UNIVERSITE PARIS XI
FACULTE DE MEDECINE DE PARIS-SUD

Année 2007-2008

MASTER PROFESSIONNALISANT DE SANTE PUBLIQUE
M2 - METHODOLOGIE ET STATISTIQUE EN RECHERCHE BIOMEDICALE

DATA MINING ET PREVENTION DES EFFETS INDESIRABLES LIES AUX MEDICAMENTS : LE PROJET EUROPEEN PSIP

Emmanuel CHAZARD

Mots clef :
PSIP, effets indésirables médicamenteux, data mining, arbres de décision

Résumé :
Les effets indésirables liés aux médicaments entraîneraient 10 000 décès par an en France. Leur prévention à l'hôpital repose sur les systèmes de prescription connectés (CPOE) couplés à des systèmes d'aide à la décision (CDSS). Les règles d'alertes de CDSS sont écrites à dire d'expert et génèreraient des alertes trop fréquentes et peu adaptées donc mal suivies. Le projet PSIP (Patient Safety through Intelligent Procedures in medication) propose entre autres de générer ces règles d'après la fouille automatisée des données (data mining), afin de constituer des règles d'alertes basées sur les erreurs passées du service. Ces règles seront ensuite validées par une revue experte de dossiers.
Ce mémoire évoque les premières phases : conception du modèle de données, analyses statistique (31 arbres de décision) et premiers résultats (223 règles).

Sommaire

Sigles utilisés

ADE	*Adverse Drug Event (effet indésirable lié aux médicaments)*
ARH	*Agence Régionale de l'Hospitalisation*
ASCII	*American Standard Code for Information Interchange (encodage de caractère américain normalisé pour l'échange d'information)*
ATIH	*Agence Technique de l'Information Hospitalière*
AVK	*Anti-vitamine K, antagoniste de la vitamine K*
CCAM	*Classification Commune des Actes Médicaux*
CDSS	*Computerized Decision Support System (système d'aide à la décision)*
CH	*Centre Hospitalier*
CHRU	*Centre Hospitalier Régional Universitaire*
CHU	*Centre Hospitalier Universitaire*
CIM	*Classification Internationale des Maladies*
CM	*Catégorie Majeure*
CMD	*Catégorie Majeure de Diagnostic*
CPOE	*Computerized Prescription Order Entry (système de prescription connectée)*
DAS	*Diagnostic Associé Significatif*
DIM	*Département de l'Information Médicale*
DMS	*Durée Moyenne de Séjour*
DP	*Diagnostic Principal*
DP	*Diagnostic Principal*
GHM, GHS	*Groupe Homogène de Malades, Groupe Homogène de Séjours*
IGS2	*Indice de Gravité Simplifié*
INR	*International Normalized Ratio*
MCO	*Médecine Chirurgie Obstétrique (= court séjour)*
OMS	*Organisation Mondiale de la Santé*
PMSI	*Programme de Médicalisation de Systèmes d'Information*
PSIP	*Patient Safety through Intelligent Procedures in medication*
RSA	*Résumé de Sortie Anonymisé*
RSS	*Résumé de Sortie Standardisé*
RUM	*Résumé d'Unité Médicale*
SGBD	*Système de Gestion de Bases de Données*
SIH	*Système d'Information Hospitalier*
SSR	*Soins de Suite et de Réadaptation (= moyen séjour)*
T3	*Triiodothyronine*
T4	*Tetraiodothyronine / Thyroxine*
TAA, T2A	*Tarification A l'Activité*
TSH	*Thyroid-Stimulating Hormone (hormone thyréotrope)*
UM	*Unité Médicale*
WHO	*World Health Organization (voir OMS)*
WP	*Workpackage (partie du projet)*

Remerciements

Monsieur le Professeur Régis Beuscart

Professeur des Universités en Biostatistiques et Informatique Médicale, université Lille 2
Praticien hospitalier, chef de service du Secteur d'Informatique et Information Médicale du CHRU de Lille
Directeur du Centre d'Etudes et de Recherche en Informatique Médicale
Officier dans l'Ordre des Palmes Académiques

Monsieur Cristian Preda

Professeur des Universités en Statistiques, université Lille 1

Madame le Docteur Béatrice Merlin

Interne en Santé Publique, CHRU de Lille

…pour son aide lors de la revue préalable de cas suspects, de la définition des politiques d'agrégation des diagnostics et des médicaments, et de l'interprétation médicale des arbres de décision

Mademoiselle Karine Wyndels

Interne en Santé Publique, CHRU de Lille

…pour son aide lors de la conception du modèle de données

Introduction

I. Le cadre de recherche du projet PSIP

Le projet PSIP (Patient Safety through Intelligent Procedures in medication)[1] est un projet européen agréé par l'European Research Council (ERC)[2] sous la référence n°216130 dans le cadre du Seventh Framework Programme (FP7)[3].

Ce projet, coordonné par le CHRU de Lille, regroupe 13 partenaires européens :

- le CHRU de Lille et l'université Lille 2
- le CHU de Rouen
- le CH de Denain
- les 10 hôpitaux de la région de Copenhague (Danemark)
- Oracle Europe
- IBM Danmark division Acure (Danemark)
- Medasys SA
- Vidal SA
- Kite Solutions (Italie)
- IDEEA Advertising (Roumanie)
- Université Aristotle de Thessalonique (Grèce)
- Université d'Aalborg (Danemark)
- UMIT université d'Innsbruck (Autriche)

Ce projet s'étale sur 40 mois, représente un budget total de 9.9 millions d'euros dont 7.3 millions d'euros sont financés par la Commission européenne. Le calendrier est le suivant :

- 19 décembre 2007 : signature
- janvier 2008 : Kick off meeting, lancement des workpackages 1, 2 et 3
- premier trimestre 2010 : expérimentations
- avril 2011 : fin du projet

II. Les effets indésirables liés au médicament

Les effets indésirables liés aux médicaments (ADE : adverse drug events) sont trop fréquents et ont souvent des conséquences sévères en terme de morbidité ou de mortalité. Les études menées en France sont partielles, c'est pourquoi il est difficile d'évaluer l'impact des ADE. Cependant, leurs résultats sont du même ordre de grandeur que dans les autres pays occidentaux. Aux USA par exemple l'estimation de référence rapporterait 98 000 ADE par an[4]. En France, plusieurs estimations révèleraient que :

- les ADE concerneraient 3% des hospitalisations soit 130 000 séjours par an
- les ADE toucheraient 10% des patients hospitalisés en court séjour médecine chirurgie obstétrique (MCO)
- le tiers des ADE auraient des effets significatifs et 9% d'entre eux seraient létaux
- 10 000 patients décèderaient chaque année d'ADE et 35 000 patients conserveraient des séquelles
- les ADE entraîneraient à eux seuls 1.2% des dépenses d'hospitalisation

- les ADE représenteraient 5 à 10% des motifs d'hospitalisation pour les patients de plus de 65 ans, 20% au-delà de 80 ans, et 50% au-delà de 95 ans

III. La prévention des ADE : innovations proposées par le projet PSIP

Les ADE qui surviennent durant une hospitalisation paraissent plus inacceptables encore que ceux survenant à domicile. Leur prévention paraît plus facile lorsque la prescription est réalisée sur une interface logicielle.

Les systèmes de prescription connectée (CPOE : Computerized Prescription Order Entry) sont de plus en plus répandus et souvent couplés avec un système d'aide à la décision (CDSS : Computerized Decision Support System). Ces systèmes sont efficaces mais la génération d'alertes lors d'une prescription à risque pose problème. Concrètement il n'est pas banal d'entendre un médecin senior expliquer à un jeune confrère : *« une fois les prescriptions saisies, tu valides puis tu cliques sur OK jusqu'à ce que toutes les alertes disparaissent »*. PSIP est un projet innovant à plusieurs égards :

- Habituellement, les règles qui permettent de générer les alertes sont écrites à dire d'expert :
    - elles se limitent donc à la connaissance académique et aux règles qu'on juge importantes en général, en arbitrant entre prévalence et gravité de l'effet. Exemple : ne pas administrer simultanément un bêta bloquant et un dérivé de l'ergot de seigle.
    - elles ne tiennent pas compte du fait que les prescripteurs maîtrisent souvent certaines précautions d'emploi. Par exemple, la prescription simultanée d'aspirine de d'héparine génère habituellement des alertes, alors que non seulement le risque hémorragique est bien connu et fait l'objet d'un arbitrage bénéfice/risque et d'une surveillance clinique et biologique, mais en plus la prescription de cette association en post-infarctus fait suite à une recommandation officielle.
    - inversement elles ne tiennent pas compte du fait que certains prescripteurs maîtrisent mal certains médicaments. Ainsi, rien n'interdit en soi de prescrire des anti-vitamine K (AVK), pourtant concrètement cette prescription induit de nombreux ADE.
- *=> Le projet PSIP se distingue des approches classiques par le fait que les règles seront initialement issues de la fouille de données (data mining, approche a posteriori), donc des erreurs passées du service, et non de la connaissance académique (approche a priori).*

- Habituellement, les règles d'alerte s'appliquent à l'identique à tous les services, nonobstant le contexte. Or il est prévisible qu'un service de cardiologie n'aura que peu de problème avec les médicaments à visée cardiologique, mais pourra rencontrer des problèmes liés aux traitement d'un diabétique, et inversement pour le service de diabétologie.
- *=> Le projet PSIP se distingue des approches classiques par la génération de jeux de règles service par service.*

- Habituellement , les alertes d'alerte se limitent à des messages de constat du type *« attention : prescription d'un anti-vitamine K (AVK) et d'un inhibiteur enzymatique : risque de surdosage »*. Peut-être un message du type *« Prescription simultanée d'un AVK et d'un inhibiteur enzymatique. L'an dernier*

dans votre service cette association a entraîné un INR trop élevé dans 27% des cas. Ajustez le dosage ou surveillez l'INR. » seraient-elles mieux acceptées et plus prises en compte par le prescripteur.

- *=> Le projet PSIP se distingue des approches classiques par l'intégration des facteurs humains, la prise en compte du processus global de la prescription, et la contextualisation des messages d'alerte.*

- Dans le même temps, la connaissance sur les effets indésirables médicamenteux résulte uniquement de deux types de sources : la déclaration volontaire, et les revues systématiques de dossiers. La dernière approche est fastidieuse, quant à la première en pratique les ADE sous sous-déclarés[5, 6]. Les déclarations spontanées à la demande améliorent peu les résultats[7]. On peut penser que seuls les effets rares ou liés au patient (manifestations allergiques par exemple) sont déclarés. Les effets fréquents (saignements modérés sous AVK) ou liés à une imprudence du prescripteur sont peu déclarés car le médecin suppose que la déclaration ne fait pas progresser la connaissance collective, ou pourrait lui porter préjudice. Le criblage automatisé des courriers de sortie (semantic mining) est une voie de recherche prometteuse[7, 8].
- *=> Le projet PSIP entend fournir un nouveau regard sur la connaissance épidémiologique des ADE en intégrant simultanément l'approche data mining et semantic mining.*

IV. Les phases du projet

Le projet se découpe en quatre phases :

- l'élicitation de la connaissance
- le développement d'un système d'aide à la décision (CDSS) couplé aux systèmes de prescription connectée (CPOE)
- l'intégration du système et les tests grandeur nature
- l'évaluation et la propagation du système

Ces quatre phases sont en réalité détaillées en treize workpackages présentés en Figure 1.

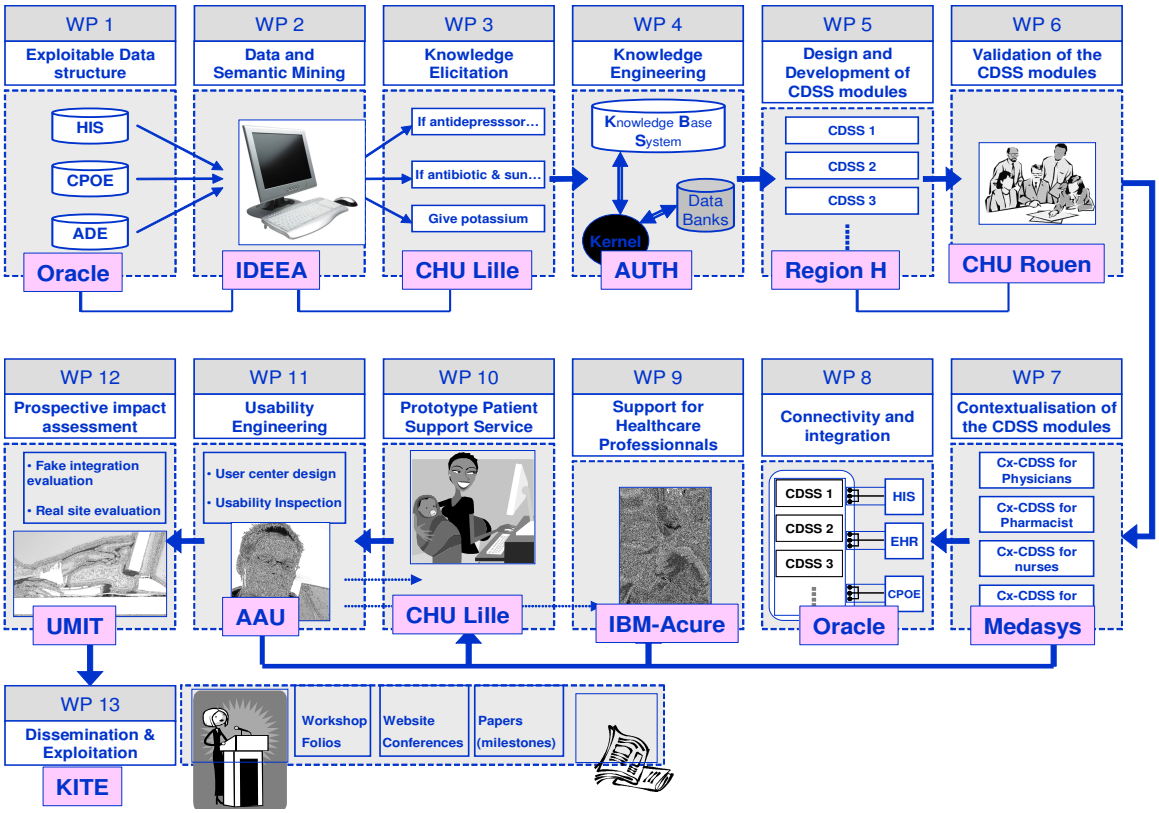


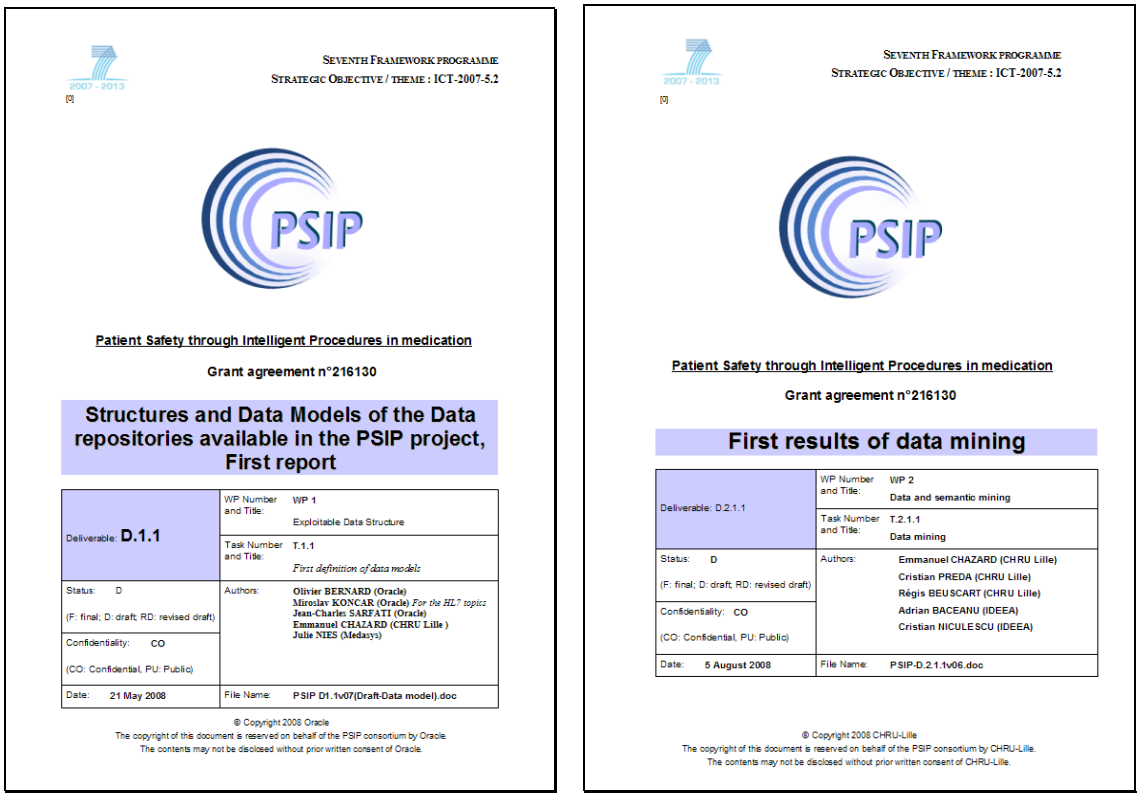
Figure 1 Les 13 workpackages du projet PSIP

V. Position du présent mémoire

Ce mémoire présente la contribution de l'auteur aux deux premiers workpackages (WP1 exploitable data structure, WP2 data and semantic mining) :

- définition d'un modèle de données
- assistance aux exports de données utilisables
- premiers résultats du data mining (NB : l'approche semantic mining sera explorée ultérieurement)

Les livrables correspondants ont été remis à la commission européenne et restent confidentiels (Figure 2). Ce mémoire présente donc un point de vue partiel.



SEVENTH FRAMEWORK PROGRAMME
STRATEGIC OBJECTIVE / THEME : ICT-2007-5.2

PSIP

Patient Safety through Intelligent Procedures in medication

Grant agreement n°216130

Structures and Data Models of the Data repositories available in the PSIP project, First report

Deliverable: **D.1.1**	WP Number and Title:	WP 1 Exploitable Data Structure
	Task Number and Title:	T.1.1 *First definition of data models*
Status: D (F: final; D: draft; RD: revised draft)	Authors:	**Olivier BERNARD (Oracle)** **Miroslav KONCAR (Oracle)** *For the HL7 topics* **Jean-Charles SAREATI (Oracle)** **Emmanuel CHAZARD (CHRU Lille)** **Julie NIES (Medasys)**
Confidentiality: CO (CO: Confidential, PU: Public)		
Date: 21 May 2008	File Name:	PSIP D1.1v07(Draft-Data model).doc

SEVENTH FRAMEWORK PROGRAMME
STRATEGIC OBJECTIVE / THEME : ICT-2007-5.2

PSIP

Patient Safety through Intelligent Procedures in medication

Grant agreement n°216130

First results of data mining

Deliverable: D.2.1.1	WP Number and Title:	WP 2 **Data and semantic mining**
	Task Number and Title:	T.2.1.1 **Data mining**
Status: D (F: final; D: draft; RD: revised draft)	Authors:	**Emmanuel CHAZARD (CHRU Lille)** **Cristian PREDA (CHRU Lille)** **Régis BEUSCART (CHRU Lille)** **Adrian BACEANU (IDEEA)** **Cristian NICULESCU (IDEEA)**
Confidentiality: CO (CO: Confidential, PU: Public)		
Date: **5 August 2008**	File Name:	**PSIP-D.2.1.1v06.doc**

Figure 2 Couvertures des deux premiers livrables de PSIP

Conception du modèle de données

I. Principes du modèle de données

Le modèle de données a été conçu ensemble par les membres des workpackages 1 et 2. Nous souhaitions proposer en schéma relationnel remplissant simultanément les qualités suivantes :

- d'un point de vue « quantité d'information » (nombre de colonnes) :
    - suffisamment simple…
        - … pour être compatible avec tous les systèmes d'information des participants
        - … pour être compatible avec les systèmes d'information d'autres hôpitaux européens chez lesquels le système pourrait être installé
    - suffisamment complexe…
        - … pour rendre compte de la complexité des données exploitables
        - … pour mettre à disposition toutes les informations nécessaires pour le traitement statistique (par exemple : non seulement la durée du séjour étudié, mais aussi la moyenne et l'écart type de la durée dans les séjours du même groupe homogène de séjours, nous reviendrons sur cela plus bas)
- d'un point de vue « complexité des relations » (nombre de tables et de jointures) :
    - suffisamment « relationnel » …
        - … pour être assez proche des données natives, et diminuer ainsi l'effort d'extraction et les erreurs
        - … pour être aussi stable que possible tout au long du projet quelles que soient les méthodes utilisées par les statisticiens
    - suffisamment « à plat » pour limiter les retraitements de données avant analyse statistique

Ces conditions nécessitaient de prendre en compte dans le même temps :

- en amont, les données disponibles, les schéma relationnels des partenaires, et les schéma relationnels que l'on pourrait s'attendre à rencontrer chez de nouveaux partenaires. Par exemple : « Quelles sont les informations rendues obligatoires par le PMSI en France, et par le système équivalent au Danemark ? », « Quelles seraient les différentes manières de représenter une prescription médicamenteuses tout au long d'un séjour ? »
- en aval, les méthodes statistiques que l'on s'apprêtait à utiliser

II. Revue de cas

Le problème le plus important est qu'une variable du type « effet indésirable médicamenteux (0/1) » est toujours absente des systèmes d'information. Ainsi, si les causes de l'ADE (Adverse Drug Event = effet indésirable lié au médicament) sont souvent lisibles, la présence de l'ADE lui-même est occulte. Il s'agissait donc d'imaginer comment l'information disponible pourrait rendre compte de la survenue d'un ADE, et

ce avant même de concevoir le modèle de données. Une des étapes préparatoires à cette conception fut une revue de cas par des internes de santé publique. Plusieurs dizaines de séjours suspects terminés par un décès furent passés en revue. La lecture de ces dossiers, associée à une bonne connaissance des données médicales disponibles, a permis d'imaginer plusieurs variables quantitatives essentiellement. Cette approche fut complétée par de entretiens informels avec des praticiens expérimentés.

Premier exemple de situation suspecte :
Monsieur Dupont, admis pour phlébite, revient 2 jours après sa sortie.

Interprétation à la lecture du dossier : ce patient s'est vu prescrire une anti-coagulation. Il a subi un surdosage en anti-coagulants et a du être ré-hospitalisé en raison de manifestations hémorragiques.

Elément suspect formalisé dans le PMSI : délai court entre deux hospitalisations
Formalisation numérique :

- récupération de la date de sortie du séjour
- récupération de la date d'entrée du séjour suivant
- calcul du délai :
    - en présence d'une hospitalisation ultérieure, [date entrée – date sortie +1]
    - sinon [+infini]
- calcul d'une quantité de distribution cohérente : 1/délai. Cette quantité vaut zéro si la personne n'est pas ré-hospitalisée, tend vers zéro pour les ré-hospitalisations tardives (donc sans doute non liées à une erreur dans le séjour précédent), et tend vers 1 pour les ré-hospitalisations précoces.

Deuxième exemple de situation suspecte :
Monsieur Durand, admis pour appendicite, décède d'un état de mal épileptique.

Interprétation à la lecture du dossier : ce patient était un épileptique connu. Hospitalisé pour appendicite, on a oublié de lui administrer son traitement suspensif, engendrant un état de mal épileptique fatal. On notera qu'il s'agit ici d'un oubli et non d'une prescription, mais l'observation de l'effet reste valable.

Premier élément suspect formalisé dans le PMSI : décès peu probable pour un séjour d'appendicectomie
Formalisation numérique :

- récupération d'une variable binaire « décès 1/0 »
- récupération du groupe homogène de séjours (GHS)
- récupération de la proportion attendue de décès pour ce GHS, la variance en découle naturellement
- calcul d'une quantité signifiant l'aspect inattendu du décès : il existe plusieurs possibilités. On peut par exemple utiliser le nombre de déviations standard par rapport à la moyenne, ce nombre sera négatif pour les séjours sans décès (d'autant plus négatif que la survie était peu probable), et positif pour les séjours avec décès (d'autant plus positif que le décès était peu probable). Pour un même GHS il n'y aura donc que 2 valeurs possibles, les séjours sont classés dans des GHS différents.

Deuxième élément suspect formalisé dans le PMSI : entrée pour appendicite et fin du séjour avec état de mal épileptique. Plus généralement, entrée pour un motif de chirurgie digestive, fin du séjour pour un motif de neurologie médicale.
Formalisation numérique :

- récupération du diagnostic principal (DP) de chaque résumé d'unité médicale (RUM) du séjour

- calcul de la catégorie majeure de diagnostics (CMD) dans laquelle le DP aurait pu orienter (NB : la moitié des DP n'orientent vers aucune CMD)
- calcul d'une liste dédoublonnée des CMD théoriques des RUM d'un même séjour
- calcul de maximum(longueur(liste des CMD), 1) : cette variable est un entier qui vaut habituellement 1, et plus le nombre augmente plus le patient a été pris en charge pour des appareils différents. Les effets indésirables liés aux médicaments se caractérisent souvent par des atteintes d'appareils différents de celui concerné par le motif d'hospitalisation.

On notera que l'information médicale utilise volontiers des variables qualitatives à nombreuses modalités. Par exemple le GHS est une variables qualitative à 780 modalités. Il est naturellement hors de question d'ajuster -ou pire, de stratifier- une analyse statistique sur le GHS. En revanche, la plupart des variables quantitatives méritent d'être interprétées en fonction du GHS. C'est le cas par exemple de la durée ou du décès. C'est la raison pour laquelle il est généralement plus habile de n'utiliser ni les variables natives, ni le GHS, mais des variables rendant compte de l'atypie par rapport à la référence du GHS.

III. Résultat

Un modèle de données a ainsi pu être conçu, intégrant en amont les possibilités des informaticiens, et en aval les besoins des statisticiens. Ce modèle a été proposé aux partenaires pour s'assurer de sa faisabilité. Au cours des 6 premiers mois il a subi des variations mineures, dans le souci toutefois de conserver une compatibilité ascendante, permettant ainsi aux statisticiens de travailler sur plusieurs versions d'exports. Les dernières modifications apparaissent en rouge.
La documentation du modèle de données comporte deux fichiers :

- une notice volontairement limitée à deux pages A4 (Annexe 1)
- un description complète champ par champ, expliquant la méthode de calcul à utiliser lorsque nécessaire (un extrait est présenté en Annexe 2)

De manière synthétique, il est demandé aux partenaires de fournir des tables et fichiers texte respectant le schéma relationnel à sept tables présenté en Figure 3.

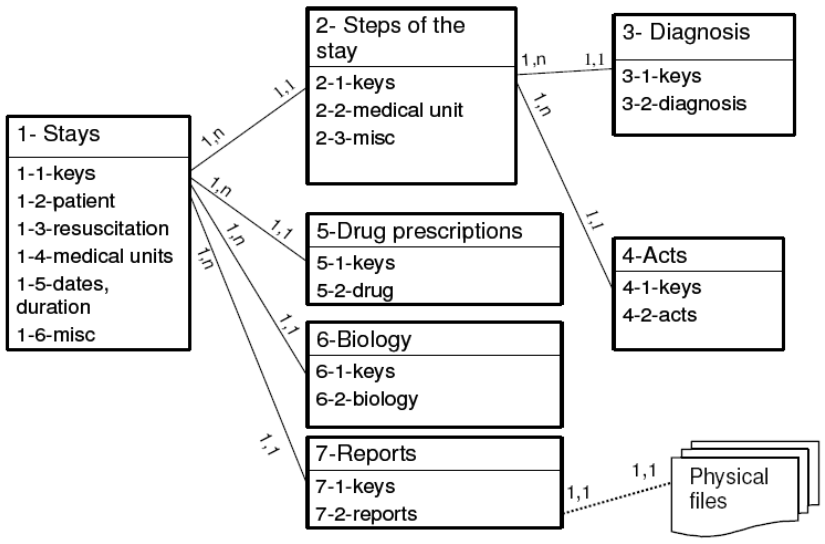


Figure 3 Schéma relationnel à 7 tables et fichiers physiques

Table 1 : Stays

Cette table contient une ligne pour chaque séjour pris en charge par l'établissement (une ligne par RSS donc). Elle contient entre autres des informations sur le patient, des données médicales (GHM, DP du séjour…), des informations sur la durée du séjour, et des informations agrégées depuis la table « steps of the stay », notamment en ce qui concerne les soins intensifs, les unités médicales traversées.

Table 2 : Steps of the stay

Cette table contient une ligne par étape de séjour (une ligne par RUM donc). Si par exemple un patient entre aux urgences puis en soins intensifs cardiologiques puis en cardiologie, à ce séjour correspondra une seule ligne dans la tables Stays mais trois lignes dans la table Steps of the stay. Elle contient notamment des informations médicales (DP du RUM) et des informations de flux (qualification de l'unité médicale, durée et position de l'étape…)

Table 3: Diagnosis of the step of the stay

Cette table contient une ligne pour chacun des diagnostics associés significatifs portés durant une étape de séjour. Les diagnostics sont encodés à l'aide de la CIM10.

Table 4: Acts of the step of the stay

Cette table contient une ligne pour chacun des actes médicaux réalisés durant une étape de séjour. En France les actes sont encodés à l'aide de la CCAM. Pour l'instant les actes ne sont pas utilisés car la CCAM est une classification purement française, à l'inverse de la CIM10. Cependant Il sera important de prendre en compte au minimum les actes nécessitant l'injection d'iode. Cet iode sera alors considéré comme un médicament.

Table 5 : Drug administration of the stay

Cette table permet de décrire les médicaments administrés durant le séjour. La manière dont la chronologie des prescriptions est prise en compte ainsi que la notion d'association de médicaments est explicitée dans l'annexe 1.

Table 6 : Biology

Cette table permet de décrire les paramètres biologiques mesurés ainsi que les résultats obtenus tout au long du séjour. Il existe une ligne par mesure et par instant de la mesure.

Table 7 : Reports & Physical files

Lorsque l'établissement est en mesure de fournir des compte-rendus anonymisés (courriers de sortie, compte-rendu opératoire, compte rendu d'hospitalisation, compte-rendu d'examen complémentaire…), il lui est demandé de fournir un fichier en texte brut ASCII par courrier. La table 7 permet de faire la jointure entre les noms de fichiers physiques et le séjour.

Les données : recueil, assurance qualité

I. Extraction des données

Les données ont été obtenues auprès de 4 partenaires :

- Le CHRU de Lille
- Le CH de Denain
- Le CHU de Rouen
- Les hôpitaux de la région de Copenhague

Tous les partenaires ne sont pas en mesure d'extraire les 7 tables en raison d'une dématérialisation variable du dossier patient. Sans rentrer dans le détail, l'extraction des données a demandé et continue à demander un travail conséquent. Les personnes chargées de l'extraction varient selon les partenaires :

- ressource interne à l'établissement
- recrutement spécifique par l'établissement
- éditeur informatique qui déploie le système d'information dans l'établissement
- ressource centrale du projet

La présente étude est réalisée à partir de 2688 séjours issus des hôpitaux de la région de Copenhague. Ces séjours correspondent à l'activité du court séjour gériatrique de l'année 2007, exception faite des séjours pour lesquels le patient entre et sort le jour même.

II. Assurance qualité des données

L'assurance qualité des données est un processus essentiel et encore très consommateur de ressources à ce jour. Chaque fois qu'un hôpital envoie un nouveau jeu de données, les fichiers sont automatique traités à l'aide d'un script développé à l'aide du langage R[11]. Ce script génère un rapport complet en HTML avec images (à l'aide de la librairie R2HTML[12]). Ce rapport décrit toutes les variables une par une :

- les variables qualitatives et binaires sont décrites à l'aide de tables de contingences et de camemberts
- les variables quantitatives sont décrites à l'aide d'histogrammes, de courbe de densité de probabilité, de boxplots et de courbes quantile-quantile de comparaison à une loi normale
- les identifiants sont écrits à l'aide d'histogrammes (on s'attend à une distribution uniforme d'intervalle connu à l'avance)

Des connaisseurs de ce type de données sont alors chargés de vérifier les données et leur cohérence avec les connaissances existantes :

- les données respectent-elles le schéma de données (i.e. les bonnes colonnes à la bonne place, les bonnes jointures…)
- les valeurs présentes dans chaque colonne sont-elles conformes aux types ou motifs attendus ? (par exemple le nombre de diagnostic doit être un entier positif, les codes diagnostiques CIM10 doivent respecter le motif d'expression rationnelle suivant : ^[A-Z]\d{2}\.?\d*$ …)
- les valeurs obtenues sont-elles compatibles avec les valeurs attendues (par exemple le taux de décès doit être voisin de 2-5%)

- les colonnes en rapport avec la même question sont-elles concordantes (par exemple : « passage en soins intensifs oui/non » et « durée du passage en soins intensifs »)

Les rapports d'erreurs sont systématiquement écrits et renvoyés aux fournisseurs de données en suivant un processus itératif (Figure 4).

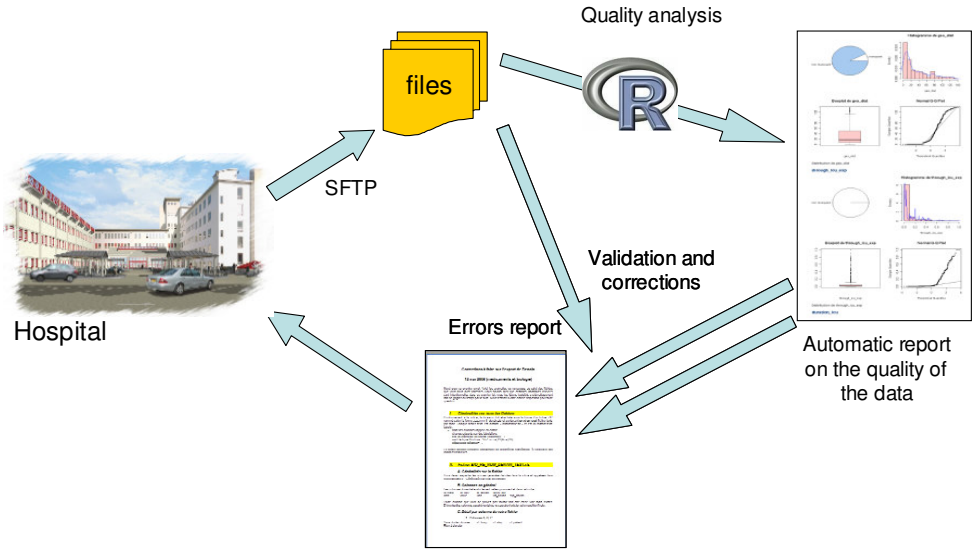


Figure 4 Processus itératif d'assurance qualité

Ce processus d'assurance qualité est important tout d'abord pour vérifier la qualité des données elles-mêmes, mais surtout afin d'améliorer les programmes d'extraction développés par chaque partenaire. Les anomalies des résultats sont les symptômes des anomalies de la mécanique sous-jacente.

Transformation des données

Le schéma relationnel à 7 tables conçu précédemment est sans doute une bonne manière d'exprimer l'information envoyée par les partenaires, néanmoins aucune méthode statistique n'est capable de tirer profit d'une telle structure. L'objectif de la transformation des données est de rendre la structure des données exploitable par les méthodes statistiques.

I. Considérations préalables sur l'information

A. Catégories d'information

Quelles que soient les tables exceptée la table des courriers, toutes les informations peuvent être classées en trois catégories :

- des variables qui pourraient constituer les causes ou contextes des ADE (par exemple l'âge, l'existence d'une insuffisance rénale, une prescription…)
- des variables qui pourraient témoigner d'un ADE (par exemple le décès, l'apparition d'une hémorragie durant le séjour…)
- les identifiants et les autres variables, que nous n'exploiterons pas

On peut appliquer cette dichotomie aux diagnostics :

- les maladies existant lors de l'admission du patient sont des causes ou contextes potentiels d'ADE. On peut identifier deux sous-catégories :
    - les maladies chroniques, existant avant l'admission
    - les maladies aiguës, qui sont alors souvent le motif d'admission
- à l'inverse les maladies qui apparaissent durant le séjour pourraient être des manifestations d'ADE

Ce raisonnement est plus facile à appliquer à la biologie :

- les anomalies biologiques qui existent à l'admission sont des causes ou contextes potentiels d'ADE (on ne sait pas si elles sont chroniques ou non, mais elles préexistent)
- les anomalies biologiques qui surviennent durant le séjour pourraient être des manifestations d'ADE

Ce raisonnement reste également le même pour les médicaments :

- les prescriptions faites durant le séjour sont naturellement des causes ou contextes potentiels d'ADE
- mais certaines prescriptions peuvent également être considérées comme des manifestations d'ADE. C'est le cas généralement des antidotes et contre-mesures. Ainsi, la prescription de vitamine K est un bon indicateur de surdosage en AVK.

La Figure 5 ci-dessous montre comment l'information disponible pourrait être classifiée. Quatre grands types d'information sont disponibles :

- l'information médicale , contenant des informations de flux et de démographie, et des diagnostics codés en CIM10. Pour prendre l'exemple des diagnostics, une partie relève des maladies chroniques, une autre des manifestations possibles d'un ADE.
- les médicaments, parmi lesquels la prescription du jour, qui est potentiellement une cause ou contexte d'ADE, et les antidotes qui seraient au contraire les manifestations d'ADE
- la biologie, parmi laquelle on retrouve encore autant les causes et contextes que les manifestations d'ADE
- et enfin les courriers et compte-rendus, dont l'exploitation automatisée relèvera du *semantic mining* et non du *data mining*. Cependant ils seront lus par les experts chargés de valider les règles issues du *data mining*.

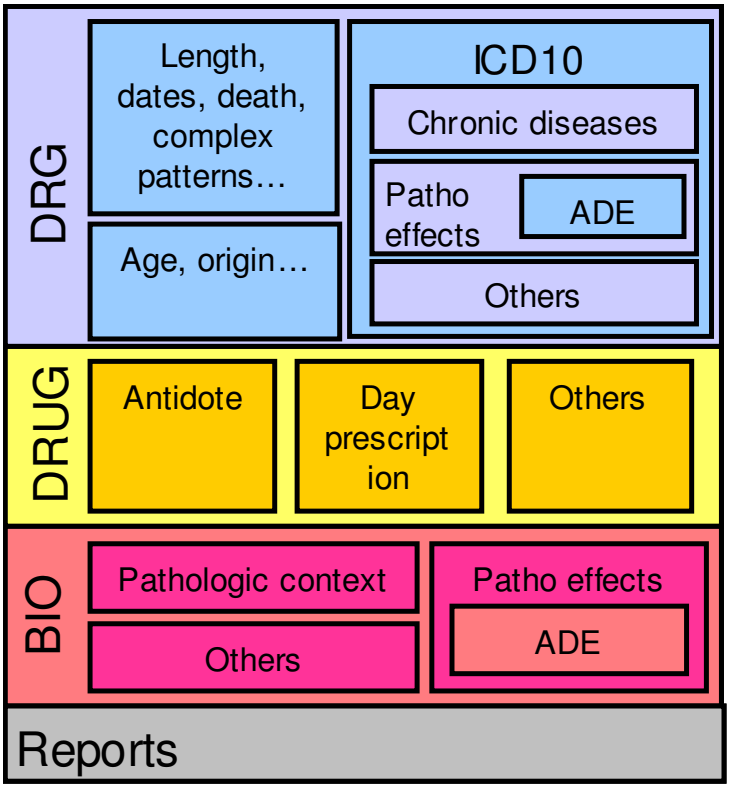


Figure 5 Classification de l'information disponible

B. Catégories d'information et génération de règles

La première étape du *data mining* sera d'utiliser les variables « manifestations potentielles d'ADE » pour identifier des profils de séjours atypiques. Dans un deuxième temps nous tenterons d'établir un lien entre les variables de « causes ou contextes » et ces profils. L'identification de ce lien permettra de générer de la connaissance qui pourra être traduite sous forme de règles d'association. Ces règles seront validées et modifiées par des experts. Plus tard, dans leur forme ultime, elles permettront de générer des alertes en fonction des variables de « causes ou contextes » disponibles au moment de la prescription.

Il est important de souligner que, lorsqu'un médecin prescrira un traitement sur le CPOE, une partie seulement des l'information de « causes ou contextes » sera disponible. On trouvera par exemple les médicaments prescrits jusqu'au jour dit, la biologie mesurée

depuis l'admission jusqu'au jour de la prescription. Pour ce qui est des diagnostics, ils sont codés après le départ du patient donc nous n'aurons pas d'information. En revanche il sera possible de récupérer une partie de l'information sur les maladies chroniques si le patient a déjà été hospitalisé préalablement.
Ce processus est représenté sur la Figure 6. Ce schéma reprend le schéma de la Figure 5, à ceci près que l'information qui n'est pas utilisée lors d'une étape apparaît en grisé.

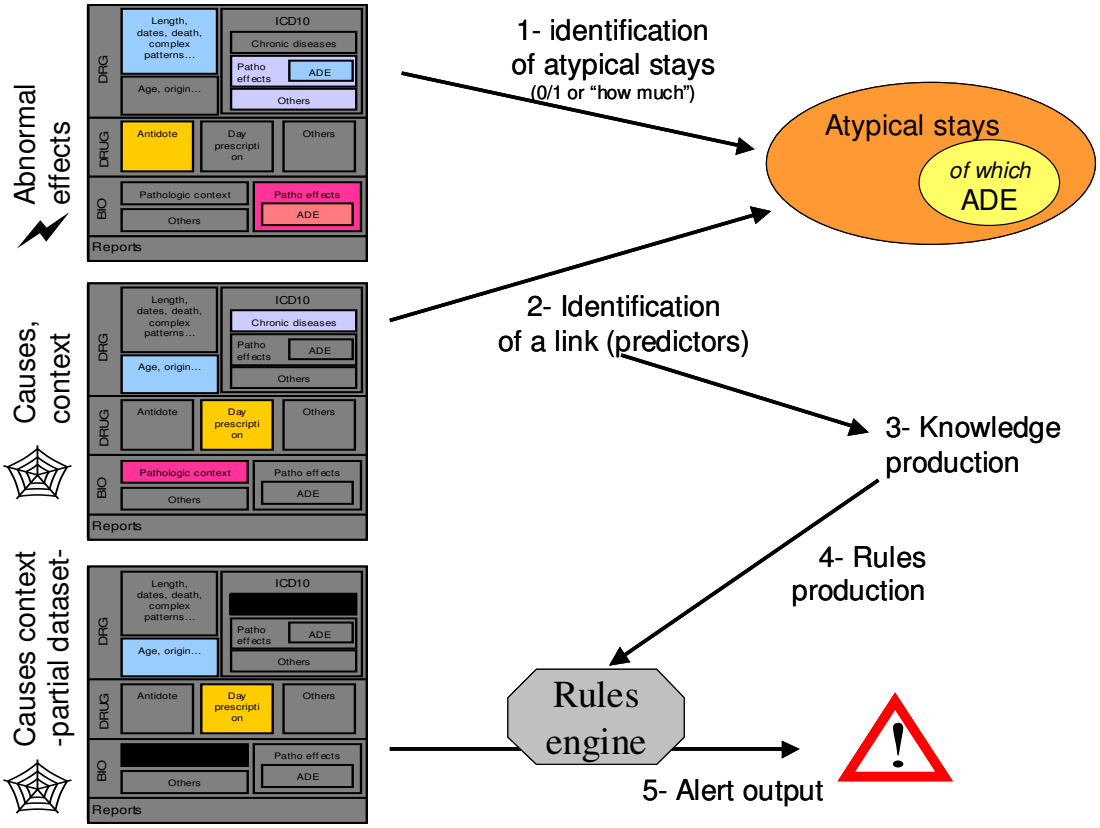


Figure 6 Comment les causes/contextes et les manifestations potentielles d'ADE permettent de générer des règles

Puisque le but semble d'établir un lien de causalité entre la prescription d'un médicament et la survenue d'un effet indésirable, il convient de rappeler qu'association statistique ne signifie pas causalité. Certes le médicament pourrait être responsable de l'ADE, mais le médicament pourrait être associé à l'ADE statistiquement pour plusieurs autres raisons :

- cette prescription est fréquemment associée avec une autre prescription
- l'utilisation du médicament est le symptôme d'une maladie chronique
- le médicament est administrée parce que l'ADE se produit
- …

II. Préparation des données

A. Transformation à plat : interprétation de variables en fonction d'une valeur de référence

Comme nous l'évoquions plus haut certaines variables ne peuvent être interprétées que par rapport à une référence, en particulier pour ce qui est des variables de la catégorie « manifestation potentielle d'un ADE » lorsque vient le moment de rechercher des

séjours atypiques. Si nous prenons pour exemple la durée du séjour, une durée de 10 jours est anormalement longue pour une appendicite. En revanche pour un séjour de soins palliatifs cela peut paraître assez court. Nous devons donc prendre en compte le groupe homogène de séjours (GHS) et la valeur attendue de la durée de séjour (espérance mathématique) dans le GHS du séjour. Le problème est le même pour le décès oui/non, le passage en soins intensifs oui/non ou la durée de séjour en soins intensifs…

Nous devons donc générer de nouvelles colonnes « à plat », c'est-à-dire dans la même table (colonne A et colonne B permettent de calculer la colonne C qui seule sera utilisée), afin de rendre compte de l'atypie de la valeur observée au lieu de rendre compte de la valeur observée elle-même. On peut utiliser plusieurs méthodes en fonction des types de variables observées (quantitatives, binaires, qualitatives), de la distribution et surtout de la sémantique de la variable initiales.

Exemples de calculs simples :
AtypieVariableQuantitative = (variable – moyenne)/ écartType
AtypieVariableQuantitative = max(0, (variable – moyenne) / écartType)
AtypieVariableQuantitative = Si(variable > moyenne + 2 * écartType, 1 , 0)
Etc…

B. Agrégation des données, mise à plat du schéma relationnel

1. Objectif

Prenons l'exemple des diagnostics associés significatifs (DAS) encodés en CIM10. Pour un séjour donné il peut y avoir de 0 à 99 DAS différents. Ces DAS se comportent comme une variable qualitative à près de 18 000 modalités différentes. Si nous souhaitions faire disparaître la jointure entre la table des séjours (ou des étapes de séjours) et celle des DAS nous devrions créer dans la table des séjours 18 000 variables binaires différentes, ce qui est naturellement impossible, et même si ça l'était une description aussi détaillée serait parfaitement inutilisable, faisant chuter l'espérance mathématique de chaque nouvelle colonne (prévalence de la valeur 1).
Le problème est similaire pour les médicaments, avec près de 9 000 codes ATC.
Le problème est encore plus complexe pour la biologie, car un même paramètre (par exemple : nombre de globules rouges) peut être mesuré plusieurs fois dans le séjour, avec à chaque fois une valeur différente à comparer avec les bornes inférieure et supérieure de normalité du paramètre si elles existent.

2. Principes généraux du processus d'agrégation

Le processus d'agrégation suit plusieurs principes :

- L'agrégation est réalisée par un « moteur d'agrégation ». Les « politiques d'agrégation » sont déclarées à l'extérieur du moteur, afin de permettre à des personnes différentes de maintenir et améliorer la politique d'agrégation. Il existe une indépendance complète entre ces entités.
- Ainsi est-il possible d'utiliser différentes politiques d'agrégation en fonction du contexte, de la profondeur et de la précision souhaitées :
    - Des politiques d'agrégation lâches permettront de générer peu de catégories et de maintenir une espérance mathématique raisonnable dans chaque colonne, ce qui peut être utile lorsqu'on analyse des jeux de données contenant peu de séjours (exemple : un service spécifique)

    - Des politiques d'agrégation précises et détaillées permettront de générer un nombre important de catégories suffisamment précises, qui nécessiteront des effectifs importants pour être utilisables.
  - Les catégories sont décidées en suivant plusieurs principes (ceci sera explicité plus bas) :
    - Toutes les catégories sont d'abord décidées en suivant un point de vue nosologique ou académique
    - Certaines catégories transversales ou redondantes permettent de regrouper de nouveau certaines modalités déjà intégrées dans des catégories académiques
- Les politiques d'agrégation ont été décidées par l'équipe dans un premier temps, mais ces politiques pourront aisément être améliorées, modifiées, mises à jour si nécessaire, y compris en fonction des résultats obtenus.

3. Le principe des catégories redondantes

Nous avons défini des catégories redondantes chaque fois que cela nous a paru nécessaire. Voici quelques exemples.

a- Exemple avec les diagnostics

Les cancers du rein peuvent être un cofacteur d'un ADE de plusieurs manières :
- la tumeur peut obstruer les voies urinaires ou détruire le tissu rénal fonctionnel
- en tant que cancer, on peut retrouver des effets généraux (syndrome paranéoplasique) ou liés aux traitements anti-cancéreux administrés
- certains effets spécifiques des cancers rénaux peuvent apparaître (atteinte des glandes surrénales…)

On pourrait décider de mettre en place, comme le font les nomenclatures, des catégories non redondantes.
- on pourrait choisir de mettre les cancers du rein dans les « anomalies morphologiques du rein », perdant alors le lien avec les cancers en général
- on pourrait choisir de créer une large catégorie des cancers, perdant alors tout lien avec l'insuffisance rénale par exemple ou les glandes surrénales
- on pourrait enfin créer une catégorie « cancer du rein », on perdrait alors tout lien avec les autres cancers, et on fragmenterait les causes d'insuffisance rénale. De plus l'effectif des séjours positifs serait trop faible.

Au contraire il est possible de définir des variables avec chevauchement :

variables générées / Contexte clinique	Anomalie morphologique du rein	cancer
Cancer du rein	**1**	**1**
Anomalies morphologiques du rein non cancéreuses (…)	**1**	**0**
Cancers ne touchant pas le rein	**0**	**1**

La combinaison de ces deux variables permettra alors d'interpréter simplement l'arbre de décision :
- si seule la condition [Anomalie morphologique du rein=1] apparaît, la branche concerne les anomalies morphologiques du rein
- si seule la condition [cancer=1] apparaît, la branche concerne les cancers en général

- si la condition [Anomalie morphologique du rein=1 & cancer=1] apparaît, la branche concerne les cancers du rein.

b- Exemple avec les médicaments

L'approche est similaire pour les médicaments. Par exemple la Rifampicine apparaît dans un sous-groupe des antibiotiques, mais aussi dans le groupe « médicaments inducteurs enzymatiques ». Cela est encore plus nécessaire pour les associations de médicaments : l'association est prise en compte simultanément dans chacune des catégories des composants princeps. Il est probable que beaucoup d'ADE apparaissent à cause d'un excès de confiance dans les associations toutes prêtes. Cet excès de confiance pourrait être lié au fait que le médicament n'est plus considéré comme une somme de principes actifs, mais comme la solution clef en main à une pathologie médicale. C'est avec les associations qu'on pense trouver le plus de redondances et de contradictions lors de la prescription. On notera pour conforter cette impression que la classification ATC elle-même n'aborde pas les médicaments par classe pharmaceutique ni par pharmacodynamique, mais bien par utilisation finale. Ainsi la classe "R05 Cough and cold preparations" inclut des combinaisons d'alcaloïdes de l'opium et d'expectorants. Les AINS quant à eux se retrouvent éparpillés dans une dizaine de catégories, selon leur utilisation et leur association.

c- Synthèse

Utilisées avec les arbres de décision, ces catégories redondantes peuvent en réalité se comporter comme des politiques d'agrégation dont la précision changerait dynamiquement : cela permettrait d'ajouter de la précision et du sens sans obtenir des effectifs trop faibles dans les classes.

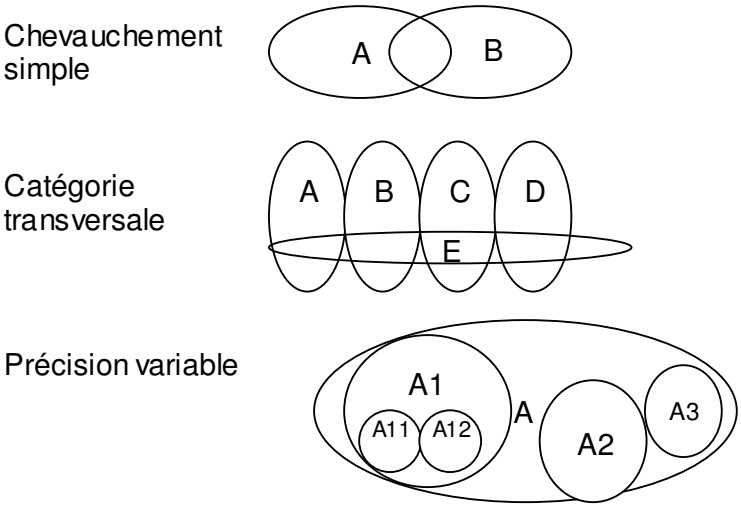


Figure 7 Exemples de catégories redondantes

C. La génération des variables binaires pas à pas

1. Transformation binaire des diagnostics

Un moteur d'agrégation est utilisé pour transformer le schéma relationnel en une seule grande table à plat. Cette table contient une ligne par séjour, elle est utilisée pour les analyses statistiques (au bas de la Figure 8).

Le moteur d'agrégation des diagnostics utilise une déclaration externe de politique d'agrégation. Cette politique d'agrégation décrit les catégories de diagnostics prises en compte. Chaque catégorie servira à générer une colonne binaire, cette colonne prendra pour valeur 1 si au moins un des diagnostics correspondants apparaît, et 0 dans les autres cas.

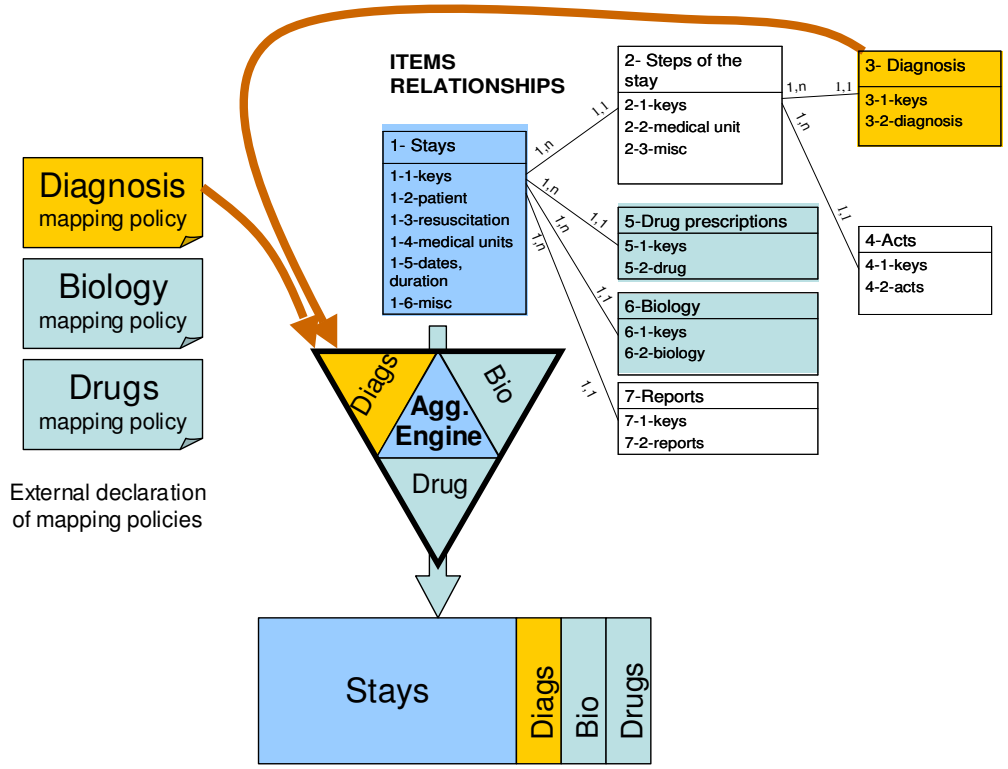


Figure 8 Processus d'agrégation des diagnostics

Voyons l'agrégation des diagnostics en détail sur un exemple simple. La table des séjours contient une ligne par séjour. Chaque séjour peut contenir une ou plusieurs étapes, et à chaque étape peuvent correspondre des diagnostics. Dans la Figure 9 le premier séjour et les informations correspondantes apparaissent en rose, le second séjour en vert.

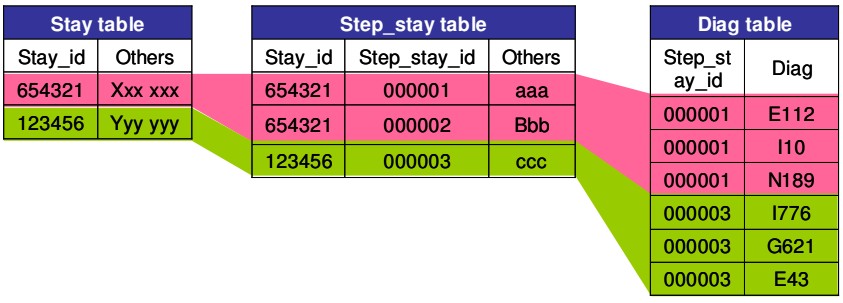


Figure 9 Extrait du schéma relationnel : du séjour aux diagnostics

Soit une politique d'agrégation (fictive) nommée « problème rénal ». Elle identifie une liste de codes CIM10 dont la présence témoigne d'une problème rénal. Cette politique permet au moteur d'agrégation d'identifier 2 diagnostics dans le premier séjour, et aucun dans le second séjour (Figure 10).

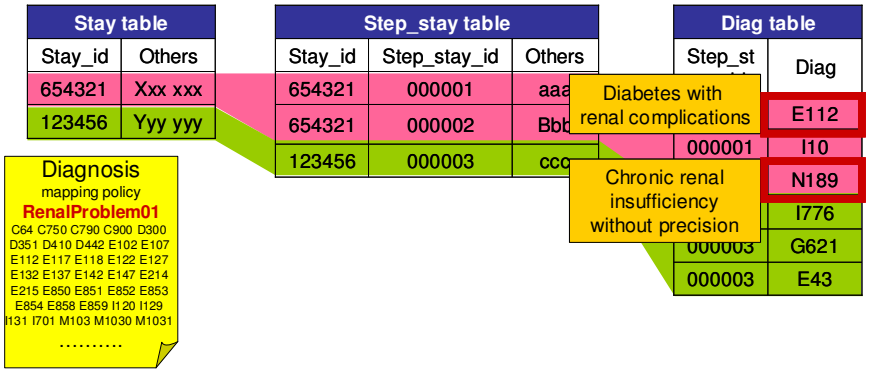


Figure 10 La politique d'agrégation « problème rénal » permet au moteur d'agrégation d'identifier 2 diagnostics compatibles.

Le moteur d'agrégation des diagnostics utilise cette information et génère la table à plat des séjours avec une colonne binaire « diagnostic de problème rénal ». Elle indiquera pour valeur 1 pour le premier séjour et 0 pour le second séjour (Figure 11).

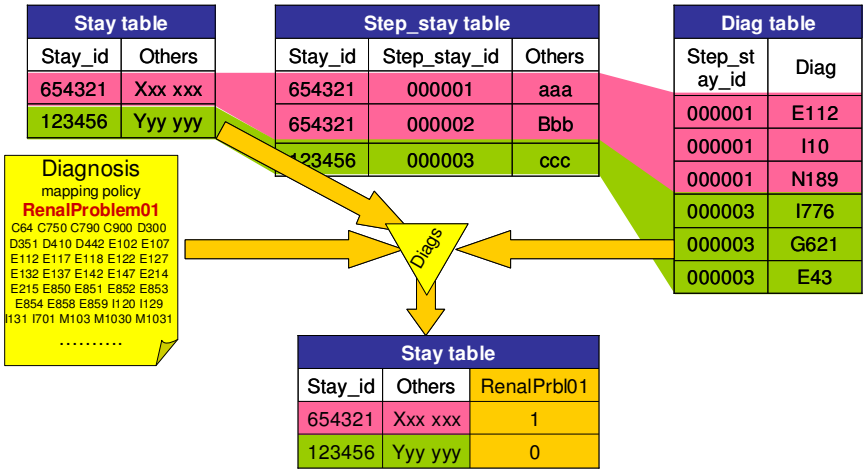


Figure 11 La politique d'agrégation « problème rénal » ajoute une colonne binaire à la grande table à plat

On voit bien dans cet exemple que le schéma de la grande table à plat (nombre et nom des colonnes) est déterminé par les politiques d'agrégation déclarées. C'est donc un schéma dynamique qui est susceptible de changer à chaque modification des politiques d'agrégation.

2. Transformation binaire des prescriptions de médicaments

Afin d'incorporer les prescriptions médicamenteuses nous avons implémenté un moteur d'agrégation des médicaments qui utilise une déclaration externe de politique d'agrégation des médicaments. Ce moteur ajoute plusieurs colonnes binaires à la grande table à plat (Figure 12).

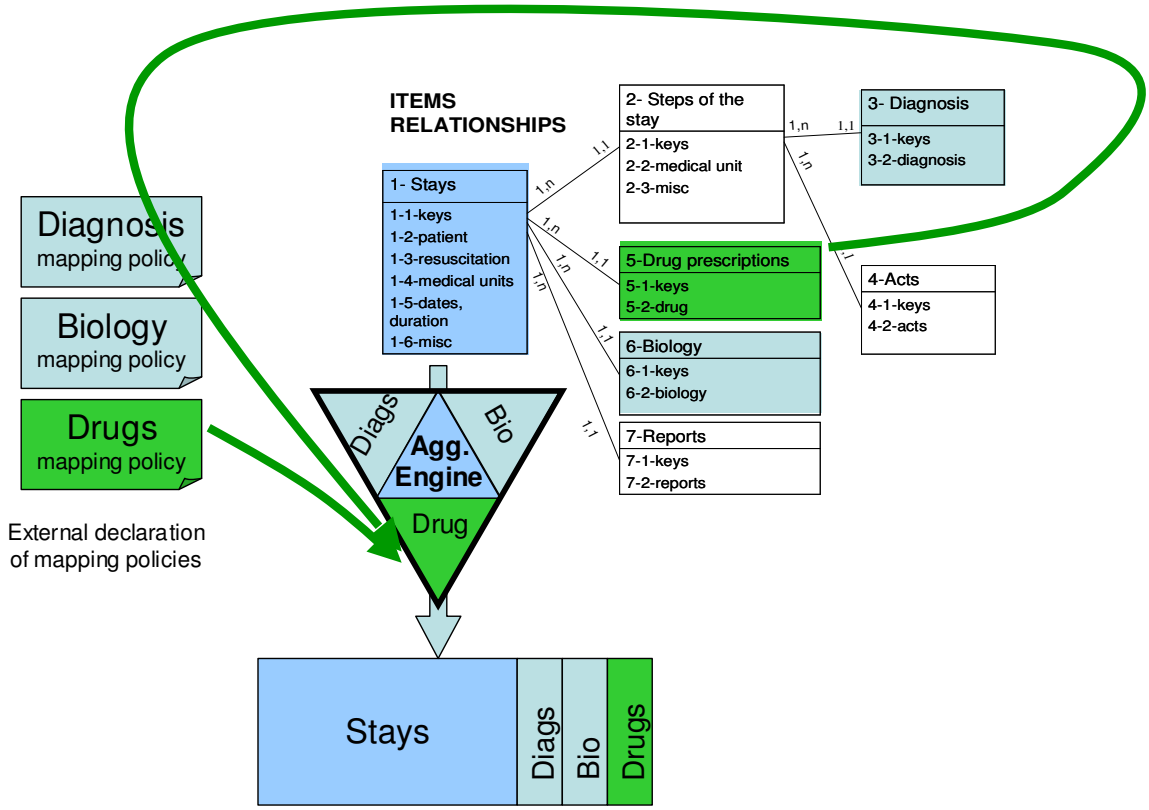


Figure 12 Processus d'agrégation des médicaments

Le processus est très proche de celui montré pour les diagnostics. La politique d'agrégation permet de définir des classes de médicaments, et d'expliciter la liste de codes ATC correspondants. Le moteur identifie ainsi toutes les prescriptions qui correspondent aux listes (Figure 13).

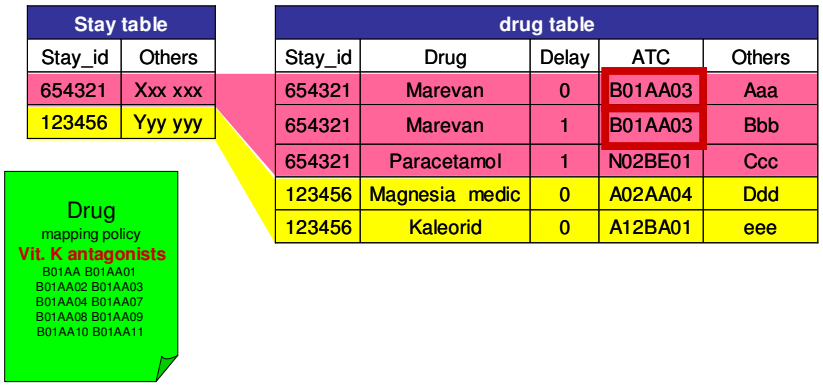


Figure 13 La politique d'agrégation « anti-vitamine K » identifie les différentes prescriptions compatibles

Le moteur d'agrégation des médicaments utilise cette information et génère dans la grande table à plat autant de colonnes binaires que de catégories, donc ici une colonne binaire pour les anti-vitamine K. Le premier séjour contient au moins un code ATC d'AVK, il prendra pour valeur 1 tandis que le deuxième séjour prendra pour valeur 0 (Figure 14). De surcroît, la date de la première prescription d'AK est mémorisée dans une autre colonnes. Cette date est exprimée en décalage par rapport à la date d'admission (offset), comme toutes les dates du séjour.

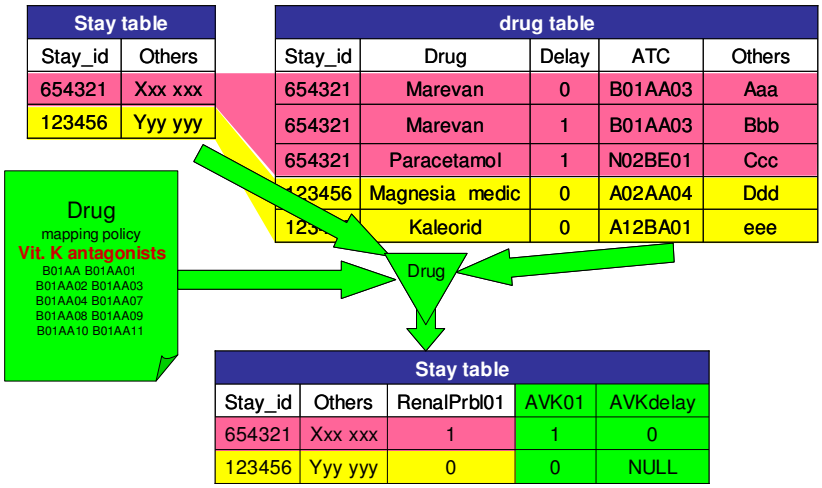


Figure 14 La politique d'agrégation « anti-vitamine K » ajoute une colonne binaire ainsi qu'une colonne de délai à la grande table à plat

3. Transformation binaire des mesures de biologie

Afin d'incorporer les mesures de biologie nous avons implémenté un moteur d'agrégation de la biologie qui utilise une déclaration externe de politique d'agrégation de la biologie. Ce moteur ajoute plusieurs colonnes binaires à la grande table à plat (Figure 15).

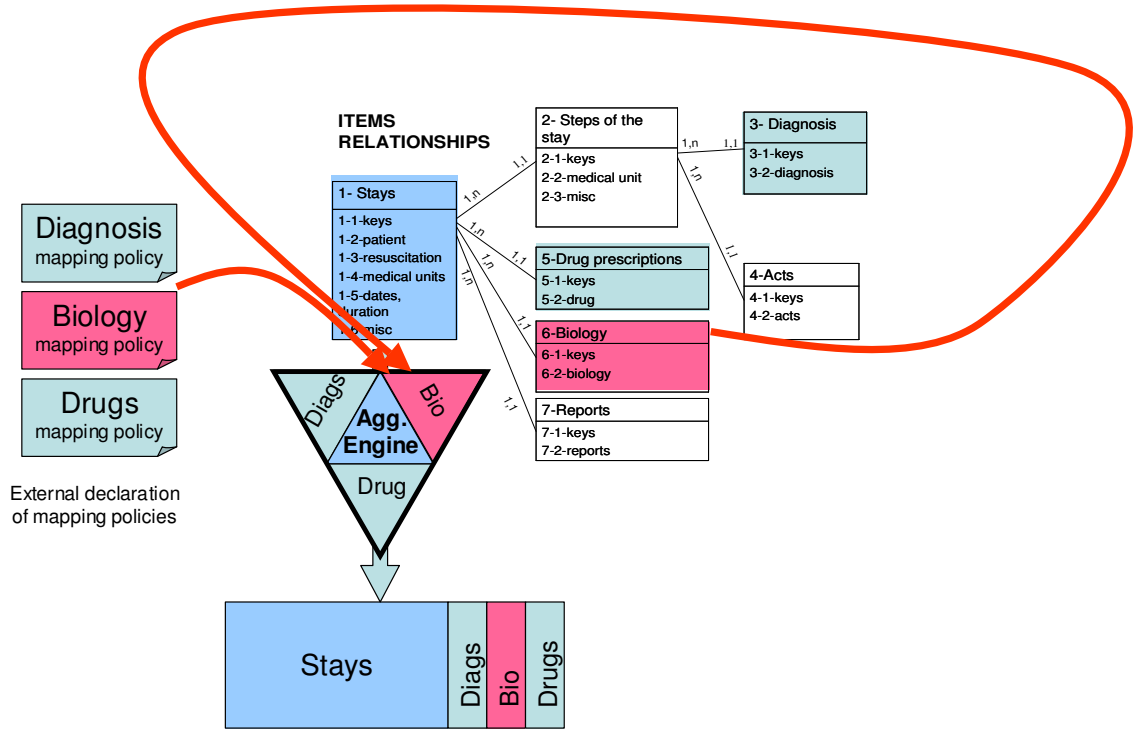


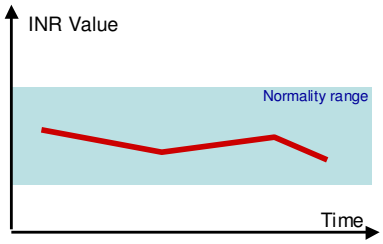
Figure 15 Processus d'agrégation de la biologie

Le processus d'agrégation est plus complexe. En fonction de sa signification médicale, un paramètre biologique peut engendrer jusqu'à 4 variables binaires. Prenons l'exemple de la mesure de l'INR (international normalized ratio) qui est un paramètre biologique qui permet d'explorer une voie de la coagulation. Son suivi permet de savoir si un patient sous anti-vitamine K est bien équilibré ou non. Une augmentation de l'INR au-delà de la

zone thérapeutique traduit un surdosage qui pourrait engendrer des hémorragies. Une diminution de l'INR en-deçà de la zone thérapeutique traduit un sous-dosage qui pourrait engendrer des thromboses. L'INR permet la création de 4 nouvelles colonnes binaires :

- causes ou contextes – INR trop élevé à l'admission (0/1)
- causes ou contextes – INR trop bas à l'admission (0/1)
- effets – survenue d'un INR trop élevé durant le séjour (0/1)
- effets – survenue d'un INR trop bas durant le séjour (0/1)

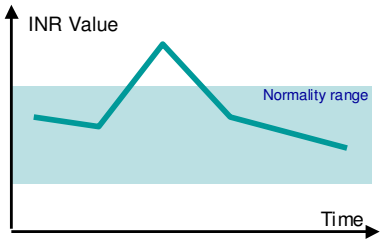
Si le paramètre n'a jamais été mesuré, compte tenu du caractère quasi-systématique que revêt son exploration dans la démarche clinique, nous pouvons considérer qu'il n'y a aucune anomalie. Il en est de même si toutes les mesures sont normales (Figure 16).



Variable	Value
Cause/Ctxt – too low INR	0
Cause/Ctxt – too high INR	0
Effect – too low INR	0
Effect – too high INR	0

Figure 16 Aucune anomalie de l'INR durant le séjour

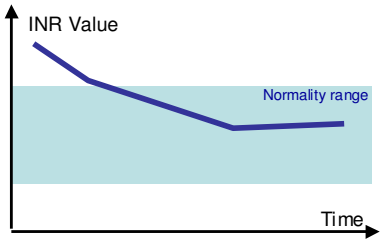
Si la première mesure connue est normale mais qu'une anomalie apparaît durant le séjour, nous considérons que c'est peu-être une manifestation d'un ADE (Figure 17).



Variable	Value
Cause/Ctxt – too low INR	0
Cause/Ctxt – too high INR	0
Effect – too low INR	0
Effect – too high INR	1

Figure 17 INR : effet indésirable

Si la première valeur connue est anormale alors nous considérons que l'anomalie existait lors de l'admission du patient (Figure 18).



Variable	Value
Cause/Ctxt – too low INR	0
Cause/Ctxt – too high INR	1
Effect – too low INR	0
Effect – too high INR	0

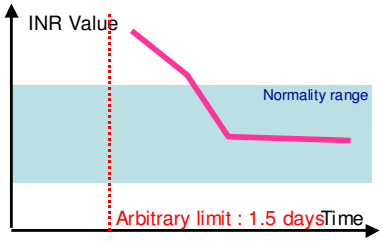
Figure 18 INR : cause ou contexte

En réalité la première mesure n'est pas forcément prise le premier jour du séjour. Nous avons fixé une limite arbitraire d'un jour et demi :

- lorsque la première mesure est anormale

- si cette mesure est prise le jour de l'admission ou le lendemain (ce qui est le cas d'un patient admis à 23h30 et dont un paramètre est mesurée 45 minutes plus tard) alors nous considérons que l'anomalie existait avant l'admission du patient (maladie chronique, motif d'hospitalisation, affection aiguë concomitante)
- si la mesure est prise après cette limite arbitraire, compte tenu du caractère quasi-systématique des mesures, on peut supposer que le paramètre n'a pas été mesuré avant parce qu'il n'y avait pas d'anomalie clinique auparavant en rapport avec ce paramètre. On peut alors supposer que la prise d'une mesure qui s'avère anormale est déclenchée par l'apparition d'une anomalie, qui pourrait être un ADE (*Figure 19*).

- Si la première mesure connue est normale, alors le cas a été traité plus haut.



Variable	Value
Cause/Ctxt – too low INR	0
Cause/Ctxt – too high INR	1
Effect – too low INR	0
Effect – too high INR	0

Figure 19 INR : effet indésirable

Dans ce premier modèle du moteur d'agrégation de la biologie nous n'exploitons que les paramètres biologiques qui satisfont deux conditions :

- le paramètre biologique est fréquemment ou systématiquement mesuré, à tel point que l'absence de mesure traduit nécessairement une absence d'anomalie clinique en rapport avec ce paramètre.
- le paramètre peut être interprété seul sans ambiguïté. Par exemple, afin d'identifier les hyperthyroïdies et les hypothyroïdies, nous utilisons uniquement la triiodothyronine (T3), la thyroxine (T4) et la T4 libre, mais pas la TSH (thyroid-stimulating hormone). Effectivement la TSH ne peut être interprétée que si l'on connaît la valeur d'une des hormones périphériques ou le contexte clinique qui a poussé le clinicien à prescrire l'examen.

Autres remarques :

- autant que possible les paramètres biologiques sont traités en utilisant la classification IUPAC lorsqu'elle est disponible auprès de l'établissement
- plusieurs paramètres biologiques peuvent être utilisés en direction du même concept. Par exemple la saturation de l'hémoglobine en oxygène, la pression artérielle en oxygène et la concentration artérielle en oxygène sont utilisées pour la même variable binaire. Le résultat le plus péjoratif est conservé : nous utilisons l'opérateur logique « OU » pour la détection d'anomalies.

4. Prochaines évolution du moteur d'agrégation

Nous prévoyons de faire évoluer le moteur d'agrégation :

- nous testerons les catégorisations déjà existantes, les ontologies…

- la date des anomalies biologiques et des prescriptions sera prise en compte dans l'analyse statistique. Ainsi, si une anomalie à j+10, une anomalie qui serait survenue à j+8 sera prise en compte comme une cause ou contexte, alors que,

en tant qu'effet, elle est pour l'instant ignorée. Ce concept de *cascade d'événements* impose une approche dynamique qui utilise des délais flottants..

- la limite arbitraire de 1.5 jours utilisée pour la biologie est la même quel que soit le paramètre. Cette valeur pourrait dépendre du paramètre biologique et en particulier de son inertie cinétique : un délai court pour les paramètres qui varient très vite, un délai long pour les paramètres qui varient lentement.

Analyse statistique, arbres de décision

I. Principes de l'analyse

A. Statistiques univariées

Nous avons systématiquement réalisé des statistiques univariées tant sur les données sources que sur les variables binaires générées par les moteurs d'agrégation, en guise d'assurance qualité.

B. Définition dynamique de la formule

La définition de la formule utilisée (effet ~cause1 + cause2 + cause3 …) ne pouvait être faite « en dur » car cela posait un réel problème de maintenance et d'évolution de l'analyse. Elle est donc écrite dynamiquement à la volée car le schéma de la grande table à plat est dynamique, les quelques 200 variables sont susceptibles de changer à chaque instant :

- la création des variables est directement commandée par la déclaration des politiques d'agrégation. Pour faciliter la maintenance du système, à aucun moment les scripts n'invoquent nominativement des variables.
- les déclarations de politiques d'agrégation ne sont pas les mêmes pour tous les partenaires. En particulier la biologie est traitée de manière très variable car les paramètres mesurés ne sont pas les mêmes dans tous les établissements. Il en résulte donc des variables différentes dans la grande table à plat.
- les colonnes constantes doivent être supprimées avant le traitement statistique, et ces colonnes changent naturellement selon les jeux de données utilisés.

Cependant le processus qui suit est entièrement automatique, et son exécution en quelques secondes n'est pas perturbée par cette grande fluidité.

Nous avons donc adopté une charte de nommage des variables. Le programme statistique parcourt donc les variables et identifie leur type en fonction de cette convention de nommage à l'aide des expressions rationnelles :

- **id_*** désigne les identifiants, qui ne seront pas utilisés pour l'analyse mais seront conservés par l'export des résultats
- **mi_cau_*** désigne une variable issue de l'information médicale et cause ou contexte potentiel d'un ADE
- **mi_eff_*** désigne une variable issue de l'information médicale et manifestation potentielle d'un ADE
- **bi_cau_*** désigne une variable issue de la biologie et cause ou contexte potentiel d'un ADE
- **bi_eff_*** désigne une variable issue de la biologie et manifestation potentielle d'un ADE
- **dr_cau_*** désigne une variable issue des prescriptions médicamenteuses et cause ou contexte potentiel d'un ADE
- **dr_eff_*** désigne une variable issue des prescriptions médicamenteuses et manifestation potentielle d'un ADE
- toutes les autres variables sont supprimées

Il est ainsi aisé d'identifier toutes les variables de causes (*_cau_*) et les variables d'effet (*_eff_*).

C. Régressions logistiques

Les premières analyses utilisaient des régressions logistiques. Cela avait plusieurs avantages :

- réalisation immédiate et interprétation accessible à de nombreuses personnes
- automatisation et réalisation à la volée rapides
- en théorie la régression logistique permet de calculer un risque relatif
- elle permet d'utiliser des variables binaires et quantitatives en tant que variables explicatives

Néanmoins la régression logistique présente plusieurs inconvénients majeurs dans ce contexte :

- utilisée avec plusieurs dizaines de variables explicatives souvent liées, elle expose au sur-ajustement et les résultats ne sont plus fiables. De manière générale, les variables explicative devraient être faiblement corrélées et en nombre nettement inférieur au nombre de séjours analysés. Il faudrait également faire apparaître les interactions dans le modèle.
- au-delà d'une dizaine de variables explicatives significativement liées à la variable à expliquer il n'est plus possible de donner une interprétation médicale sérieuse du résultat : les causes ne s'enchaînent plus mais participent à un magma ininterprétable
- l'interprétation du risque relatif est illusoire car elle suppose en réalité que les risques liés à chaque variable ne font que s'additionner, alors que l'expérience montre que la potentialisation est nettement supérieure : il existe une interaction très forte entre certaines variables.

D. Procédures A et B d'analyse en arbres de décision

Nous avons choisi d'utiliser des arbres de décision. Le principe est décrit plus bas (II Les arbres de décision utilisés dans la procédure B).

Nous envisageons deux procédures. La procédure A est celle vers laquelle nous nous orientons pour les mois qui viennent, la procédure B est celle mise en place pour fournir des résultats dans le livrable soumis à la commission européenne.

La procédure A comprend deux phases (Figure 20) :

- premièrement utiliser toutes les variables « effet » pour identifier des séjours atypiques ou des clusters de séjours atypiques, dont certains seront clairement des ADE
- deuxièmement, une fois ces groupes caractérisés, utiliser toutes les variables de causes ou contextes pour établir un lien avec cette atypie

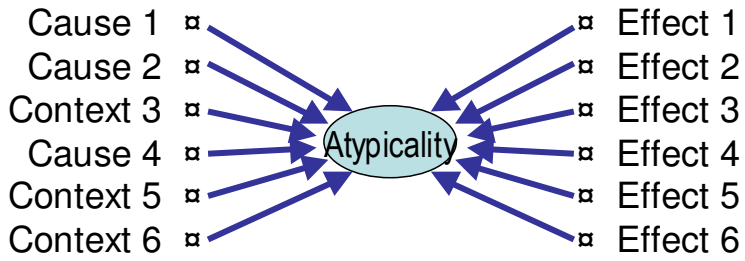


Figure 20 Procédure A de recherche de règles d'association

La procédure B nous a permis d'obtenir des résultats plus rapidement en considérant les effets un par un. Ainsi, pour chaque variable d'effet prise séparément des autres, nous essayons d'établir une relation avec toutes les variables de causes ou contextes (Figure 21).

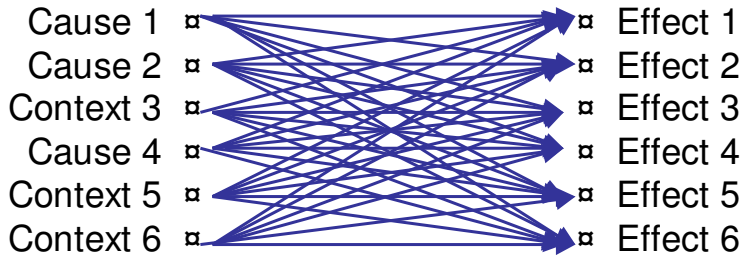


Figure 21 Procédure B de recherche de règles d'association

Le principal défaut de la procédure B est de considérer qu'une anomalie biologique isolée est intéressante à expliquer, ce qui n'est pas nécessairement le cas. Il est vrai que toute hyperkaliémie mérite impérativement d'être prévenue tant le risque vital est avéré, en revanche tel n'est pas forcément le cas d'une diminution modérée de l'INR. De plus on peut s'attendre à ce que certaines diminutions de l'INR soient accompagnées de manifestations thrombotiques, et d'un allongement du séjour voire d'un décès, il est alors un peu dommage de considérer tous ces effets séparément les uns des autres.
Quoi qu'il en soit cette première approche a déjà permis d'obtenir des résultats intéressants, et nous permet d'affiner le développement de la procédure A. Pour chaque variable d'effet nous obtenons donc un arbre de décision. Chaque arbre a été analysé par des médecins afin de savoir quelles règles retenir.

II. Les arbres de décision utilisés dans la procédure B

Dans la procédure B nous cherchons l'association entre chaque variable d'effet prise isolément et toutes les variables de cause / effet. La grande table à plat est analysée à l'aide du logiciel de programmation en statistiques R [11]. Pour chaque variable d'effet nous produisons un arbre de décision à l'aide de la bibliothèque[13] de R, cette bibliothèque utilise l'algorithme CART[14].

Les méthodes d'arbres de décision sont devenues très populaires en raison de leur utilisation simple flexible et puissante. Plus encore, les résultats sont faciles à interpréter et ont une application pratique directe. Ces méthodes sont amplement utilisées en recherche biomédicale car le résultat s'apparente à la démarche diagnostique utilisée en médecine. L'implémentation informatique des règles issues de ces méthodes est alors relativement simple. L'avantage majeur dans notre cas est qu'un nombre de variables

explicatives élevé ne perturbe nullement le déroulement de l'analyse, contrairement aux modèles de régression qui exposent au risque de sur-ajustement.

La démarche de construction des arbres se détaille ainsi. Dans une premier temps il faut trouver à la racine la variable la plus liée statistiquement à la variable d'effet à expliquer (la méthode choisie est un des points qui discrimine les différentes méthodes [15]). Admettons que cette variable soit binaire, il en découle naturellement une branche dans laquelle la prévalence de la variable à expliquer augmente, une autre dans laquelle la prévalence diminue. On dit alors qu'on a diminué l'impureté du nœud. On poursuit ensuite dans chaque branche, conditionnellement au choix précédent (v1=0 dans une branche, v1=1 dans l'autre) donc sur un effectif restreint, de manière à diminuer à chaque fois l'impureté des nœuds. La variable choisie à chaque étape n'est donc pas nécessairement la même dans chaque banche, et chaque branche pourra avoir une longueur variable. Il existe également une méthode pour déterminer quand le processus itératif prend fin :

- soit dans une approche prospective : dans chaque branche on décide de poursuivre ou non, on parle de pré-élagage
- soit rétrospectivement : l'arbre est construit intégralement tant qu'il existe une manière de segmenter chaque branche, que cette segmentation paraisse pertinente ou non, puis l'arbre subit un post élagage (« pruning »)

Les méthodes les plus utilisées sont les méthodes CART et CHAID. La méthode CHAID utilise le test du Chi2 (ou plus exactement le t de Tschuprow qui tient compte du nombre de degrés de liberté) pour déterminer à chaque nœud la variable la plus pertinente. Dans cette logique, c'est ce même critère qui est utilisé en pré-élagage pour déterminer quand les segmentations s'arrêtent. La méthode CART quant à elle utilise le concept d'impureté, et réalise un post-élagage en comptant les individus mal classés dans un échantillon de test. Cette méthode est réputée plus robuste du fait de l'existence d'un échantillon de test. Informatiquement la procédure n'est pas vraiment plus lourde.

Selon les méthodes, il est possible de prendre en compte différents types de variables explicatives :

- les variables binaires sont toujours utilisables
- les variables qualitatives sont parfois utilisables
- les variables quantitatives peuvent être utilisées lorsqu'elles sont binarisées. Le seuil utilisé peut être la médiane dans l'échantillon, ou un seuil déterminé automatiquement afin de maximiser la statistique de test utilisée pour la segmentation.

Quelle que soit la méthode utilisée, il en résulte des règles de décision du type :
SI $X_1=0$ ET $X_5=1$ ET $X_{13}=0$ ALORS $Y=1$
(où Y est la variable à expliquer et $X_{1...n}$ sont les variables explicatives)
Cette règle qualifie une branche dont on peut également définir :

- **le support :**
  rapport entre l'effectif présent dans la branche et l'effectif total
  $\#(X_1=0 \cap X_5=1 \cap X_{13}=0) / \#(\Omega)$
  (ou parfois plus simplement l'effectif présent dans la branche)
- **la confiance :**
  estimation de la probabilité que Y égale 1 si les conditions sont réunies
  $p(Y=1 \mid X_1=0 \cap X_5=1 \cap X_{13}=0)$
  autrement dit plus généralement lorsque Y est une variable quantitative, la moyenne de Y dans l'échantillon

III. Résultats

La procédure B a permis de générer 31 arbres correspondant à 31 variables biologiques d'effet, et 223 règles en rapport avec ces règles. Ces règles font appel à 76 variables différentes sur les 173 variables explicatives soumises aux arbres.

IV. Un exemple détaillé de résultat de la procédure B : apparition d'un INR trop bas durant le séjour

Dans cet exemple nous cherchons l'association entre le seul effet « apparition d'un INR trop bas durant le séjour » et toutes les variables de cause / effet. Un INR trop bas est une anomalie biologique qui témoigne d'un sous-dosage en anti-vitamine K, il pourrait en résulter des thromboses.
La grande table à plat est analysée à l'aide du logiciel de programmation en statistiques R [11]. Les arbres de décision sont générés à l'aide de la bibliothèque de R, cette bibliothèque utilise l'algorithme CART. Nous supprimons préalablement toutes les colonnes constantes.

On obtient alors l'arbre de décision présenté en Figure 22.

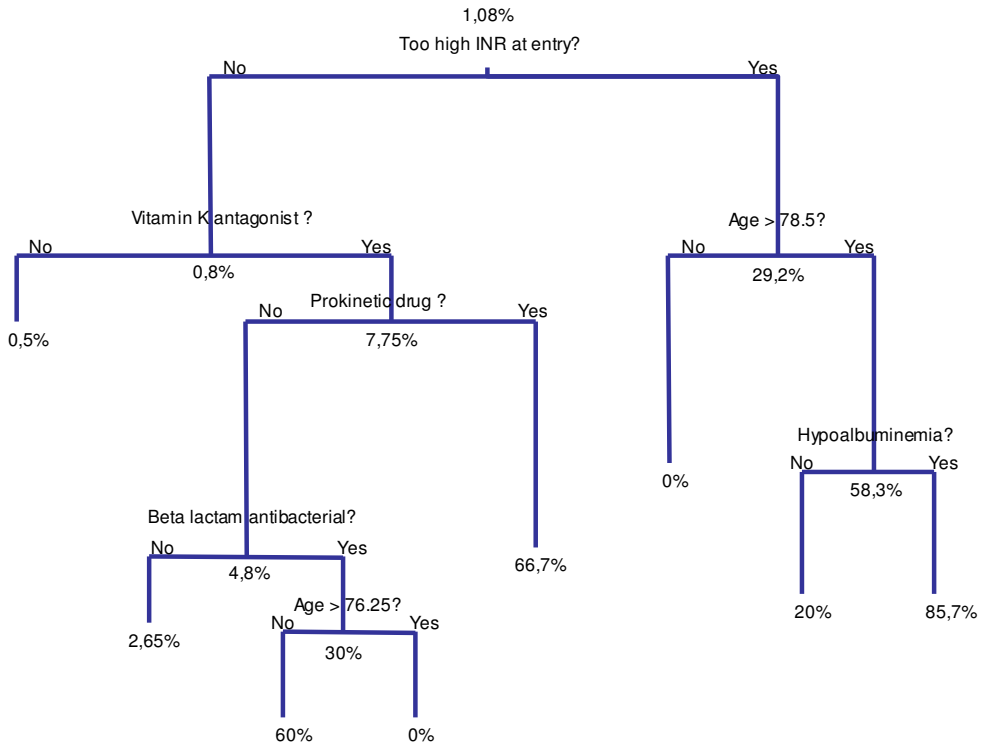


Figure 22 Arbre de décision « apparition d'un INR trop bas »

L'arbre est une succession de tests logiques dont le but est de diminuer l'impureté de chaque nœud. La variable la plus liée à l'apparition d'un INR trop bas est la variable « INR trop élevé à l'admission » (Figure 23). Dans la population totale la proportion de séjours avec apparition d'un INR trop bas est de 1.08%, elle devient 29.2% pour les patients avec un INR trop élevé à l'admission, et 0.8% dans les autres cas.

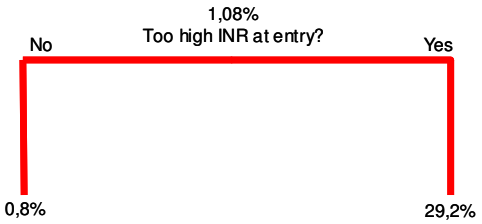


Figure 23 Premier test

Nous nous intéressons uniquement aux branches qui permettent d'augmenter la proportion de survenue de l'effet. Ce premier test permet de générer une règle (Figure 24).

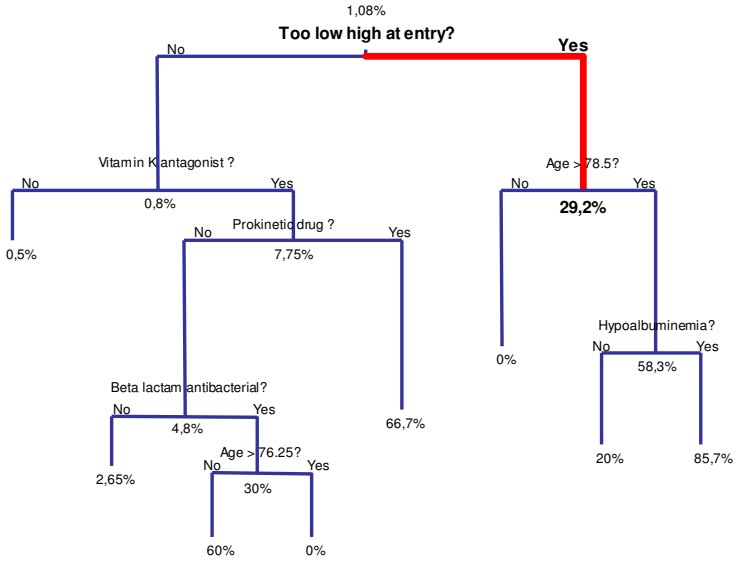


Figure 24 Première règle

Règle N° 1 : *INR trop élevé à l'admission => apparition d'un INR trop bas*
Confiance : *29.2%*
Support : *24*
Interprétation : *dans ce service, les patients qui sont admis avec un INR trop élevé sont souvent sur-corrigés*
Gravité : *0% de décès, durée moyenne de séjour=6.4 jours*

Une fois dans une branche il est possible de continuer. La confiance des règles augmentera et leur support diminuera (Figure 25).

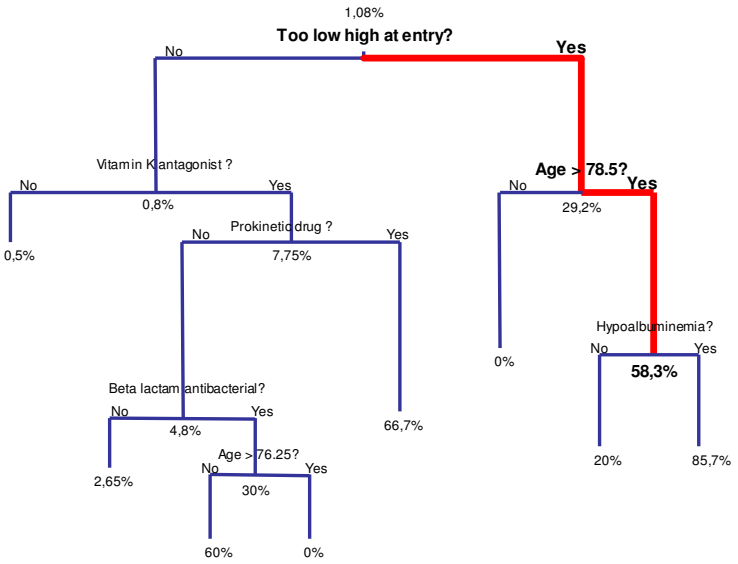


Figure 25 Deuxième règle

Règle N°2 : *INR trop élevé à l'admission ET âge>78.5 => apparition d'un INR trop bas*
Confiance : *58.3%*
Support : *12*
Interprétation: *dans ce service, les patients âgés qui sont admis avec un INR trop élevé sont très souvent sur-corrigés*
Gravité : *0% de décès, durée moyenne de séjour=9.3 jours*

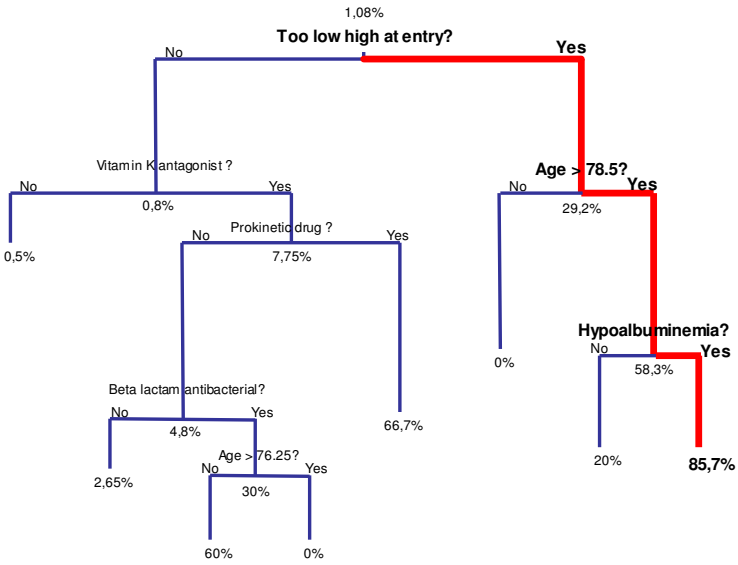


Figure 26 Troisième règle

Règle N°3 : *INR trop élevé à l'admission ET âge>78.5 ET hypoalbuminémie à l'admission => apparition d'un INR trop bas*
Confiance : *85.7%*
Support : *7*
Interprétation: *dans ce service, les patients âgés qui sont admis avec un INR trop élevé sont très presque toujours sur-corrigés si cet INR trop élevé est lié à une hypoalbuminémie. L' hypoalbuminémie augmente la biodisponibilité des*

anti-vitamine K (AVK). La correction du surdosage en AVK devrait être assortie de précautions car l'hypoalbuminémie augmente l'effet de la correction.

Gravité : *0% de décès, durée moyenne de séjour=13.4 jours*

On peut également étudier les branches « négatives » mais seules les règles qui augmentent la prévalence de l'effet nous intéressent, c'est pourquoi nous passons outre l'interprétation de la première étape (Figure 27, en vert).

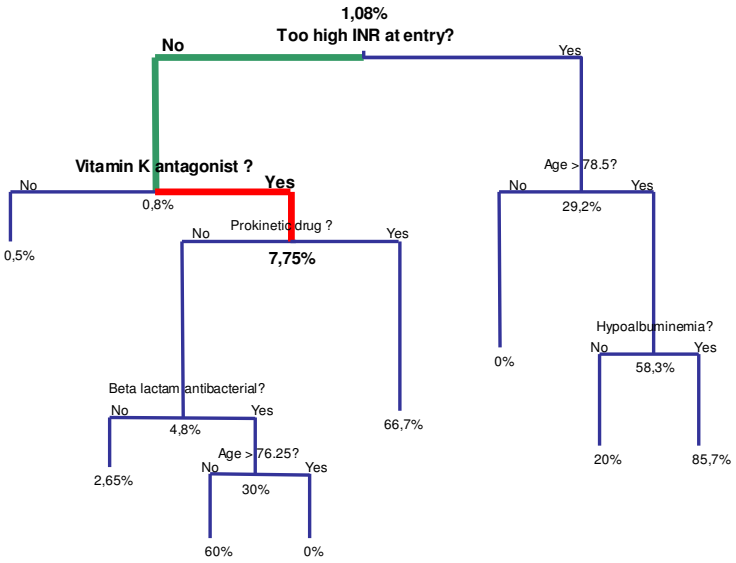


Figure 27 Quatrième règle

Règle N°4 : *Pas d'INR trop élevé à l'admission ET AVK => apparition d'un INR trop bas*

Confiance : *7.75%*

Support : *129*

Gravité : *1.55% de décès, durée moyenne de séjour=7.5 jours*

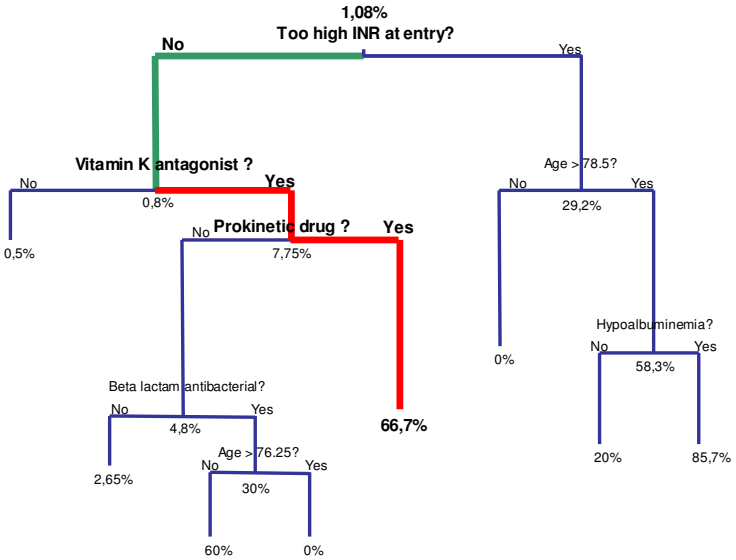


Figure 28 Cinquième règle

Règle N°5 : *Pas d'INR trop élevé à l'admission ET AVK ET prokinétique => apparition d'un INR trop bas*
Confiance : *66.7%*
Support : *6*
Interprétation: *Dans ce service, lorsqu'on administre un AVK et un prokinétique à un patient, le dosage des AVK n'est pas suffisamment adapté. Les prokinétiques sont connus pour modifier la flore bactérienne digestive et donc la production de vitamine K2.*
Gravité : *16.67% de décès, durée moyenne de séjour=15 jours*

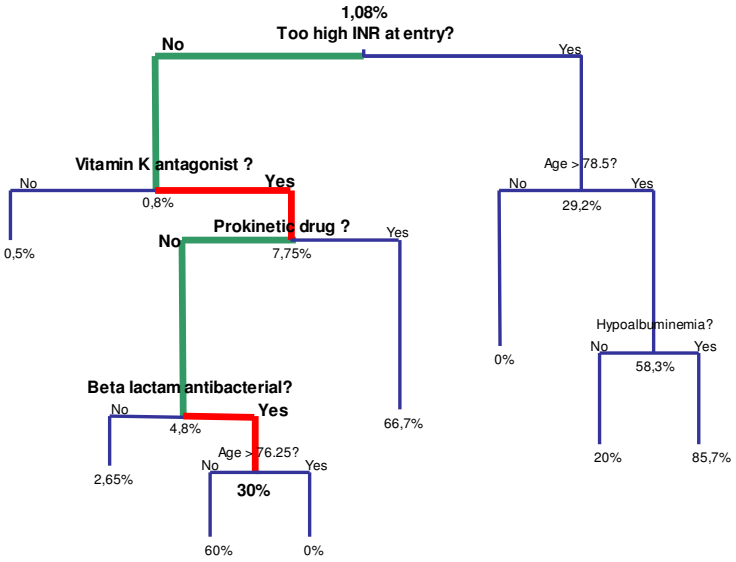


Figure 29 Sixième règle

Règle N°6 : *Pas d'INR trop élevé à l'admission ET AVK ET absence de prokinétique ET bêta-lactamine => apparition d'un INR trop bas*
Confiance : *30%*
Support : *10*
Interprétation : *dans ce service, lorsqu'on administre à un patient des AVK et une bêta-lactamine (pénicilline ou céphalosporine), le dosage des AVK n'est pas bien adapté. Les bêta-lactamines sont connues pour modifier l'activité de la flore bactérienne digestive et donc la production de vitamine K2. De surcroît les infections qui poussent à prescrire ces antibiotiques sont susceptibles d'induire un hypermétabolisme qui pourrait augmenter le catabolisme hépatique des AVK et diminuer leur activité biologique.*
Gravité : *0% de décès, durée moyenne de séjour=13.3 jours*

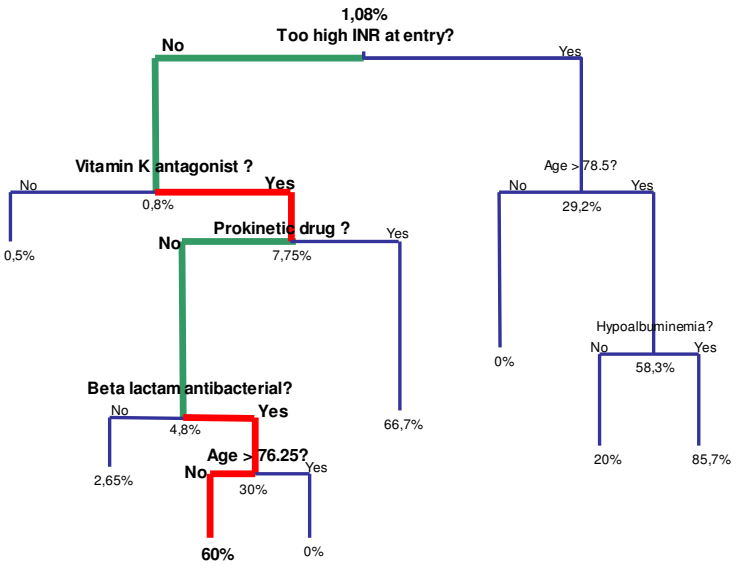


Figure 30 Septième règle

Règle N°7 : *Pas d'INR trop élevé à l'admission ET AVK ET absence de prokinétique ET bêta-lactamine ET âge<76.25=> apparition d'un INR trop bas*
Confiance : *60%*
Support : *5*
Interprétation: *l'effet décrit plus haut semble être plus important chez les patients les moins âgés*
Gravité : *0% de décès, durée moyenne de séjour=12.6 jours*

V. Perspectives…

Quelle que soit la procédure utilisée, une revue d'expert de plusieurs dossiers est indispensable, c'est l'objet du Workpackage 3. Ses objectifs sont les suivants :

- sélectionner les règles pertinentes
  - sur des arguments pharmacologiques et cliniques : les associations statistiques devront permettre d'affirmer des liens de causalité en tenant compte de la validité intrinsèque (pharmaco-dynamique et pharmaco-cinétique des produits) et de la validité extrinsèque (bibliographie, rapports d'ADE) mais aussi de l'imputabilité (arguments chronologiques)
  - sur des arguments numériques, en prenant en compte la gravité de l'effet, la confiance et le support des règles. Une règle qui mène à un effet grave avec une confiance de 20% sera conservée, alors qu'une telle confiance ne sera pas suffisante pour un effet bénin.
- en expliquant les règles fournies par le data-mining, donner des arguments ou poser des questions permettant d'affiner les éléments qui ont permis de générer ces règles, en modifiant les déclarations de politiques d'agrégation tout comme les moteurs d'agrégation
- aller au-delà de la simple alerte, et proposer un message plus approprié. Exemple : non pas *« attention : risque de surdosage en AVK »* mais *« une surveillance accrue de l'INR est recommandée »* voire une prescription automatique de l'examen biologique.

La procédure A décrite dans le chapitre [I- Principes de l'analyse] fait l'objet des travaux actuels. Nous espérons inclure ainsi des éléments de gravité liés aux effets biologiques.

Rappelons à ce propos que les ADE sont des événements qui ont des conséquences sur le déroulement du séjour du patient ou en terme de séquelles. On peut citer par exemple :

- allongement de la durée de séjour
- réalisation d'un acte médical qui n'aurait pas été nécessaire autrement
- passage en soins intensifs
- décès
- séquelles ayant une répercussion sur l'état de santé du patient, handicap
- …

En particulier un effet biologique mineur dont la correction routinière n'entraîne aucun de ces effets ne devrait pas être pris en compte.

Conclusion

Le projet PSIP est une approche novatrice dans l'univers des effets indésirables liés aux médicaments. Tous domaines confondus, dans de nombreuses études la force des hypothèses a priori et de l'argumentation bibliographique est très appréciée. Cette approche a montré ses limites dans l'études des effets indésirables médicamenteux « de la vraie vie ». Dans l'écriture de règles de contrôle, cette approche induit un nombre trop important d'alertes peu pertinentes pour les prescripteurs. Dans la surveillance des effet indésirables au contraire elle induit une sous-détection en ne prenant en compte que les cas de figure déjà décrits. Les premiers résultats montrent au contraire comment le Data Mining peut être une approche nouvelle et originale.

Les techniques statistiques utilisées en data mining n'ont rien en soi de spécifique. Leur choix tient en revanche compte de contraintes différentes par rapport aux technique conventionnelles, et on pourra citer par exemple la nécessité de pouvoir supporter un nombre important de lignes mais surtout un nombre très important de colonnes dont on ne peut assurer l'absence de corrélation entre elles. Les approches de schéma dynamique sont particulièrement appréciées car ce nombre important de colonnes et la mouvance des processus qui les génèrent poseraient un réel problème de maintenance des scripts, problème que l'on perçoit déjà dans les projets plus classiques. L'aphorisme selon lequel le data management représenterait 80% du travail d'une étude statistique semble ici un euphémisme.

Mais penser qu'une seule méthode pourrait être la solution à un problème est sans doute une erreur. Le data mining est un outil comme d'autres au service de la résolution d'un problème. En cela aussi PSIP se distingue des approches classiques d'étude de l'effet indésirable médicamenteux. La nouveauté de PSIP dans ce domaine est d'utiliser simultanément les compétences (l'ordre est ici aléatoire) des médecins, des pharmacologues, des informaticiens, des statisticiens et des ergonomes.

A peine six mois après son lancement, les premiers résultats présentés dans le livrable et dans ce mémoire sont en tout cas très prometteurs.

Bibliographie

References

1- http://www.psip-project.eu (last visited: August. 15, 2008)
2- http://erc.europa.eu/ (last visited: August. 15, 2008)
3- http://cordis.europa.eu/fp7/home_en.html (last visited: August. 15, 2008)
4- Kohn, L. T., Corrigan, J. & Donaldson, M.S. (2000): To Err Is Human: Building a Safer Health System. National Academies Press.
5- Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform. 2003 Feb-Apr;36:131-43.
6- Morimoto T, Gandhi TK, Seger AC, Hsieh TC, Bates DW. Adverse drug events and medication errors: detection and classification methods. Qual Saf Health Care. 2004 Aug;13:306-14.
7- Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting adverse events using information technology. J Am Med Inform Assoc. 2003 Mar-Apr;10:115-28.
8- Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. J Am Med Inform Assoc. 2003 Jul-Aug;10:339-50.
9- http://www.php.net (last visited: August. 18, 2008)
10- http://www.atih.sante.fr/?id=000240000DFF (last visited: August. 15, 2008)
11- R Development Core Team. R: A Language and Environment for Statistical Computing. . 2006 ;:.
12- Eric Lecoutre. R2HTML: HTML exportation for R objects. . 2006 ;:.
13- Ripley, B. D. (1996): Pattern Recognition and Neural Networks. Cambridge University Press.
14- Breiman, L. (1993): Classification and regression trees. .
15- http://eric.univ-lyon2.fr/~ricco/doc/tutoriel_arbre_revue_modulad_33.pdf (last visited: August. 15, 2008)

Tables & Figures

Annexe 1 : Notice du modèle de données destinée aux fournisseurs de données

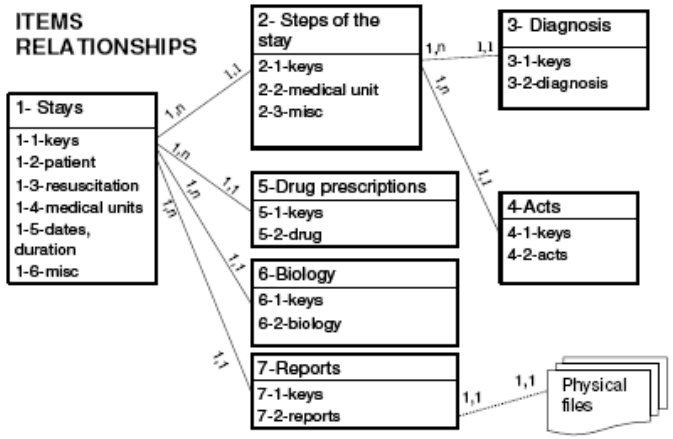
File version: 2008-06-04

Description of the tables

The following data scheme is expected by the WP2 from the WP1. This is naturally not a specification of storage schema. That's why those tables are not normalized. We suggest to first read the description of the tables 2 3 and 4. Then the table n°1 will be easier to understand.

1. Table1: Stays
   - One row = one stay = a discharge standardized case record (RSS in France)
   - Some data correspond to aggregated data from the others tables.
   - Please read table 2 fields description before trying to read table 1 fields description.
2. Table2: Steps of the stay
   - One row = one step of one stay = one medical unit case report (RUM in France)
3. Table 3: Diagnosis of the step of the stay
   - One row = one of the associated diagnosis recorded in a medical unit case report.
   - 0 to k rows per step of stay.
4. Table 4: Acts of the step of the stay
   - One row = one of the acts recorded in a medical unit case report.
   - 0 to k rows per step of stay.
5. Table 5: Drug administration of the stay
   - One row = one drug administered during one day of one stay.
   - One drug could appear several times during the stay. Please refer to the "important notes" section.
     0 to k rows per stay.
6. Table 6: Biology
   - One row = one biology datum recorded during the stay. (If the same sample is given using several units, you can provide us one or the other measure: we will simply compare it to the bounds you will provide us.)
   - One kind of biology record could appear several times during the stay.
     0 to k rows per stay
7. Table 7: Reports
   - One row = one report related to the stay. A physical file should correspond to this row
   - 0 to k rows per stay

Selection criteria:
Please select **all the stays** of your hospital where ("A and B and C"):
- the stay is in year 2007 (depending on your system you can use either the admission or the exit date)
- the duration of the stay is greater than 1 (exit_date – admission_date > 1)
- the stay is in medicine surgery or obstetrics (excluding psychiatrics, rehabilitation care…)



Important notes

Output format
Please use upload your files using the following format:

- each table is a tabulated text file:
  - a first line contains the names of the fields
  - fields values are not delimited using quotes nor double quotes
  - fields are separated using the tab character (\t)
  - lines are separated using the Windows end of line sequence (\r\n) or car(13)+car(10)
  - the decimal separator is the dot (.)
  - the NULL values are written using the (\N) convention. Do not use (NULL) nor ().
  - the name of the file is *name_of_the_table* + ".txt"
- the tables are grouped together into a ZIP file
  - the name of the ZIP file is "*aaaa-mm-jj*_lille.zip" or "*aaaa-mm-jj*_rouen.zip" or "*aaaa-mm-jj*_denain.zip" or "*aaaa-mm-jj*_copenhagen.zip" where *aaaa-mm-jj* should be replaced with the date of data extraction
- another ZIP file contains the physical files of the reports if possible
  - the name of the ZIP file is "*aaaa-mm-jj*_lille_files.zip" or "*aaaa-mm-jj*_rouen_files.zip" or "*aaaa-mm-jj*_denain_files.zip" or "*aaaa-mm-jj*_copenhagen_files.zip" where *aaaa-mm-jj* should be replaced with the date of data extraction
  - The files should be present at the root of the ZIP file. If you really need to create sub-directories the complete pathway should be reported in the "filename" field of the "reports" table. See complete table description for details.

ICU Intensive Care Units (about Resuscitation, Intensive Care and Continuous Monitoring Units, and about "cold units"):
We have chosen to distinguish the Intensive Care Units (ICU) from the other medical units.
For the French hospitals:

- [ICU=1] will include both "resuscitation units" and "intensive care units"
- [ICU=0] will concern every other units, including "continuous monitoring units". Those units are sometimes called "cold units".

For Danish hospitals:

- [ICU=1] will include both Danish levels of intensive care units
- [ICU=0] will concern every other units. Those units are sometimes called "cold units".

Life basins and health territory
Those fields have been suppressed.

Simplified acute physiological score (SAPS)
A gravity score can be computed during days spent in resuscitation units: the SAPS (simplified acute physiological score) (IGS2 in France). If this score is not available for the current stay or if it is never computed in your hospital, please leave blank.

ICD10 typography
Please don't forget to use the following typography for ICD10 codes (as well as for joints with the tables we provide): delete the dots "." or any existing space and use upper letters.

Delays and durations
The durations will always be computed using the following formula:
exit_date - entry_date +1 (equals 1 when the patient arrives and exits on the same day)
The delays will always be computed using the following formula:
event_date – reference_date (equals zero when the event happens on the same day)

Drug records
A new line should appear in the drug table each day a drug is administered to the patient.
Example : the patient arrives on Monday; the drugA is prescribed and is delivered 10mg twice a day on Tuesday and Wednesday. The records are:

patient_id	*name*	*atc*	*delay_drug*	*dose*	*unit*	*route*
123	drugA	atcA	1	20	mg	oral
123	drugA	atcA	2	20	mg	oral

Now let us suppose that drugA (10mg) corresponds to 2 ATC codes atcA (4mg) and atcB (6mg). Then:

patient_id	*name*	*atc*	*delay_drug*	*dose*	*unit*	*route*
123	drugA	atcA	1	8	mg	oral
123	drugA	atcB	1	12	mg	oral
123	drugA	atcA	2	8	mg	oral
123	drugA	atcB	2	12	mg	oral

Annexe 2 : Extrait de la description des champs destinée aux fournisseurs de données (19 champs/76)

Table	Heading	Field (short name)	Field (long name)	kind	origin	unit	description
1- Stays {stay}	1- Keys	id_hosp	Hospital ID number	constant field	constant	an ID number	Please use these identifiers : 1 Denain 2 Lille 3 Rouen 4 (to 13 if needed): Capital Region of Copenhagen
		id_stay	Stay ID number	datum existing as this	database	an ID number	
		id_patient	Patient ID number	datum existing as this	database	an ID number	
	2- Patient	age	Age	simple computation or computation on a simple joint	database	years (float)	the age of the patient at the beginning of the stay : (entry_date-birth_date)/365.25
		sex	Sex	simple computation or computation on a simple joint	database	0/1	- 1 for men - 0 for women
		drg	DRG (Diagnosis Related Group)	datum existing as this	database	DRG code	
		death_01	Death during the stay	simple computation or computation on a simple joint	database	0/1	- 1 if the patient dies - 0 in the other case
		death_exp	Expected proportion of death in this DRG	external joint	the proportion in the whole hospital for each DRG	proportion, float between 0 and 1	You should compute the proportion of death in the whole hospital and not only for the selected stays. Example of computation assuming you have a [stay] table in which the [death_01] field tells you if the patient dies during the stay: SELECT drg, AVG(death_01) FROM stay GROUP BY drg ;
		geo_dist	Distance from the hospital	external joint	geographic reference	kilometers	the distance between the patient's home and the hospital (as the crow flies) in kilometers. The center of the ZIP code area can be used. If not available the value is \N.
		geo_state_01	Does the patient comme from the hospital's country (state) ?	external joint	constant	0/1	If the patient usually lives in the state (Denmark or France) where the hospital stands then 1. Otherwise 0.
		geo_region_01	Does the patient comme from the hospital's region ?	external joint	geographic reference	0/1	If the patient usually lives in the administrative region where the hospital stands then 1. Otherwise 0.
		geo_dpt_01	Does the patient comme from the hospital's department ?	external joint	geographic reference	0/1	If the patient usually lives in the administrative department where the hospital stands then 1. Otherwise 0.

	p_diag	Principal diagnosis	datum existing as this	database	ICD10 code	Please pay attention to the typography : dots (.) and spaces should be deleted.
3- Resuscitation	through_icu_01	Taken care of in intensive care/resuscitation unit ?	complex computation	database	0/1	-1 if the patient went through a resuscitation unit or an intensive care unit at least once during this stay - 0 in the other cases Example of computation once you got the [step_stay] table containing an [icu_01] field telling you if the step of the stay occured in an ICU unit or not : SELECT id_hosp, id_stay, max(icu_01) as through_icu_01 FROM step_stay GROUP BY id_hosp, id_stay ;
	through_icu_exp	expected proportion of stays with intensive care/resuscitation for this DRG	complex computation	the proportion computed in the whole hospital for each DRG	proportion, float between 0 and 1	You should compute this proportion in your hospital. The reference should be the whole hospital and not only the stays that are selected for PSIP. Example of computation assuming you have a [stay] table containing an [through_icu_01] field telling if the step of the stay occured in an ICU/resuscitation unit or not : SELECT drg, avg(through_icu_01) FROM stay GROUP BY drg
	duration_icu	Duration in an intensive care/resuscitation unit	complex computation	database	days (integer)	- IF the patient went through an intensive care or resuscitation unit, = date_of_exit-date_of_entry+1 in that unit (or the sum if several units). - ELSE, then 0 Example of computation once you got the [step_stay] table containing an [icu_01] field telling if the step of the stay occured in an ICU/resuscitation unit or not, and a [duration] field : SELECT id_hosp, id_stay, sum(icu_01*duration) FROM step_stay GROUP BY id_hosp, id_stay ; Nota bene: a tiny overestimation could occur using that formula if a patient is transferred from an ICU to another. This approximation is acceptable.
	duration_icu_exp	Expected duration in an intensive care/resuscitation unit	complex computation	the average duration computed in the whole hospital for each DRG	days (float)	You should compute the average duration in intensive care per DRG for your whole hospital and not only for the stays that are selected for PSIP. See the fields above for examples of computation.
	saps	Gravity score	simple computation or computation on a simple joint	database	integer	A gravity score is computed during days spent in ICU : IGS2 (indice de gravité simplifié) or SAPS (simplified acute physiological score). The maximum gravity score during the stay should be used. If the SAPS is not available for this stay or for the whole hospital, then use the **\N** value.
	delay_icu	Delay before ICU/resuscitation step	complex computation	database	integer	- If the patient is in a "cold" unit and is transferred into an intensive care/resuscitation unit, fill with the delay before transfer (ICU_entry_date - stay_entry_date). - if the patient directly enters the hospital into an ICU, then the value is zero (0) - in other case, if the patient doesn't go through an ICU, then the value is **\N**