au carrefour des thématiques de financement et d'exploitation des données de santé, les méthodes de data mining appliquées aux bases PMSI faisant partie des outils susceptibles d'être utilisés dans cette optique.

Il est intéressant de noter qu'en contraste avec les méthodes informatiques et statistiques d'évaluation des données PMSI et d'exploitation des données de santé, l'amélioration de la valorisation des séjours des patients demeure en grande partie empirique ; il serait alors légitime, en plus de la partie experte, d'introduire des notions formelles dans la valorisation T2A du financement des séjours.

.V. Contrôle qualité des données médico-administratives

A. Généralités

En réponse à un système de rémunération prospectif, dynamique, concurrentiel et imposant des contraintes de performance aux offreurs de soins, il était légitime et attendu que les établissements développent des stratégies afin d'optimiser la rémunération des prestations réalisées. Cette activité, sous la responsabilité du médecin chargé de l'information médicale, prend la forme d'un ensemble de procédures dites de « contrôle qualité », visant à assurer non seulement l'exhaustivité du recueil, mais également à améliorer son efficience, notamment en recueillant les informations de santé de telle manière que la rémunération allouée soit maximale.

Certaines pathologies et certaines situations cliniques sont ainsi particulièrement lucratives en matière de rémunération, il est donc essentiel que le recueil prenne en compte ces pathologies dès lors qu'elles existent et inclue des stratégies afin de les repérer, tout défaut de recensement ayant forcément des répercussions fâcheuses sur les finances des établissements.

Plusieurs approches se sont développées, de manière plus ou moins empirique et experte (62–64), dans la démarche de contrôle qualité : codage des actes classants pour l'attribution de forfaits liés à des activités spécifiques, correction de DP inappropriés orientant vers des GHM-GHS mieux valorisés, contrôle des séjours EXB et EXH etc. Le codage de CMA représente un volet primordial de ces procédures.

B. Cas particulier des CMA

L'approche la plus courante en matière de contrôle qualité, car plus simple et somme toute logique, consiste à améliorer le codage des DAS qui sont des CMA (65). Cette approche a en plus l'avantage d'améliorer la qualité du codage en termes d'information médicale, ce qui est remarquable si l'on considère l'utilisation de plus en plus fréquente des bases PMSI pour des applications autres que tarifaires ; la fiabilité des bases PMSI ainsi accrue en fait un outil précieux dans le domaine de l'information hospitalière et de la santé publique.

Le codage des CMA se fait le plus souvent à partir du courrier de sortie du patient, parfois à partir du dossier médical. Des procédés informatiques d'aide à ce codage, basés sur l'analyse continue et régulière du recueil PMSI des établissements, ont parfois été développés, notamment par les éditeurs de logiciels d'aide et d'accompagnement au codage. A titre d'exemple, une recherche des séjours exempts de CMA est la forme la plus basique de ce genre de procédure. Nous pouvons aussi citer une procédure se basant sur le repérage des séjours dont la durée est compatible avec une augmentation potentielle du niveau du GHM.

D'autres procédures de valorisation ciblent plus précisément certaines CMA particulièrement fréquentes et/ou valorisantes, en identifiant les situations cliniques et les patients qui sont plus susceptibles de présenter ces CMA (par exemple la dénutrition et les escarres, potentielles CMA de niveau 3, chez les personnes âgées) (66,67).

Il convient de rappeler, comme défini précédemment, que la détermination des CMA par l'ATIH est statistique sur le critère d'augmentation de la Durée Moyenne de Séjour (DMS) d'au moins 2 jours et d'au moins 25 % de la DMS de référence, avec au moins 55 % des séjours avec ce DAS concernés par cette augmentation de la durée de séjour par rapport à leur médiane de référence.

La liste des CMA est de ce fait actualisée pour chaque version de l'algorithme de groupage et publiée chaque année dans le volume 1 du Manuel des GHM édité par l'ATIH (22).

C. Amélioration de la qualité du codage par data mining

Paradoxalement, on observe un contraste entre les méthodes de définition des CMA par l'ATIH, formalisées selon une méthodologie informatique et statistique, et l'identification des CMA les plus valorisantes par les professionnels du codage, qui se base essentiellement sur une méthodologie experte et empirique. Partant de ce constat, il était judicieux de tenter de définir des règles de prédiction de CMA utilisables dans le cadre du contrôle qualité, en se basant sur des méthodes statistiques et informatiques renforçant la couche experte déjà existante, l'élaboration de ces règles pouvant notamment se faire à partir des bases de données de santé existantes, disponibles, importantes et fiables.

A la lumière de ce qui précède, l'utilisation des méthodes de data mining appliquées aux bases PMSI se révélait être une approche intéressante.

L'objectif de ce travail n'était pas de s'affranchir de l'approche experte du contrôle qualité ; comme l'explicite par la suite la méthodologie de ce travail, l'approche experte est indispensable, tant pour le choix des CMA à prédire que pour la validation des règles de prédiction générées.

L'expérience des professionnels du codage demeure indispensable dans la procédure de contrôle qualité des séjours, d'autant plus que le codage doit se faire selon les consignes et les préconisations régulièrement mises à jours du Guide méthodologique (21).

La finalité des règles produites serait de servir de support au contrôle expert des séjours, permettant un repérage ciblé et plus rapide des CMA les plus intéressantes.

.VI. Objectif de l'étude

L'objectif principal de ce travail était donc de construire, à partir des enregistrements de la base nationale PMSI MCO, des règles de prédiction à type de règles d'association et de règles séquentielles, pour la prédiction de CMA définies et sélectionnées selon leur intérêt en termes de valorisation des séjours, dans le cadre du contrôle qualité T2A.

MATERIELS ET METHODES

Le déroulement de notre étude pouvait être schématiquement structuré en plusieurs étapes :

- Sélection des codes diagnostiques à prédire ;
- Construction des règles par data mining ;
- Sélection des règles construites par filtre statistique ;
- Réexécution des règles et sélection de séjours à contrôler ;
- Revue experte des séjours et calcul d'une confiance « qualité ».

Figure 2 – Etapes de travail

.I. Sélection des codes diagnostiques à prédire

Comme indiqué précédemment, les codes diagnostiques prédits devaient bien sûr avoir la propriété de CMA dans l'algorithme de groupage, mais cette propriété nécessaire n'était pas suffisante.

En effet, les codes prédits devaient également être fréquents. D'une part, il était logique de chercher à prédire, dans une logique de valorisation optimale, des CMA fréquentes. D'autre part, les contraintes méthodologiques liées à l'outil informatique

imposaient d'axer la prédiction sur des motifs fréquents ; faute d'évènements nécessaires, il eût été impossible de construire des règles de prédiction valides sur des codes diagnostiques rares, quels que soient leurs niveaux de CMA.

En plus de la fréquence, les codes diagnostiques prédits devaient concerner des pathologies chroniques. La finalité de notre étude étant de construire non seulement des règles d'association mais également et surtout des règles séquentielles, la prise en compte de la dimension temporelle s'avérait plus judicieuse en choisissant de prédire des codes exprimant des pathologies chroniques. Comme ces codes étaient potentiellement présents au long cours et dans plusieurs séjours successifs pour un même patient, il était plus aisé de construire des règles les supportant.

En résumé, les codes diagnostiques à prédire devaient être des CMA fréquentes et chroniques.

Certains travaux se sont attachés à recenser les CMA les plus fréquemment codées et les plus valorisantes ; l'étude Valodiag est une illustration intéressante de ces travaux (68).

L'étude Valodiag avait pour objectif d'attribuer un tarif moyen à chaque diagnostic utilisé en position de DAS, en affectant à ces DAS un rang pondéré par leur fréquence ; ainsi, un DAS très valorisé mais rarement codable ou rarement pris en compte dans le groupage s'éloignait d'autant plus de la première place. Inversement, un DAS, même de valeur moyenne, s'il était très fréquent, remontait dans le classement. Les résultats de cette étude sont résumés dans le tableau en annexe [annexe 1].

Pour la sélection des CMA ciblées par notre étude, nous nous sommes donc basés sur les résultats de l'étude Valodiag ; en effet, le rang attribué par cette étude

aux CMA avait l'avantage de tenir compte non seulement de la valorisation moyenne induite par les CMA mais également de leur fréquence.

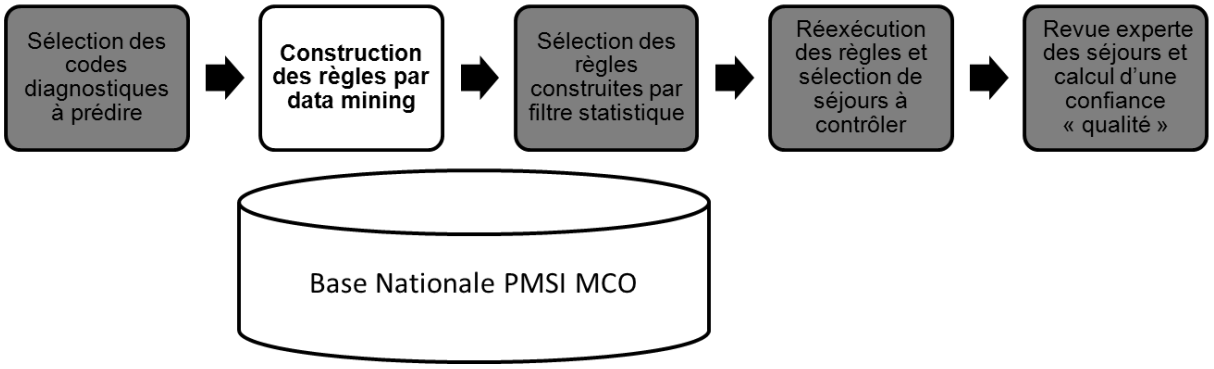
En appliquant le critère de chronicité aux CMA sélectionnées par l'étude Valodiag, nous pouvions retenir 3 CMA à prédire : E1190 « diabète sucré non insulinodépendant insulinotraité sans complications », I48 « fibrillation atriale », I501 « insuffisance ventriculaire gauche ».

Une dernière contrainte nous a imposé de restreindre notre travail à la prédiction de catégories de codes ; pour rappel, la CIM-10 classe les codes en catégories, à savoir des codes constitués de trois caractères alphanumériques, parfois dénommés « codes pères », qui sont eux-mêmes subdivisés en sous-catégories ou « codes fils », codés avec quatre caractères, voire plus selon les différentes mises à jour. Or, la CIM-10 compte actuellement 2049 codes pères, contre 33816 codes fils. De ce fait, baser la prédiction sur les codes fils aurait entraîné une perte d'information et réduit le nombre d'événements dans notre base de données, d'où une réduction de la puissance.

Les déclinaisons de ces codes étant des CMA, pour toutes les déclinaisons du code I50 et pour la plupart de celles du code E11, et ces CMA étant toutes de même niveau de sévérité, ce dans toutes les versions concernées de la classification des GHM, il était donc admissible de baser notre prédiction sur les codes pères.

Au final, les codes sélectionnés étaient donc E11, I48 et I50.

.II. Construction des règles par data mining



A. Echantillon de travail

Nous avons construit notre base de travail à partir des extraits de la base nationale PMSI MCO pour les années 2007, 2008, 2009 et 2010.

En effet, le Centre d'Etudes et de Recherche en Informatique Médicale (CERIM, Equipe d'Accueil 2694, Université Lille2) et le Service de l'Information et des Archives Médicales (SIAM, Clinique de Santé Publique, Centre Hospitalier Régional Universitaire de Lille) ont fait l'acquisition des bases de données PMSI MCO pour les années 2007 à 2010 ; l'autorisation CNIL accordée pour l'exploitation de ces bases inclut le chaînage des séjours d'un même patient, ce qui était indispensable pour la réalisation de notre travail.

La construction de ces bases dérivait de la structure de la base nationale et reprenait certaines variables nécessaires à notre étude ; chaque ligne de la base correspondait à un séjour hospitalier, les variables résumées dans les colonnes étant :

- Un identifiant de séjour, numéro non signifiant crée selon l'ordre chronologique des séjours dans la base ;
- Le numéro FINESS de l'établissement ;
- Le numéro de chaînage anonyme, identifiant patient ;
- L'âge du patient, en années révolues ;

- Le sexe du patient (masculin codé 1, féminin codé 0) ;
- Le GHM du RSA ;
- Le nombre de RUM dans le RSA ;
- La durée de séjour (durée PMSI, exprimée en nombre de nuitées) ;
- Le DP du RSA ;
- Le DR du RSA, éventuellement ;
- Les DAS du RSA ;
- Le nombre de DAS ;
- Les actes CCAM du RSA ;
- Le nombre d'actes CCAM ;
- Le mois de sortie du RSA ;
- L'année de sortie du RSA.

A partir des bases PMSI MCO des années 2007 à 2010, nous avons réalisé par une requête SQL un extrait de séjours afin de constituer notre échantillon de travail.

Cette requête permettait d'extraire, au format texte tabulé, tous les séjours d'un même patient, à condition que le patient ait au moins deux séjours, ceci en prévision de la construction des règles séquentielles.

L'autre condition de la requête concernait les CMA ciblées par notre étude ; au moins un des codes ciblés devait être codé au moins une fois dans au moins un des séjours du patient, quelle que soit la position du code (DP, DR ou DAS).

Il convient de signaler que la variable DAS prenait la forme d'un champ unique dans lequel les codes diagnostiques étaient agrégés en une seule modalité constituée de tous les DAS du RSA séparés par des points-virgules (;). Il en était de même pour la variable actes CCAM. Ceci étant incompatible avec l'analyse que nous projetions de réaliser, il a fallu, à l'aide d'un script automatisé, reformater ces deux

variables afin que chaque diagnostic et chaque acte soit considéré comme une modalité distincte au sein d'une variable distincte (variables DAS1, DAS2, DAS3, acte1, acte2, acte3...).

L'extrait obtenu a ensuite été scindé en deux bases distinctes selon une proportion de 0.7/0.3, pour construire respectivement un échantillon d'apprentissage pour la construction des règles de prédiction et un échantillon de test pour la validation a posteriori des règles construites.

B. Construction des règles

La construction des règles de prédiction reposait sur les méthodes de data mining. Ces méthodes produisent des règles d'association et des règles séquentielles, elles recourent pour cela à des algorithmes d'analyse et de traitement des données.

1. Règles d'association

Les méthodes de data mining permettent d'extraire les règles d'association à partir d'un ensemble de données ou base de données transactionnelles constituant le domaine d'application, base formée par une liste limitée d'éléments qu'on appelle items ; on définit alors une transaction comme un ensemble d'items. L'ensemble des transactions est l'ensemble d'apprentissage permettant la détermination des règles d'associations.

Les règles ainsi créées expriment les possibilités d'association entre différents items.

La création des règles d'association se fait par des algorithmes qui reposent sur les notions de support et de confiance pour évaluer la pertinence des règles.

Si l'on considère T comme une base de données transactionnelle, et A un item issu de T, le support de A se définit comme le nombre de transactions contenant le motif A divisé par le nombre de transactions contenues dans T ou :

$$\text{support}\,(A) = \frac{\text{cardinal}\,(A)}{\text{cardinal}\,(T)}$$

Si l'on considère de même une règle stipulant une association entre un motif A et un item B, la confiance de cette règle d'association est le nombre de transactions contenant l'association A et B divisé par le nombre de transactions contenant le motif A ou[1] :

$$\text{confiance}\,(A => B) = \frac{\text{support}\,(A \cup B)}{\text{support}\,(A)}$$

L'utilisateur détermine préalablement et empiriquement pour ces deux paramètres des seuils minimaux, définissant un support minimal minSupp et une confiance minimale minConf. Les règles d'association générées ne seront retenues que si leur support et leur confiance sont supérieurs aux seuils respectifs. Les items concernés par ces règles seront alors considérés comme des ensembles d'items fréquents.

De nombreux algorithmes sont utilisés pour l'élaboration de règles d'association ; nous présenterons celui qui est à l'origine de cette approche de data mining et qui demeure la référence de cette méthode, à savoir l'algorithme APriori, décrit en 1993 par R. Agrawal (69).

APriori est un algorithme classique de recherche de règles d'association qui repose sur des bases de données transactionnelles, et qui utilise les notions de support et confiance minimaux tels que définis précédemment.

[1] La notation « U » peut prêter à confusion dans son utilisation pour l'expression des règles d'association ; par « U », il est entendu intersection.

L'algorithme démarre en constituant la liste des items les plus fréquents dans la base de données et respectant les seuils de support. Un ensemble de règles, ou candidats, est créé à partir de cette liste. Les candidats sont ensuite testés sur la base de données et ceux ne respectant pas les conditions de minSupp et minConf sont exclus. L'algorithme répète ce processus en augmentant à chaque étape la dimension des candidats d'une unité, ce tant que des règles pertinentes sont découvertes. L'algorithme s'achève par la fusion de l'ensemble des règles découvertes.

Les règles d'association produites s'expriment alors selon la syntaxe {B,C}=>{A}, où {A} constitue l'item prédit et {B,C} le motif prédictif pour la même transaction.

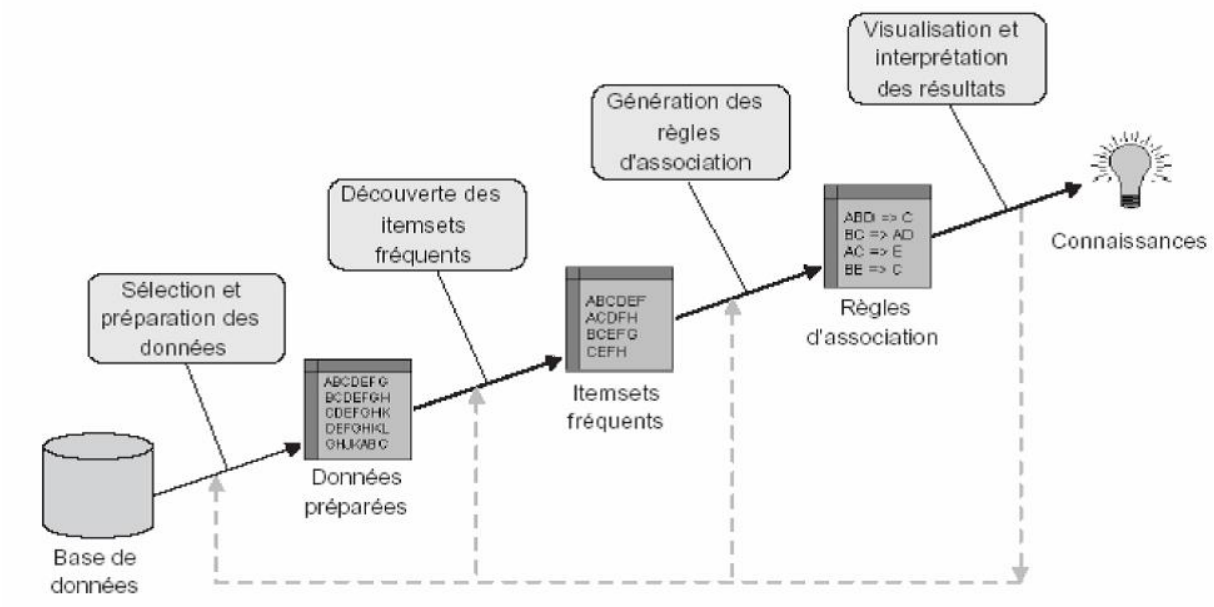


Figure 3 – Etapes d'extraction de règles d'association (d'après Abdellali M. (55))

Dans la logique d'application de l'algorithme APriori à la prédiction de codes diagnostiques, chaque séjour était considéré comme une transaction et chaque variable du séjour comme un item de cette transaction.

L'objectif était donc de détecter des associations fréquentes d'items, les items en question étant les diagnostics (DP, DR et DAS) et les actes CCAM. Ainsi, nous cherchions à retrouver pour chaque code diagnostique ciblé (E11, I48, I50) des associations fréquentes de codes diagnostiques et d'actes CCAM, constituant des motifs prédictifs du code ciblé.

Pour pouvoir appliquer l'algorithme APriori, il fallait donc construire, à partir de notre échantillon d'apprentissage, un tableau de transaction ; dans ce tableau, chaque ligne, à savoir chaque séjour, correspondait à une transaction et contenait l'ensemble des diagnostics et des actes codés dans ce séjour.

Il était nécessaire que nous définissions par avance les valeurs seuils de support et de confiance (minSupp et minConf) ; pour mémoire, le support correspond à la proportion des séjours contenant le motif prédictif sur l'ensemble des séjours de la base, tandis que la confiance correspond à la proportion des séjours contenant le motif prédictif parmi l'ensemble des séjours contenant le code prédit.

Nous avons fixé le seuil de confiance minConf à 0.5, ce seuil nous paraissant raisonnable en matière de fiabilité des règles construites, tandis que le seuil de support minSupp à était fixé à 0.00075, de manière à obtenir des règles concernant au moins 30 séjours.

En retour, l'algorithme APriori devait indiquer, pour chaque motif produit, le support et la confiance de ce motif. Ces valeurs devaient nous être utiles car elles nous permettraient d'opérer une deuxième sélection parmi les règles produites, en ne conservant que celles avec des valeurs optimales de support et de confiance.

La dernière valeur exprimée par l'algorithme APriori pour chaque règle était le lift, défini comme la proportion entre le support de la règle et le support prévu sous l'hypothèse d'indépendance des motifs constituant la règle, soit :

$$\text{lift}\ (A => B) = \frac{\text{support}\ (A \cup B)}{\text{support}\ (A)\ \text{support}\ (B)}$$

Ce paramètre reflète la valeur informative d'une règle, la force d'association entre les items d'une règle étant proportionnelle à la valeur du lift de cette règle (70). Cette valeur était particulièrement intéressante à utiliser pour réaliser dans un second temps un nettoyage de notre échantillon de travail.

2. Règles séquentielles

Une règle séquentielle est une règle d'association complétée par un facteur temporel. En effet, une séquence utilise le principe de précédence, c'est-à-dire que le motif prédictif précède dans le temps le motif prédit (55).

Si l'on considère deux motifs {ae} et {bc} dans une liste de transactions, le fait que le motif {ae} prédit le motif {bc}, ou {ae}=>{bc}, implique que le motif {ae} survienne avant le motif {bc}, d'où le caractère prédictif implémenté à la règle édictée.

Ainsi, la construction de tels motifs repose sur l'extraction d'ensembles d'items couramment associés sur une période de temps spécifiée ; ce procédé dévoile des associations inter-transactions, contrairement à celui des règles d'association qui extrait des combinaisons intra-transactions.

Les notions de support et de confiance minimaux sont également employées pour la génération des règles séquentielles, les seuils de minSupp et minConf étant là aussi fixés préalablement par l'utilisateur.

L'extraction des motifs séquentiels est plus ardue que celle des règles d'association classiques, la notion de temporalité prise en compte dans ces règles améliorant la précision et l'utilité des règles créées, mais impliquant aussi une plus grande difficulté d'implémentation pour introduire la notion d'ordre chronologique.

Plusieurs techniques algorithmiques permettent la création de règles séquentielles, l'algorithme le plus commun, « pionnier » dans le domaine, étant l'algorithme SPADE (Sequential PAttern Discovery using Equivalence classes), conçu en 2001 par M.J. Zaki (71).

Les règles séquentielles produites s'expriment toujours selon la même syntaxe, {B,C}=>{A}, {A} constituant l'item prédit et {B,C} le motif prédictif, en gardant à l'esprit que si le motif {B,C} se situe au niveau d'une transaction donnée, l'item {A} est au niveau d'une transaction ultérieure.

La logique d'application de l'algorithme SPADE était similaire à celle de l'algorithme APriori, chaque séjour étant considéré comme une transaction constituée d'items représentés par les codes diagnostiques et d'actes, cet algorithme se distinguant cependant par la prise en compte de la dimension temporelle dans la construction des règles prédictives.

En adéquation avec notre travail, les règles séquentielles devaient être construites de telle manière que la présence d'un motif prédictif, à savoir une association de codes diagnostiques et/ou d'actes, dans un séjour donné n, indiquerait la présence du motif prédit, l'une des CMA ciblées, dans le séjour suivant n+1. Il convient de signaler que le code diagnostique prédit pouvait entrer dans la constitution du motif prédictif, en association avec d'autres codes diagnostiques et/ou d'actes ; de même, il était prévisible d'obtenir des règles séquentielles ne se basant que sur le motif prédit, et évaluant le pouvoir prédictif d'une CMA, seule considérée dans un séjour, sur sa présence dans le séjour suivant.

Afin de prendre en compte la temporalité des règles séquentielles, il était dès lors nécessaire d'inclure au tableau de transaction une nouvelle variable qui indiquait, pour un même patient, l'ordre chronologique des séjours. Cette variable

d'ordre a été construite à partir d'une concaténation de l'identifiant de séjour, du mois et de l'année de sortie du séjour.

Les notions de confiance et de support, avec les seuils afférents de minConf et de minSupp définis à l'avance par l'utilisateur, s'appliquaient également aux règles séquentielles, avec cependant quelques particularités. La confiance dans ce contexte se définissait comme la proportion des séjours contenant le motif prédictif et pour lesquels le code ciblé se retrouvait dans le séjour suivant, par rapport à l'ensemble des séjours contenant le motif prédictif. De même, il fallait distinguer deux supports, le support du motif prédictif suppX, proportion des séjours contenant le motif prédictif sur l'ensemble des séjours et le support du motif prédit suppY, proportion des séjours contenant le motif prédit et associés au motif prédictif sur l'ensemble des séjours.

Nous avons fixé le seuil de confiance minConf à 0.5 et les seuils de supports minSuppX et minSuppY à 0.00075 chacun.

Comme pour l'algorithme APriori, l'algorithme SPADE devait indiquer pour chaque règle séquentielle les supports et la confiance.

C. Outils de travail

Les règles de prédiction devaient être élaborées sur le logiciel R© version 3.0.2. (72) et les scripts d'analyse écrits sur le logiciel Notepad++© version 6.5.1.

La construction des règles d'association se faisait par la fonction APriori du package arules (73) qui se base sur l'algorithme du même nom précédemment décrit.

La construction des règles séquentielles devait se faire par la fonction SPADE du package arulesSequences (74) reposant sur l'algorithme éponyme.

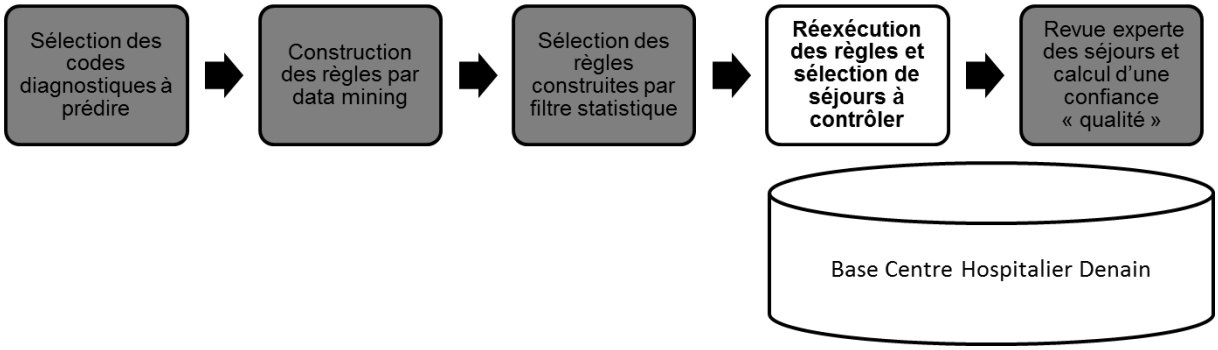
.III. Sélection des règles construites par filtre statistique

Nous nous attendions à extraire de notre échantillon d'apprentissage un certain nombre de règles non informatives, car issues de consignes de codage rendant obligatoire le codage d'associations de diagnostics ou d'actes, et par cet état de fait particulièrement fréquentes, mais sans intérêt car artificiellement induites, et masquant de plus les règles informatives. A titre d'exemple, nous pouvons citer le codage de l'accouchement en obstétrique (association des codes O80 « accouchement unique et spontané » et Z37 « résultat de l'accouchement ») ou le codage des suppléments pour injection de produits de contraste pour les actes de radiologie.

Plutôt que d'exclure les codes concernés selon une procédure experte, potentiellement discutable car source d'erreurs, nous avons projeté de réaliser une première extraction de règles d'association, puis d'exclure de l'échantillon d'apprentissage les codes diagnostiques et actes associés aux règles ayant une valeur de lift basse, en prenant pour seuil de lift le premier quartile de l'ensemble des lifts ; une fois ce « bruit de fond » éliminé, la construction de nouvelles règles d'association à partir de l'échantillon épuré devait permettre d'obtenir des résultats plus informatifs.

L'application du filtre statistique devait ensuite se baser sur les valeurs de confiance et de support des règles ; afin d'obtenir le meilleur compromis entre ces deux paramètres, nous avons décidé d'opérer la sélection à partir des valeurs optimales du produit (confiance * support) associé aux règles.

.IV. Réexécution des règles et sélection de séjours à contrôler



Cette partie de notre méthode avait pour objet de valider les règles d'association et les règles séquentielles produites. Elle devait être réalisée à partir de l'échantillon de test, le but étant d'évaluer la pertinence des règles produites en appréciant leur confiance réelle et surtout en les confrontant aux données des courriers médicaux issus des séjours analysés.

Dans un premier temps, nous prévoyions de calculer, pour chaque règle, la confiance sur l'échantillon de test. Il était ainsi prévu, via un script R, de reconstruire chaque règle et de confronter la confiance « test » à la confiance « apprentissage » associée à la règle évaluée.

La deuxième partie de cette étape d'évaluation portait uniquement sur les règles séquentielles et nécessitait un retour aux courriers de sortie des séjours ; l'objectif de cette démarche était de parvenir à une appréciation réelle et pratique de la valeur des règles séquentielles en termes de repérage des codes ciblés par notre étude, dans une perspective d'application de ces règles au contrôle qualité T2A des séjours.

Il est judicieux de rappeler à ce stade l'objectif principal de ce travail, à savoir la construction de règles pour la prédiction de CMA, applicables dans le cadre du contrôle qualité pour la valorisation des séjours. Nous avions décidé de choisir les

CMA prédites sur, entre autres, le critère de chronicité. En effet, à partir du moment où une CMA était représentée par une pathologie chronique, il apparaissait légitime et licite, pour un patient donné, de pouvoir coder cette CMA pour tous les séjours ultérieurs de ce patient. Le fait de recourir à des règles séquentielles pour la prédiction des CMA s'avérait alors particulièrement intéressant, puisque l'application de la règle à un séjour donné permettait d'espérer une valorisation, non pas sur un séjour uniquement, mais sur tous les séjours suivants, y compris les potentiels séjours à venir de ce patient.

Pourtant, s'agissant de pathologies chroniques, il apparaissait, selon toute vraisemblance et toute logique, qu'une CMA donnée suffisait par sa seule présence dans un séjour à prédire son existence au sein des séjours suivants.

Il s'agissait donc non seulement d'évaluer la qualité des règles séquentielles en matière de valorisation, mais également de confronter et de comparer la qualité et le pouvoir prédictif de ces règles au pouvoir prédictif de la CMA codée seule.

Il était dès lors indispensable de pouvoir accéder aux courriers de sortie des séjours afin de procéder à cette évaluation.

Nous disposions d'une autorisation pour l'exploitation des courriers de sortie de séjours s'étant déroulés au sein d'un centre hospitalier de test, pour les années 2007 à 2014, ces courriers étant dûment anonymisés.

Nous prévoyions donc, par une requête SQL sur la base du centre hospitalier de test, de réaliser pour chaque règle prédictive évaluée une requête d'extraction de séjours à contrôler afin d'apprécier la pertinence des règles.

.V. Revue experte des séjours et calcul d'une confiance qualité

L'évaluation du pouvoir prédictif des règles devait se faire en extrayant pour chaque règle séquentielle une liste de séjours à contrôler, de telle sorte que pour une séquence de séjours liée à un patient donné, le motif prédictif soit présent au niveau d'un séjour donné n et le motif prédit absent au niveau du séjour suivant n+1, c'est-à-dire {B,C}≠>{A}.

Parmi ces séjours, il faudrait alors, en se référant au courrier de sortie, rechercher ceux qui avaient réellement la CMA prédite mais qui n'était finalement pas codée. Dès lors, le pouvoir prédictif de la règle de contrôle pouvait s'apprécier par la proportion de séjours pour lesquels la CMA était légitimement codable par rapport à l'ensemble des séjours dans lesquels la CMA était absente, cette proportion étant équivalente à une confiance « qualité ».

Comme on peut le constater, il s'agissait là d'une démarche pratique et pragmatique s'inscrivant en droite ligne du contrôle qualité T2A, à savoir identifier grâce à une règle de contrôle des séjours potentiellement valorisables, puis rechercher la présence de la CMA potentiellement manquante dans le courrier de sortie.

L'étape ultime de cette évaluation consistait à réaliser une évaluation du pouvoir prédictif de la CMA seule selon la même démarche, c'est-à-dire repérer pour tous les séjours n ayant une CMA donnée, les séjours n+1 n'ayant pas cette même CMA, c'est-à-dire {A}≠>{A}.

Parmi ces séjours, la recherche de ceux ayant réellement la CMA prédite mais non codée et le calcul de la proportion de ces séjours, permettait d'estimer la confiance « qualité » de la CMA seule en matière de règles de contrôle.

Il ne resterait plus alors qu'à confronter la confiance « qualité » des motifs séquentiels à celle des CMA seules et d'évaluer le gain potentiel en matière de pourcentage de séjours recodés par la formule :

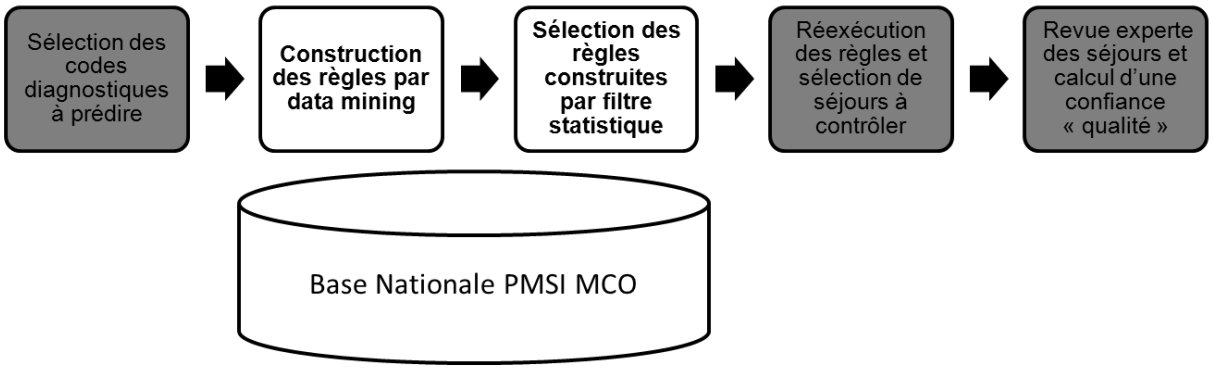
$$\text{gain de recodage} = \frac{(\text{confiance qualité règle}) - (\text{confiance qualité CMA})}{\text{confiance qualité CMA}}$$

Cette dernière étape de validation devait permettre d'opérer une sélection finale des règles prédictives, en ne conservant, par rapport à la CMA seule, que celles avec un gain substantiel de recodage de CMA.

Pour chaque CMA ciblée, en l'occurrence E11, I48 et I50, nous espérions, par l'ensemble de cette procédure, aboutir à la construction de règles de prédiction validées, efficientes et applicables au contrôle qualité et à la valorisation des séjours dans le cadre de la Tarification à l'Activité.

RESULTATS

.I. Construction des règles par data mining et sélection par filtre statistique



A. Echantillon de travail

Suite à l'application des conditions précédemment énoncées, nous avons construit un échantillon de travail à partir de la base nationale PMSI MCO pour les années 2007 à 2010.

Cet échantillon comptait 59170 séjours pour 12125 patients ; ses caractéristiques au niveau « patient » et au niveau « séjour » sont résumées dans le tableau suivant.

		Patients	Séjours
Nombre		12125	59170
Age moyen (ans)		51	50.7
Sexe	Homme	5134 (42.3 %)	26866 (45.4 %)
	Femme	6991 (57.7 %)	32304 (54.6 %)
Nombre moyen séjours par patient			5
Année	2007		8623 (14.6 %)
	2008		14089 (23.8 %)
	2009		16990 (28.7 %)
	2010		19468 (32.9 %)
Durée moyenne séjour PMSI (jours) *			4.4
Durée moyenne séjour réelle (jours)			5.4
Nombre moyen diagnostics par séjour			4
Nombre moyen actes par séjour			4

Tableau 1 – Descriptif de l'échantillon de travail

Notons par ailleurs que les deux GHM les plus fréquents ex aequo étaient 14Z02A « accouchements par voie basse sans complication significative » et 15Z05A « nouveau né de 2500 gr et plus sans problème significatif », avec 1594 séjours (2.7 %) pour chaque GHM. Le DP le plus fréquent était Z380 « Enfant unique, né à l'hôpital », avec 1802 séjours (3 %).

Nous avons par la suite scindé notre échantillon en base d'apprentissage et en base de test, selon une proportion 0.7/0.3, en respectant le chaînage des séjours par patient.

Nous avons ainsi constitué notre échantillon d'apprentissage (40876 séjours) et notre échantillon de test (18294 séjours). Nous avons par ailleurs, ainsi que prévu, restreint les codes diagnostiques aux codes pères.

B. Règles d'association

Conformément à la procédure prévue, nous avons commencé par appliquer la fonction APriori à l'échantillon d'apprentissage, ce afin d'extraire un premier jeu de

règles d'association pour chaque CMA ciblée, en se focalisant sur les valeurs de lift, afin d'exclure les codes diagnostiques et d'actes non informatifs.

A l'issue de cette première étape, les codes diagnostiques et d'actes exclus étaient les suivants :

- K29 « Gastrite et duodénite » ;
- O70 « Déchirure obstétricale du périnée » ;
- O80 « Accouchement unique et spontané » ;
- Z37 « Résultat de l'accouchement » ;
- Z38 « Enfants nés vivants, selon le lieu de naissance » ;
- AFLB010 « Anesthésie rachidienne au cours d'un accouchement par voie basse » ;
- DEQP007 « Surveillance continue de l'électrocardiogramme par oscilloscopie et/ou télésurveillance, avec surveillance continue de la pression intraartérielle et/ou de la saturation artérielle en oxygène par méthodes non effractives, par 24 heures » ;
- HEQE002 « Endoscopie oesogastroduodénale » ;
- HZHE002 « Biopsie et/ou brossage cytologique de la paroi du tube digestif ou de conduit biliopancréatique, au cours d'une endoscopie diagnostique » ;
- JMCA002 « Suture immédiate de déchirure obstétricale du vagin, de la vulve et/ou du périnée [périnée simple] » ;
- YYYY030 « Supplément pour réalisation d'un examen radiographique à images numérisées » ;
- YYYY467 « Supplément pour injection intraveineuse de produit de contraste au cours d'un examen radiographique ou scanographique » ;

- ZBQK002 « Radiographie du thorax » ;
- ZZLP025 « Anesthésie générale ou locorégionale complémentaire niveau 1 ».

Une fois cette étape d'épuration réalisée, nous avons extrait, pour chaque CMA ciblée, les règles d'association correspondantes.

Nous avons ainsi obtenu six règles d'association pour E11, 57 règles d'association pour I48 et neuf règles d'association pour I50.

Pour sélectionner les règles selon nos critères statistiques, nous avons choisi de ne conserver, pour les étapes ultérieures de notre travail, qu'une règle d'association par CMA ciblée, en se basant comme prévu sur la valeur du produit (confiance * support).

Au final, les 3 règles sélectionnées sont résumées dans le tableau suivant.

CMA prédites	Motifs prédictifs	Libellés	Support	Confiance apprentissage
E11	N08	Glomérulopathies au cours de maladies classées ailleurs	0.0022	0.56
I48	DERP003	Choc électrique cardiaque transcutané, en dehors de l'urgence	0.0017	0.93
I50	DZQM006 ; J90	Échographie transthoracique du coeur et des vaisseaux intrathoraciques Épanchement pleural, non classé ailleurs	0.0017	0.50

Tableau 2 – Règles d'association sélectionnées

C. Règles séquentielles

Toujours à partir de l'échantillon d'apprentissage après exclusion des codes diagnostiques et d'actes non informatifs, nous avons extrait par la fonction SPADE, pour chaque CMA ciblée, les règles séquentielles correspondantes.

Nous avons ainsi obtenu 28 règles séquentielles pour E11, 41 règles séquentielles pour I48 et huit règles séquentielles pour I50.

Pour sélectionner statistiquement les règles à conserver pour la suite de notre travail, en se basant toujours sur la valeur du produit (confiance * support) et en

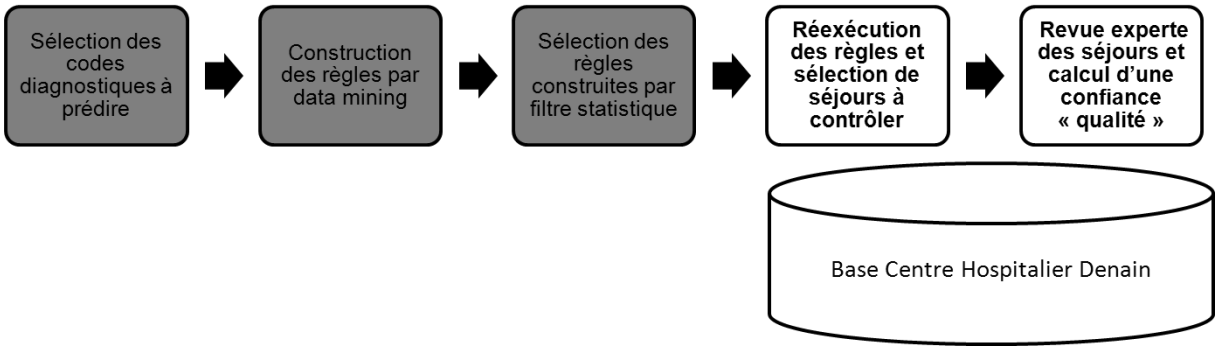
tenant compte de la proportion du nombre de règles allouées à chaque CMA sur l'ensemble des règles produites, nous avons décidé de conserver deux règles séquentielles pour E11, trois règles séquentielles pour I48 et une règle séquentielle pour I50.

Les six règles sélectionnées sont résumées dans le tableau suivant. Nous présentons également les règles séquentielles propres à chaque CMA seule en termes de motif prédictif ; bien que ne satisfaisant pas au seuil de confiance minimal, ces règles fournissaient un point de comparaison utile pour l'évaluation des règles séquentielles.

CMA prédites	Motifs prédictifs	Libellés	Support X	Support Y	Confiance apprentissage
E11	E11	Diabète sucré non insulinodépendant	0.0399	0.0717	0.55
	I10 ; DZQM006 ; E11	Hypertension essentielle (primitive) Échographie transthoracique du coeur et des vaisseaux intrathoraciques Diabète sucré non insulinodépendant	0.0049	0.0069	0.71
	I10 ; I48 ; E11	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Diabète sucré non insulinodépendant	0.0037	0.0052	0.72
I48	I48	Fibrillation et flutter auriculaires	0.0292	0.0571	0.51
	I10 ; I48 ; E78	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Anomalies du métabolisme des lipoprotéines et autres lipidémies	0.0021	0.0035	0.60
	I10 ; I48 ; Z95	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Présence d'implant et de greffe cardiaques et vasculaires	0.0021	0.0035	0.60
	I48 ; I69	Fibrillation et flutter auriculaires Séquelles de maladies cérébrovasculaires	0.0023	0.0037	0.62
I50	I50	Insuffisance cardiaque	0.0121	0.0321	0.37
	I10 ; I48 ; I50	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Insuffisance cardiaque	0.0018	0.0037	0.50

Tableau 3 – Règles séquentielles sélectionnées

.II. Réexécution des règles et revue experte des séjours



Comme prévu par notre méthode, cette partie a été réalisée sur l'échantillon de test, soit un échantillon de 18294 RSA.

Les résultats du calcul pour chaque règle sélectionnée de la confiance « test » sont résumés dans le tableau suivant.

	CMA prédites	Motifs prédictifs	Libellés	Confiance apprentissage	Confiance test
Règles d'association	E11	N08	Glomérulopathies au cours de maladies classées ailleurs	0.56	0.67
	I48	DERP003	Choc électrique cardiaque transcutané, en dehors de l'urgence	0.93	0.75
	I50	DZQM006 ; J90	Échographie transthoracique du coeur et des vaisseaux intrathoraciques Épanchement pleural, non classé ailleurs	0.50	0.50
Règles séquentielles	E11	I10 ; DZQM006 ; E11	Hypertension essentielle (primitive) Échographie transthoracique du coeur et des vaisseaux intrathoraciques Diabète sucré non insulinodépendant	0.71	0.38
		I10 ; I48 ; E11	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Diabète sucré non insulinodépendant	0.72	0.30
	I48	I10 ; I48 ; E78	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Anomalies du métabolisme des lipoprotéines et autres lipidémies	0.6	0.33
		I10 ; I48 ; Z95	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Présence d'implant et de greffe cardiaques et vasculaires	0.60	0.44
		I48 ; I69	Fibrillation et flutter auriculaires Séquelles de maladies cérébrovasculaires	0.62	0.36
	I50	I10 ; I48 ; I50	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Insuffisance cardiaque	0.50	0.50

Tableau 4 – Contrôle des règles sélectionnées

Finalement, dans le cadre de la dernière partie de notre travail, nous avons, pour chacune des six règles séquentielles sélectionnées, extrait une liste de couples

de séjours n et n+1 issus du centre hospitalier de test, de telle sorte que pour les séjours n ayant un motif prédictif, les séjours n+1 correspondants n'aient pas le code prédit. Parmi ces séjours, en se référant au courrier de sortie, nous avons alors déterminé ceux qui avaient réellement la CMA prédite, mais qui n'était finalement pas codée, et calculé la proportion de ces séjours pour lesquels la CMA était légitimement codable.

Nous avons ensuite, pour chaque CMA ciblée, répété la même démarche en utilisant uniquement la CMA au niveau des séjours n comme motif prédictif de sa propre présence au niveau des séjours n+1.

Enfin, nous avons déterminé le gain potentiel en termes de recodage de CMA apporté par l'application des règles séquentielles, par rapport aux CMA seules utilisées comme motifs prédictifs.

Cette dernière évaluation nous a permis d'opérer la sélection ultime de trois règles séquentielles, en ne conservant que les règles engendrant une augmentation significative du recodage de CMA.

Le tableau suivant résume les résultats de cette dernière étape et les règles séquentielles retenues à l'issue de notre travail.

Motifs prédictifs	Libellés	N séjours contrôlés	Confiance qualité	Gain de recodage
E11	**Diabète sucré non insulino-dépendant**	117	0.53	
E11 ; I10 ; DZQM006	Hypertension essentielle (primitive) Échographie transthoracique du coeur et des vaisseaux intrathoraciques Diabète sucré non insulinodépendant	32	0.69	29.7 %
E11 ; I10 ; I48	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Diabète sucré non insulinodépendant	20	0.75	41.5 %
I48	**Fibrillation et flutter auriculaires**	92	0.30	
I48 ; I10 ; E78	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Anomalies du métabolisme des lipoprotéines et autres lipidémies	16	0.25	- 17.8 %
I48 ; I10 ; Z95	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Présence d'implant et de greffe cardiaques et vasculaires	25	0.24	- 21.1 %
I48 ; I69	Fibrillation et flutter auriculaires Séquelles de maladies cérébrovasculaires	23	0.39	28.6 %
I50	**Insuffisance cardiaque**	70	0.21	
I50 ; I10 ; I48	Hypertension essentielle (primitive) Fibrillation et flutter auriculaires Insuffisance cardiaque	37	0.21	0.9 %

Tableau 5 – Pouvoir prédictif des règles séquentielles

DISCUSSION

En accord avec l'objectif principal de notre étude, nous avons suite à notre travail extrait trois règles séquentielles de prédiction de CMA, applicables au contrôle qualité pour la valorisation des séjours ; deux de ces règles concernent le code E11 et une règle le code I48. Ces règles ont montré un gain considérable dans la proportion de séjours pouvant être recodés.

Nous avons choisi de constituer notre échantillon de travail à partir de la base nationale PMSI MCO, puisqu'elle offrait une source de données disponible, fiable, abondante, et pour laquelle nous disposions des autorisations CNIL accréditant son utilisation, y compris et surtout en termes d'informations de chaînage des séjours par patient, condition préalable indispensable à l'aboutissement de notre projet.

La constitution de la base nationale n'est bien sûr pas exempte de failles, sources potentielles de biais et d'erreurs pour notre étude.

Rappelons en premier lieu que l'exhaustivité et la fiabilité de la base nationale sont avant tout tributaires de la qualité du codage des diagnostics et des actes réalisé au niveau des établissements de santé. Le guide méthodologique précise bien que la réalisation de ce codage revient au praticien responsable des soins du patient, le médecin chargé de l'information médicale veillant à l'organisation du recueil, au respect de l'exhaustivité et des règles et consignes de codage, et accompagnant les différents intervenants dans leurs missions. La mise en place de la T2A ne s'est pas faite sans rencontrer des réticences (6,67), liées non seulement aux questions de financement et de confidentialité des données, mais également à la charge de travail supplémentaire imputée au personnel soignant, ceci dans un

contexte préalable d'augmentation des tâches administratives pour les soignants (75,76). Il va sans dire que cet état des lieux n'était pas favorable à la production d'un codage et d'informations de qualité.

Néanmoins, la mise en place progressive de la T2A a permis le renforcement des services chargés de l'information médicale, tant en termes de moyens humains que techniques ; dès lors, le déploiement des missions du médecin chargé de l'information médicale, secondé par les techniciens d'information médicale, avec le soutien de nouvelles applications informatiques, ont considérablement amélioré la qualité du codage, de même que l'évolution des règles de codage régulièrement mises à jour par l'ATIH en réponse aux difficultés rencontrées et aux évolutions du paysage sanitaire et hospitalier. Au bout de sept années de financement total des activités MCO par la T2A, il est raisonnable de considérer les informations de la base nationale comme fiables et de qualité.

Il serait bien sûr illusoire de prétendre à une exhaustivité complète et exempte de toute erreur. Toutefois, la base nationale PMSI MCO constituait bien une source intéressante de données pour la réalisation de notre étude. De plus, rien n'indiquait que les codes ciblés par notre étude puissent être particulièrement manquants et sous codés, ce qui aurait été à l'origine d'un biais important. Tout ceci a conforté notre décision de recourir à cette base pour notre travail.

Le chaînage des séjours par patient pouvait également s'avérer imparfait, tel que le rappelle le guide de la procédure de chaînage édité par l'ATIH ; la proportion de séjours concernés reste toutefois très faible. De plus, nous n'avons pas constaté a posteriori d'incohérences flagrantes entre les différents séjours d'un même patient lors des procédures de contrôle des règles produites, ce qui nous rassure à propos de la qualité des données de chaînage.

Nous nous sommes limités dans la construction de notre extrait de travail aux données des années 2007 à 2010 ; il aurait été préférable d'étendre notre étude aux années ultérieures, ce qui aurait augmenté la quantité de données et donc la puissance de notre étude, mais nous ne disposions pas, au moment du lancement de l'étude, des bases pour les années ultérieures.

Compte tenu des limites de nos outils de travail en matière de capacité de traitement des données, il serait de plus malaisé d'étendre notre étude à d'autres années. Nous avons en effet exploité au maximum les performances du logiciel R dans la réalisation de notre étude. Augmenter la taille de l'échantillon traité nécessiterait de recourir à d'autres outils et d'autres méthodes de traitement de l'information.

Concernant ces mêmes outils de travail, l'originalité de notre travail résidait dans le recours aux méthodes de data mining pour l'élaboration de règles de contrôle qualité. Cette approche se voulait inédite, et avait le mérite de contraster avec les approches classiques de contrôle T2A, notamment en matière de détection de CMA, qui consistent essentiellement à rechercher les CMA sur le courrier de sortie des patients.

Les méthodes de data mining nous semblaient particulièrement appropriées dans ce contexte, car fiables, rapides, susceptibles de traiter de grandes quantités de données, relativement économes en termes de ressources et pour lesquelles nous disposions d'outils informatiques que nous maîtrisions bien ; par ailleurs, la structure même de la base nationale et les informations que nous comptions utiliser, à savoir une répétition d'associations de codes, évoquant spontanément la structure d'un tableau de transactions, se prêtaient parfaitement à la construction de règles d'association. Signalons que d'autres méthodes d'apprentissage auraient pu être envisagées (arbres de décision, méthodes de clustering etc.) (54).

Il apparaissait judicieux pour l'élaboration des règles de prédiction de recourir à des fonctions se basant sur les algorithmes de références, en l'occurrence les algorithmes APriori et SPADE. Il convient de noter que les méthodes de data mining peuvent s'appuyer sur d'autres algorithmes (54) (ECLAT, GSP, VPSP...), dont certains réputés efficaces et offrant des possibilités supplémentaires. Toutefois, les algorithmes APriori et SPADE, références dans les méthodes de data mining, constituaient des outils éprouvés et adaptés à notre travail pour lesquels nous pouvions raisonnablement opter.

Un aspect critiquable de notre méthodologie était de restreindre la constitution des règles aux seuls codes pères des diagnostics. Ce choix réduisait forcément la quantité d'information apportée par notre étude, certains codes pères étant imprécis avec des codes fils sensiblement différents, tels les codes en Z ou en T à titre d'exemple. Ces derniers codes ne sont cependant pas les plus fréquents de la CIM-10, et la majorité des codes pères sont porteurs d'une information satisfaisante ; ajoutons à cela que pour les codes ciblés par notre étude, les déclinaisons en codes fils étaient de même niveau de CMA et porteuses d'une information équivalente et concordante avec les codes pères. Surtout, cette restriction aux codes pères était indispensable pour avoir un nombre suffisant de motifs dans notre base, et par là même créer des règles informatives et pertinentes. Au vu de cet objectif, la perte d'informations induite par cette troncature était acceptable.

Le même raisonnement prévaut en ce qui concerne les exclusions de codes diagnostiques et d'actes associés à des règles non informatives ; il est plausible que ces exclusions aient entraîné la perte de règles associées à ces mêmes diagnostics et actes, mais cette étape nous avait également semblé incontournable, la conservation des codes en question produisant un grand nombre de règles sans

intérêt réel et masquant les règles informatives. Le gain en termes de qualité des résultats et de lisibilité compensait là aussi la perte d'information induite.

Les critères de choix des CMA ciblées pouvaient porter à réflexion. L'étude Valodiag a été d'une aide précieuse dans ce choix, puisque le classement des CMA qu'elle proposait tenait compte à la fois du tarif attribué à un diagnostic et de sa fréquence. L'argument de fréquence était particulièrement important à prendre en compte pour la faisabilité de l'analyse comme nous l'avons décrit plus haut, d'où le recours aux résultats de l'étude Valodiag ; nous aurions souhaité intégrer les résultats de travaux similaires dans notre procédure, mais le nombre d'études antérieures se focalisant sur l'évaluation de CMA s'est révélé particulièrement restreint (77–79), ces études ne répondant pas à nos attentes, en particulier en matière de classement des CMA.

Le critère de chronicité découlait quant à lui d'un objectif de rentabilisation optimale des règles produites, en partant du principe qu'une CMA définie par une pathologie chronique devenait valable pour tous les séjours d'un patient ultérieurs à l'apparition de cette CMA, comprenant les séjours futurs de ce patient. Il s'avérait alors particulièrement intéressant et rentable de pouvoir cibler, par l'application d'une règle, non pas un, mais plusieurs séjours valorisables.

Ces notions de chronicité et de dimension temporelle étaient par ailleurs, mérite supplémentaire, rattachées au concept de parcours de soins ; il se trouve que les approches intégrant le parcours de santé font partie des options envisagées dans les projets de réforme des systèmes de financement hospitaliers et de la T2A, intronisant la notion de « tarification au parcours » ; cette nouvelle approche est notamment développée dans le rapport sur la réforme de la Tarification à l'Activité, réalisé en 2012 par la Mission d'Evaluation et de Contrôle de la Sécurité Sociale (MECSS) (80).

Cela nous a incité à centrer notre étude sur la production et la validation de règles séquentielles, rendant possible l'exploitation du critère de chronicité par la prise en compte de la dimension temporelle associée à ces règles.

Le critère de CMA était bien entendu incontournable. Les CMA choisies étaient toutes de niveau 2. Il aurait été envisageable d'axer la prédiction de codes sur des CMA de niveaux 3 ou 4. Ces CMA sont cependant moins fréquentes et ne satisfont pas toutes au critère de chronicité. De plus, bien que de niveau plus élevé, et donc plus valorisantes considérées individuellement, l'argument de fréquence réduisait l'influence de ces CMA sur la valorisation des séjours considérée globalement, ce dont rendaient parfaitement compte les résultats de l'étude Valodiag.

Le principal écueil dans la valeur des codes en tant que CMA tient lieu à la définition même des CMA et de leurs conditions d'application. De fait, la présence d'une CMA, quel que soit son niveau, ne garantit pas le passage du GHM vers un niveau supérieur.

Le frein le plus évident à l'influence d'une CMA est l'existence, dans le même séjour, d'une autre CMA de niveau équivalent ou supérieur, rendant l'influence de la CMA ciblée par la règle de contrôle complètement obsolète.

Rappelons de plus que d'autres conditions s'appliquent pour tenir compte d'une CMA dans un séjour : position du diagnostic (DP, DR ou DAS), seuils de durée de séjour, listes d'exclusion des CMA.

Notre travail ne tenait pas compte de ces limites et s'est focalisé sur la prédiction de CMA en se référant aux codes diagnostiques et d'actes associés. Il aurait été envisageable de compléter le travail par la prise en compte des durées de séjours et des niveaux de GHM entre autres, dans les limites des possibilités offertes par nos outils ; nous avions néanmoins choisi, par souci d'efficience, de procéder à une extraction brute des règles prédictives, ceci cadrant également mieux avec les

méthodes de data mining utilisées. L'intégration d'autres critères de prédiction est une piste intéressante à prendre en compte dans des travaux ultérieurs, avec possiblement des outils plus appropriés.

Nos résultats en eux-mêmes pouvaient sembler décevants sous certains aspects. Ainsi, la confiance « test » s'est avérée inférieure à la confiance « apprentissage » pour quasiment toutes les règles séquentielles, ce qui peut faire discuter un problème de surajustement des règles produites, à savoir que ces règles s'ajustent trop exactement à l'échantillon d'apprentissage (52,81), cet ajustement optimisé entravant leur application à un autre échantillon. C'est en prévision de ce biais que nous avons procédé à l'évaluation des règles sur un échantillon distinct de celui d'apprentissage.

Par ailleurs, à l'issue de l'étape de sélection finale reposant sur l'évaluation du potentiel de recodage de CMA de chaque règle, nous n'avons pu retenir de règles pour le code I50.

Ceci peut en partie s'expliquer par le fait que ce code était moins souvent codé dans notre base, d'où la production de règles moins fiables car portant sur un échantillon restreint.

De plus, le codage de l'insuffisance cardiaque et l'utilisation du code I50 et de ses subdivisions semble porter à confusion.

La définition de l'insuffisance cardiaque, à savoir « une maladie générale résultant de l'incapacité du cœur à adapter le débit sanguin aux besoins métaboliques et fonctionnels des différents organes dans des conditions de pressions de remplissage ventriculaire normales » (82), laisse à penser qu'il s'agit bien d'une pathologie chronique. Cependant, les consignes de codage qui figurent dans le guide méthodologique à propos de l'œdème pulmonaire incitent à réserver le codage du code I50 à la décompensation cardiaque aiguë (21). Notons qu'aucune

autre consigne de codage n'est stipulée dans ce guide à propos du codage de l'insuffisance cardiaque, de même qu'aucun fascicule de codage, comme il en existe pour certaines disciplines (83), n'est consacré au codage spécifique des pathologiques cardiovasculaires. Une recherche effectuée sur le forum Agora de l'ATIH (84), forum dévolu aux questions des professionnels du codage, n'a pas apporté de réponse concluante. Une seule question abordait spécifiquement le problème du distinguo entre les caractères aigu et chronique de l'insuffisance cardiaque ; la réponse à cette question s'est hélas avérée des plus ambiguës [annexe 2].

Si l'on considère, en plus de ces interrogations sur le codage, les difficultés mêmes de définition de l'insuffisance cardiaque, ou plutôt des insuffisances cardiaques, en tant qu'entités cliniques (85), on comprend dès lors que le code I50 ne soit pas toujours codé dans certaines situations, notamment en cas d'insuffisance cardiaque connue mais stable cliniquement et ne se manifestant pas bruyamment durant un séjour hospitalier motivé par une autre cause.

Considérant ces éléments, le choix du code I50 comme CMA chronique à prédire s'avérait a posteriori discutable.

Dans le même registre, le codage de la fibrillation atriale pouvait également susciter des questions.

En termes cliniques, il est habituel de distinguer fibrillation atriale paroxystique et permanente. Cependant, la fibrillation atriale, même lorsqu'elle est paroxystique, nécessite un traitement au long cours, et notamment une anticoagulation efficace (86). Si l'on considère la définition d'un DAS par le guide méthodologique, à savoir une pathologie obligatoirement prise en charge à titre diagnostique ou thérapeutique ou majorant l'effort de prise en charge du patient, il semble licite de coder la fibrillation atriale, y compris paroxystique, lors du séjour d'un patient, étant donné

l'administration de traitements et la charge induite par la gestion du traitement anticoagulant.

La maladie rythmique atriale, définie comme l'association d'une fibrillation atriale et de troubles de la conduction atriale, soulevait le même type d'interrogations. Cette pathologie relève d'un traitement par antiarythmiques complété par l'implantation d'un stimulateur cardiaque (87) ; ce traitement ne doit cependant pas faire suspendre l'anticoagulation, le risque thrombotique lié à la fibrillation demeurant.

Signalons qu'il n'existe pas de code CIM-10 propre à cette pathologie, dont le codage devrait se faire par l'association des codes I48 et I455 « Autre bloc cardiaque précisé ».

Il était concevable, dans ces deux situations, que certains professionnels n'aient pas codé le code I48 devant un électrocardiogramme normal (fibrillation atriale paroxystique) ou électroentrainé (maladie rythmique appareillée). Il était pourtant légitime de procéder à ce codage, vue la définition d'un DAS précédemment donnée ; les indications postées sur le forum Agora vont d'ailleurs dans ce sens.

Signalons enfin que le problème de codage lié à la fibrillation atriale paroxystique a été résolu par la dernière mise à jour du manuel des GHM (Version 11f - 2014), qui inclut désormais des subdivisions du code I48 concordant à la classification clinique de la fibrillation atriale (I480 « fibrillation auriculaire paroxystique », I481 « fibrillation auriculaire persistante » et autres subdivisions correspondantes).

Le codage du diabète non insulinodépendant quant à lui ne semblait pas soulever de difficultés, le caractère chronique de cette affection étant indéniable et les subdivisions du code E11 rendant parfaitement compte des différentes présentations cliniques de cette pathologie.

Pourtant, il était manifeste lors de notre analyse que ce code n'était pas systématiquement présent. Plusieurs causes à ce constat peuvent être évoquées.

Rappelons en premier lieu que certaines déclinaisons du code E11 ne sont pas des CMA, les déclinaisons de ce code ayant connu de nombreuses mises à jour au fil des versions successives du manuel des GHM. Il est possible que certains codeurs aient omis de renseigner les déclinaisons n'ayant pas l'attribut de CMA. Peut-être faut-il y voir également une méconnaissance de la définition des DAS, le code E11 n'étant pas renseigné bien que la majoration de l'effort de soin induite par le diabète lors d'un séjour soit patente. Reste enfin l'éventualité des simples oubli ou négligence de codage, explication au demeurant valable pour les omissions des autres codes.

L'énoncé des limites de notre étude ne doit cependant pas occulter la valeur de ses résultats. Nous sommes en effet parvenus à extraire trois règles séquentielles, simples, fiables et efficaces, avec une amélioration substantielle du recodage de CMA, supérieure à 28 % pour ces trois règles, par rapport aux CMA utilisées seules comme motifs prédictifs.

La valeur ajoutée de ces règles, en plus de leur application dans la valorisation des séjours, réside dans leur méthode de construction ; nous avons expérimenté pour l'extraction de ces règles des méthodes de data mining sur un extrait considérable de la base nationale PMSI MCO, ce qui constitue une approche originale pour la création de règles de contrôle qualité, se basant sur des méthodes formelles et scientifiques plutôt qu'empiriques.

Notre travail avait également l'avantage d'allier à ce volet technique une partie experte, concernant notamment le choix des CMA ciblées par les règles.

Enfin, l'appréciation du pouvoir de recodage de chaque règle, par le retour aux courriers de sortie des séjours et la confrontation aux CMA seules comme

prédicteurs, permettait une évaluation précise et objective des règles, avec le mérite d'une mise à l'épreuve pratique de ces règles en tant qu'outils de recodage en situation réelle de contrôle qualité des séjours.

Dernier point important en faveur de notre travail, nous tenons à signaler que toutes les étapes de cette étude se sont faites dans le respect des conditions d'anonymat et de confidentialité conformément aux autorisations CNIL accordées pour l'exploitation des bases de données concernées, y compris la partie dévolue à la consultation des courriers de sortie des séjours, ces courriers étant dûment anonymisés.

Malgré les critiques et limites de la Tarification à l'Activité (7,61), ce système de financement a rempli une partie de ces objectifs de maitrise des dépenses de santé (58,88) et semble, après sept années d'existence, devoir se pérenniser.

Le maintien de la T2A passera vraisemblablement par l'instauration d'un ensemble de réformes, indispensables pour remédier aux limites de ce système et pour prendre en compte les évolutions du paysage sanitaire français ; parmi les mesures adaptatives à venir, la notion de « financement au parcours » (80), renvoyant au concept de parcours de soins, fait partie des options envisagées.

Cette éventualité fait écho à notre travail, les règles séquentielles trouvant une application toute désignée dans l'évaluation des parcours de soins des patients.

Il va sans dire que les règles séquentielles extraites ne sauraient avoir la même pérennité que la Tarification à l'Activité, ne serait-ce que du fait des évolutions annuelles du manuel des GHM et de la liste des CMA, avec la possibilité de déclassement des codes ciblés par ces règles.

Ainsi, afin d'actualiser nos règles de prédiction avec les évolutions du système de financement, plusieurs perspectives peuvent s'envisager en continuité de notre étude. Une réalisation de ce même travail sur un échantillon plus important, prenant

en compte les années ultérieures à 2010, accroîtrait certainement la quantité de règles produites ainsi que la robustesse et la validité de ces règles. Il serait alors envisageable, pour contourner la limite imposée par les capacités de traitement de nos applications, de recourir à des outils informatiques innovants et plus adaptés à la gestion de bases de données volumineuses, tel le NoSQL (89).

Etendre la prédiction à d'autres CMA est également une option très intéressante, en particulier en ne se restreignant plus aux seules CMA chroniques, la prédiction de CMA aiguës relevant plus du domaine des règles d'association plutôt que séquentielles.

Enfin, la perspective la plus séduisante pour la mise en valeur de notre travail serait d'éprouver les règles construites en situation réelle de contrôle qualité ; cette application pratique permettrait en outre de corriger les limites de notre méthode, en intégrant des conditions supplémentaires sur les séjours ciblés par le contrôle et non prises en compte initialement par notre méthode, portant entre autres sur les durées de séjours et les niveaux des GHM.

Surtout, cette ultime mise à l'épreuve permettrait d'apprécier réellement et objectivement les capacités de valorisation de ces règles de prédiction en estimant sur une période donnée le montant exact de la valorisation produite, ce que notre étude ne permettait pas.

Nous espérons, comme aboutissement de notre travail, parvenir à intégrer ces règles dans un environnement de contrôle qualité T2A, en partenariat avec les services d'information médicale qui ont collaboré à notre étude.

CONCLUSION

Notre étude nous a permis d'extraire à partir de la base nationale PMSI MCO, par des méthodes de data mining, des règles de prédiction de CMA valides, fiables et simples d'application dans le cadre du contrôle qualité des séjours. Ces règles sont :

- ✓ {E11,I10,DZQM006}=>{E11}
- ✓ {E11,I10,I48}=>{E11}
- ✓ {I48,I69}=>{I48}

Cette approche a le mérite d'être innovante et de reposer sur une méthodologie formelle, complétant la couche experte sur laquelle se base usuellement le contrôle qualité des données médico-administratives dans le cadre de la T2A.

REFERENCES BIBLIOGRAPHIQUES

1. Albouy V, Bretin E, Carnot N, Deprez M. Les dépenses de santé en France : déterminants et impact du vieillissement à l'horizon 2050. Dir Générale Trésor Polit Econ DGTPE. 2009;(11). Disponible sur : http://www.eic.minefi.gouv.fr/directions_services/dgtpe/etudes/doctrav/pdf/cahiers-2009-11.pdf

2. Bac C, Cornilleau G. L'évolution des dépenses totales de santé depuis 1970. DREES; 2002. Disponible sur : http://www.drees.sante.gouv.fr/IMG/pdf/er175.pdf

3. Rapport sur la santé dans le monde 2000. OMS; 2000. Disponible sur : http://apps.who.int/gb/archive/pdf_files/WHA53/fa4.pdf

4. Grignon M. Les conséquences du vieillissement de la population sur les dépenses de santé. Quest Déconomie Santé. 2003;(66):1-6.

5. Raynaud D. Les déterminants individuels des dépenses de santé. Doss Solidar Santé. 2002;(1):29-58.

6. Nisand G. PMSI et systèmes d'informations hospitaliers. 2009. Disponible sur : http://www-ulpmed.u-strasbg.fr/medecine/cours_en_ligne/e_cours/medecine_sociale/Systemes_information_013.pdf

7. Krief N. L'impact de la tarification à l'activité sur le lien social à l'hôpital : étude du « Plan Hôpital 2007 ». Communication, 16ème conférence de l'AGRH-Paris Dauphine. 2005. Disponible sur : http://www.reims-ms.fr/agrh/docs/actes-agrh/pdf-des-actes/2005krief089.pdf

8. Fetter RB, Shin Y, Freeman JL, Averill RF, Thompson JD. Case mix definition by diagnosis-related groups. Med Care. 1980;i - 53.

9. Kervasdoué J. Politique de régulation et systèmes d'information. Rev Déconomie Financ. 2004;76(3):267-78.

10. Cash R. La tarification à l'activité : première année de mise en oeuvre. Rev Déconomie Financ. 2004;76(3):209-22.

11. Aballea P, Bras P-L, Seydoux S. Mission d'appui sur la convergence tarifaire public privé. Inspection Générale des Affaires Sociales IGAS; 2006. p. 2. Disponible sur : http://www.robertholcman.net/public/documents_institutionnels/rapports/convergence_public_prive.pdf

12. Milcent C, Dormont B. Comment évaluer la productivité et l'efficacité des hôpitaux publics et privés ? Les enjeux de la convergence tarifaire. Econ Stat. 2013;(455-456):143-73.

13. Or Z, Renaud T, Com-Ruelle L. Les écarts des coûts hospitaliers sont-ils justifiables? Rélexions Sur Une Converg Tarif Entre Sect Public Privé En Fr IRDES. 2009;(25). Disponible sur :

http://www.irdes.fr/EspaceRecherche/DocumentsDeTravail/DT25EcartsCoutHospitaliers.pdf

14. Rapport 2011 au Parlement sur la convergence tarifaire. Ministère du Travail, de l'Emploi et de la Santé; 2011. Disponible sur : http://www.sante.gouv.fr/IMG/pdf/Rapport_convergence_au_Parlement_2011_4_1_191011.pdf

15. Forfait hospitalier journalier. Ministère des Affaires Sociales, de la Santé et des Droits des Femmes. 2009. Disponible sur : http://www.sante.gouv.fr/IMG/pdf/Forfait_journalier_hospitalier-2.pdf

16. ATIH. Dispositifs médicaux pris en charge en sus. 2014. Disponible sur : http://www.atih.sante.fr/dispositifs-medicaux-pris-en-charge-en-sus

17. ATIH. Arrêtés prestations et tarifaires 2014. 2014. Disponible sur : http://www.atih.sante.fr/arretes-prestations-et-tarifaires-2014

18. ATIH. Dotations annuelles MIGAC MERRI. 2014. Disponible sur : http://www.atih.sante.fr/dotations-annuelles-migac-et-fir

19. Bonnet L. Le contrôle externe de l'Assurance Maladie. 2010. Disponible sur : http://fulltext.bdsp.ehesp.fr/Ehesp/Memoires/edh/2009/bonnet.pdf

20. ATIH. Présentation de l'ATIH. Disponible sur : http://www.atih.sante.fr/l-atih/presentation

21. ATIH. Guide méthodologique MCO 2014. 2014. Disponible sur : http://www.atih.sante.fr/guide-methodologique-mco-2014

22. ATIH. Manuel des GHM - Version 11f. 2014. Disponible sur : http://www.atih.sante.fr/manuel-des-ghm-version-11f

23. ATIH. CCAM - version 33. 2014. Disponible sur : http://www.atih.sante.fr/ccam-version-33

24. ATIH. Aide à l'exploitation des bases MCO 2013. 2014. Disponible sur : http://www.atih.sante.fr/aide-l-exploitation-des-bases-mco-2013

25. Guigue L, Richard C. Le Big data en Santé préfigure-t-il la «médecine 3.0» ? HEGEL ISSN 2115-452X 2014 3. 2014; Disponible sur : http://documents.irevues.inist.fr/handle/2042/54093

26. Fender P, Weill A. Epidémiologie, santé publique et bases de données médico-tarifaires. Rev Médicale Assur Mal. 2005;36(2):163-8.

27. Olive F, Gomez F, Schott A-M, Remontet L, Bossard N, Mitton N, et al. Analyse critique des données du PMSI pour l'épidémiologie des cancers : une approche longitudinale devient possible. Rev DÉpidémiologie Santé Publique. 2011;59(1):53-8.

28. Paty A-C, Suzan F. Analyse de la morbidité d'une maladie rare à partir des données du programme de médicalisation des systèmes d'information (PMSI) : exemple de la drépanocytose. Rev DÉpidémiologie Santé Publique. 2012;60:S30.

29. Neumann A, Tuppin P, Danchin N, Weill A, Ricordeau P, Allemand H. Facteurs associés aux ré-hospitalisations et au décès tardif à 30 mois après un infarctus du myocarde. Une analyse à partir des données chaînées du PMSI MCO et du SNIIRAM. Rev DÉpidémiologie Santé Publique. 2010;58:S20.

30. Gerbier S, Bouzbid S, Pradat E, Baulieux J, Lepape A, Berland M, et al. Intérêt de l'utilisation des données du PMSI pour la surveillance des infections

nosocomiales aux Hospices Civils de Lyon. Rev DÉpidémiologie Santé Publique. 2011;59(1):3-14.

31. Neumann A, Weill A, Ricordeau P, Fagot J-P, Alla F, Allemand H. Étude de cohorte sur le risque de cancer de la vessie chez les personnes diabétiques traitées par pioglitazone à partir des données chaînées du SNIIRAM et du PMSI. Rev DÉpidémiologie Santé Publique. 2012;60:S14.

32. Jay N, Egho E, Nuemi G, Kohler F, Napoli A, Quantin C. Apports des méthodes de fouille de données pour l'étude des trajectoires de prise en charge du cancer du poumon en région Bourgogne. Rev DÉpidémiologie Santé Publique. 2012;60:S13-4.

33. Touati M, Rahal M, Quantin C, Leteuff G, Limam M, Afonso F, et al. Analyse de trajectoires hospitalières de patients atteints d'un infarctus aigu du myocarde. Actes Atelier «Fouille Données Temporelles» Sous Dir René Quiniou Georges Hebrail EGC2006 Lille Janvier 2006. 2006; Disponible sur : http://hal.archives-ouvertes.fr/hal-00477353/

34. Blum D. Anonymat du patient dans le Programme de médicalisation des systèmes d'information : quel leurre est-il ? Rev DÉpidémiologie Santé Publique. 2011;59:S54.

35. ATIH. Commande de bases 2014. Disponible sur : http://www.atih.sante.fr/bases-de-donnees/commande-de-bases

36. ATIH. Aide à l'utilisation des informations de chaînage. Disponible sur : http://www.atih.sante.fr/aide-lutilisation-des-informations-de-chaînage

37. ATIH. Système National d'Information sur l'Hospitalisation SNATIH. 2014. Disponible sur : http://www.atih.sante.fr/snatih

38. ATIH. Hospi Diag. 2014. Disponible sur : http://www.atih.sante.fr/hospidiag

39. Faggionato D. Le Système National d'Information Inter Régimes de l'Assurance Maladie SNIIRAM. DREES; 2014. Disponible sur : http://www.ameli.fr/fileadmin/user_upload/documents/Presentation_du_SNIIRAM.pdf

40. Eco-Santé Base de données en licence ouverte. IRDES. 2014. Disponible sur : http://www.ecosante.fr/

41. Riou C, Fresson J, Serre JL, Avillach P, Leneveut L, Quantin C. Guide to good practices to ensure privacy protection in secondary use of medical records. Rev Epidemiol Sante Publique. 2014; Disponible sur : http://www.ncbi.nlm.nih.gov/pubmed/24889912

42. data.gouv.fr. Plateforme ouverte des données publiques françaises. 2014. Disponible sur : https://www.data.gouv.fr/fr/

43. IDS. Institut des Données de Santé. 2014. Disponible sur : http://www.institut-des-donnees-de-sante.fr/

44. Elghazel H. Classification et prévision des données hétérogènes: application aux trajectoires et séjours hospitaliers. Lyon 1; 2007. Disponible sur : http://www.theses.fr/2007LYO10325

45. Jay N. Découverte et représentation des trajectoires de soins par analyse formelle de concepts. Université Henri Poincaré-Nancy I; 2008. Disponible sur : http://hal.archives-ouvertes.fr/tel-00585411/

46. Goldberg M, Zins M. Santé et Big Data : beaucoup de questions, encore peu de solutions. 2014; Société Française de Statistique SFdS. Disponible sur : http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=1796

47. Salzano G. Réutilisation de données publiques. 2013. Disponible sur : http://inforsid.fr/actes/2013/2013_6_1%20Salzano.pdf

48. CADA Commission d'accès aux documents administratifs. La réutilisation des informations publiques. 2014. Disponible sur : http://www.cada.fr/la-reutilisation-des-informations-publiques,2.html

49. El Fadly N, Lucas N, Lastic P-Y, Verplancke P, Daniel C. Projet REUSE : utilisation du dossier patient informatisé (DPI) en recherche clinique. Risques, Technologies de l'Information pour les Pratiques Médicales. Springer; 2009. p. 227-38. Disponible sur : http://link.springer.com/chapter/10.1007/978-2-287-99305-3_21

50. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data : A transnational perspective. Int J Med Inf. 2013;82(1):1-9.

51. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment : enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-51.

52. Chazard E. Adapter les méthodes statistiques aux Big Data. 2012; Faculté de médecine Lille2. Disponible sur : http://www.chazard.org/emmanuel_/contenuperso/support_cours/chazard_L2BMQ_S10_bigdata.pdf

53. Kodratoff Y. Knowledge discovery in texts : a definition, and applications. Foundations of Intelligent Systems . Springer; 1999. p. 16-29. Disponible sur : http://link.springer.com/chapter/10.1007/BFb0095087

54. Calas G. Etudes des principaux algorithmes de data mining. Artic Rech Éc Ing En Inform EPITF Fr. 2006; Disponible sur : http://guillaume.calas.free.fr/data/Publications/DM-Algos.pdf

55. Laurent A, Teisseire M, Allia MR, Bouadi T, El Moutaouakil S, Keira M. Fouille de données : Règles séquentielles. Disponible sur : http://www2.lirmm.fr/~leclere/enseignements/TER/2008/Rapport/31.pdf

56. ARS. Parcours de soins, parcours de santé, parcours de vie. 2014. Disponible sur : http://www.ars.sante.fr/Parcours-de-soins-parcours-de.148927.0.html

57. Ferrari A. Parcours de soins, parcours de santé, parcours de vie. ARS; 2012. Disponible sur : http://www.ars.sante.fr/fileadmin/PORTAIL/image/organisation/PPT/Presentation_parcours.ppt

58. Rapport 2011 au Parlement sur la tarification à l'activité (T2A). Ministère du Travail, de l'Emploi et de la Santé; 2011. Disponible sur : http://www.sante.gouv.fr/IMG/pdf/Rapport_T2A_au_Parlement_2011_transmis_1509_11.pdf

59. Pascal P, Coutard J, Dupuy E, Varnier F, Welter G, Olivier M, et al. Evaluation de la tarification des soins hospitaliers et des actes médicaux. IGAS; 2012. Disponible sur : http://www.ladocumentationfrancaise.fr/rapports-publics/134000556/index.shtml

60. Olivier P. Perspectives du financement des parcours de soins. ARS Ile-de-France, Direction de la Stratégie; 2013. Disponible sur : http://www.numerique-sante.fr/wp-content/uploads/Vend5_14h-14h30_P_OLIVIER.pdf

61. Or Z, Häkkinen U. Qualité des soins et T2A : pour le meilleur ou pour le pire ? IRDES. 2012;(53):1-20.

62. Maravic M, Le Bihan C, Boissier M-C, Landais P. Valorisation de l'activité rhumatologique en France au crible du programme de médicalisation du système d'information (PMSI). Étude d'un exemple. Rev Rhum. 2003;70(3):274-80.

63. Guidet B, Taright N. Comment optimiser le codage en réanimation ? 2010; Disponible sur : http://sofia.medicalistes.org/spip/IMG/pdf/Comment_optimiser_le_codage_en_reanimation__.pdf

64. Roattino N, Sartor C, Romain F, La Scola B, Sambuc R, Durif L. Collaboration entre le Clin et le DIM–optimisation de la valorisation et du système d'information, hôpital de la Conception, Marseille. Rev DÉpidémiologie Santé Publique. 2010;58:S9.

65. Berthier F, Daideri G, Gendreike Y, Brocker P, Quaranta J-F, Staccini P. Influence de la qualité du codage d'une CMA sur la valorisation de l'activité d'un établissement : exemple des escarres. J Déconomie Médicale. 2005;(2):73-81.

66. Brocker P. Impact médico-économique de la dénutrition chez les sujets âgés. Rev Gériatrie. 2008;33(7):619-26.

67. Crouzet C, Chaput B, Grolleau J-L. Optimisation de la cotation dans la prise en charge des escarres : oui, mais à quel prix ? Annales de chirurgie plastique esthetique . Elsevier; 2013. p. 183-7. Disponible sur : http://www.sciencedirect.com/science/article/pii/S0294126013000277

68. Ficheur G, Genty M, Chazard E, Flament C, Beuscart R. Proposition d'une méthode automatisée calculant la valeur moyenne d'un diagnostic associé significatif. Rev DÉpidémiologie Santé Publique. 2013;61:S18-9.

69. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record. ACM; 1993. p. 207-16. Disponible sur : http://dl.acm.org/citation.cfm?id=170072

70. Hahsler M, Grün B, Hornik K. A computational environment for mining association rules and frequent item sets. 2005; Disponible sur : http://epub.wu.ac.at/132/

71. Zaki MJ. SPADE : An efficient algorithm for mining frequent sequences. Mach Learn. 2001;42(1-2):31-60.

72. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. Disponible sur : http://www.R-project.org/.

73. Hahsler M, Buchta C, Gruen B, Hornik K, Borgelt C. arules : Mining Association Rules and Frequent Itemsets. 2014. Disponible sur : http://cran.r-project.org/web/packages/arules/index.html

74. Hahsler CB and M, Diaz D. arulesSequences : Mining frequent sequences . 2014. Disponible sur : http://cran.r-project.org/web/packages/arulesSequences/index.html

75. Loriol M. Accréditation, protocolisation et évolutions du travail soignant. L'individu social : autres réalités, autre sociologie, XVIIème congrès international des sociologues de langue française GT « Sociologie de la santé ». 2004. p. 225-32. Disponible sur : http://halshs.archives-ouvertes.fr/halshs-00368391

76. Guex P, Stiefel F. De la souffrance du patient à celle des équipes. Médecine Palliat Soins Support-Accompagnement-Éthique. 2010;9(1):32-5.

77. Patris A, Gomez S, Noury JF, Mendelshon M. Comparaison de quatre mesures de l'«effet CMA» d'un diagnostic associé. Interprétation sur quelques exemples. Rev DÉpidémiologie Santé Publique. 2008;56(1):16.

78. Patris A, Gomez S, De Mey P. Indicateurs de la qualité du codage des complications et morbidités associées (CMA). Rev DÉpidémiologie Santé Publique. 2012;60:S30.

79. Patris A, De Mey P. Évolution du codage des CMA et de leurs effets pendant dix ans dans les bases PMSI françaises. Impact de la T2A et des modifications de groupage. Rev DÉpidémiologie Santé Publique. 2013;61:S19.

80. Budet J-M, Le Menn J, Milon A. L'évolution du financement des hôpitaux. Refonder la tarification hospitalière au service du patient. Gest Hosp. 2012;(521):590-8.

81. Monari G. Sélection de modèles non linéaires par « leave-one-out » : étude théorique et application des réseaux de neurones au procédé de soudage par points. Université Pierre et Marie Curie-Paris VI; 1999. Disponible sur : http://pastel.archives-ouvertes.fr/pastel-00000676/

82. Pousset F, Isnard R, Komajda M. L'insuffisance cardiaque: problème de santé publique. Rev Médecine Interne. 2005;26(11):843-4.

83. ATIH. Fascicules de conseils de codage CIM-10. 2014. Disponible sur : http://www.atih.sante.fr/fascicules-de-conseils-de-codage-cim-10

84. ATIH. Agora et supports utilisateurs. 2014. Disponible sur : http://www.atih.sante.fr/support-utilisateurs

85. Jondeau G, Aumont MC, Aupetit JF, Cohen-Solal A, Davy JM, Degroote P, et al. Insuffisance cardiaque et cardiomyopathies. Arch Mal Coeur Vaiss. 2006;99(2 Suppl):3-79.

86. Abaléa J, Mansourati J. Fibrillation atriale. Rev Prat. 2013;63(9):1287-92.

87. Brembilla-Perrot B. Fibrillation auriculaire. 2010; Disponible sur : http://www.em-consulte.com/en/article/263247

88. Le Garrec M-A, Bouvet M. Les Comptes nationaux de la santé en 2013. DREES; 2014. Disponible sur : http://www.drees.sante.gouv.fr/IMG/pdf/er890.pdf

89. Cattell R. Scalable SQL and NoSQL data stores. ACM SIGMOD Rec. 2011;39(4):12-27.

ANNEXES

Annexe 1 : Classement des CMA par l'étude Valodiag

Rang	Code	Libellé	Valeur moyenne
1	E440	Malnutrition protéino-énergétique modérée	699 €
2	E43	Malnutrition protéino-énergétique grave, sans précision	666 €
3	E559	Carence en vitamine D, sans précision	310 €
4	J690	Pneumopathie due à des aliments et des vomissements	1 951 €
5	L892	Ulcère de décubitus de stade III	1 562 €
6	R33	Rétention d'urine	230 €
7	L97	Ulcère du membre inférieur, non classé ailleurs	714 €
8	E441	Malnutrition protéino-énergétique légère	403 €
9	E1190	Diabète sucré non insulinodépendant insulinotraité, sans complication	256 €
10	I48	Fibrillation et flutter auriculaires	108 €
11	U801	Agents résistant à la méthicilline	1 829 €
12	B965	Pseudomonas (P. aeruginosa), cause de maladies classées dans d'autres chapitres	1 795 €
13	T801	Complications vasculaires consécutives à une injection thérapeutique, une perfusion et une transfusion	356 €
14	I802	Phlébite et thrombophlébite d'autres vaisseaux profonds (des membres inférieurs)	628 €
15	L891	Ulcère de décubitus de stade II	372 €
16	R02	Gangrène, non classée ailleurs	1 150 €
17	J960	Insuffisance respiratoire aiguë	390 €
18	T810	Hémorragie et hématome compliquant un acte à visée diagnostique et thérapeutique, non classés ailleurs	441 €
19	I501	Insuffisance ventriculaire gauche	293 €
20	B956	Staphylococcus aureus, cause de maladies classées dans d'autres chapitres	368 €
21	K632	Fistule de l'intestin	5 534 €
22	L893	Ulcère de décubitus de stade IV	1 267 €
23	R410	Désorientation, sans précision	74 €
24	B962	Escherichia coli, cause de maladies classées dans d'autres chapitres	75 €
25	T814	Infection après un acte à visée diagnostique et thérapeutique, non classée ailleurs	570 €
26	R2630	État grabataire	152 €
27	E6691	Obésité, sans précision, avec indice de masse corporelle égal ou supérieur à 40kg/m² et inférieur à 50kg/m²	166 €
28	E6601	Obésité due à un excès calorique, avec indice de masse corporelle égal ou supérieur à 40 kg/m² et inférieur à 50 kg/m²	219 €
29	E109	Diabète sucré insulino-dépendant, sans complication	334 €
30	D509	Anémie par carence en fer, sans précision	147 €
31	R630	Anorexie	214 €
32	B964	Proteus (P.mirabilis) (P.morganii), cause de maladies classées dans d'autres chapitres	530 €
33	D611	Aplasie médullaire médicamenteuse	1 107 €
34	T818	Autres complications d'un acte à visée diagnostique et thérapeutique, non classées ailleurs	392 €
35	N179	Insuffisance rénale aiguë, sans précision	292 €
36	A46	Érysipèle	603 €
37	U88	Agents résistant à de multiples antibiotiques	374 €
38	I500	Insuffisance cardiaque congestive	183 €
39	R296	Chutes à répétition, non classées ailleurs	321 €
40	E6602	Obésité due à un excès calorique, avec indice de masse corporelle égal ou supérieur à 50 kg/m²	876 €
41	E6692	Obésité sans précision, avec indice de masse corporelle égal ou supérieur à 50 kg/m²	619 €
42	A415	Sepsis à d'autres microorganismes Gram négatif	2 995 €
43	I509	Insuffisance cardiaque, sans précision	147 €
44	K560	Iléus paralytique	1 644 €
45	T827	Infection et réaction inflammatoire dues à d'autres prothèses, implants et greffes cardiaques et vasculaires	868 €
46	J961+0	Insuffisance respiratoire chronique obstructive	199 €
47	L890	Zone de pression et ulcère de décubitus de stade I	72 €
48	J154	Pneumopathie due à d'autres streptocoques	2 157 €
49	G819	Hémiplégie, sans précision	224 €
50	D500	Anémie par carence en fer secondaire à une perte de sang (chronique)	167 €

Annexe 2 : Copie écran extraite du forum Agora concernant le codage de l'insuffisance cardiaque

Agora > CIM > CIM-10

Insuffisance cardiaque gauche

Suivre ce fil

Ajouter une nouvelle contribution à ce sujet

décembre 2009

Bonjour,

Après avoir pris connaissance sur le forum des différents échanges concernant le codage de l'insuffisance cardiaque je souhaiterais une précision concernant le code I50.1.

Je le réservais au codage de l'IVG décompensée. A la lecture des récentes modifications je m'interroge sur la signification à donner à la note placée sous ce code et libellée de la façon suivante : "insuffisance ventriculaire gauche".

L'IVG peut être décompensée ou non. Dans ce dernier cas (situation non décompensée) peut-on utiliser ce code I50.1.

(ex. Fraction d'éjection du ventricule gauche diminuée , Dysfonction ventricule gauche à l'écho ... Le code I50.1 convient-il ?

Je vous remercie par avance de votre réponse, et vous adresse mes salutations les meilleures

Intervention référent demandée

janvier 2010

Bonjour,

Je ne comprends pas de quelles « récentes modifications » vous parlez à propos de l'insuffisance cardiaque. « Insuffisance ventriculaire gauche » n'est pas une note d'inclusion mais l'intitulé même de I50.1. Les critères que vous vous imposez constituent une interprétation personnelle. Ils ne se fondent sur aucune précision contenue dans la CIM-10 ni dans le guide méthodologique. Oui, une IVG peut être « décompensée » ou non, mais la CIM-10 ne tient pas compte de cette distinction. Il existe des définitions échographique, hémodynamique, scintigraphique, clinique... de l'insuffisance ventriculaire gauche. La CIM-10 n'en tient pas compte non plus. I50.1 est le code de l'insuffisance ventriculaire gauche, quelle que soit la méthode de diagnostic et quelles que soient ses manifestations.

Cordialement

AUTEUR : Nom : DJENNAOUI **Prénom :** Mehdi

Date de Soutenance : 21 novembre 2014

Titre de la Thèse :

Construction et évaluation de règles de prédiction de diagnostics à partir des bases de données hospitalières : application au contrôle qualité des données médico-administratives

Thèse - Médecine - Lille 2014

Cadre de classement : Santé Publique et Médecine Sociale

DES + spécialité : Santé Publique et Médecine Sociale

Mots-clés : tarification à l'activité, contrôle qualité, base nationale PMSI, data reuse, data mining, règles de prédiction

Résumé

Contexte : Dans le cadre du Programme de Médicalisation du Système Informatique (PMSI), l'instauration de la Tarification à l'Activité (T2A) a incité les établissements de santé à établir des procédures de contrôle qualité optimisant la rémunération des séjours, le repérage et le recodage de Complications et Morbidités Associées (CMA) manquantes constituant une de ces procédures. Les méthodes de data mining suscitent beaucoup d'intérêt dans l'analyse des bases de données. Ces méthodes produisent des règles prédictives simples, intuitives et faciles à appliquer. Nous avions pour objectif de produire des règles de prédiction de CMA applicables au contrôle qualité des séjours en soumettant les données de la base nationale PMSI aux méthodes de data mining.

Méthode : Notre échantillon de travail était constitué à partir de la base nationale PMSI pour les années 2007 à 2010. Les CMA à prédire devaient être fréquentes et relever d'une pathologie chronique, la prédiction se faisant à partir des codes diagnostiques et d'actes des séjours. Notre étude se poursuivait par l'évaluation des règles produites en calculant la confiance réelle des règles et en appréciant le gain en termes de recodage de CMA.

Résultats : Notre échantillon comportait 59170 séjours. Les CMA ciblées étaient les codes E11 « diabète sucré non insulinodépendant », I48 « fibrillation atriale » et I50 « insuffisance cardiaque ». Nous avons extrait trois règles d'association et six règles séquentielles, et validé à l'issue de la procédure de contrôle trois règles prédictives, deux pour le code E11 et une pour le code I48. Les trois règles validées ont toutes une confiance supérieure à 0.60 et un gain de recodage de CMA supérieur à 28 %.

Conclusion : Notre étude nous a permis d'extraire à partir de la base nationale PMSI, par data mining, des règles de prédiction de CMA valides, fiables et simples d'application dans le cadre du contrôle qualité des séjours.

Composition du Jury :

Président : Monsieur le Professeur Jean-Louis SALOMEZ

Assesseurs : Monsieur le Professeur Régis BEUSCART

Monsieur le Docteur Emmanuel CHAZARD

Monsieur le Docteur Grégoire FICHEUR