Université de Lille 2 DROIT ET SANTÉ

Faculté de médecine
Henri Warembourg

Santé Publique, Epidémiologie
et Qualité des Soins

Réutilisation et fouille de données massives de santé produites en routine au cours du soin

Mémoire présenté en vue de l'obtention de
l'Habilitation à Diriger des Recherches
en biostatistiques, informatique médicale et technologies de communication, par

Emmanuel Chazard

Université de Lille
EA2694, CERIM, pôle recherche
Faculté de Médecine de Lille
F-59045 Lille Cedex

Je souhaite exprimer ma profonde gratitude envers :

les Membres du jury,

qui acceptent de juger le présent travail,

les Professeurs Régis Beuscart, Alain Duhamel et Cristian Preda,

dont l'exigence et la bienveillance m'ont sans cesse aidé à progresser,

Frédérique,

ma petite fille adorée, qui me donne tant et à laquelle je donnerai tout,

mes chers parents, frère et sœur,

pour cette affection et ce soutien mutuels qui nous lient à jamais,

et mes chers étudiants,

dont la gentillesse et le sérieux récompensent chaque jour mes efforts.

Sommaire

Préambule

Médecin de santé publique, je suis actuellement maître de conférences des universités praticien hospitalier (MCU-PH) à l'Université de Lille et au CHU de Lille depuis le 1er septembre 2011. Cette activité se tient :

- au Centre d'Etudes et de Recherche en Informatique Médicale (CERIM), rattaché à l'EA2694 « Santé Publique : Epidémiologie et Qualité des Soins », pour la partie recherche,
- au Département de biostatistiques et informatique médicale, à la faculté de Médecine Henri Warembourg de l'université de Lille pour la partie enseignement,
- au Service « Méthodologie, Biostatistiques, Gestion de données, Archives », au sein du pôle de Santé Publique, Pharmacie et Pharmacologie (S3P) du CHU de Lille pour la partie hospitalière.

J'ai l'honneur de soumettre le présent mémoire intitulé « Réutilisation et fouille des données massives de santé produites en routine à l'occasion des soins », en vue de l'obtention de l'Habilitation à Diriger des Recherches, dans la sous-section 4604 du CNU (biostatistiques, informatique médicale et technologies de communication).

Ce mémoire suivra le plan suivant :

- Présentation des activités de recherche
- Présentation des activités d'encadrement
- Prospective de recherche
- Liste des publications
- Références
- Annexes

Présentation des activités de recherche

Cette partie sera structurée comme suit :

- une première partie (section 1.1, page 8) présentera les termes utilisés dans le titre de ce mémoire, puis le concept et le processus de *data reuse*. Elle nous amènera à décrire un projet type de *data reuse* à travers cinq phases.
- une deuxième partie (section 2, page 26) présentera mes travaux de recherche relatifs aux méthodes mises en œuvre.
- Une troisième partie (section 3, page 68) présentera mes travaux de recherche relatifs aux applications de ces méthodes.

1 Thématique de recherche

1.1 Définition des termes, présentation des concepts

Ce mémoire traite de la réutilisation et la fouille des données massives de santé produites en routine à l'occasion des soins. Dans cette première section, nous définirons les termes employés dans le titre du mémoire et certains termes connexes :

- La découverte de connaissances dans les bases de données, ou *knowledge discovery in databases* (KDD)
- La réutilisation de données, ou *data reuse*
- Les données massives, ou *big data*
- La fouille statistique de données, ou *data mining*
- Enfin, nous présenterons les données de santé produites en routine à l'occasion des soins, en hospitalisation.

1.1.1 Découverte de connaissances dans les bases de données, « knowledge discovery in databases »

Le KDD (*knowledge discovery in databases)* combine les statistiques, l'intelligence artificielle, et le data management pour faire apparaître des motifs d'associations dans des bases de données, et en extraire de nouvelles connaissances [1,2]. Ce terme fut introduit pour la première fois en 1989 durant un workshop [3]. D'un point de vue opérationnel, le KDD peut être défini comme un processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données [1]. D'un point de vue plus abstrait, le KDD permet de transformer des données de bas niveau (c'est-à-dire décrites sous forme de données individuelles et adaptée aux transactions, normalisées et essentiellement « neutres »), en d'autres formes plus compactes (par exemple un rapport), plus abstraites (par exemple un modèle explicatif) ou plus utile (par exemple un modèle

prédictif) [1]. Au cœur de ce processus se trouvent des méthodes de fouille statistique de données, ou *data mining* [1], que nous développerons ci-après.

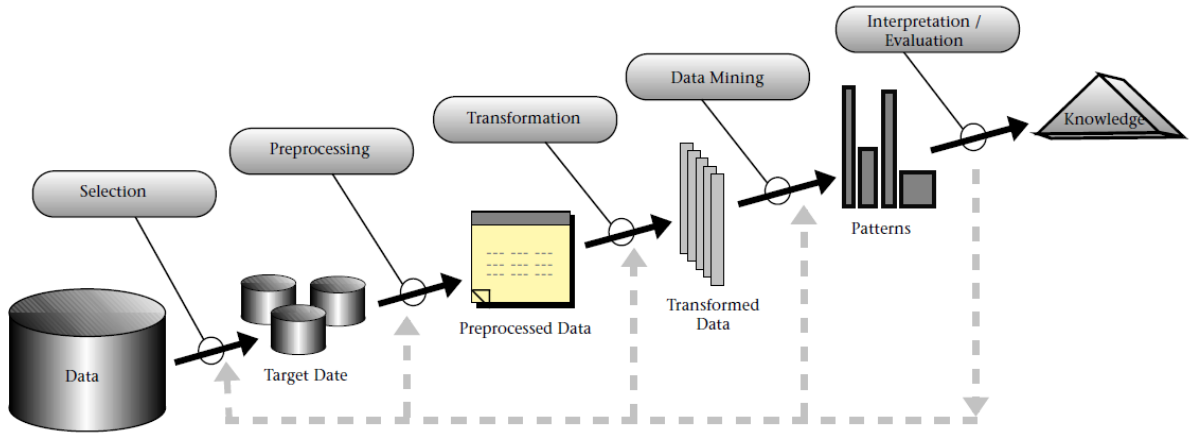


Figure 1. Vue globale des étapes d'un processus de KDD, proposée par Fayyad et a. [1]

Le processus de KDD est constitué de cinq étapes selon Fayyad et al. [1] (voir Figure 1) :

- Une étape de sélection des données pertinentes pour l'étude
- Une étape de prétraitement des données, qui consiste notamment à nettoyer les données, gérer les données manquantes et prendre en compte l'évolution de la représentation des connaissances en fonction du temps, car les données ont pu être collectées sur une période étendue
- La transformation de données, qui consiste essentiellement en une projection et réduction des données, de manière notamment à les simplifier et les rendre compatibles avec les étapes analytiques suivantes
- La fouille statistique de données (*data mining*), qui permet de mettre en évidence des associations (*patterns*) dans les données
- L'interprétation qui permet d'extraire de véritables connaissances de ces associations. Comme l'indique Fayyad [1], cette interprétation amène à itérer le processus complet, car elle soulève des problèmes qui peuvent être corrigés à toutes les étapes précédentes.

Dans le monde industriel, le terme de *business intelligence* désigne l'exploitation de données afin d'en extraire un avantage concurrentiel lié à l'information [4]. Ce terme désigne habituellement deux sous-ensembles :

- d'une part la surveillance au jour le jour voire plusieurs fois par jours d'indicateurs descriptifs (chiffre d'affaire, etc.), qu'on peut qualifier « d'indicateurs sentinelles », à l'aide de méthodes numériques très simples (décompte, somme, moyenne, etc.)
- et d'autre part l'analyse plus fine faisant appel notamment à des méthodes de *data mining* : ce deuxième cadre correspond au KDD tel que nous l'évoquons dans ce mémoire.

1.1.2 Réutilisation de données, « data reuse »

En décrivant le processus de KDD [3], les pionniers n'avaient cependant pas encore individualisé le concept de réutilisation de données (*data reuse*), alors qu'une partie de leurs travaux en relevaient déjà. Le terme *data reuse* ou *secondary use of data*

traduit le fait que des données aient été collectées pour une finalité précise, et soient dans un second temps analysées pour une autre finalité [5]. Ce terme est utilisé de manière croissante depuis peu [6]. Au moment de la rédaction de ce mémoire, il est encore absent du MESH. Depuis 2015 néanmoins, la réutilisation de données est enfin évoquée à travers le terme « data curation », défini comme suit [7] : « *Management activities required to maintain research data to ensure that they are fit for contemporary use and available for discovery and reuse* ». Toujours au moment de la rédaction de ce mémoire, seulement 16 articles incluent le terme « *data reuse* » dans leur titre dans la base Medline.

Dans les études traditionnelles, pour répondre à une question scientifique, les chercheurs collectent de nouvelles données prospectivement (cohortes, essais thérapeutiques, etc.) ou rétrospectivement (cas-témoins, etc.). Après une longue et coûteuse phase de collecte des données, l'analyse statistique répond précisément à la question scientifique posée (voir Figure 2).

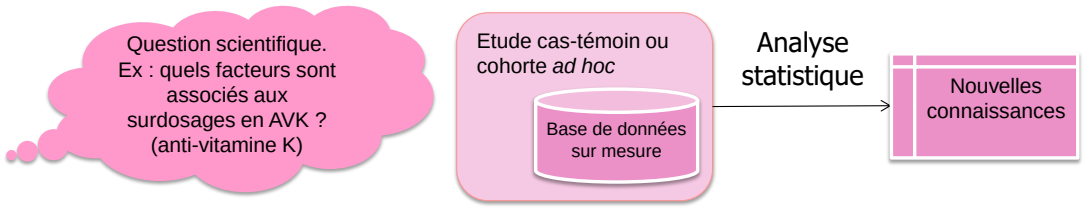


Figure 2. Acquisition et traitement des données dans les études traditionnelles

D'autre part, il existe souvent des activités transactionnelles, de routine, qui génèrent au fil de l'eau des données dont la vocation première est de servir ces activités. Ces données, généralement massives, peuvent parfois être réutilisées pour répondre aux questions scientifiques. Il s'agit d'une réutilisation à des fins essentiellement différentes des fins initialement définies lors du recueil. On peut parler de « seconde vie » des données. Ce processus est illustré ci-dessous (voir Figure 3).

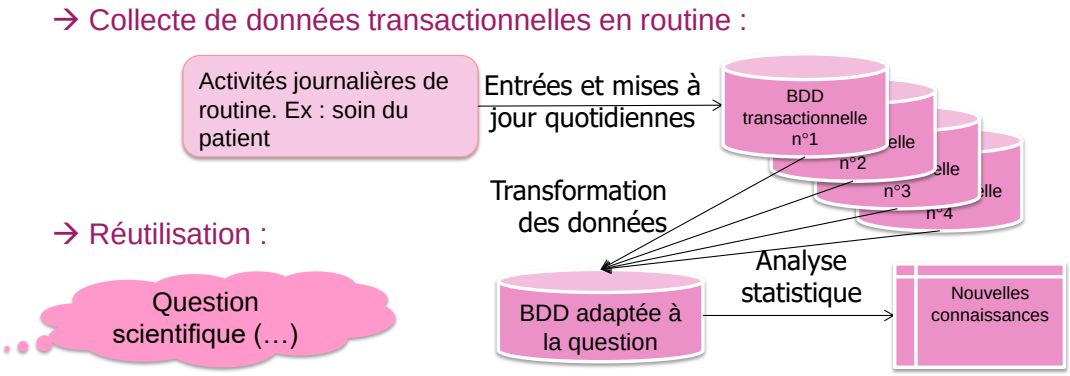


Figure 3. Acquisition et traitement des données en data reuse (BDD = base de données)

Les exemples historiques de *data reuse* concernent le secteur de la Banque, de l'Assurance et de la Grande distribution. Le processus est illustré ci-dessous dans le cas des compagnies d'assurance (voir Figure 4). Les activités transactionnelles de la compagnie lui permettent d'enregistrer et mettre à jour chaque jour des données sur leurs clients, les recettes qu'ils fournissent (primes d'assurance, cotisations) et les

dépenses qu'ils engendrent (fréquence et montant des sinistres). L'essence même de l'Assurance étant la prédiction et l'individualisation du risque, l'analyse de ces données fut immédiatement réalisée afin de produire des modèles prédictifs de risque individuel. Par la suite, pour chaque nouveau client, en fonction de ses caractéristiques propres, des primes d'assurance personnalisées purent être proposées, ces primes devant couvrir le risque moyen prédit du client, ainsi que la marge réalisée par la compagnie d'assurance (charges de fonctionnement et bénéfices) (voir Figure 4).

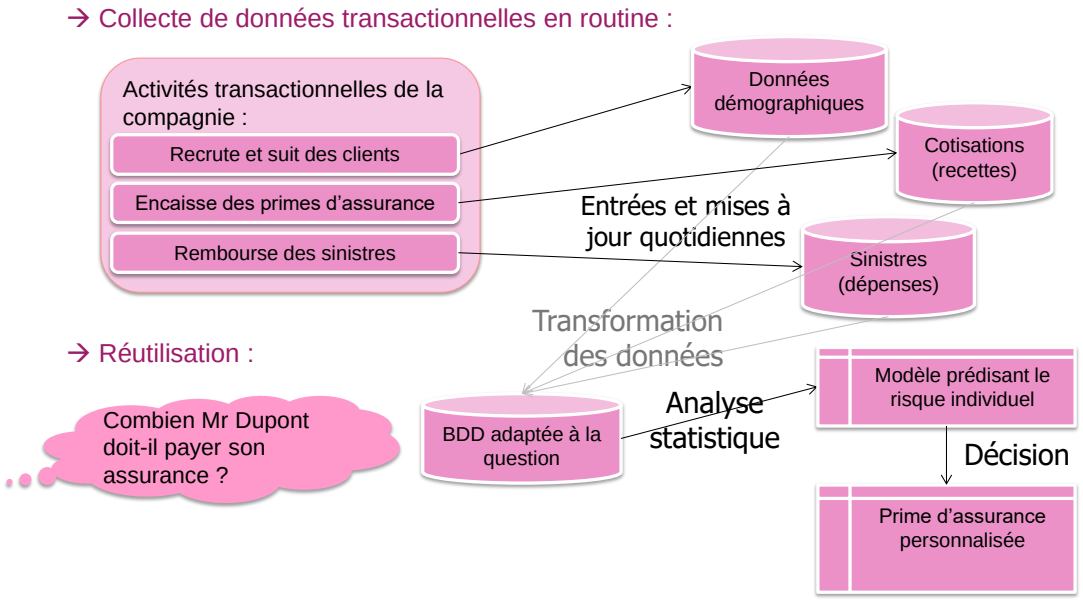


Figure 4. Exemple de data reuse dans le domaine des assurances

Cet exemple relevait donc du *data reuse*, mais utilisait des méthodes inférentielles traditionnelles telles les régressions et analyses de survie. Par la suite, des méthodes de *data mining* (ce terme sera défini ci-après, section 1.1.4, page 16) furent utilisées dans la grande distribution. C'est ainsi qu'Agrawal découvrit des associations fréquentes dans le « panier de la ménagère » [8], telles les associations « couches → bières » ou « oignons & pommes de terre → burger ». De telles associations permirent une réorganisation des rayons ou la mise en place de promotions ou ventes couplées visant à augmenter les ventes. Ce procédé est aujourd'hui au cœur des suggestions de produits dans le domaine de la vente en ligne (par exemple Amazon) ou de la publicité ciblée proposée par des moteurs de recherche (par exemple Google).

Cette démarche s'est plus tardivement développée dans le champ de la santé et en particulier de la prise en charge des patients, qu'il s'agisse d'hospitalisation ou de prise en charge ambulatoire [9–11].

Dans le champ de l'hospitalisation (voir Figure 5), l'admission et même la préadmission du patient permettent de collecter des données administratives et démographiques. Le soin quotidien du patient permet de collecter des données sur les médicaments administrés, sur les résultats structurés d'examens paracliniques (biochimie, hématologie, etc.). S'y ajoutent les comptes rendus d'examens, d'actes thérapeutiques et d'hospitalisation en texte libre. Enfin, les diagnostics et actes sont codés dans le cadre du PMSI. Ces types de données seront détaillés plus bas

(section 1.1.5, page 17). Si ces données servent avant toute chose à prendre en charge le patient, à assurer la continuité des soins et enfin à recevoir le paiement correspondant, leur relative exhaustivité permet dans certains cas une réutilisation en vue de répondre à des questions scientifiques, qu'il s'agisse de preuve formelle ou même simplement d'étude préliminaire.

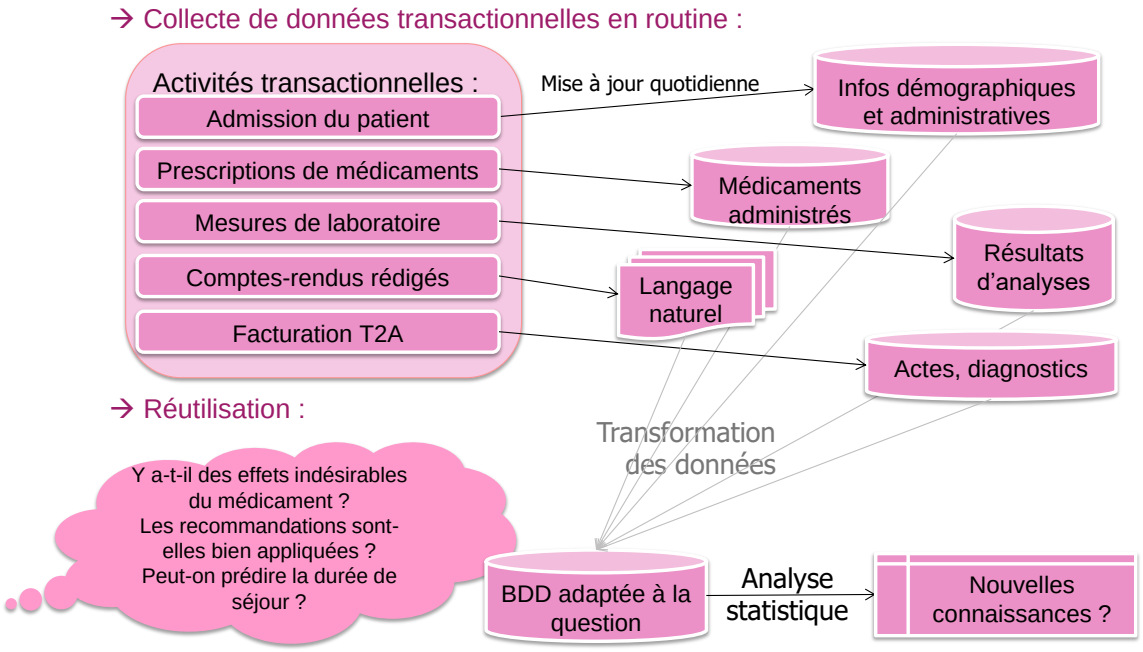


Figure 5. Application du data reuse au champ de l'hospitalisation

Dans le champ de la médecine ambulatoire, une démarche similaire peut être entreprise, avec des données néanmoins plus limitées et moins standardisées. Les travaux que nous présenterons dans ce mémoire ne portant que sur l'hospitalisation, nous n'évoquerons plus le sujet de la médecine ambulatoire.

La démarche de réutilisation des données d'hospitalisation présente des avantages [5,10,11] :

- Elle permet de réaliser des études avec une durée marginale faible et donc un coût marginal faible.
- Elle permet généralement d'obtenir des effectifs élevés et donc une forte puissance statistique.
- Elle permet d'obtenir des designs proches des cohortes rétrospectives, avec un recrutement de patients antérieur à l'observation de la variable d'intérêt.
- Elle permet de découvrir des facteurs de risque insoupçonnés, que l'on n'avait pas prévu de tester initialement, sous réserve que ces facteurs de risque soient collectés en routine.

Cette démarche présente naturellement des inconvénients [5,10,11] :

- Elle apporte une réponse imparfaite à la question scientifique, essentiellement parce que les données ne comprennent pas précisément les variables telles qu'on aurait aimé les avoir, et ce parce que le recueil de données poursuivait un objectif différent de l'analyse proposée.

- Elle rend difficile le contrôle des variables de confusion, tout simplement parce que ces variables n'ont pas nécessairement été recueillies (par exemple le régime alimentaire du patient), ou avec une qualité insuffisante (par exemple le tabagisme est parfois codé en CIM10 mais, en l'absence d'impact tarifaire, ce codage n'est pas fiable). Comme dans toutes les études non-expérimentales, les facteurs latents amenant le médecin à choisir un traitement spécifique plutôt qu'un autre constitueront un redoutable biais d'indication.
- Enfin, cette démarche est d'une grande complexité technique et méthodologique.

L'approche de réutilisation de données en santé a ainsi pu être utilisée avec succès dans plusieurs champs [5,10], comme par exemple :

- la mesure et l'amélioration de la **qualité des soins** [12,13]
- la détection et la prévention des **effets indésirables liés aux médicaments** [14,15]
- l'optimisation des processus de soins et en particulier des **flux de patients** [16–18]
- la prévention des **infections nosocomiales** [19]
- ou encore le **recrutement de patients** pour les essais thérapeutiques [20].

Selon les cas, les études de réutilisation de données peuvent se suffire à elles-mêmes, ou constituer des études préliminaires dont les résultats doivent être confirmés par une étude traditionnelle.

1.1.3 Données massives en santé, « big data »

Sans nommer les *big data*, Fayyad et al. évoquaient déjà les enjeux liés au volume des données en 1996 [1]. Ce caractère massif des données est désormais évoqué par le terme *big data*. Ce terme est encore absent du MESH, mais son utilisation croît de manière exponentielle [6,21]. Dans le monde de la santé, une revue exhaustive de la littérature médicale a permis de proposer une définition claire [21]. Les *big data* sont définies comme étant des données de grande dimension. Cela peut s'entendre de plusieurs manières, comme représenté ci-dessous en Figure 6 :

- de **nombreux individus** statistiques (nombre de lignes) :
  - c'est généralement le cas en santé publique et réutilisation de données (voir Figure 7). *Ainsi par exemple, la base nationale du PMSI (programme de médicalisation des systèmes d'information) contient près de 24 à 27 millions de séjours MCO (médecine chirurgie obstétrique) par année [22].*
  - La plupart des méthodes statistiques s'accommodent théoriquement de cela, mais la puissance de calcul (mémoire vive, processeur) peut devenir un obstacle.
- de **nombreuses variables** (nombre de colonnes) :
  - c'est généralement le cas en réutilisation de données, et ce nombre explose dans les sciences dites « –omiques » (voir Figure 7). *Ainsi par exemple, en génomique, on peut récolter jusqu'à 5 milliards de paires de bases par individu.*
  - Les méthodes traditionnelles de régression multiple ne peuvent fonctionner dans de telles conditions.
- de **nombreuses modalités** pour les variables qualitatives :

    - c'est généralement le cas lorsqu'on s'intéresse aux données médicales codées. *Ainsi par exemple, dans le PMSI, un diagnostic CIM10 peut prendre près de 40 000 valeurs possibles.*
    - Les méthodes statistiques utilisant des variables qualitatives ne peuvent généralement utiliser plus d'une dizaine de modalités.
- de **nombreuses tables et relations** :
    - c'est presque toujours le cas des bases transactionnelles. *Ainsi par exemple, un patient a un ou plusieurs séjours, eux-mêmes constitués d'un ou plusieurs diagnostics, de zéro à plusieurs actes, de zéro à plusieurs mesures biologiques, etc.*
    - Aucune méthode statistique ne peut nativement travailler sur plus d'une table à la fois, c'est-à-dire avec plusieurs types d'individus statistiques.
- de **nombreuses mesures répétées** (formellement, il s'agit d'un cas particulier du précédent item) :
    - c'est presque toujours le cas dans les données cliniques et paracliniques. *Ainsi par exemple la glycémie d'un patient pourra-t-elle être mesurée chaque jour durant tout un séjour, avec des résultats variables.*
    - Il existe des méthodes statistiques lorsqu'on s'intéresse à un seul paramètre, mais pas lorsque plusieurs paramètres présentent cette caractéristique.

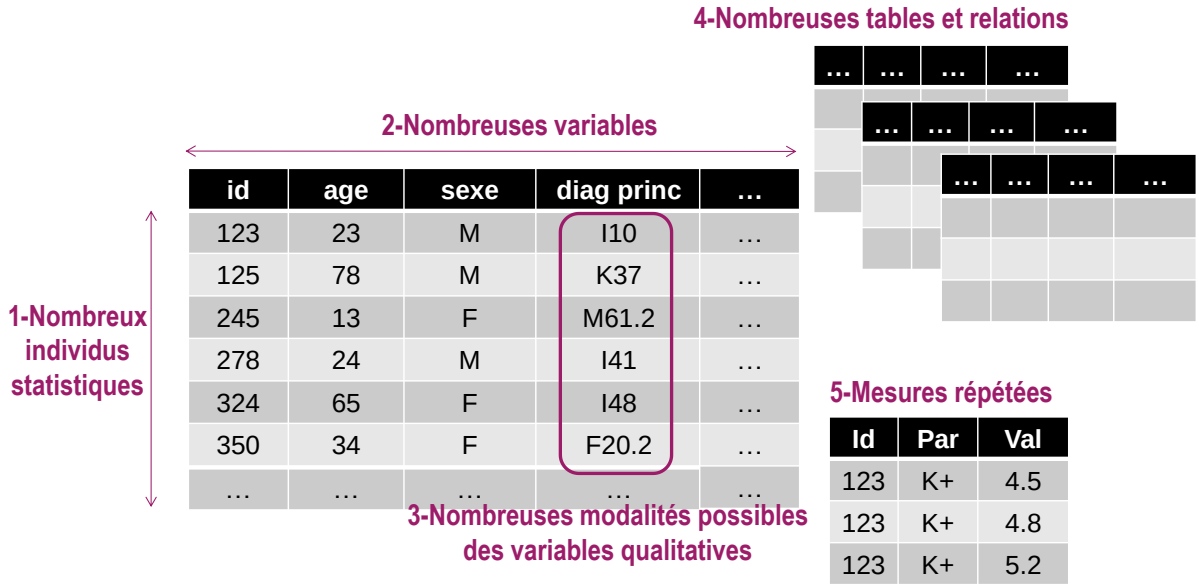


Figure 6. Cinq aspects du caractère massif des données de santé

Lorsque Fayyad et al. évoquaient les enjeux liés au volume des données en 1996 [1], le caractère massif de ces données se limitait aux caractères 1, 2 et 4 illustrés sur la Figure 6.

La Figure 7 positionne les jeux de données qualifiés de *big data* par leurs auteurs [21]. Les études de santé publique et *data reuse* exploitent généralement un nombre très élevé d'individus statistiques, tandis que les études dites « –omiques » exploitent plutôt un nombre de variables très élevé. A l'avenir, les études basées sur la réutilisation de données massives issues des dispositifs médicaux (figurés en vert

sur la Figure 7) devraient avoir à la fois un nombre élevé d'individus statistiques et de variables. Ces dispositifs peuvent être implantables ou externes. On peut citer à titre d'exemple les appareils de surveillance continue (ex : scope ECG en réanimation, saturation artérielle), les appareils cardiologiques implantés (ex : pacemaker, défibrillateur implantable), les appareils de surveillance à domicile (ex : pèse-personne implanté dans le lit et connecté, porte de réfrigérateur connectée), etc. Ces appareils contribuent à l'accumulation de données massives. Ils sont de ce point de vue-là le pendant médical des appareils utilisés à des fins de « bien-être » ou de suivi sportif, définissant le concept de « quantified self » (ex : podomètre intégré dans un bracelet, montre GPS, etc.).

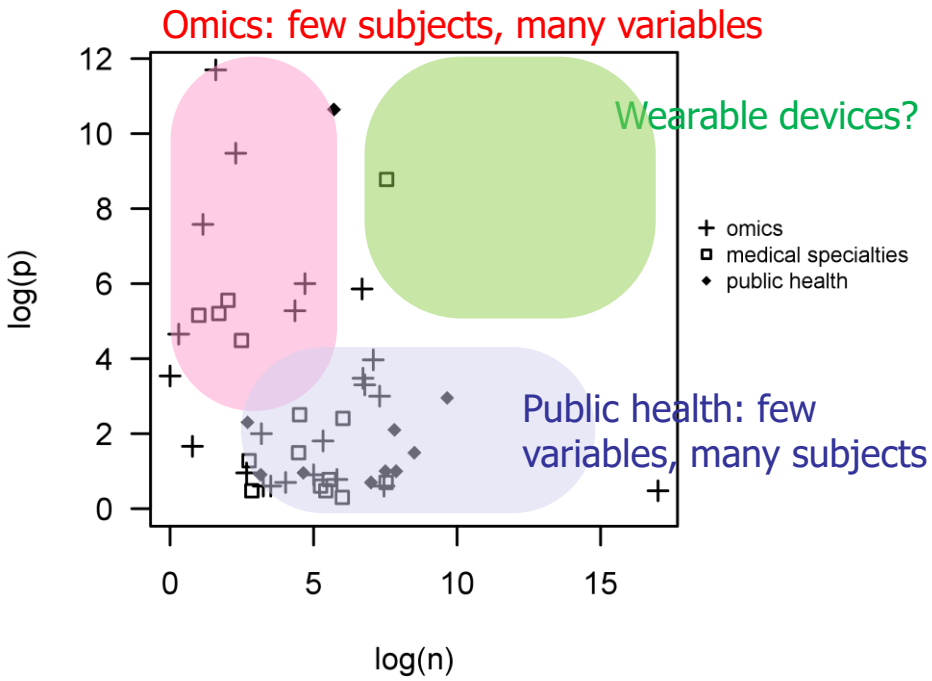


Figure 7 Classification des jeux de données qualifiés de "big data" et publiés, d'après [21] (n=nombre d'individus, p=nombre de variables, zones colorées = champs proposés)

Ainsi, les *big data* en santé peuvent être caractérisées par les cinq V [21] : *volume*, *variety*, *velocity*, *veracity*, et *valorization* :

- ***Volume* - volume et complexité :** la principale caractéristique des big data est leur taille et leur complexité [23–32]. Le caractère massif de ces données est souligné en premier lieu [26,30,32–37]. Tantôt ce caractère s'entend en termes de taille de la base de données (terabytes 10^{12} octets, petabytes 10^{15} octets, exabytes 10^{18} octets), tantôt un seuil de « big » est évoqué sans précision [32,38]. Dans le domaine de la santé, Baro observe néanmoins que la moitié des jeux de données publiés est caractérisé par un produit *(nombre d'individus * nombre d'attributs)* supérieur à 10^7 [21]. Ce seuil concerne le jeu de données finalement étudié, tandis que les volumes cités précédemment peuvent concerner la taille de la base de données initialement analysée, y compris les données non nettoyées, les index, les métadonnées liées à chaque mesure, etc.
- ***Variety* - diversité :** les *big data* sont également caractérisées par leur diversité [32,35–37,39–42]. Elles proviennent de diverses sources [34,43]

parfois indépendantes [32]. Les données non-structurées, tels les documents en texte libre [26,32,44] et les images [38,45–47] représentent un enjeu particulier. Les dossiers patients électroniques sont des sources de *big data* et posent le problème de la transformation de données brutes en information utilisable à des fins d'amélioration de la santé publique [48]. Cette diversité des sources entraîne une complexité des données, qui doivent souvent être représentées par des arbres semi-structurés comme des fichiers XML, parfois peu compatibles avec les bases de données relationnelles traditionnelles [37,40].

- ***Velocity* - volatilité et vélocité :** certaines sources de données sont volatiles [34]. Ainsi par exemple, une mesure en continu d'un paramètre en réanimation sera disponible avec une mesure par seconde durant 24h, mais rapidement les données seront effacées et seule une mesure par minute sera conservée, afin de libérer l'espace disque du serveur. Les *big data* sont également caractérisées par une mise à jour à haut débit [32–35,37,39,49], parfois au point de parler de données en temps réel [39].
- ***Veracity* - véracité :** il est néanmoins parfois délicat d'attester la fiabilité des *big data* [31,35,50,51]. De plus, en les analysant, il est aisé d'aboutir à des conclusions erronées [52–54], et ce pour différentes raisons dont par exemple l'inflation du risque de première espèce, ou l'augmentation de la significativité statistique pour une taille d'effet donnée : "*big size is not enough for credible epidemiology.*" [53].
- ***Valorization* - enjeux liés à la valorisation :** les différentes opérations sont impactées, en particulier la gestion des données [55–59], les traitements [23,26,35,60–66], et l'analyse des données [32,37,38,40,46,55–59,63,64,67–73]. Un enjeu consiste également à adapter (ou découvrir) les méthodes d'analyse statistique [40,56,63,69,72,74,75] et de visualisation de données [38,56,66,67]. Plusieurs auteurs insistent sur la nécessité et la difficulté d'extraire des informations pertinentes depuis de telles bases de données [36,67,76,77]. Cette difficulté tient en partie à l'élimination du bruit, et des valeurs aberrantes [60]. La valorisation scientifique notamment de ces données est également difficile [35,78]. Selon certains auteurs, il manque de professionnels disposant à la fois de l'expertise clinique et de la compétence d'analyse des données [36,79]. Ces profils sont parfois appelés *data scientists* [80]. Enfin, les *big data* annoncent d'intéressantes perspectives de *data reuse* dans des champs encore insoupçonnés [28,33,48,55,81–85].

1.1.4 Fouille statistique de données, « data mining »

La fouille statistique de données, ou *data mining*, constitue une phase centrale du processus de KDD [1,3]. Elle permet, en combinant des méthodes informatiques et statistiques, d'identifier des associations (patterns) au sein des données. L'interprétation de ces associations permet d'extraire des connaissances [1].

Les méthodes de *data mining* peuvent être classées de différentes manières selon le point de vue : les objectifs, les utilisations possibles, le type d'induction, etc. On peut par exemple les classer ainsi [1] :

- Les méthodes non-supervisées sont purement exploratoires. Elles permettent d'organiser et résumer l'information disponible. On peut distinguer :
    - Les analyses factorielles (ex : analyses en composantes principales, analyses des correspondances multiples)

    - Les classifications non-supervisées, dont les partitionnements (ex : K-means) et les classifications hiérarchiques (ex : classification hiérarchique ascendante)
    - Les recherches d'associations (ex : règles d'association « A Priori »)
    - Les analyses de corrélations
- Les méthodes supervisées permettent d'expliquer une variable d'intérêt à l'aide des autres variables. Dans un deuxième temps elles permettent, lorsque la valeur de cette variable d'intérêt est inconnue, de prédire cette valeur avec naturellement une erreur de prédiction :
    - On parlera de classification supervisée (ou discrimination) si cet attribut est une variable qualitative ou binaire (ex : régression logistique, arbres de décision, réseaux de neurones, réseaux bayésiens, analyses discriminantes)
    - Dans les autres cas, on parlera généralement de régressions. Dans la phrase qui suit, nous citerons à chaque fois un exemple additif et un exemple non-additif. La variable à expliquer ou à prédire peut être quantitative (régression linéaire, arbre de régression), un décompte (régression de Poisson, arbre de Poisson), ou un événement couplé à un délai d'apparition (modèle de Cox, arbre de survie).

Dans le cadre des travaux de recherche détaillés ci-après, nous avons été amenés à examiner et tester plusieurs méthodes sur les données structurées de plusieurs centaines de milliers de séjours hospitaliers. Nous avons plus particulièrement utilisé des méthodes supervisées, telles les arbres de décision [86], les règles d'association dans leur version supervisée [8], ou le modèle de Cox avec covariables temps-dépendantes et événements répétés [87].

1.1.5 Données de santé produites en routine à l'occasion des soins, en hospitalisation

Au cours de chaque hospitalisation de patient, en France et dans la plupart des pays développés, les données suivantes sont recueillies et numériquement disponibles (par ordre chronologique) [22,88] :

- des **données administratives** liées aux mouvements du patient (identité, dates, lieux, etc.), démographiques (âge, sexe, lieu de résidence, etc.) et assurancielles (couverture maladie, etc.)
- des **résultats d'analyses biologiques**, généralement prélevées par des infirmières et analysées par des professionnels ou par des automates
- des données médicales **produites automatiquement par des dispositifs médicaux autonomes**. Ces dispositifs peuvent être implantables ou externes, comme évoqué plus haut.
- des données relatives aux **médicaments administrés** au patient, généralement par des infirmières ou par des médecins, éventuellement dans le cadre d'un acte diagnostique ou thérapeutique
- des données relatives aux **dispositifs médicaux implantés** au patient au cours d'actes chirurgicaux
- des données relatives aux **actes médicaux**, qu'ils soient diagnostiques ou thérapeutiques. Ces données sont généralement codées par le réalisateur, parfois par la machine qui les réalise.
- des observations en **texte libre**, éventuellement formalisées en courriers ou comptes rendus

- des **diagnostics médicaux**, codés a posteriori par les médecins qui ont soigné le patient, ou par des techniciens spécialisés à la lecture des courriers.

En France par exemple, les données administratives, les actes et diagnostics médicaux sont obligatoirement recueillis pour chaque hospitalisation dans le cadre du PMSI (programme de médicalisation des systèmes d'information).

Les données fréquemment disponibles et sur lesquelles nous réalisons la plupart de nos travaux sont illustrées dans le cas d'un séjour de pose de prothèse totale de hanche ci-après (Figure 8).

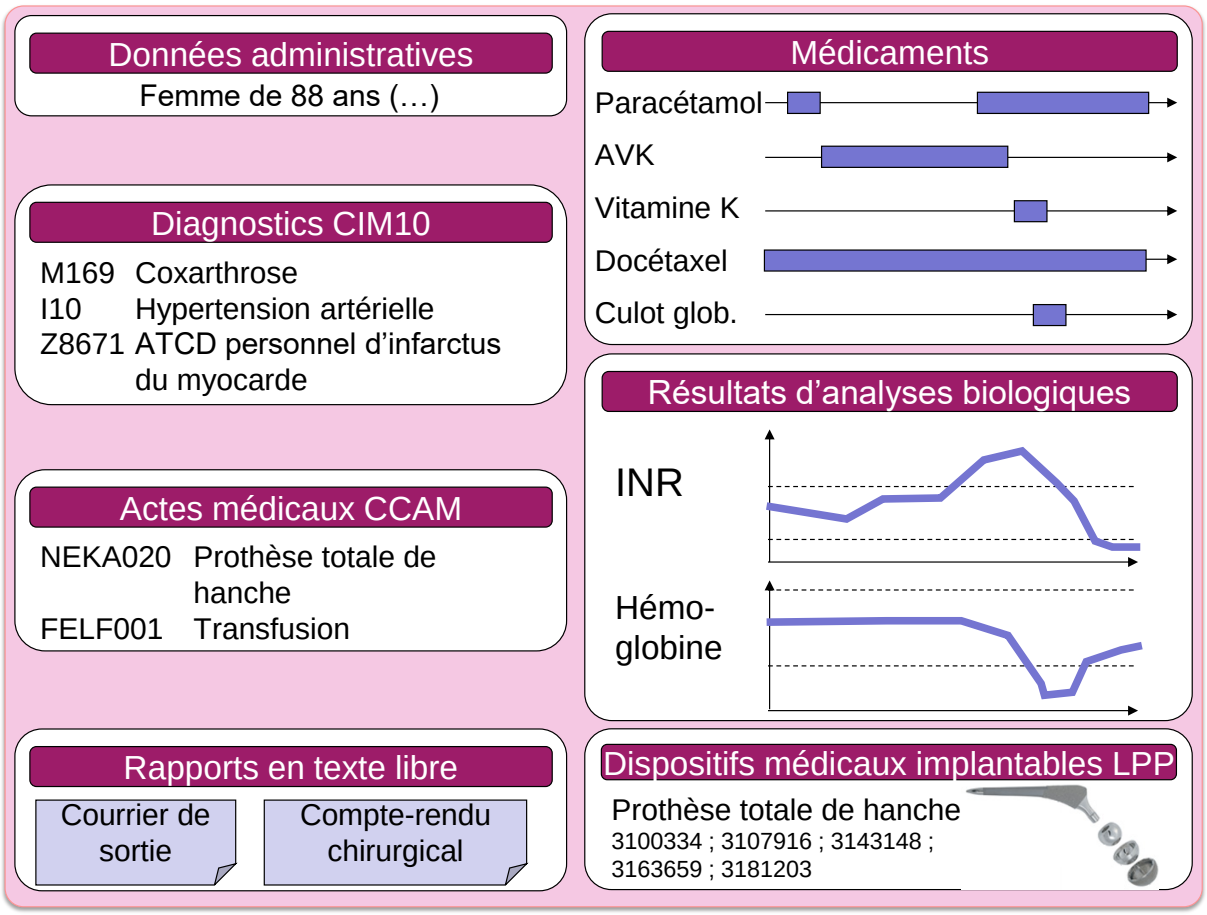


Figure 8. Illustration des données disponibles en hospitalisation

Dans le cadre du PMSI, chaque établissement doit collecter une partie de ces données. Elles sont ensuite colligées et constituent les bases nationales du PMSI [22]. Il est intéressant de noter que les bases nationales du PMSI comportent une part importante de ces données. Ainsi, chaque année, on peut disposer d'un jeu exhaustif de près de 27 millions de séjours, comportant les données du PMSI : informations administratives et démographiques, diagnostics, actes, dispositifs médicaux implantables (secteur public et privé non-lucratif également) et même certaines administrations de médicaments (molécules onéreuses et anti-thrombotiques).

Le tableau ci-après (voir Table 1) décrit ces éléments de données à travers plusieurs caractéristiques : la temporalité de production, le type de données, le mécanisme de production, la finalité supérieure et la disponibilité numérique dans les hôpitaux français.

Table 1. Typologie des données disponibles en hospitalisation

	Temporalité de production	Données structurées	Mécanisme de production	Finalité supérieure	Disponibilité numérique
Données administratives	Avant et pendant le séjour	oui	Saisie manuelle (peu d'interprétation)	Soin	Constante (obligatoire PMSI)
Résultats d'analyses biologiques	Au fil du séjour (juste après la mesure réelle, et avant sa réception par le service)	oui	Très souvent écriture directe par l'automate. Parfois saisie humaine suite à une interprétation	Soin	Quasiment constante de fait
Données produites par des dispositifs médicaux autonomes	Au fil du séjour (juste après la mesure réelle et avant son interprétation)	oui	Ecriture directe par l'automate	Soin	Très faible mais rapidement croissante
Médicaments administrés	Au fil du soin (après la prescription, et en théorie avant la prise du médicament)	oui	Saisie manuelle et validation par le professionnel de santé	Soin	Intermédiaire, croissante
Dispositifs médicaux implantés	Au fil du soin (après l'implantation) ou après le séjour	oui	Scan ou recopie par le professionnel de santé	Traçabilité	Constante (obligatoire PMSI)
Actes médicaux	Selon le cas, au fil du soin ou après le séjour	oui	Généralement encodage et saisie humains par l'opérateur. Rarement écriture directe par l'appareil.	Facturation	Constante (obligatoire PMSI)
Texte libre	Selon le cas, au fil du soin (après un acte, après une interprétation) ou après le séjour.	NON	Rédaction complexe par un médecin	Soin ou continuité du soin	Quasi-constante
Diagnostics médicaux	A la fin du séjour	oui	Encodage et saisie humains.	Facturation	Constante (obligatoire PMSI)

Comme l'illustre la colonne « finalité supérieure » du tableau précédent (voir Table 1), ces données sont le plus souvent collectées à des fins de soin immédiat du patient, ou à des fins de traçabilité ou facturation individuelle. Elles sont difficiles à analyser du fait de leur finalité initiale (problème de réutilisation), de leur forme (problème de données massives) : une analyse faisable et pertinente nécessite de

maîtriser à la fois les aspects informatiques, de traitement de données, d'analyse statistique de données, et d'analyse médicale des données.

1.2 Processus de *knowledge discovery in databases*, par réutilisation de données

Dans cette partie, nous proposerons un schéma général de *KDD* par *data reuse*, en nous appuyant sur le schéma de Fayyad et al. [1]. Ce nouveau schéma sera par la suite utilisé pour présenter l'ensemble de nos travaux.

1.2.1 Schéma proposé par Fayyad et al.

Fayyad et al. ont illustré le processus de KDD (voir Figure 1, page 9) en identifiant cinq étapes [1] :

- La sélection des données pertinentes pour l'étude
- Le prétraitement des données (nettoyage, gestion des données manquantes, etc.)
- La transformation des données
- La fouille statistique (*data mining*) ou extraction d'associations (*pattern extraction*)
- L'interprétation des associations

Ces auteurs ont en outre insisté sur l'aspect itératif de ce processus [1].

Nous proposons ici une mise à jour de ce schéma afin d'**illustrer le KDD dans le contexte précis de *data reuse***. Nous proposons donc les modifications suivantes :

- De notre expérience, puisque les données utilisées n'ont pas été recueillies pour une finalité d'étude particulière, un nombre important d'études pourra s'appuyer sur une même source de données, présentées de manière « neutre », c'est-à-dire indépendante de la finalité. Ce parti-pris rend possible (et nécessaire afin d'accélérer les recherches) la mise en place d'un entrepôt de données. Nous proposons donc de supprimer la phase de ciblage des données, et de la remplacer par la mise en place d'un entrepôt de données destiné au *data reuse*.
- Le lecteur pourrait confondre les termes de « preprocessing » et de « transformation » utilisés dans deux étapes successives. Comme il s'avère que cette transformation vise toujours à réduire les données, ainsi que l'indiquent les auteurs [1], nous proposons le terme d'« agrégation ».
- Toujours afin de les rendre plus intelligibles, nous proposons de matérialiser les termes de « preprocessed data » et « transformed data », en parlant d'« entrepôt » et de « table simplifiée », car en pratique il s'agit toujours d'une table prête à être analysée par une méthode statistique.
- Enfin, comme les auteurs [1] le faisaient dans le texte, nous ajoutons une dernière étape d'« action ».

1.2.2 Proposition de schéma en cinq phases

Ainsi, un projet de KDD s'appuyant sur la fouille statistique de données dans un contexte de *data reuse* se déroule en cinq phases illustrées ci-après (voir Figure 9).

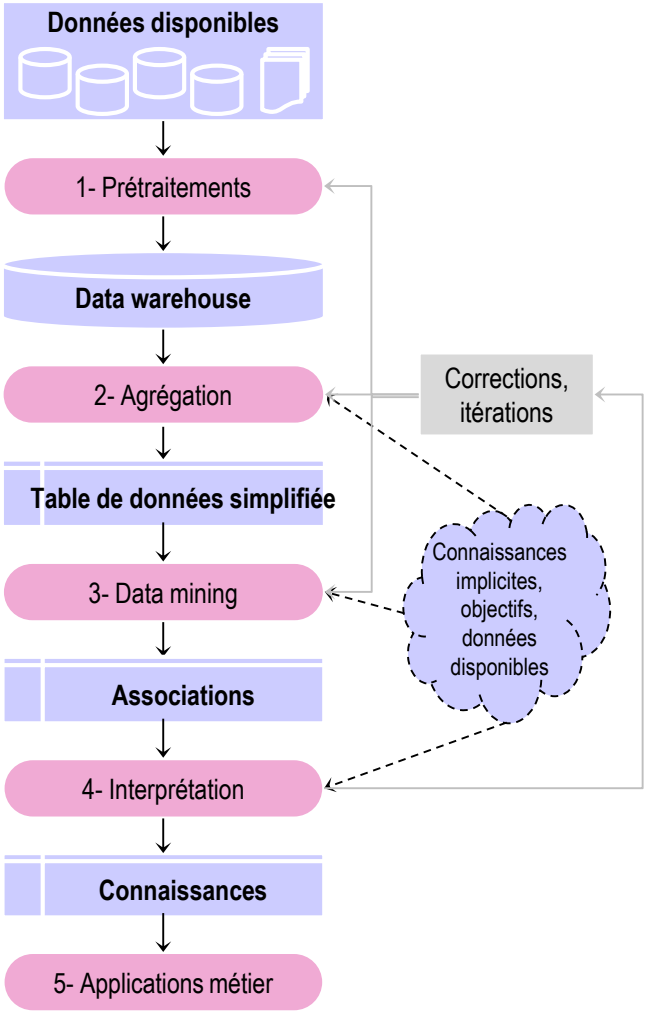


Figure 9 Les cinq phases d'un projet de KDD (avec data mining) dans un contexte de data reuse

Ces cinq phases sont les suivantes :

- **La première phase est le prétraitement des données**. Elle consiste à transformer les données afin de les intégrer dans un entrepôt de données, les nettoyer, éliminer les valeurs aberrantes, dans certains cas inférer les données manquantes, et aligner ou rétablir les terminologies.
- **La deuxième phase est l'agrégation de données.** Elle consiste à transformer les données brutes en information porteuse de sens, et facilement exploitable. *Un exemple simple consiste, dans l'étude des facteurs de risque d'infarctus, à remplacer la taille et le poids par l'indice de masse corporelle.* Cette phase est donc guidée par une connaissance métier importante. Elle représente de notre expérience 80% du temps de traitement. En outre, elle a un impact considérable sur le succès de l'étude.
- **La troisième phase est la fouille statistique de données**. Elle utilise des procédures dérivées des statistiques, de l'informatique et de l'intelligence artificielle. Elle génère des signaux d'associations statistiques en grand nombre, pour lesquels il est prudent de considérer que la plupart d'entre eux sont des associations résultant de la présence de facteurs de confusion non-explicités ou, pire, d'une inflation du risque de première espèce dépassant très largement la seule étape d'inférence statistique, rendant la phase suivante indispensable.

- **La quatrième phase est l'interprétation.** Il est prudent de filtrer ces signaux à dire d'expert, en s'appuyant notamment sur la littérature, et de réorganiser ces connaissances de manière simple et exploitable.
- **Enfin, la cinquième phase consiste à mettre en œuvre les actions appropriées** consécutives à cette production de connaissance.

Les phases d'agrégation, de data mining et d'interprétation sont fortement influencées par les connaissances métier liées à l'étude menée. En outre, les résultats obtenus à l'issue de l'interprétation amènent à réaliser de nombreuses corrections de chacune des étapes précédentes, rendant le processus itératif. Le caractère itératif et les connaissances métier sont représentés sur la droite de la Figure 9.

1.2.3 Décomposition des 5 phases

Il est possible de détailler un peu plus ce schéma comme suit (voir Figure 10 page 24). Ce schéma détaille les cinq phases précédentes, en déterminant des sous-ensembles qui seront repris dans la suite de ce mémoire pour positionner nos travaux :

- **La phase 1, de prétraitement des données** soulève notamment deux catégories de problèmes :
    - les problèmes liés à l'incorporation des données dans un entrepôt :
        - la fusion des données, c'est-à-dire l'extraction multi-source de toutes les données relatives à un même patient ou épisode, sur la base d'un identifiant commun, et leur insertion dans un modèle de données déterminé
        - le contrôle qualité de ces données
        - l'alignement terminologique, c'est-à-dire la transformation de codes d'une terminologie en codes d'une autre terminologie, ou l'ajout de codes d'une terminologie standard en lieu et place d'identifiants maison, ou même l'ajout de codes déduits d'un libellé textuel
    - les problèmes liés à la conservation de données dans un entrepôt de données, et notamment la sécurisation de cet entrepôt
- **La phase 2, d'agrégation des données** soulève notamment trois catégories de problèmes : le filtrage, l'agrégation et la simplification des données
- **La phase 3, de *data mining*** nous permettra de présenter trois points :
    - la fouille statistique de données stricto sensu, utilisant des méthodes de *data mining*
    - la fouille visuelle de données
    - certains problèmes méthodologiques soulevés par la prise en compte du temps notamment
- **La phase 4, d'interprétation** sera évoquée à travers les applications proposées dans ce mémoire
- **La phase 5, d'application métier,** permettra d'évoquer 7 domaines d'applications dans lesquelles nous avons été impliqués :
    - les effets indésirables des médicaments
    - les prescriptions inappropriées
    - l'adhésion aux recommandations
    - la gestion hospitalière

- la télécardiologie
- l’épidémiologie des soins
- et d’autres applications plus hétérogènes

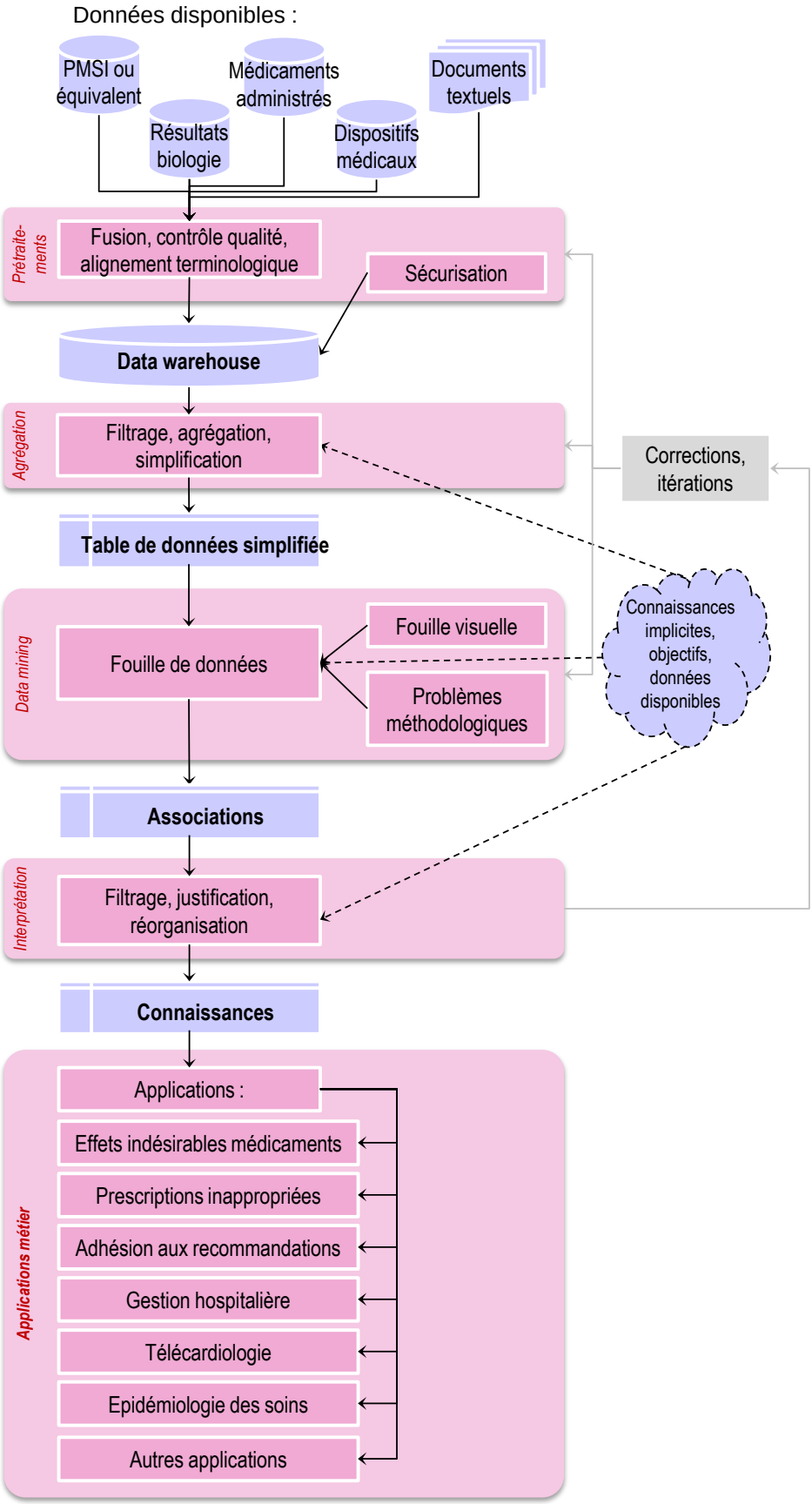


Figure 10 Les cinq phases d'un projet de KDD (avec data mining) dans un contexte de data reuse, détaillées en sous-ensembles

1.3 Synthèse de l'exposé de la thématique, objectif du mémoire

La découverte de connaissance dans les bases de données (KDD) est un processus complexe alliant informatique, analyse statistique et connaissances métiers, dans l'objectif d'extraire des connaissances et de prendre des décisions ou mettre en place des actions spécialisées (voir section 1.1.1, page 8). Ce KDD peut être appliqué à des données recueillies pour une autre finalité, on parle alors de *data reuse* (voir section 1.1.2, page 9). Le KDD par *data reuse* se développe en santé, et s'appuie notamment sur des données collectées en routine pour le soin individuel (voir section 1.1.5, page 17). Ces données peuvent être des *big data*, ce qui complexifie leur traitement (voir section 1.1.3, page 13).

Nous avons, au terme de cette partie, proposé un schéma de KDD par réutilisation de données en cinq phases (Figure 9, page 21). Nous avons également développé ce même schéma en sous-ensembles (Figure 10, page 24).

L'objectif de ce mémoire est de présenter nos travaux de recherche en les positionnant sur ce schéma général. Nous présenterons tout d'abord les travaux liés aux méthodes employées (voir section 2, page 26), puis les travaux liés aux applications métier (voir section 3 page 68).

2 Principaux travaux liés aux méthodes

2.1 Positionnement des travaux relatifs aux méthodes

Cette section présente les travaux que je réalisai dans le cadre de ma thématique de recherche, la « **réutilisation et fouille des données massives de santé produites en routine à l'occasion des soins** », en suivant le cheminement d'une étude type de réutilisation de données hospitalières. Nous avions déjà présenté les cinq phases d'une étude de *knowledge discovery in databases* avec *data reuse* dans la Figure 9 en page 21, puis nous avions détaillé certains sous-ensembles de ce processus dans la Figure 10 en page 24. Nous suivrons ce même cheminement.

Nous exposerons tout d'abord les principaux résultats obtenus dans le champ des méthodes. La Figure 11 positionne les publications auxquelles je fus associé sur ces quatorze sous-ensembles : chaque pastille jaune y indique le nombre de publications. Le détail de ces publications est disponible en annexe.

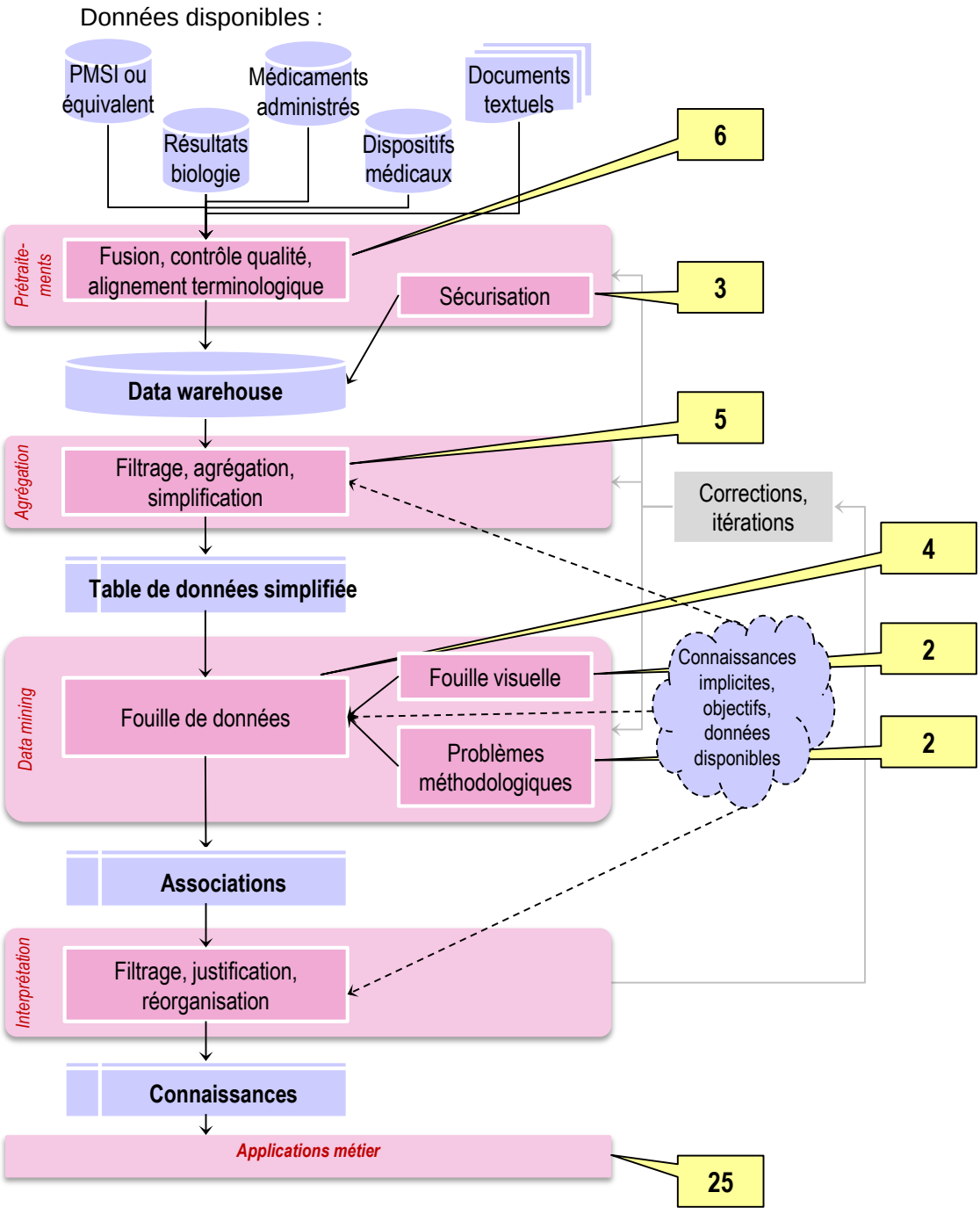


Figure 11. Thématique détaillée : travaux relatifs aux méthodes (étiquettes jaunes : nombre de publications Pubmed, la même figure avec les titres complets des articles est visible en annexe)

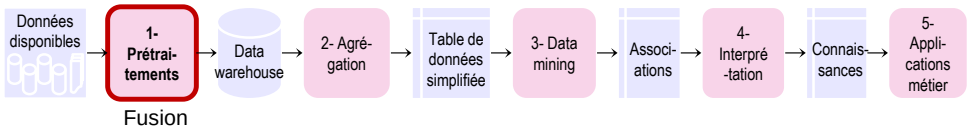
2.2 Acquisition, fusion et contrôle qualité des données

Les données que nous utilisons de manière habituelle sont :

- les données du PMSI ou des systèmes équivalents rencontrés dans la majorité des pays développés :
  - données démographiques
  - données de mouvements
  - autres données administratives dont parfois la couverture assurancielle
  - données médicales codées qu'il s'agisse de diagnostics, d'actes diagnostiques ou d'actes thérapeutiques
- les résultats de laboratoire
- les médicaments administrés
- les courriers de sortie et comptes-rendus d'actes en texte libre

Nous cherchâmes à améliorer la qualité de certaines de ces données, afin que leur traitement apportât des résultats plus fiables. Nous définîmes également le terme de « big data », qui faisait l'objet de nombreux faux-sens.

2.2.1 Définition du terme « big data » en santé

Position :	
Publication :	*Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. Biomed Res Int 2015;2015. doi:10.1155/2015/639021.* [texte intégral en ligne] *[21]*

Nous nous intéressâmes tout d'abord à la notion de *big data* en santé : ce terme était utilisé de manière croissante, et pourtant il était toujours absent du MESH (c'est encore le cas au moment où ce mémoire est rédigé), et n'avait à l'époque pas fait l'objet d'une définition claire en littérature blanche dans le domaine de la santé. Cette notion étant partagée par plusieurs champs disciplinaires, et parfois confondue avec celle de *data reuse.* Nous réalisâmes donc en 2014 une revue exhaustive de la littérature [21] incluant tous les articles de la base Pubmed mentionnant le terme « big data » dans leur résumé ou titre. Ainsi, 196 articles furent inclus. Certains de ces articles publiaient des analyse de jeux de données qu'ils qualifiaient de *big data*. Nous nous appuyâmes sur le jeu de données finalement analysé, et non la base de données initiale. En tenant compte de ces articles seulement, la médiane de la quantité *Log(n.p)* était de 7, *n* étant le nombre d'individus et *p* le nombre de variables. Autrement dit, le « nombre de cases » du tableau de données analysé excédait 10 millions dans la moitié des études. Nous observâmes également par une approche quantitative la séparation du monde des *big data* en deux sous-groupes cités plus haut (voir Figure 7 en page 15). Nous pûmes également retrouver les propriétés consensuelles des *big data*, à savoir leur variété, leur vélocité, les

problèmes de véracité, et les besoins de nouvelles méthodes. Sur la base de la littérature, nous pûmes réaffirmer que le concept de *data reuse* était clairement différent, contrairement à l'utilisation couramment faite par les décideurs et journalistes. Ainsi, un projet pouvait réutiliser des données sans qu'elles fussent massives et, inversement et comme l'illustraient les études pan-génomiques, des données pouvaient être massives sans avoir été collectées pour d'autres finalités que la recherche.

2.2.2 Le contrôle qualité des données

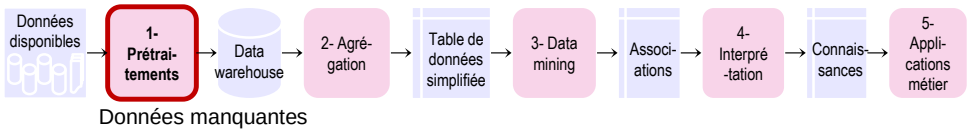
Le contrôle qualité des données est une étape essentielle de tout travail de KDD. Nous passerons rapidement dessus car les principes présentés ci-dessous ont été mis en œuvre dans tous nos travaux, mais n'ont pas fait l'objet de publication propre.

Nous illustrerons le contrôle qualité à travers plusieurs contrexemples, qui permettent de mieux comprendre ce qu'est une donnée de qualité :

- Les données doivent être extraites selon le bon format. Chaque valeur doit unitairement être valide :
    - La valeur doit suivre le type attendu.
      Contrexemple 1 : âge= « vieux » (non numérique)
    - La valeur doit être possible.
      Contrexemple 2 : âge=834 (hors limite)
      Contrexemple 3 : diagnostic= « HHFA001 » (on attend une seule lettre)
    - Pour les variables qualitatives, la valeur, en plus d'être syntaxiquement plausible, doit appartenir à la liste des codes autorisés par la terminologie
      Contrexemple 4 : diagnostic= « B990 » (plausible mais inexistant)
- Les données doivent également suivre une distribution univariée correcte
  Contrexemple 5 : 0<âge<100 mais moyenne=80 dans un hôpital (trop élevée)
  Contrexemple 6 : 0<âge<100 mais SD(âge)= 0 (âge constant)
- Enfin, les données doivent suivre une distribution bivariée (ou conditionnelle) correcte
    - Il peut s'agit d'une relation déterministe induite par une dépendance fonctionnelle ou redondance des données
      Contrexemple 7 : durée=2, entrée="2013-05-14", sortie="2013-05-14"
      Contrexemple 8 : moyenne(âge | lieu=« gériatrie »)=21 (trop faible)
    - Il peut également s'agit d'une relation probabiliste
      Contrexemple 9 : moyenne(durée | entrée = « transfert »)=1 (trop faible)
      Contrexemple 10 : correlation(âge, durée) = -0.3 (devrait être positive)

En pratique, le contrôle qualité est itérativement mêlé au débogage des programmes. De notre expérience, le contrôle qualité doit concerner toutes les étapes du processus, et le contrôle qualité des résultats d'une méthode permet généralement de détecter des anomalies de la méthode mais également des anomalies des données source.

2.2.3 Amélioration de la qualité des codes diagnostiques

Position :	
Publication :	*Djennaoui M, Ficheur G, Beuscart R, Chazard E. Improvement of the quality of medical databases: data-mining-based prediction of diagnostic codes from previous patient codes. Stud Health Technol Inform 2015;210:419–23.* [texte intégral en ligne] *[89]*

Nous avions observé que la qualité du codage des diagnostics était parfois insuffisante, en particulier pour les maladies chroniques qui n'étaient pas le motif d'hospitalisation. Certaines de ces maladies chroniques étaient utiles à notre champ de recherche, comme par exemple l'insuffisance rénale ou l'insuffisance hépatique, qui pouvaient contribuer à la survenue d'effets indésirables des médicaments. Ce défaut de qualité persistait néanmoins lorsque pourtant ces maladies pouvaient augmenter la valeur d'un séjour en tarification à l'activité, en étant valorisées comme CMA (complications et morbidités associées).

Partant du principe que de nombreuses maladies étaient irréversibles, nous cherchâmes à savoir s'il était licite alors de reconduire automatiquement leur codage d'un séjour sur l'autre, en s'appuyant sur le numéro d'identifiant unique du patient dans un hôpital [89]. Sur 94 millions de séjours issus de la base nationale du PMSI, nous testâmes donc si certaines pathologies (le diabète de type 2, la fibrillation auriculaire et l'insuffisance cardiaque) pouvaient être automatiquement reconduites. Face à des résultats imparfaits, à l'aide de méthodes de data mining, nous découvrîmes des facteurs améliorant la reconductibilité de ces diagnostics. Ces facteurs formaient ainsi des règles (voir Table 2). Nous appliquâmes ensuite les règles découvertes dans un établissement de court séjour français, et calculâmes leur valeur prédictive positive en relisant en détail les courriers relatifs à ces patients. Nous obtînmes des règles de prédiction. La Table 2. Montre l'exemple de règles prédisant un code de fibrillation atriale.

Table 2. Règles permettant de reconduire un code de fibrillation atriale (FA, code I48)

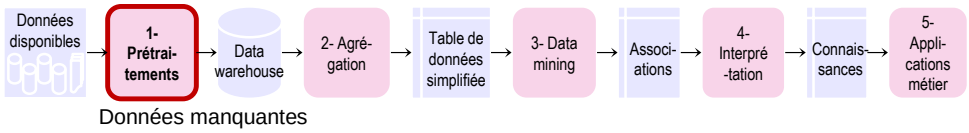
Conditions (séjour n-1)	Confiance
FA (I48) → FA (I48)	0.51
HTA essentielle (I10) & FA (I48) & anomalie lipidique (E78) → FA (I48)	0.60
HTA essentielle (I10) & FA (I48) & implant cardiaque ou vasculaire (Z95) → FA (I48)	0.60
FA (I48) & séquelle de maladie cérébrovasculaire (I69) → FA (I48)	0.62

Ces résultats nous enseignèrent plusieurs leçons. La première leçon est qu'il n'était pas licite de reconduire le codage d'une pathologie au prétexte qu'elle était chronique et avait été codée précédemment (dans la Table 2, les trois règles constituées d'une seule condition ont une confiance faible). Les apparentes

contradictions entre un séjour avec code et le séjour suivant sans code pouvaient aussi bien être liées à un mauvais codage du premier séjour et non du second, ou encore au fait qu'une pathologie pût persister mais ne pas être prise en charge et ne pas avoir d'incidence sur la conduite du soin. Depuis cette étude, **nous ne préconisons plus de rechercher les maladies chroniques dans un séjour précédent** sans discernement.

La deuxième leçon est que les méthodes de data mining identifient des facteurs de risque statistiques, mais que ces facteurs n'auraient pas pu être prédits par un humain. Leur validité statistique ne tient pas à des relations de cause à effet, mais bien à des associations purement statistiques, tel le syndrome métabolique, ou à des parcours de soins qui petit-à-petit renforcent la certitude diagnostique, et ce notamment par des bilans systématiques de pathologies associées.

2.2.4 Amélioration de la qualité des codes de médicaments

Position :	
Publication :	*Merlin B, Chazard E, Pereira S, Serrot E, Sakji S, Beuscart R, et al. Can F-MTI semantic-mined drug codes be used for adverse drug events detection when no CPOE is available? Stud Health Technol Inform 2010;160:1025–9.* [texte intégral en ligne] *[90]*

La majorité de nos travaux porte sur la prescription médicamenteuse, qu'il s'agisse de conformité des prises en charge ou de détection des effets indésirables. Or, comme nous l'indiquions plus haut (section 1.1.5 en page 17), ces médicaments administrés sont inconstamment présents dans les bases de données hospitalières. Pourtant, le traitement habituel et le traitement de sortie sont presque toujours mentionnés dans le courrier de sortie. Vint l'idée de tenter d'utiliser les courriers de sortie pour inférer ces données manquantes. Nous utilisâmes pour ce faire le F-MTI (French Multi-Terminology Indexer) développé par le laboratoire d'Informatique, de traitement de l'information et des systèmes (EA 4108, Pr Stefan Darmoni) de l'Université et du CHU de Rouen [90].

Dans un premier temps, nous évaluâmes l'aptitude du F-MTI à extraire des codes ATC de médicaments de courriers de sortie en texte libre. Les résultats obtenus étaient satisfaisants (voir Table 3).

Table 3. Aptitude de F-MTI à trouver les mêmes codes ATC qu'un expert dans les courriers de sortie

Hôpital	Nb courriers	Précision (VPP)	Rappel (Se)	F-mesure
A	50	0.84	0.93	0.88
B	32	0.88	0.88	0.88

Le problème néanmoins était que les médicaments cités dans le courrier n'étaient pas tous ceux administrés durant le séjour. Nous testâmes ensuite si les codes ATC prédits par F-MTI en analysant les courriers étaient bien ceux enregistrés dans le système d'information au décours d'une réelle prescription en ligne. Les résultats s'avérèrent moins bons (voir Table 4), principalement parce que les courriers

rappelaient les traitements à l'admission et à la sortie, mais évoquaient rarement les médicaments administrés temporairement (antalgiques, AINS, médicaments agissant sur le transit, etc.), pourtant fréquemment impliqués dans des effets indésirables du médicament.

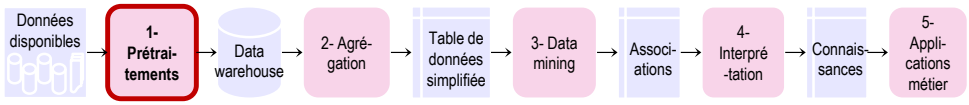
Table 4. Aptitude de F-MTI à prédire les codes ATC réellement administrés en analysant les courriers de sortie

Hôpital	Nb courriers	Précision (VPP)	Rappel (Se)	F-mesure
A	37	0.73	0.37	0.49

Une troisième étape nous permit néanmoins d'observer que la force de l'association entre les médicaments ainsi inférés et la survenue d'un événement biologique (par exemple « anti-vitamine K → élévation de l'INR ») était la plupart du temps similaire entre un établissement avec de vraies données médicamenteuses et un établissement avec extraction des codes de médicaments depuis les courriers.

Ces résultats nous montrent qu'il semble peu réaliste d'utiliser un outil de *semantic mining* sur les courriers de sortie en espérant ainsi remplacer une base de données de médicaments prescrits. Néanmoins, et en particulier si les médicaments d'intérêt sont des traitements au long cours, il peut être intéressant de cibler des séjours à l'aide d'un outil comme F-MTI, sous réserve contrôle humain. Cette approche peut donc s'avérer très utile en l'absence de base de données de médicaments administrés.

2.2.5 Alignement terminologique et conversion des résultats de biologie

Position :	 Alignement terminologique
Publication :	*Ficheur G, Chazard E, Schaffar A, Genty M, Beuscart R. Interoperability of medical databases: construction of mapping between hospitals laboratory results assisted by automated comparison of their distributions. AMIA Annu Symp Proc 2011;2011:392–401.* [texte intégral en ligne] *[91]*

Intégrer les données d'un nouvel hôpital dans un entrepôt de données existant nécessite souvent la production de mappings pour les données biologiques car les terminologies internationales telles LOINC ou IUPAC sont rarement utilisées. En pratique, chaque établissement propose des libellés maison. Ces libellés posent plusieurs problèmes. Quatre sont énumérés ci-dessous.

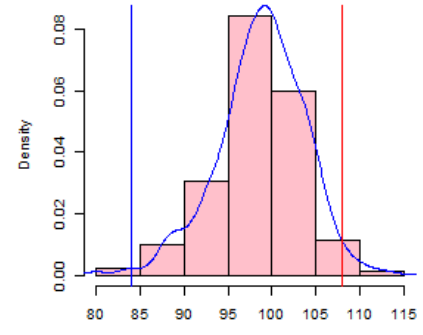
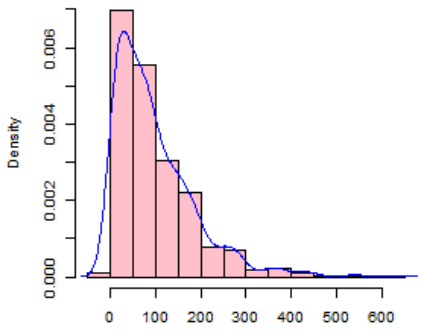
Plusieurs libellés possibles par paramètre : ainsi par exemple, la kaliémie peut être nommée « Potassium », « Potassium (sang) », « Kaliémie » ou encore « K1 » (ces quelques exemples sont réels). De plus, au sein d'un établissement, il est fréquent que les libellés changent lorsque l'automate de biologie est mis à jour. Enfin, il est même fréquent que deux automates différents d'un même laboratoire ne produisent pas les mêmes libellés.

Plusieurs paramètres possibles par libellé : deux libellés identiques peuvent désigner deux paramètres différents, comme par exemple « calcium » pour la calcémie et la calciurie (cet exemple est celui de la Table 5 lisible plus bas).

Intelligibilité du libellé : la structure des bases de données de biologie est fortement marquée par le standard HPRIM [92] et la tradition d'impression sur papier. Ainsi par exemple, on peut retrouver ce type de libellé : « soit en g/l : ». Ce libellé est en soi incompréhensible. Il ne l'est plus lorsqu'il est imprimé sur papier et que la ligne supérieure a pour libellé « glycémie (mmol/l) : ». Cette manière de ranger l'information entre en contradiction avec les règles de normalisation des schémas relationnels, dans lesquelles il n'existe pas de dépendance fonctionnelle entre deux lignes d'une même table [93].

Unités erronées : il est fréquent que les unités déclarées dans une base de données de biologie soient fausses. Nous avons ainsi pu observer des décomptes de globule rouge annoncés en *millions par mm^3*, alors que la moyenne des valeurs observées valait aux alentours de 5000 et non 5, comme cela aurait dû être le cas. Il s'agissait donc de *milliers par mm^3*. Pourtant, ce type d'erreur n'a aucune conséquence clinique pour les paramètres fréquents, car les médecins déduisent l'unité en fonction de l'ordre de grandeur des valeurs elles-mêmes, et des bornes de normalité fournies par le laboratoire. Ces bornes étant affectées par la même erreur, le positionnement des valeurs par rapport aux bornes est inchangé, tant en termes de « trop bas / normal / trop haut » qu'en termes de « x fois la borne supérieure ».

Table 5. Exemple réel de données de biologie (même établissement, même période, même libellé, même unité). A gauche, vraisemblablement la calcémie. A droite, vraisemblablement la calciurie.

Libellé :	Calcium	Calcium
Unité :	mg/l	mg/l
Bornes :	[84 ; 108]	[NA ; NA]
Distribution obervée :		

Ces observations rendent nécessaire la constitution de mappings de résultats de biologie ad hoc, incluant non seulement un alignement terminologique mais aussi une conversion d'unités. Nous mîmes en place la procédure suivante [91] :

- Regroupement de toutes les valeurs correspondant à un quadruplet commun : libellé, unité si elle existe, borne inférieure et borne supérieure si elles existent
- Définition de la loi empirique de distribution du paramètre dans l'échantillon analysé
- Transformation de cette distribution à l'aide de coefficients multiplicateurs de la forme $i*10^k$, où :
    - *k* vaut *0* ou un multiple (négatif ou positif) de *3*

    - *i* prend une valeur dépendant uniquement du paramètre. Par exemple, pour la glycémie, *1* ou *5,5* (*5,5* permet de passer des *g/l* aux *mmol.l*)
- Confrontation de toutes les distributions transformées aux distributions d'une base de données de référence, déjà nettoyée et utilisant la terminologie cible. La Figure 12 illustre sur la gauche deux paramètres alignables, et sur la droite deux paramètres non-alignables.
- Proposition automatisée d'une liste de paramètres possibles avec le coefficient de conversion approprié

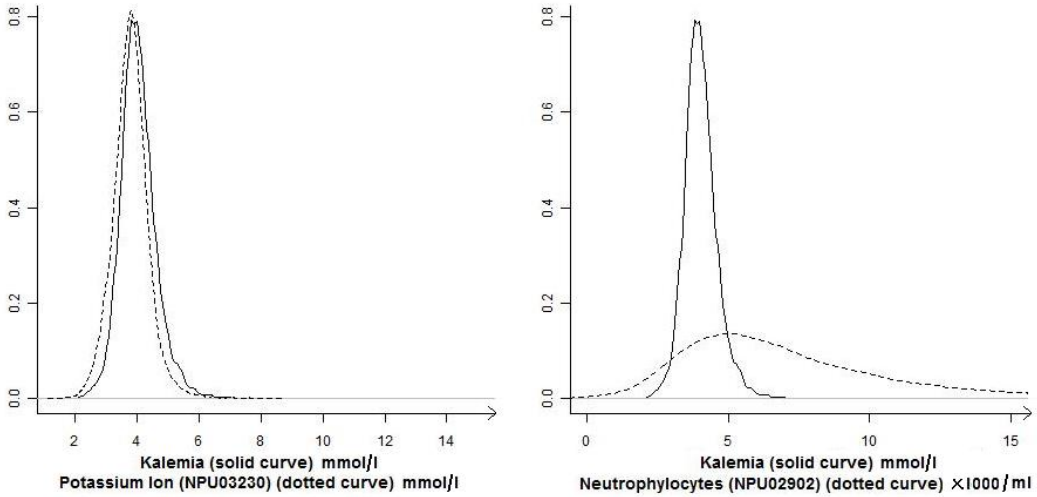
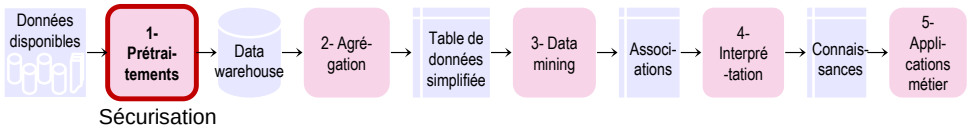


Figure 12. Mapping automatisé de données biologiques :
Gauche : exemple de concordance automatique
Droite : exemple de discordance automatique

Cette procédure, contrairement aux procédures existantes, n'utilise pas l'analyse textuelle des libellés. Pour chaque nouveau paramètre, l'outil développé propose les meilleurs paramètres assortis des facteurs de conversion qui conviennent. L'évaluation de l'outil conclut que, sur 15 nouveaux paramètres, le bon paramètre de référence (avec le bon coefficient) fut proposé parmi les 5 premiers candidats à 14 reprises, alors que la base de référence décrivait 70 paramètres différents. A l'aide d'un tel outil, il est alors très aisé pour la personne qui finalise le mapping de trancher en lisant simplement les libellés.

2.3 Sécurisation des données

2.3.1 Tatouage de bases de données

Position :	
Publication :	*Coatrieux G, Chazard E, Beuscart R, Roux C. Lossless watermarking of categorical attributes for verifying medical data base integrity. Conf Proc IEEE Eng Med Biol Soc 2011;2011:8195–8. doi:10.1109/IEMBS.2011.6092021.* [94] *Franco Contreras J, Coatrieux G, Chazard E, Cuppens F, Cuppens-Boulahia N, Roux C. Robust lossless watermarking based on circular interpretation of bijective transformations for the protection of medical databases. Conf Proc IEEE Eng Med Biol Soc 2012;2012:5875–8. doi:10.1109/EMBC.2012.6347330.* [95]

Nous nous intéressâmes également à la sécurisation des données, par tatouage notamment. Le tatouage est plus habituellement utilisé pour la protection des images de radiologie. Prosaïquement, il consiste à calculer une chaîne de contrôle d'intégrité d'après les données réelles, puis à insérer cette chaîne dans les données elles-mêmes, en les perturbant de manière imperceptible, mais de telle manière qu'un algorithme permette d'extraire cette chaîne de tatouage et la comparer aux données réelles [96]. Le tatouage remplit alors deux objectifs.

Le premier objectif est d'assurer l'intégrité de l'image. Lorsque l'image est modifiée avec un logiciel de retouche de photographie, la chaîne devient incohérente avec les données, et révèle ainsi la manipulation. Un cas d'utilisation peut être le suivant : *un radiologue ne détecte pas une tumeur du sein sur une mammographie. Poursuivi par la patiente ayant subi un retard de diagnostic, il modifie l'image pour rendre la tumeur moins visible et ainsi dégager sa responsabilité.* Un deuxième cas d'utilisation peut être à l'inverse : *un radiologue est poursuivi par une patiente l'accusant de ne pas avoir vu une tumeur évoluée. S'appuyant sur la technologie de tatouage, le radiologue peut prouver que la tumeur était bien invisible à l'époque de la mammographie, en produisant une image dont il est certifié par le tatouage qu'elle n'a pas été modifiée entre-temps. Il ne peut donc pas être accusé de falsification*.

Le deuxième objectif est de connaître l'origine d'une image qui a été divulguée illégalement, caractérisant une infraction au secret médical. Ainsi par exemple, chaque export d'image peut engendrer un tatouage *ad hoc* intégrant notamment le login de la personne qui réalise l'extraction, ainsi que la date, l'heure et le lieu de l'extraction. De la sorte, lorsqu'un magazine *people* publie l'image, la saisie de l'image permet aux enquêteurs de connaître l'origine de la fuite. Selon l'intensité de la perturbation appliquée, ces techniques de tatouage peuvent résister à des pertes d'information importantes, comme une impression suivie d'une numérisation.

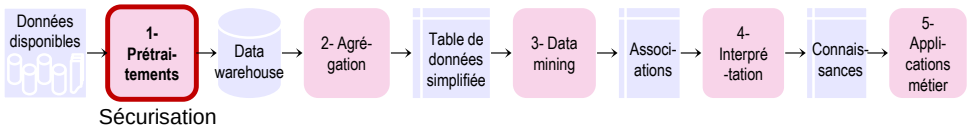
Ces techniques sont très utilisées en imagerie. L'objectif des travaux menés sous la direction de Gouénou Coatrieux (LATIM, laboratoire de traitement de l'information médicale, INSERM U1101, Brest) était d'appliquer ces techniques à des bases de données constituées d'attributs qualitatifs, tels les codes du PMSI [94,95]. Nous

pûmes ainsi tatouer les données issues d'un grand extrait de la base nationale 2011 du PMSI MCO, et montrer que cette stratégie permettait de retrouver le tatouage, et altérait peu les données, au sens où la simulation de groupage en GHM était conservée à 100%.

Cette technique suggère deux applications dans le cadre de la diffusion de bases nationales pour la recherche (par exemple bases du PMSI) :

- Une application de tatouage : il serait possible de tatouer les bases de données avant chaque cession à un tiers. De la sorte, en cas de divulgation même partielle de la base, il serait possible de retrouver le destinataire initial de la base de données.
- Une application de perturbation des données : au lieu de réduire considérablement les informations des bases, il est possible de les perturber de manière à ce qu'aucun enregistrement précis ne corresponde exactement à un individu réel, sans néanmoins perdre l'essentiel de l'information en termes d'interprétation médicale populationnelle.

2.3.2 Anonymisation de courriers en texte libre

Position :	
Publication :	*Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inform 2014;83:303–12. doi:10.1016/j.ijmedinf.2013.11.005.* [texte intégral en ligne]*[97]*

Dès le début de nos travaux de recherche, nous observâmes que l'analyse automatisée des données n'était pas toujours suffisante, et qu'il fallait **revoir les cas analysés un à un**. Ce constat s'imposa pour deux raisons principalement. Tout d'abord, la qualité des données est généralement imparfaite, et ce notamment parce que le codage sert à facturer le séjour. Le courrier de sortie, à l'inverse, vise à améliorer la prise en charge du patient. Ensuite, les analyses automatiques se basent généralement sur la présence de quelques attributs, prévus par le chercheur, mais ignorent les autres attributs. L'humain, en lisant les courriers, peut se faire une opinion plus fine de la prise en charge.

La confidentialité des données fit rapidement obstacle à toute initiative de revue de cas. En l'absence d'outil existant en langue française, il nous fallut donc mettre au point et évaluer une méthode d'anonymisation, ou plus précisément de dé-identification, de courriers médicaux francophones.

Nous mîmes au point puis évaluâmes la méthode FASDIM, première méthode accessible sous licence libre pour la dé-identification de textes médicaux francophones [97]. FASDIM signifie « fast and simple deidentification method ». L'approche de FASDIM constitue une rupture avec les approches précédentes, pour des raisons détaillées dans l'article. La méthode FASDIM suit le synopsis présenté en Figure 13.

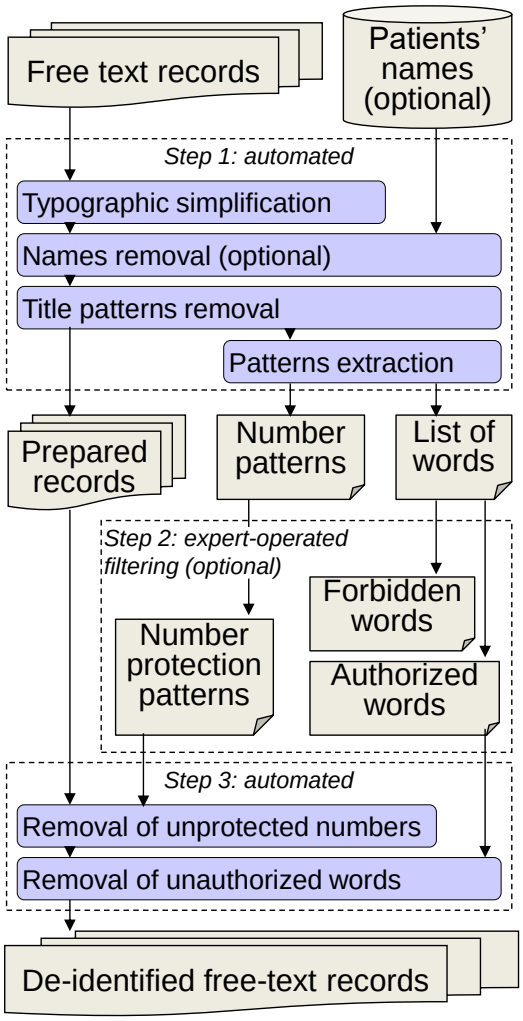


Figure 13. Synopsis de la méthode FASDIM

Les étapes peuvent être décrites comme suit.

Une première étape est automatique, et vise à préparer le texte. Elle débute par une simplification typographique (voir Figure 14). Cette simplification altère le texte au sens où il n'est plus lexicalement et syntaxiquement valide, mais n'empêche pas le lecteur de le comprendre, tout en diminuant fortement le nombre de mots différents.

$$é, è, ê, ë \rightarrow e \qquad œ \rightarrow oe \qquad ç \rightarrow c$$

Figure 14. Exemples de simplification typographique

Ensuite, lorsque le nom du patient auquel se rapporte le courrier est disponible, son nom et son prénom sont séparés en autant de mots qui sont recherchés et supprimés du texte (cette étape est tout à fait facultative, contrairement à ce qu'on pourrait penser). Enfin, nous supprimons les mots détectés par 48 expressions rationnelles [98] correspondant à des civilités, comme illustré en Figure 15. Ces civilités sont des titres qui introduisent un nom, tels « madame », « docteur », etc. Cette approche est satisfaisante dans un contexte de langage écrit soutenu.

Regular expression:	*\bmr\.\s+\w+\s+\w+\b (case insensitive)*
which means	*[WB]"mr."[WS][word] [WS] [word] [WB]*
with	*WB=word boundary, WS=whitespace character(s)*
Original string:	*"I have examined Mr. James Jones."*
Transformed string:	*"I have examined @ @ @."*

Figure 15. Exemple de suppression de civilité

Puis nous extrayons automatiquement les différents mots rencontrés dans tous les courriers analysés, ainsi que les mots qui précèdent ou suivent des nombres. Chaque élément est présenté accompagné d'un effectif d'occurrences.

La deuxième étape, à réaliser de temps à autres, consiste à filtrer les motifs extraits précédemment, afin de constituer des listes :

- Une liste de mots autorisés, qui sont jugés sans danger par l'expert quel que soit le contexte d'utilisation
- Une liste de motifs qui permettent d'identifier des nombres sans danger

La particularité de cette méthode est de considérer que l'aspect identifiant n'est pas le seul fait des noms propres. Ainsi par exemple, on peut penser que les mots « place » et « liberté » sont dangereux car ils peuvent souvent correspondre à des adresses postales, alors que ce sont bien des noms communs. Inversement, « Guillain » est bien un nom propre, mais dans un contexte médical il s'agira plus souvent du syndrome de Guillain-Barré.

La troisième et dernière phase est entièrement automatique. Elle consiste à protéger tout d'abord les nombres placés dans un contexte anodin (par exemple deux suivis d'une unité de mesure), à protéger les mots identifiés sur la liste des mots à conserver, et à supprimer tous les autres nombres et mots.

Nous testâmes la méthode FASDIM sur plus de 27 000 courriers. Tout d'abord, nous évaluâmes son efficacité versus un expert, de manière traditionnelle (voir Table 6). La F-mesure atteignit 87,9%, un très bon résultat en langue française.

Table 6. Résultats de l'évaluation de FASDIM en termes de rappel et précision (PHI = information personnelle à protéger)

Mesure	Valeur
Nombre de courriers	508
Nombre moyen de mots par courrier	510
Nombre moyen de PHI par courrier	20
Rappel (sensibilité)	98,1%
Précision (valeur prédictive positive)	79,6%
F-mesure	87,9%

Durant cette évaluation, nous observâmes que la plupart des mots supprimés à tort (sur 5 mots supprimés, 1 l'était à tort) n'étaient en réalité pas indispensables à la compréhension du texte. Nous évaluâmes donc la perte d'information en termes de concepts correspondant à des codes CIM10, CCAM ou ATC, perdus dans le courrier du fait de la dé-identification (voir Table 7). Il s'avéra que seulement 1% des

instances de concepts médicaux étaient affectées. De plus, la plupart des concepts étant répétés dans les courriers, la perte réelle est nettement inférieure.

Table 7. Conservation de l'information avec FASDIM

Catégorie d'information médicale	Taux de conservation
Toutes catégories	99,0%
CCAM : actes diagnostiques et thérapeutiques	99,7%
CIM10 : - Maladies, symptômes et motifs d'accès aux soins - Actes - Résultats anormaux de biologie	 99,5% 98,9% 97,0%
ATC : médicaments	98,8%

Enfin, la méthode FASDIM a la particularité de permettre un démarrage très rapide de l'anonymisation, avec inversement une tâche non-négligeable de filtrage de mots au cours de sa mise en œuvre. La définition de la méthode partit initialement d'un arbitrage entre la charge de travail initiale et la charge de travail marginale. Nous définîmes FASDIM en espérant proposer des temps de mise en œuvre représentés sur la Figure 16.

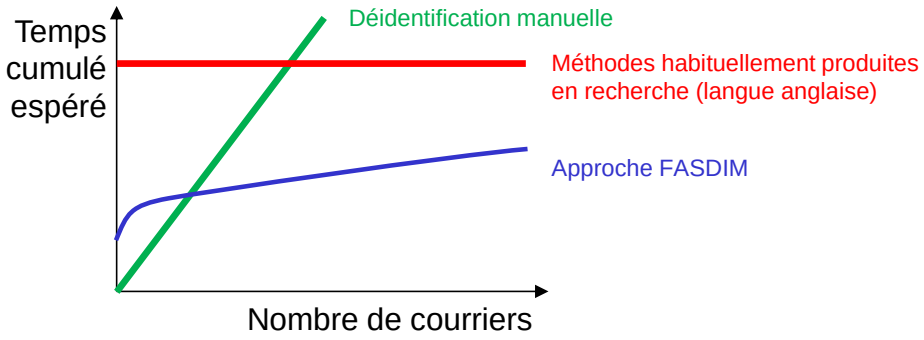


Figure 16. Comparaison des temps de traitement espérés (incluant mis en œuvre et traitement des courriers, hormis le temps machine)

Nous obtînmes un temps de mise en œuvre représenté en Figure 17, en partant de rien. Il faut néanmoins comprendre que, si une équipe reprenait le code tel qu'il a été diffusé, elle se situerait sur la partie linéaire de la courbe et non sur la partie gauche, essentiellement parce que les listes de mots actuellement utilisées contiennent déjà 512 motifs de nombres à protéger, et 28 325 mots à protéger, or le temps passé consiste essentiellement à produire ces listes.

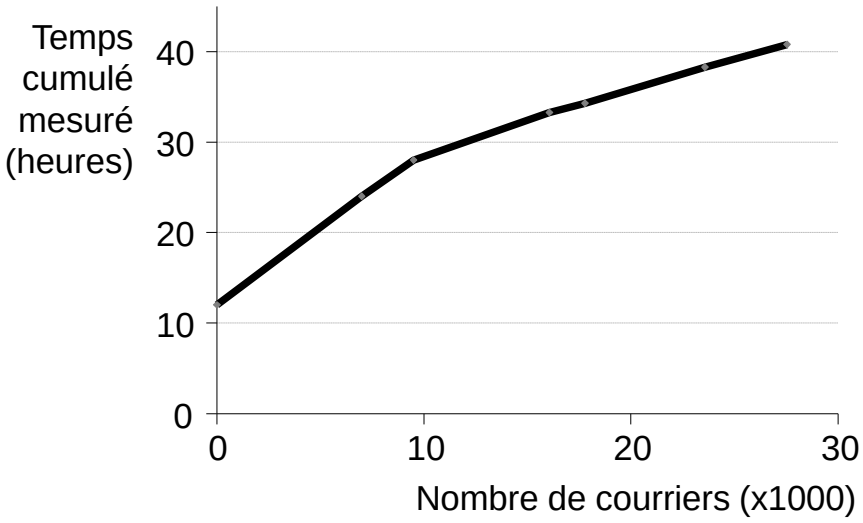
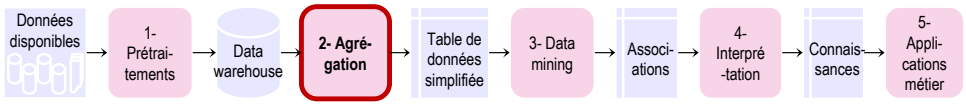


Figure 17. Temps de mise en œuvre de FASDIM en heures, en fonction du nombre de courriers en milliers (12h pour 0 courriers,40h pour 30 000 courriers)

Au total, nous pensons que FASDIM est l'approche idéale pour un jeu de courriers moyen, c'est-à-dire compris entre 100 courriers et 200 000 courriers. En-deçà de 100 courriers, une méthode purement manuelle sera efficace. Au-delà de 200 000 courriers, l'adaptation de méthodes anglophones à la langue française pourrait être à investiguer.

2.4 Agrégation des données (préalable aux analyses statistiques)

Position :	
Publication :	*Chazard E, Ficheur G, Merlin B, Genin M, Preda C, PSIP consortium, et al. Detection of adverse drug events detection: data aggregation and data mining. Stud Health Technol Inform 2009;148:75–84.* [texte intégral en ligne] *[99]* *Chazard E, Ficheur G, Merlin B, Serrot E, PSIP Consortium, Beuscart R. Adverse drug events prevention rules: multi-site evaluation of rules from various sources. Stud Health Technol Inform 2009;148:102–11.* [texte intégral en ligne] *[100]* *Caron A, Chazard E, Muller J, Perichon R, Ferret L, Koutkias V, et al. IT-CARES: an interactive tool for case-crossover analyses of electronic medical records for patient safety. J Am Med Inform Assoc 2016. doi:10.1093/jamia/ocw132.* [texte intégral en ligne] *[101]* *Ficheur G, Caron A, Beuscart J-B, Ferret L, Jung Y-J, Garabedian C, et al. Case-crossover study to examine the change in postpartum risk of pulmonary embolism over time. BMC Pregnancy Childbirth 2017;17:119. doi:10.1186/s12884-017-1283-y.* [texte intégral en ligne] *[102]* *Ficheur G, Caron A, Beuscart J-B, Ferret L, Putman S, Beuscart R, et al. The risks of pulmonary embolism and upper gastrointestinal bleeding beyond 35days after total hip replacement for coxarthrosis among middle-aged patients: A cross-over cohort. Prev Med 2016. doi:10.1016/j.ypmed.2016.09.010.* *[103]*

2.4.1 Utilité de l'agrégation des données

Dans le contexte du KDD, l'agrégation de données est un préalable indispensable à l'analyse statistique. Elle est nécessaire pour les raisons suivantes :

- Les méthodes statistiques, pour leur grande majorité, doivent être alimentées par un tableau présentant **une ligne par individu statistique** (séjour ou patient) et une colonne par variable d'intérêt. Il faut donc s'affranchir des nombreuses tables liées, et en particulier du détail des mesures répétées.
- Le **nombre de modalités** des variables qualitatives doit être **réduit**, et l'effectif de chaque modalité doit être augmenté.
- Les résultats des méthodes statistiques doivent **correspondre aux concepts et aux seuils** utilisés par les médecins et les experts du domaine.
- Il est nécessaire de **clarifier les concepts** intriqués sous-entendus par certaines données.
- Il est utile de **préparer les variables** explicatives pour optimiser leur relation avec les variables à expliquer.

En revanche, l'agrégation des données telle que nous l'entendons en KDD ne réduit pas toutes les dimensions d'un jeu de données :

- Elle **conserve strictement le nombre d'individus** statistiques
- Elle ne cherche **pas à réduire le nombre de variables** disponibles à l'échelle de l'individu. En général, ce nombre de variables augmente au contraire.

On voit finalement que cette agrégation a notamment pour vocation de trois des cinq caractères des données massives, tels que définis dans la Figure 6 en page 14. Elle permet en quelque sorte de passer des *complex big data* aux *flat big data*.

Nous allons illustrer cette nécessité dans le cas précis de la détection d'effets indésirables des médicaments.

Dans le cadre de nos travaux de recherche [15,101,104–107], nous examinâmes et testâmes plusieurs méthodes sur les données structurées de plus de 175 000 séjours hospitaliers (PMSI, biologie et médicaments) ou même de plusieurs millions (base nationale du PMSI). Nous utilisâmes des méthodes supervisées qui apportaient un résultat sous la forme de règles du type :

cause 1 & … & cause k → effet

Dans le domaine des ADE, un formalisme de règles simple permet une validation par des experts et un portage des règles dans n'importe quel langage (programmation, SQL, et même tableur). Ce type de règle peut notamment représenter la sortie de techniques comme les arbres de décision [86] et les règles d'association dans leur version supervisée [8].

Nous définîmes un modèle de données approprié et compatible avec les données d'une vingtaine d'hôpitaux danois, français et bulgares [88], permettant l'incorporation de près de 175 000 dossiers électroniques de patients issus de plusieurs établissements. Il s'agit ensuite simultanément de détecter des cas d'EIM, et d'identifier des motifs expliquant ou prédisant leur survenue.

Il est important de comprendre que les analyses que nous menâmes ne s'appuyèrent pas sur des variables informatives directement disponibles, contrairement à ce qu'on peut trouver par exemple dans les études épidémiologiques menées à partir d'un formulaire. Les données disponibles en KDD et en particulier en *data reuse* sont

généralement plus complexes mais également plus « neutres ». A notre sens, **un enjeu important du KDD et du data reuse est de simplifier les données mais également de mettre en évidence l'information implicite qu'elles portent en elles**. Nous illustrerons l'importance de cette démarche en revenant à l'exemple présenté initialement en Figure 8, page 18.

Dans cette figure, nous avons représenté des **données**, qui sont essentiellement **neutres** : des données administratives et démographiques, des diagnostics, des actes, des médicaments, des résultats de biologie et des prothèses. Aucune d'entre elles n'indique explicitement que la patiente en question a subi une hémorragie. Néanmoins, un médecin remarque immédiatement des **informations implicites** (voir Figure 8, page 18). Dans les résultats de biologie, nous observons une élévation de l'INR suivie d'une diminution de l'hémoglobine. Cela suggère une augmentation de l'activité anticoagulante d'un anti-vitamine K, suivie d'un saignement. Dans les médicaments administrés, nous observons une administration d'anti-vitamine K, suspendue le jour où l'INR atteint son pic, et remplacée par de la vitamine K, qui contrecarre son effet. Nous observons que l'INR se normalise immédiatement. Dans ces mêmes médicaments, nous observons également l'administration de culots globulaires, suivie d'une normalisation de l'hémoglobine. Cette administration se retrouve également dans les actes médicaux (FELF001). Nous retrouvons également une administration de paracétamol, une molécule susceptible d'interagir avec l'anti-vitamine K en potentialisant son effet.

Les éléments énumérés ci-dessus nous font fortement suspecter un effet indésirable du médicament correctement pris en charge. Pourtant, à aucun moment un tel diagnostic n'a été codé explicitement. Cela montre que, bien souvent, les bases apportent des données explicites très nombreuses mais neutres, tandis qu'un humain sait en extraire une information implicite synthétique ciblée sur ce qui nous intéresse. La machine étant incapable d'un tel raisonnement, il nous parut utile de développer des mécaniques de **transformation des données neutres en information ciblée**.

2.4.2 Procédé générique de l'agrégation de données

L'agrégation des données utilise en particulier trois opérations : la **binarisation**, l'**agrégation** et la **jointure**. L'ensemble du procédé est fortement guidé par une **expertise métier**.

D'un point de vue informatique, cette phase aboutit généralement à :

- La réintroduction de **redondances** dans les données
- La **perte d'informations** jugées peu utiles, et la conservation des autres
- La **réduction du schéma relationnel** à une seule table contenant une ligne par individu statistique (l'épisode de soins ou, plus rarement, le patient)

2.4.2.1 Binarisation

La binarisation consiste remplacer une variable native (qualitative ou quantitative) par une variable binaire fonctionnellement dépendante de cette première. Elle constitue une perte d'information ciblée, visant à conserver (et donc mettre en valeur) l'information que l'expert juge utile pour la suite de l'analyse.

Pour une variable quantitative, on utilisera volontiers les seuils communément admis au-delà ou en-deçà duquel la variable prend une valeur méliorative ou péjorative.

Pour une variable qualitative, on cherchera à classer et regrouper les modalités. Cette phase suppose la création d'une table de correspondance, ou « mapping ». Ces deux situations sont illustrées dans la Table 8.

Table 8. Exemples simples de binarisation. La colonne calculée est dans les deux cas la colonne de droite. (Exemple de gauche : variable quantitative. Exemple de droite : variable qualitative)

Id	Paramètre	Valeur	Hyperkaliémie
1	K+ sanguin	3.5	0
2	K+ sanguin	6.2	1
3	K+ sanguin	1.9	0

Id	Motif_admission	Diabète
1	Diabète type 1	1
2	Diabète type 2	1
3	Entorse	0

Nous énoncions plus haut que la binarisation permettait d'extraire l'information la plus important d'une variable, mais cette notion d'importance dépend naturellement de la finalité de l'analyse statistique. Ainsi par exemple, si on s'intéresse à la Rifampicine dans le cadre des effets indésirables des médicaments, c'est sans doute son caractère d'inducteur enzymatique qui est le plus important, alors que d'un point de vue bactériologique c'est le fait qu'il s'agisse d'un antibiotique antituberculeux qui importe le plus.

2.4.2.2 Agrégation

L'agrégation consiste à partir d'une table de données décrivant une entité, et de générer une table de données décrivant une entité de niveau supérieur c'est-à-dire, prosaïquement, avec moins de lignes. On agrège ainsi par exemple des mesures de biologies vers des séjours, ou des séjours vers des patients.

Les fonctions d'agrégation sont :

- Dans tous les cas : le dénombrement, le dénombrement des valeurs distinctes
- Pour les variables quantitatives (numériques) : la somme d'un groupe, le produit d'un groupe, le minimum, le maximum, la moyenne, le mode, la médiane, les autres quantiles, l'écart type, la variance
- Pour les variables qualitatives (textuelles) : le minimum, le maximum, la médiane et autres quantiles, la concaténation de groupe, la concaténation de groupe des valeurs distinctes
- Pour les variables booléennes : les opérateurs « et », « ou » et « ou exclusif »

Dans l'exemple ci-dessous (voir Table 9), la table d'origine présente des mesures répétées de kaliémie pour deux séjours hospitaliers. L'identifiant du séjour est une clef étrangère. La table de droite est calculée par simple agrégation de la précédente. Elle présente successivement le nombre de mesures, la mesure maximale et le nombre d'hyperkaliémie.

Table 9. Exemple simple d'agrégation
(Source à gauche : mesures de biologie ; Destination à droite : séjours)

Table source **=>** ***Table résultat***

Id_sej	Id_mes	Paramètre	Valeur	HyperK
A	1	K+	3.5	0
A	2	K+	6.2	1
A	3	K+	5.9	1
B	1	K+	3.4	0
B	2	K+	2.9	0

Id_sej	Nb_mes	K+_max	Nb_hyperK
A	3	6.2	2
B	2	3.4	0

2.4.2.3 Jointure

La jointure permet « d'accoler » deux tables sur la base d'une même clef, alors qu'éventuellement le nombre de lignes de ces deux tables peut différer. Nous illustrerons ici la **jointure externe gauche** (« left outer join » en SQL), la plus fréquemment utilisée.

Dans l'exemple de la Table 10, on dispose pour chaque séjour de l'âge du patient au moment du séjour, mais on souhaite y ajouter le sexe. Le sexe étant une information constante pour le patient, elle est disponible dans une table distincte, qui est une table de personnes physiques et non de séjours.

Table 10. Exemple de jointure vers l'entité supérieure (en « left outer join »)

Séjours

Id_sej	Id_pat	Age
1	A	53
2	A	54
3	B	62
4	Z	16

Patients

Id_pat	Sexe
A	H
B	F
C	H
D	F
E	H

=> ***Séjours v2***

Id_sej	Id_pat	Age	Sexe
1	A	53	H
2	A	54	H
3	B	62	F
4	Z	16	***NA***

Nous illustrerons maintenant comment ces trois techniques peuvent être mises en œuvre pour agréger des données et préparer leur analyse. Nous montrerons à travers ces exemples comment la connaissance experte peut guider le choix d'agrégation pour permettre la génération de connaissance depuis des données trop volumineuses au départ.

2.4.3 Développement de moteurs d'agrégation

Afin d'automatiser l'agrégation de données de nature différente vers un individu statistique unique (le séjour hospitalier), nous développâmes un moteur pour chaque type d'information (administrative et démographique, diagnostique, biologique et relative aux administrations de médicaments) [15,99]. Ces développements supposaient d'avoir une connaissance poussée, en amont, des données disponible et de leur signification médicale et, en aval, des méthodes d'analyse statistique capables de les traiter. Les moteurs développés poursuivaient deux buts (voir Figure 18) :

- un but sémantique : transformer les données neutres en information ciblée et synthétique
- un but structurel : transformer un schéma relationnel normalisé en une table unique entièrement dénormalisée mais plus aisée à traiter, comprenant des événements binaires associés à des dates

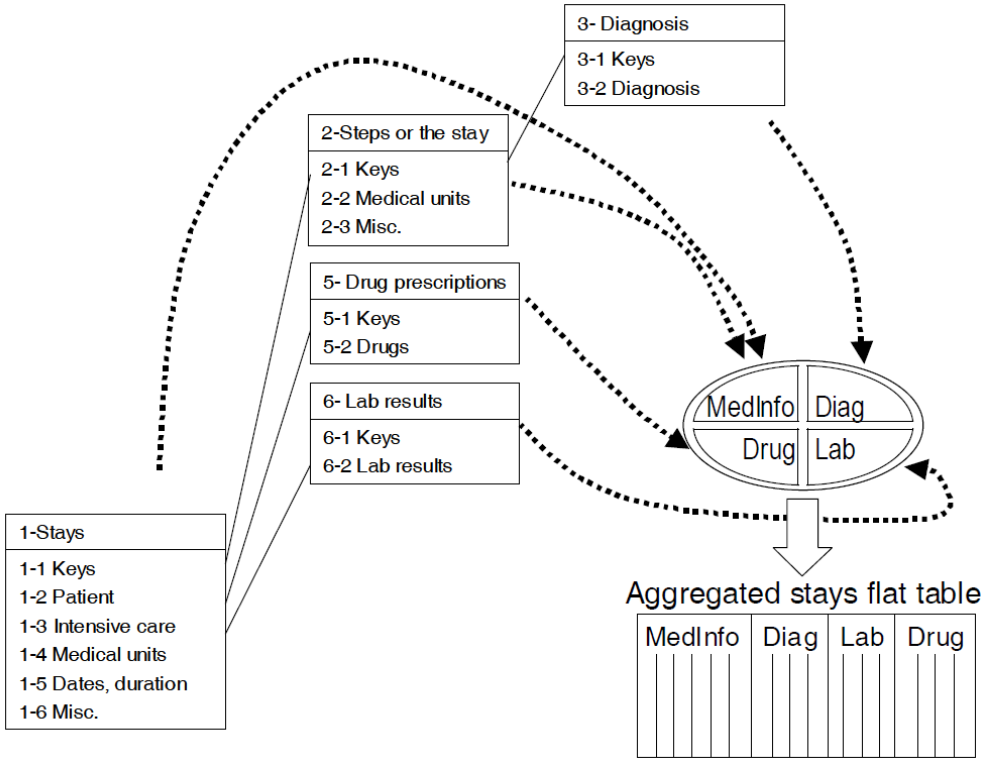


Figure 18. Schématisation des 4 moteurs d'agrégation de données hétérogènes

Le principe général de l'agrégation de données fut de mettre en évidence des « événements » suivant le même formalisme, bien que les données initiales fussent de forme et de sémantique très différente. Nous définîmes un événement comme une variable binaire assortie d'une date de début et une date de fin en cas de réalisation de l'événement (voir Figure 19). Nous prîmes le parti de générer tous les événements pertinents qu'il était possible d'extraire des données, à condition que leur fréquence fût suffisante.

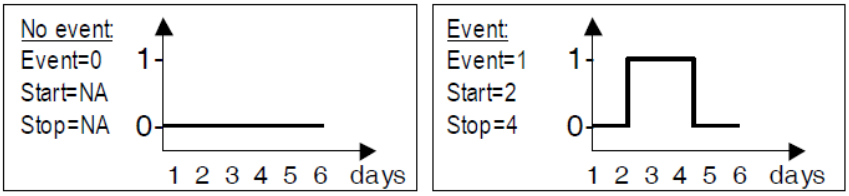


Figure 19. Formalisme générique d'un événement temps-dépendant.
Gauche : l'événement ne se produit pas. Droite : un événement se produit.

Nous développâmes ainsi quatre moteurs d'agrégation :

- un moteur d'agrégation des diagnostics, prenant les quelques 39 000 codes CIM10 en entrée pour produire **48** types d'événements (les diagnostics n'étant

pas datés, nous ne pûmes prendre en compte que les grandes catégories de maladies chroniques)
- un moteur d'agrégation des médicaments, prenant les quelques 5 400 codes ATC pour créer **284** types d'événements « exposition », avec également **284** événements de type « arrêt du médicament »
- un moteur d'agrégation des résultats de biologie, capable de créer **35** types d'événements différents
- un moteur d'agrégation des données démographiques et administratives, capables de créer **15** types d'événement différents

Ces moteurs d'agrégation permirent de définir ainsi **666** types d'événements différents.

Nous illustrerons plus précisément le fonctionnement du moteur d'agrégation des données de biologie, et des médicaments administrés.

2.4.4 Exemple de données biologiques

2.4.4.1 Présentation des données initiales

Les résultats d'analyses biologiques sont généralement représentés sous le modèle « **entité-attribut-valeur** ». Ce modèle permet d'intégrer dans une base de données des mesures de nature encore inconnue au moment où le modèle de données a été conçu. De plus, les mesures sont de plus répétées dans le temps. Un séjour donné peut ainsi être caractérisé par zéro à plusieurs milliers de lignes dans les tables de biologie. Dans les données dont nous disposons, selon les hôpitaux, un séjour hospitalier est ainsi caractérisé par une moyenne de **100 à 150 lignes** de résultats de biologie [15,99].

Dans l'exemple qui suit, on cherche à résumer les dosages de natrémie et de kaliémie de plusieurs patients, afin de disposer d'une ligne par séjour in fine. Les données d'origine sont représentées partiellement dans la Table 11. Leur représentation graphique correspond à la Figure 20.

Table 11. Exemple de données biologiques : kaliémie et natrémie d'un séjour, puis kaliémie d'un deuxième séjour

IdSej	Date	Parametre	Valeur
123	0	Potassium	4
123	1	Potassium	4
123	…	…	…
123	0	Sodium	140
123	…	…	…
528	0	Potassium	3.2
528	…	…	…

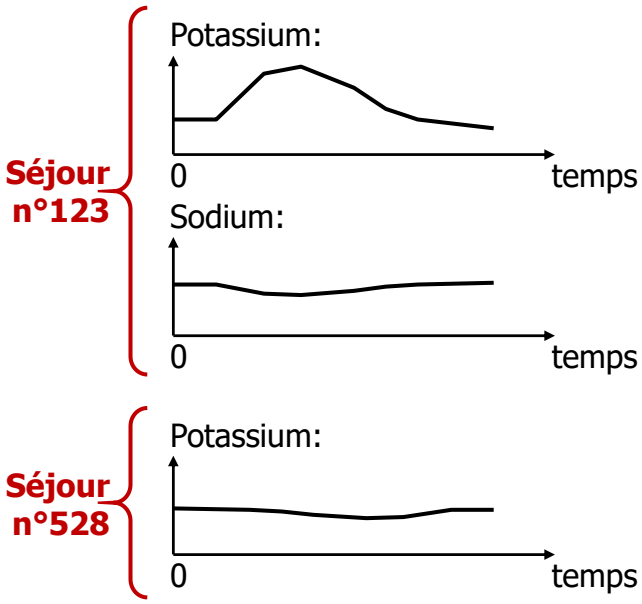


Figure 20. Représentation graphique des données de kaliémie et natrémie de l'exemple

2.4.4.2 Approche naïve

L'approche la plus simple consiste à calculer le minimum et le maximum de chaque paramètre par séjour (agrégation), puis à les réintégrer dans la table des séjours (jointure), comme illustré dans la Table 12. En l'absence de valeur mesurée, comme c'est le cas ici de la natrémie dans le séjour n°528, une valeur manquante apparaîtra.

Table 12. Exemple d'agrégation naïve des données biologiques

IdSej	Potassium.Min	Potassium.Max	Sodium.Min	Sodium.Max
123	3.8	6.2	136	142
528	2.8	3.2	*NA*	*NA*

2.4.4.3 Approche par seuils experts

Pourtant, l'expression la plus commune de la connaissance médicale est la notion d'**écart à la norme**, en l'occurrence l'hyperkaliémie, l'hypokaliémie, l'hypernatrémie et l'hyponatrémie. En ajoutant les bornes de normalité (en pointillés sur la Figure 21), les données de notre exemple deviennent ainsi celle de la Figure 21 : seules les valeurs rouges sont anormales, tandis que toutes les valeurs bleues sont normales.

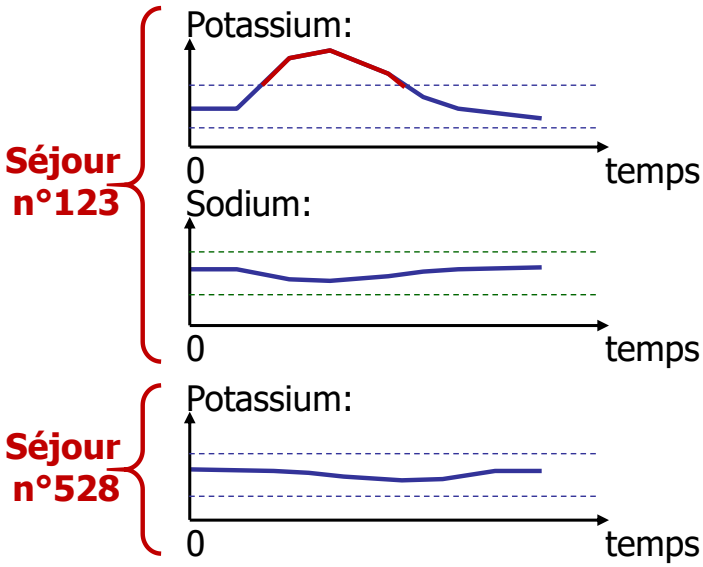


Figure 21. Données de kaliémie et natrémie de l'exemple : interprétation par rapport à des seuils

Ainsi, on peut inférer que le séjour est déviant si au moins une ses mesures est déviante. On obtient alors la Table 13. Là encore, les paramètres non mesurés font place à des données manquantes. Pour les mesures les plus fréquentes, en hospitalisation complète, il est peut également être admis qu'une valeur non-mesurée est une valeur normale. Ainsi par exemple, les patients chez lesquels on ne mesure pas la glycémie ont généralement une glycémie normale : en cas de signe d'appel ou d'antécédent personnel, ce paramètre est mesuré en routine.

Table 13. Exemple d'agrégation des données biologiques basée sur des seuils experts

IdSej	Hyperkaliemie	Hypokaliemie	Hypernatremie	Hyponatremie
123	1	0	0	0
528	0	0	NA… ou 0	NA… ou 0

Cette deuxième approche est moins naïve (moins neutre) car elle intègre une **connaissance experte**. Elle permet notamment la définition d'une date d'anomalie, exploitable avec des méthodes gérant les événements temporels.

2.4.4.4 Approche mono-variable complexe

Il existe des **mappings mono-variable plus complexes**, là aussi basés sur la connaissance experte, prenant en compte la cinétique d'un paramètre biologique. Nous citerons l'exemple du critère KDIGO [108], qui définit l'insuffisance rénale aiguë comme la présence d'au moins un de ces critères :

- Une augmentation de la créatininémie d'au moins 0.3 mg/dl (26.5 µmol/l) en au plus 48h
- Une augmentation de la créatininémie d'un facteur supérieur ou égal à 1.5 durant une période d'au plus 7 jours
- Une production d'urine inférieure ou égale à 0.5 ml/kg/h pendant 6 heures

Si l'on s'intéresse aux seuls premier et deuxième critère, ces critères nécessitent une analyse complexe de la créatininémie dans le temps. Il est nécessaire de

programmer deux fenêtres mobiles dans le temps, une de 48h et l'autre de 168h, qui glissent sur les valeurs de chaque patient pris individuellement. Une interpolation de type LOCF (« last observation carried forward », l'interpolation en marches d'escalier bleues de la Figure 22) sera utilisée. A chaque position de la fenêtre mobile, l'écart entre la valeur minimale et la valeur maximale capturées sera évaluée au regard des deux critères. La Figure 22 montre les valeurs d'origine, tandis que la légende indique les trois critères retenus.

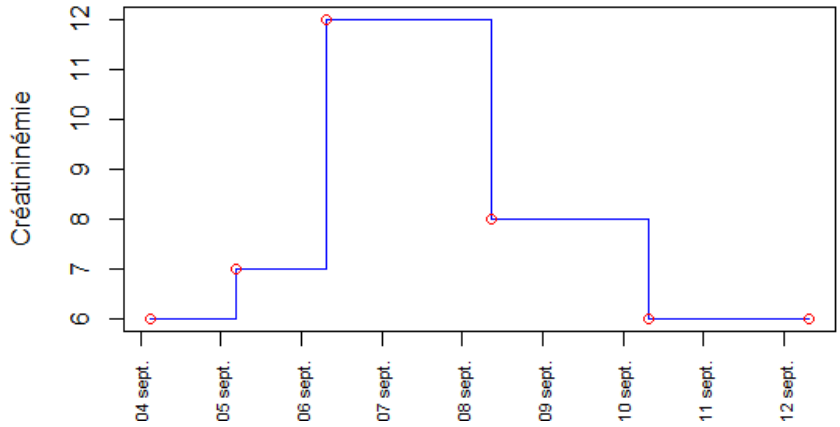


Figure 22. Exemple de mesures de créatininémie durant un séjour (points rouges). Critères validés :
Critère « x1.5 » à partir du 4 septembre à 03:59 : la créatininémie passe de 6 à 12
Critère « +3 » à partir du 5 septembre à 05:25 : la créatininémie passe de 7 à 12
Critère « x1.5 » à partir du 5 septembre à 05:25 : la créatininémie passe de 7 à 12

2.4.4.5 Approche multi-variable

On peut aller plus loin dans l'intégration de connaissances expertes dans le processus de mapping, en définissant des variables agrégées multi-sources [15,99].

On peut ainsi définir par exemple :

- l'insuffisance rénale biologique :
  augmentation de l'urémie **OU** augmentation de la clairance de la créatinine **OU** augmentation de la créatininémie (avec des seuils adaptés)
- la pancytopénie :
  diminution du nombre de globules rouges **ET** diminution du nombre de globules blancs **ET** diminution du nombre de plaquettes

2.4.5 Exemple de l'agrégation des médicaments administrés

2.4.5.1 Présentation des données

L'agrégation des médicaments suit un procédé similaire, mais il existe trois différences majeures :

- Les médicaments sont administrés ou non. Contrairement aux données de biologie, il n'est pas indispensable d'utiliser des seuils. De plus, un médicament non-administré n'est pas une donnée manquante, mais l'absence d'exposition.
- Ensuite, on peut également tenir compte de l'**arrêt du médicament** sous la forme d'un événement spécifique.

- Enfin, l'effet d'un médicament, qu'il soit pharmacodynamique ou pharmacocinétique, ne dépend pas d'une seule propriété, mais d'**un ensemble de propriétés**. En outre, réciproquement, ces propriétés sont **partagées** pour tout ou partie avec d'autres médicaments. Enfin, un même principe actif peut être décrit dans **plusieurs items** de la classification ATC.

Illustrons ce troisième point avec trois exemples, autour de l'aspirine.

Tout d'abord, l'**aspirine** peut être utilisée pour des indications thérapeutiques et sur des appareils différents. De ce fait, on retrouve ce médicament référencé sous **8 codes ATC** différents, eux-mêmes relatifs à **5 appareils** différents :

- A alimentary tract and metabolism
  - A01AD other agents for local treatment
    - A01AD05 …aspirin…
- B blood and blood forming organs
  - B01AC platelet aggregation inhibitors
    - B01AC06 …aspirin…
- C cardiovascular system
  - C10BX (…) other combinations
    - C10BX01 & C10BX02 …aspirin…
- M musculo-skeletal system
  - M01BA anti-inflammatory
    - M01BA03 …aspirin…
- N nervous system
  - N02BA salicylic acid and derivatives
    - N02BA01, N02BA51, N02BA71 …aspirin…

Ensuite, quelle que soit l'indication thérapeutique, l'aspirine possède **plusieurs propriétés** d'importance dans le champ qui nous concerne :

- Des propriétés pharmacodynamiques d'importance thérapeutique : anti-inflammatoire, antalgique, antipyrétique, antiagrégant plaquettaire
- Des propriétés pharmacocinétiques, augmentant les interactions possibles : molécule se liant à l'albumine, métabolisme hépatique, excrétion rénale, etc.
- De nombreux effets indésirables qui découlent directement des deux catégories ci-dessus, mais également : augmentation directe de l'acidité gastrique, risque de manifestations allergiques diverses, syndrome de Reye, etc.

Enfin, chacune des propriétés ci-dessus est **partagée par d'autres molécules**. Prenons l'exemple de deux propriétés :

- Antiagrégant plaquettaire : c'est également le cas des anti-inflammatoires non stéroïdiens, des inhibiteurs de la voie de l'ADP (dipyridamole, ticlopidine, clopidogrel) et des inhibiteurs des récepteurs GPIIbIIIa (abciximab).
- Liaison à l'albumine : c'est le cas de très nombreux acides faibles, tels les diurétiques, barbituriques, hypocholestérolémiants, anti-vitamines K, etc.

2.4.5.2 Approche par table de correspondance (mapping)

Nous définîmes un **mapping des médicaments** afin de rendre compte des principales propriétés, tout en réduisant au maximum le nombre de catégories, et en augmentant le nombre d'événements dénombrés dans chaque catégorie. Ce

mapping fut défini **avec redondance**, c'est-à-dire qu'un code ATC donné permlettait de déclencher simultanément plusieurs types d'événements, tandis qu'un événement donné pouvait être déclenché par différents codes ATC. Cette redondance pouvait être la manifestation de différents concepts de représentation des connaissances, comme illustré en Figure 23.

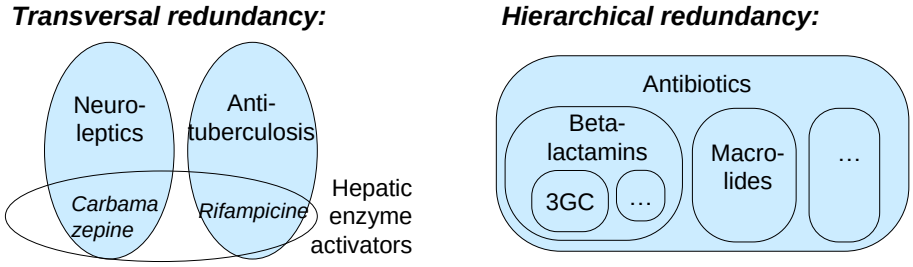


Figure 23. Exemples de redondance dans le mapping des médicaments : redondance liée à une propriété transversale (à gauche), et redondance liée à une classification hiérarchique (à droite)

Nous mîmes au point une matrice contenait **5383 lignes** (autant de codes ATC en entrée) et **284 colonnes** (autant de types d'événements en sortie), soit 1,5 millions de cellules, décrivant **8028 relations**, notées « 1 » dans cette matrice. L'exemple de la catégorie ATC N02BA est montré en Figure 24.

atc	name	plateletAgginhib	NSAI	NSAI_without_aspirin	NSAI_aceticAcidDerivate	NSAI_aspirin	potassium_ion	otherAntalgic	psycholeptic
N02BA01	Acide acétylsalicylique	1	1	0	1	1	0	0	0
N02BA02	Aloxiprin	0	1	0	1	1	0	0	0
N02BA03	Salicylate de choline	0	1	0	1	1	0	0	0
N02BA04	Salicylate de sodium	0	1	0	1	1	0	0	0
N02BA05	Salicylamide	0	1	0	1	1	0	0	0
N02BA06	Salsalate	0	1	0	1	1	0	0	0
N02BA07	Éthenzamide	0	1	0	1	1	0	0	0
N02BA08	Salicylate de morpholine	0	1	0	1	1	0	0	0
N02BA09	Dipyrocétyl	0	1	0	1	1	0	0	0
N02BA10	Bénorilate	0	1	0	1	1	0	0	0
N02BA11	Diflunisal	0	1	0	1	1	0	0	0
N02BA12	Salicylate de potassium	0	1	0	1	1	1	0	0
N02BA14	Guacétisal	0	1	0	1	1	0	0	0
N02BA15	Carbasalate calcium	0	1	0	1	1	0	0	0
N02BA16	Salicylate d imidazole	0	1	0	1	1	0	0	0
N02BA51	Acide acétylsalicylique, associations sans psycholeptiques	1	1	0	1	1	0	1	0
N02BA55	Salicylamide, associations sans psycholeptiques	0	1	0	1	1	0	0	0
N02BA57	Éthenzamide, associations sans psycholeptiques	0	1	0	1	1	0	0	0
N02BA59	Dipyrocétyl, associations sans psycholeptiques	0	1	0	1	1	0	0	0
N02BA65	Carbasalate calcium associations sans psycholeptiques	0	1	0	1	1	0	0	0
N02BA71	Acide acétylsalicylique, associations avec des psycholeptiques	0	1	0	1	1	0	0	1
N02BA75	Salicylamide, associations avec des psycholeptiques	0	1	0	1	1	0	0	1
N02BA77	Éthenzamide, associations avec des psycholeptiques	0	1	0	1	1	0	0	1
N02BA79	Dipyrocétyl, associations avec des psycholeptiques	0	1	0	1	1	0	0	1

Figure 24. Extrait de matrice permettant un mapping des codes ATC (lignes) vers des propriétés (colonnes). Extrait correspondant à la catégorie ATC N02BA (Salicylés). Au total : 5383 lignes et 284 colonnes.

Ensuite, **certains actes CCAM impliquaient l'administration** d'un médicament, qui n'est pas toujours référencé tel quel. En voici trois exemples :

- L'acte de transfusion de concentré de globules rouges (code CCAM FELF001) impliquait de fait l'administration d'érythrocytes (code ATC B05AX01)
- L'acte de scintigraphie du corps entier à l'iode 131 (code CCAM ZZQL003) impliquait de fait l'administration d'un composé à l'iode 131 (code ATC V09XA)
- L'acte d'anesthésie pour appendicectomie par voie iliaque (code CCAM HHFA001-0-4) impliquait de fait l'administration d'anesthésiques par voie générale (code ATC N01A)

Afin de prendre en compte les médicaments portés par ces actes, nous dûmes définir un **mapping CCAM-ATC** en amont de la transformation par la matrice des médicaments.

Enfin, certaines administrations de médicaments étaient représentées par un libellé, qui pouvait par exemple résulter de la concaténation des composés d'une perfusion sanguine, et étaient **dépourvues de code ATC**. Dans un exemple récent, la base de données étudiée comportait 36% de lignes d'administrations sans code ATC. Après prise en compte de seulement 387 mots clefs dans un **mapping mot clef-ATC**, cette proportion de données manquantes chuta à 5.5% sur près d'un million de lignes.

2.4.6 Evénements temps-dépendants : causes ou effets ?

Les événements définis précédemment à l'aide des moteurs d'agrégation purent ensuite être tantôt utilisés pour tenter d'expliquer (ou prédire) des événements péjoratifs, tantôt constituer eux-mêmes des événements péjoratifs à expliquer. Les deux règles ci-dessous illustrent par exemple comment l'insuffisance rénale aiguë peut être tantôt une cause, tantôt un effet :

Insuffisance rénale aiguë & médicament A → effet indésirable B

Médicament C & médicament D → insuffisance rénale aiguë

Afin de déterminer les effets potentiels des médicaments à recherche, nous analysâmes les résumés des caractéristiques du produit (RCP) produites par l'AFSSAPS, afin d'en extraire la liste des manifestations possibles des EIM, en faisant abstraction des causes qui, selon les RCP, auraient pu induire ces manifestations. Parmi ces manifestations possibles, nous pûmes surveiller l'apparition dans les données de **83 types événements**, soit directement (exemple : hyperkaliémie) soit indirectement (exemple : l'administration d'un thrombolytique à distance de l'admission du patient peut témoigner de la survenue d'une thrombose durant le séjour).

Nous pûmes ainsi réutiliser certains des événements définis plus haut comme des « effets potentiels » (variables à expliquer, *dependant variables*), tandis que tous les autres événements étaient considérés comme des « causes potentielles » (variables explicatives, *independant variables*). Ainsi, les quatre moteurs d'agrégation cités précédemment purent tantôt créer des événements à expliquer (« effets »), tantôt des événements potentiellement explicatifs (« causes »). Ainsi par exemple :

- moteur d'agrégation des diagnostics :
  - *Exemple de cause : insuffisance rénale chronique*

    - *Effet : il ne fut pas possible d'en créer dans la mesure où les diagnostics ne sont pas datés*
- moteur d'agrégation des médicaments :
    - *Exemples de cause : antivitamine K, arrêt d'une butyrophénone*
    - *Exemple d'effet : administration de vitamine K (elle peut être le témoin d'un surdosage en AVK)*
- moteur d'agrégation des résultats de biologie :
    - *Exemple de cause : hypoalbuminémie*
    - *Exemple d'effet : élévation de l'INR*
- moteur d'agrégation des données démographiques et administratives :
    - *Exemple de cause : âge>80*
    - *Exemple d'effet : décès*

Ainsi, tandis que seules certaines causes purent être recyclées en effets, réciproquement, tous les effets furent utilisables comme causes potentielles. Cette approche originale permit de potentiellement découvrir des effets domino, que nous baptisâmes « adverse drug event domino effect », comme par exemple :

Médicament A & Âge>70 → insuffisance rénale aiguë (élévation créatininémie)

PUIS insuffisance rénale aiguë & Médicament B → hémorragie

2.4.7 Conclusion sur l'agrégation des données

L'agrégation de données a un impact majeur sur la possibilité d'obtenir des résultats probants [1]. Cette agrégation nécessite la synergie de plusieurs compétences, par ordre chronologique.

Il faut premièrement **connaître la nature des données médicales et leur signification**, afin de sacrifier de l'information en termes statistique, en gardant l'essentiel de l'information en termes médicaux, et parfois même en en faisant apparaître (par exemple le BMI, la propriété d'induction enzymatique de médicaments, etc.).

Ensuite, il est nécessaire de comprendre quelle forme de données les **méthodes statistiques** sont capables de traiter. L'étape d'agrégation de données est également fortement influencée par la **finalité du traitement**.

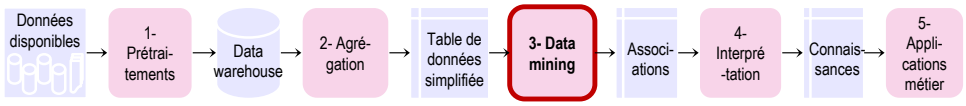
Enfin, ces transformations de données, une fois définies, nécessiteront le recours à des **compétences de programmation**, car le data management nécessaire pour les mettre en œuvre dépassera souvent les opérations couramment implémentées dans des fonctions standard.

Une dernière remarque est que la chose parait assez claire lorsqu'il s'agit de mettre en évidence une exposition et un résultat, lorsque la question est clairement posée et que l'exposition comme le résultat sont connus au moment d'analyser les données. Ce fut le cas par exemple lorsqu'Irène Frachon et al. investiguèrent le lien entre l'administration de Mediator® (benfluorex) et la survenue d'insuffisance mitrale [109,110]. Les investigateurs savaient avant d'agréger les données quelle exposition et quel résultat (*outcome*) exprimer sous forme binaire au niveau de l'individu statistique (ici le patient au sens personne physique).

En data mining, les choses sont moins claires. La plupart du temps, certes, le résultat est clairement défini (ce n'était néanmoins pas le cas dans nos travaux [15]). Néanmoins, l'objectif est souvent de découvrir des **facteurs de risque encore**

inconnus. A cet effet, il faut donc le plus souvent faire apparaître sous forme de variables binaires une multitude d'expositions, qu'il s'agisse de pathologies, d'états biologiques, de médicaments administrés, d'actes réalisés, d'antécédents personnels, ou même de variables de contexte. Ce choix est donc un **choix d'expert** guidé tout d'abord par ce qu'on s'attend à retrouver (connaissance du domaine) et par l'expérience, puis généralisé à titre systématique à l'ensemble des données disponibles.

2.5 Fouille de données : induction supervisée de règles et filtrage automatisé

Position :	
Publication :	*Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed 2011;15:823–30. doi:10.1109/TITB.2011.2165727.* [texte intégral en ligne] *[15]* *Chazard E, Preda C, Merlin B, Ficheur G, PSIP consortium, Beuscart R. Data-mining-based detection of adverse drug events. Stud Health Technol Inform 2009;150:552–6.* [texte intégral en ligne] *[107]* *Ficheur G, Chazard E, Merlin B, Ferret L, Luyckx M, Beuscart R. Supervised analysis of drug prescription sequences. Stud Health Technol Inform 2013;192:293–7 .* [texte intégral en ligne] *[111]* *[1]Chazard E, Bernonville S, Ficheur G, Beuscart R. A statistics-based approach of contextualization for adverse drug events detection and prevention. Stud Health Technol Inform 2012;180:766–70.* [texte intégral en ligne] *[105]*

Nous illustrerons la fouille statistique de données dans le cas des effets indésirables des médicaments. Nous appliquâmes cependant des techniques similaires à d'autres domaines, cités en articles, mais qui ne seront pas développés ici.

Afin de prédire les résultats potentiels à l'aide des causes potentielles (définis précédemment), nous utilisâmes des méthodes d'induction supervisée de règles, principalement les arbres de décision et les règles d'associations supervisées [1,8,86]. Nous exécutâmes ces techniques de manière itérative, en testant chaque variable qualifiée d'effet potentiel, en relation avec toutes les variables qualifiées de cause potentielle, proposées simultanément au modèle. Nous introduisîmes des contraintes temporelles de manière à ce que les règles produites respectassent nécessairement l'antériorité des causes par rapport aux effets. Pour ce faire, lors de chaque itération, nous modifiâmes automatiquement et artificiellement le jeu de données de manière à « désactiver » les causes chronologiquement incompatibles avec l'effet étudié.

L'arbre en Figure 25 illustre comment il est possible, dans un service donné, d'isoler des sous-groupes d'individus qui présentent une diminution de l'INR durant leur séjour. A partir de cet arbre, la Figure 26 montre comment trois règles peuvent être extraites. On observe dans ce cas que les facteurs impliqués sont l'âge, le niveau de

l'INR à l'admission du patient et les médicaments administrés. A ce stade, les règles découvertes ne sont pas généralisables.

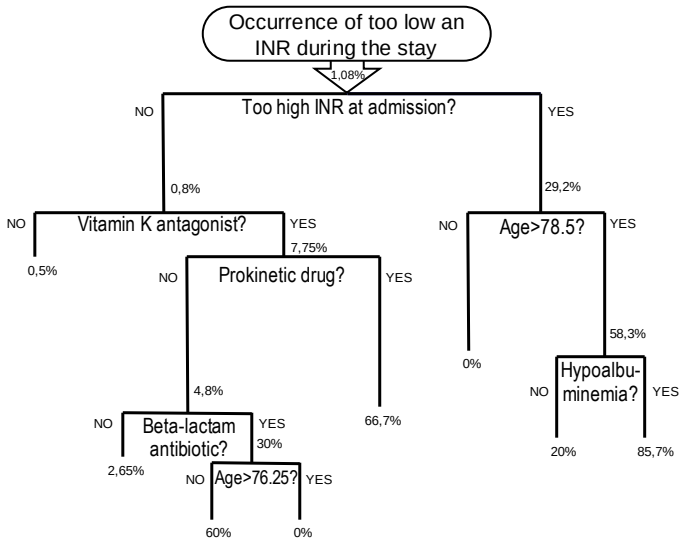


Figure 25. Exemple d'un arbre de décision visant à prédire la survenue d'un sous-dosage en AVK

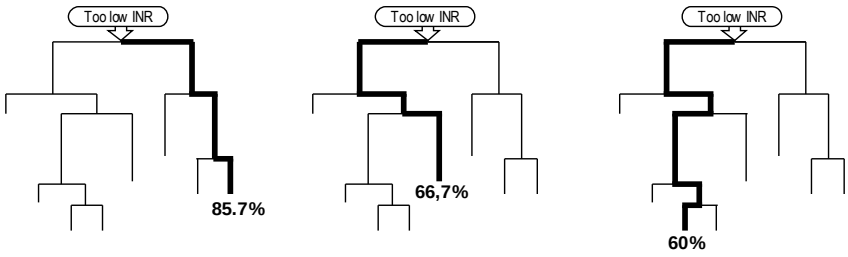


Figure 26. Trois exemples de règles associées à une augmentation de probabilité dans l'arbre précédent

Nous mîmes ensuite en place une procédure informatique pour automatiquement incorporer les résultats des méthodes statistiques dans un entrepôt de règles (voir Figure 27). Nous décrivîmes pour ce faire un schéma de données compatible avec les règles, indépendamment de la méthode statistique dont elles pouvaient être issues (voir Figure 28). Ce même schéma de données nous permit d'inclure manuellement des règles issues de la connaissance académiques, et que nous souhaitions tester. Il permit également d'incorporer les résultats issus de la réexécution des règles dans d'autres jeux de données : ainsi, lorsqu'une règle était découverte dans un service et y obtenait par exemple une certaine confiance, il était possible de savoir quelle confiance cette règle obtenait lorsqu'elle était exécutée de force dans un autre service, y compris lorsque cette règle n'y avait pas été découverte.

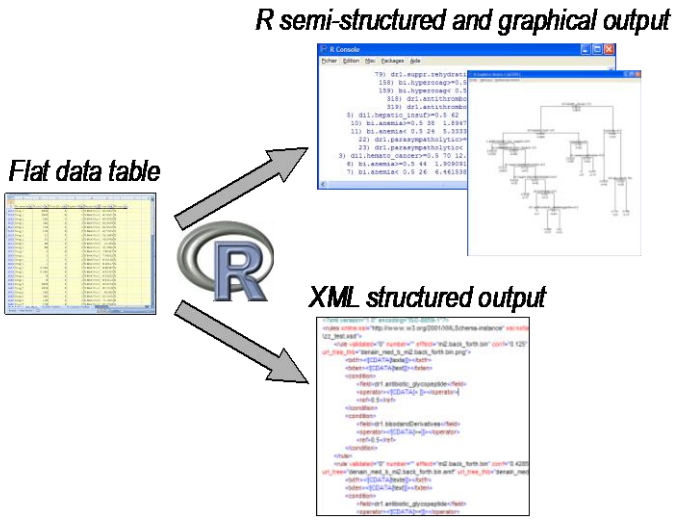


Figure 27. Transcription automatique des sorties semi-structurées (en haut à droite) dans un entrepôt de règles (en bas à droite)

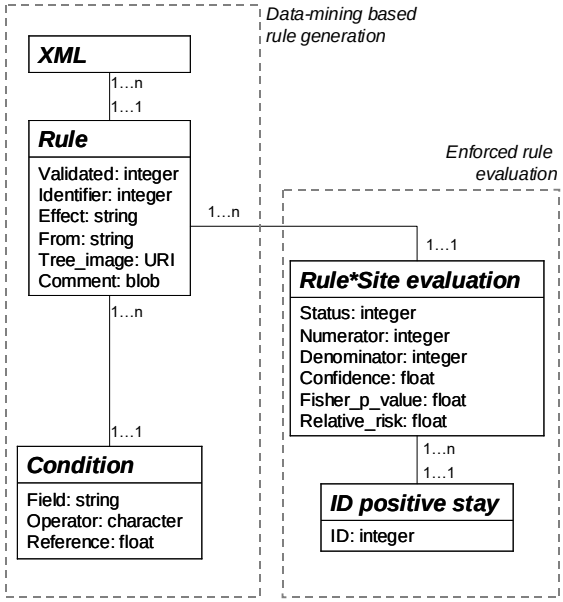


Figure 28. Description du schéma de données permettant de stocker les règles

Ces règles durent ensuite être filtrées automatiquement. En effet, lorsque le patient subit un événement indésirable (médicamenteux ou non), c'est le plus souvent lié à sa maladie et non aux médicaments, ce qui est exactement l'inverse de la définition des EIM (voir Figure 29). Nous procédâmes donc à un filtrage automatisé des résultats produits par les procédures de data mining. Seules les règles contenant au moins un médicament, ou l'arrêt d'un médicament, ou encore un résultat de biologie implicitement lié à un médicament (exemple : digoxinémie, INR, etc.) furent conservées.

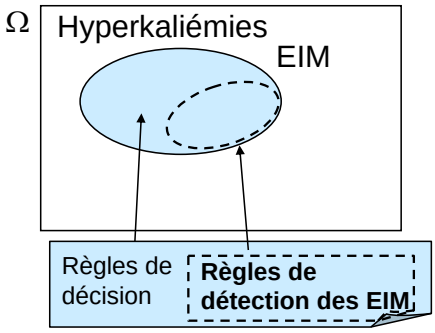
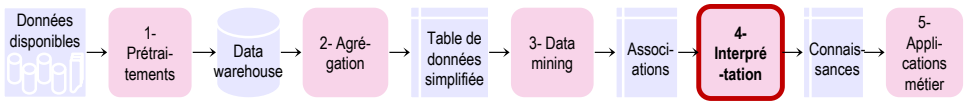


Figure 29. Parmi les règles prédisant une hyperkaliémie, il fallait filtrer celles en rapport avec un EIM

Nous obtînmes ainsi plusieurs centaines de règles, qui furent filtrées validées et réorganisées par des experts, comme développé ci-après.

2.6 Filtrage expert, validation, réorganisation

Position :	
Publication :	*Cette section suit directement la précédente. Elle réfère aux mêmes publications, dans lesquelles génération et validation de règles sont indissociables.*

Nous organisâmes des sessions de filtrage expert, validation et réorganisation de ces règles en présence de médecins, pharmaciens, pharmacologues et statisticiens. Ces sessions permirent de valider les règles, les réorganiser pour limiter les recouvrements, et les accompagner de commentaires circonstanciés et de références bibliographiques [99].

Enfin, nous décrivîmes les règles sous la forme d'une base structurée objet, et nous développâmes un environnement de test automatisé des règles dans tous les services disponibles, afin de calculer des statistiques contextualisées, c'est-à-dire calculées pour chaque règle dans chaque service (voir Figure 30) [100,104].

Rule: $C_1 \cap ... \cap C_k \rightarrow O$

Time constraints:

For $(C_1 \cap ... \cap C_k)$: all the conditions are present in the same time (they can start at different times)

For $(O \cap C_1 \cap ... \cap C_k)$: the same as above, and all the conditions are present before the outcome starts.

Support: $Sup = P(O \cap C_1 \cap ... \cap C_k)$

Confidence: $Conf = P(O / C_1 \cap ... \cap C_k)$

Risk ratio: $RR = \frac{P(O / C_1 \cap ... \cap C_k)}{P(O / \overline{(C_1 \cap ... \cap C_k)})}$

P value: p value of the Fisher's exact test for independency between the outcome O and the set of conditions $(C_1 \cap ... \cap C_k)$

Delay: median delay between $Time(C_1 \cap ... \cap C_k)$ and $Time(O)$ (when both events occur).

Figure 30. Exemple de statistiques contextualisées calculées pour chaque règle dans chaque service

Nous produisîmes *in fine* un ensemble de 236 règles de détection des EIM. Ces règles étaient formalisées de manière structurée, accompagnées de commentaires textuels et références bibliographiques et caractérisées dans une vingtaine de service par des statistiques contextualisées.

Parmi les 236 règles produites par ce travail, 72% apportèrent de nouvelles connaissances (voir Table 14). Ces nouvelles connaissances furent le plus souvent modestes mais utiles (ex : l'effet de telle interaction n'est perceptible que chez les plus de 70 ans) et parfois totalement novatrice mais néanmoins validées (ex : l'arrêt d'halopéridol chez un patient sous anti-vitamine K peut induire un surdosage en AVK).

Table 14. Les 236 règles de détection des EIM

Type de règle	Nombre
Règles présentes dans les résumés des caractéristiques du produit (RCP), insérées mais non découvertes spontanément	40
Règles découvertes par data mining et par ailleurs présentes dans les RCP	25
Règles découvertes par data mining, apportant de nouveaux éléments par rapport aux RCP	127
Règles entièrement nouvelles mais validées	44
Total	***236***

Chacune des 236 règles fut en outre caractérisée par des statistiques détaillées, calculées dans chacun des services disponibles (voir Table 15 ; les données ne sont détaillées que sur 2 des 5 hôpitaux disponibles).

Table 15. Exemple de statistiques détaillées calculées pour une règle donnée.

VKA & amoxicillin and clavulanic acid & age ≥ 70 → VKA overdose (INR>4.9)

Department	Confidence (PPV)	Support (frequency)	Median delay	Relative risk	Fisher's test P value
H1_all	15/73=**20.6%**	15/6110=**2.5‰**	6j	16.54	0
H1_chir	0/3=**0%**	0/1150=**0‰**		0	1
H1_geriatrics	5/12=**41.7%**	5/358=**14‰**	5j	14.42	0
H1_gynobs	*No stay*				
H1_med_a	2/13=**15.4%**	2/1337=**1.5‰**	4j	9.7	0.0197
H1_med_b	3/17=**17.7%**	3/1026=**2.9‰**	3j	11.13	0.0031
H1_pneumo	6/32=**18.8%**	6/881=**6.8‰**	9.5j	6.63	0.0005
H2_all	1/10=**10%**	1/11923=**0.1‰**	6j	33.09	0.0306
H2_apoplexy	*No stay*				
H2_cardio_endocrino	1/2=**50%**	1/1967=**0.5‰**	6j	51.71	0.0202
H2_geriatrics	0/2=**0%**	0/493=**0‰**		0	1
H2_gynecology	*No stay*				
H2_icu	*No stay*				
H2_internal_med	0/5=**0%**	0/1514=**0‰**		0	1
H2_obstretrics	*No stay*				

Le formalisme proposé pour les règles de détection des EIM était classique, de même que le fait d'utiliser les médicaments administrés, les variables démographiques et les résultats de biologie [14,112–117]. En revanche, le fait d'associer indifféremment tous ces types de variables dans une règle donnée fut une valeur ajoutée de ce travail, chaque système étant le plus souvent limité à 1 ou 2 types de variables par règle. De plus, ce travail introduisit des types de conditions ignorés jusqu'à présent : l'arrêt d'un médicament, et des facteurs liés au déroulement du séjour (exemple : antécédent d'INR trop bas). La plupart des règles produites ici rentrèrent dans la catégorie « AA5 – Advanced alerts – Complex prescribing alerts » définie par Honigman [118].

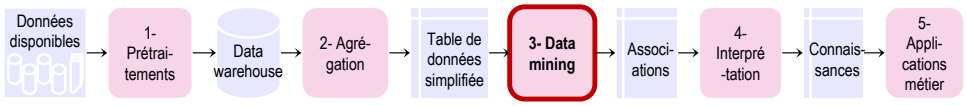
On peut considérer que le formalisme relativement abstrait des événements permettrait d'insérer d'autres types de données. Si par exemple des résultats d'ECG étaient disponibles, la méthode employée ici n'aurait aucune difficulté à les intégrer. Pour ce qui concerne la méthode de production des règles, présentement le data mining, il n'existait pas dans la littérature d'exemple similaire. Le data mining est fréquemment employé sur les rapports d'EIM, mais jamais sur les dossiers électroniques eux-mêmes. Par cette approche nous évitâmes le biais lié à la forte sous-déclaration des EIM.

Cette méthode d'induction de connaissances soufre néanmoins de faiblesses qui sont celles, plus généralement, de la réutilisation de données. Tout d'abord, il n'est pas possible d'identifier des situations qui ne sont pas observables dans les données (exemple : contrindications absolues correctement respectées). De plus, il n'est pas possible de détecter et d'exploiter des événements qui ne sont pas explicitement décrits dans les données, comme par exemple certains événements cliniques (nausées, rash, etc.). Enfin, la qualité des données est parfois insuffisante, et on

espère, sans pouvoir le vérifier, que le biais de sous-codage soit aussi peu différentiel que possible.

Enfin, même si des contraintes temporelles purent être introduites dans la phase d'induction de règles, cette approche souffrit d'une prise en compte incomplète du temps. Pour y remédier, nous prolongeâmes ces travaux d'un point de vue applicatif lors de ma mobilité à Boston, et d'un point de vue méthodologique lors de ma mobilité à Inria. Les travaux en découlant seront présentés par la suite.

2.7 Fouille visuelle de données

Position :	
Publication :	*Chazard E, Beuscart R. Graphical representation of the comprehensive patient flow through the hospital. AMIA Annu Symp Proc 2007:110–4.* *[texte intégral en ligne]* *[18]* *Chazard E, Puech P, Gregoire M, Beuscart R. Using Treemaps to represent medical data. Stud Health Technol Inform 2006;124:522–7.* *[texte intégral en ligne]* *[119]*

Revenons aux méthodes de fouille visuelle de données, comme un préalable à la fouille statistique.

Les méthodes de data mining permettent de découvrir des associations dans une table « à plat », c'est-à-dire représentant chaque individu statistique sous la forme d'une ligne, et chaque variable sous la forme d'une colonne. Nous avons vu précédemment (voir section 2.4 page 40) que ce type de table était produite par l'agrégation de données, et que cette agrégation était réalisée à dire d'expert. Le problème est qu'il existe de nombreuses situations où, en réalisant cette agrégation, l'expert ne « pense pas » à générer telle ou telle variable, notamment parce que les données initiales sont trop complexes. L'objectif de la fouille visuelle de données est dans certains cas d'**explorer les données avant leur agrégation**, de manière à **inspirer une agrégation** la plus pertinente possible.

Nous nous intéressâmes tout d'abord aux Treemaps [119,120], essentiellement utilisés à l'époque pour représenter l'allocation des volumes (en octets) dans un disque dur. Nous n'inventâmes pas cette méthode, néanmoins nous en développâmes une implémentation libre à l'époque où il n'en n'existait pas, et nous améliorâmes notamment la représentation du texte dans les rectangles. Nous développâmes ce programme en PHP 5 [121], avec une sortie au format SVG [122,123], rendue interactive par des effets de *mouseOver* programmés en *actionScript* [122], qui permettaient d'obtenir des informations supplémentaires. Nous implémentâmes l'algorithme *squarrified* [124], qui permet par simulations itératives de trouver une disposition dans laquelle le ratio grande longueur / petite longueur des rectangles est le plus faible possible. Nous développâmes également des échelles de couleur permettant la transformation d'un nombre en code couleur non-ambigu, et imprimable en niveaux de gris [119]. Nous pûmes discuter l'intérêt de cette méthode comparé à différentes approches traditionnelles dans le cas de l'analyse d'activité PMSI par exemple (voir Figure 31).

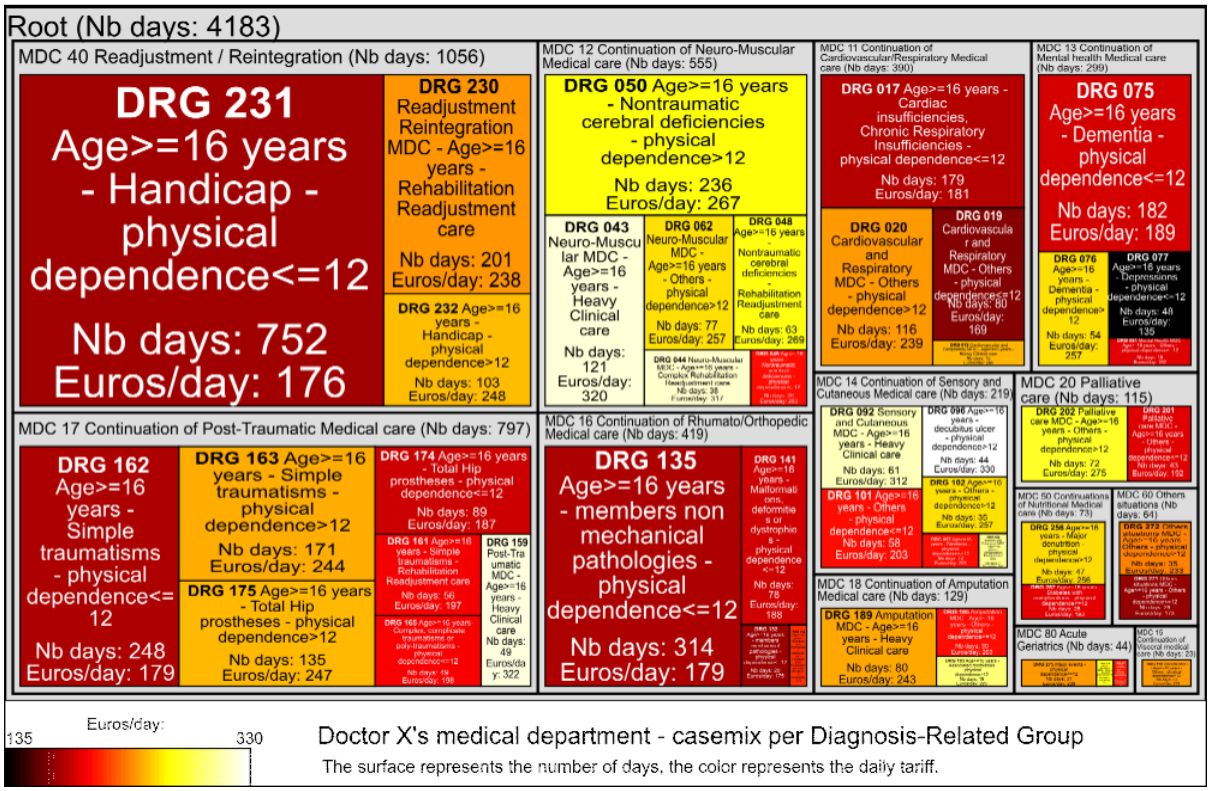


Figure 31. Exemple de Treemap représentant l'activité d'un centre hospitalier de moyen séjour. La taille de chaque rectangle représente le nombre de journées. La couleur représente le tarif journalier. La hiérarchie représente le groupage SSR des séjours[1].

Nous nous intéressâmes ensuite au cas plus complexe du parcours intra-hospitalier d'un patient, mélangeant des états qualitatifs et des transitions, tenant compte de la notion de temps. Nous développâmes une représentation adaptée à ce contexte [18]. Il n'existe pas à notre connaissance d'autre solution à ce jour (voir Figure 32). Nous développâmes là aussi ce programme en PHP 5, SVG et *actionScript*. Une fois qu'un tel graphique permet de faire un diagnostic humain, il est toujours possible de catégoriser les séjours plus simplement à l'aide de règles expertes simples, puis d'appliquer alors des méthodes traditionnelles sur les données simplifiées.

[1] A l'époque le groupage SSR classait les journées d'hospitalisation en CMC (catégories majeures cliniques) et GHJ (groupes homogènes de journées).

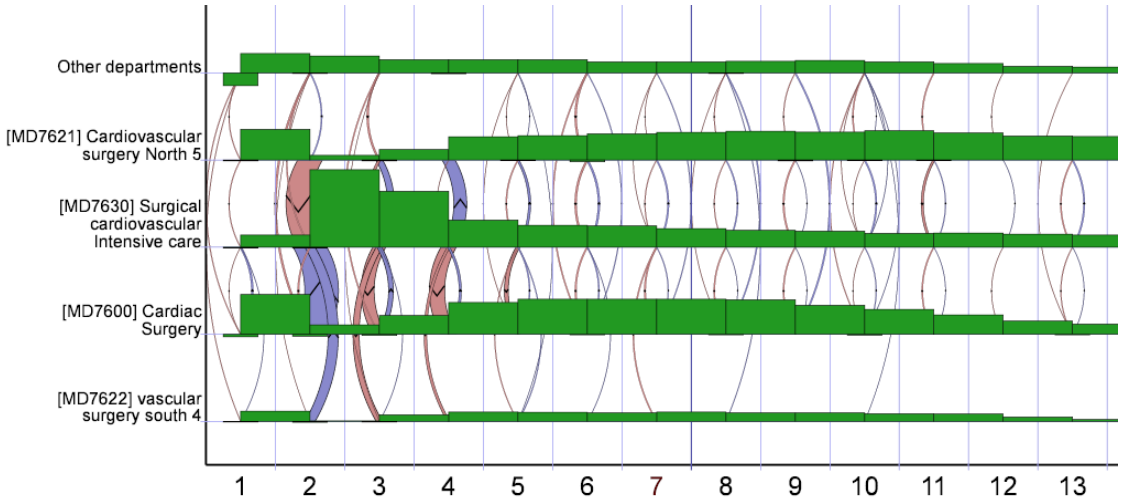


Figure 32. Représentation exhaustive des séjours passant par l'unité MD7630 (X=temps, Y=lieux). La plupart proviennent des unités 7621, 7600 et 7622 le lendemain de leur admission, restent quelques jours et retournent progressivement dans leur unité d'origine pour une longue durée (graphique tronqué).

2.8 Problèmes méthodologiques

L'utilisation de données réelles amène à rencontrer des situations dans lesquelles certains tests statistiques ne sont plus valides, ou fournissent un résultat qui pourrait être mal interprété. La diversité des données réutilisable entraîne également une certaine diversité dans les problèmes méthodologiques à explorer.

La durée de séjour est importante en santé publique, et ce à deux égards.

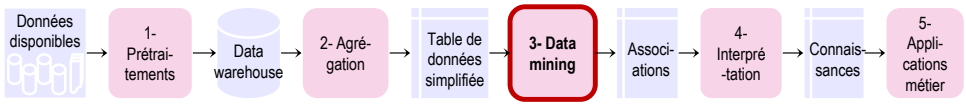
Tout d'abord, un événement indésirable est susceptible d'allonger la durée de séjour. Cette variable est alors un *outcome*, au même titre que le décès, le passage en réanimation, etc. La durée de séjour présente néanmoins l'avantage d'être quantitative alors que la plupart des variables d'*outcome* sont binaires avec une modalité rare (ex : le décès). L'utilisation de la durée de séjour permet donc d'obtenir une puissance statistique supérieure.

Ensuite, une durée de séjour longue peut à son tour être un facteur de risque d'événement, comme par exemple les infections nosocomiales (respiratoires, urinaires), les thromboses (phlébite, embolie pulmonaire), ou les escarres.

Autour de cette variable, nous présenterons ici deux problèmes en particulier :

- la validité des tests statistiques pour comparer des durées de séjour, du fait des particularités de distribution de cette variable
- la difficulté d'évaluer l'augmentation de durée de séjour imputable à un événement, si cet événement est lui-même temps-dépendant.

2.8.1 Comparer des durées de séjour

Position :	
Publication :	*Chazard E, Ficheur G, Beuscart J-B, Preda C. How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests. Value in Health 2017. doi:10.1016/j.jval.2017.02.009.* [125]

Le cadre d'étude est le suivant : on dispose de deux groupes de patients (par exemple les patients exposés ou non à un facteur à l'admission), on mesure la durée de séjour (notée *LOS* pour *Length Of Stay*) de chaque patient, et on souhaite savoir si la LOS moyenne du premier groupe est significativement différente de celle du deuxième groupe. Cette situation, très fréquente, ne fit pas l'objet de publication méthodologique auparavant. La popularité respective de la durée de séjour, des risques alpha et bêta, et de ces risques dans le domaine des durées de séjour, est illustrée en Table 16. Nous décidâmes d'investiguer cette question [125].

Table 16. Nombre d'articles référencés Pubmed de janvier 2006 à décembre 2015

Concept	Citation en titre	Citation en titre ou abstract
Durée de séjour (LOS) [a]	1 704	22 548
Erreur de type 1 [b]	100	2 180
Erreur de type 2 [c]	223	4 925
Erreur de type 1 [b] **et** LOS [a]	0	4
Erreur de type 2 [c] **et** LOS [a]	0	12

[a] Recherche : "length of stay"
[b] Recherche : "type 1 error", "type I error", "first type error", or "alpha risk"
[c] Recherche : "type 2 error", "type II error", "second type error", "beta risk", or "statistical power"

Nous comparâmes donc les risques alpha et bêta de différentes méthodes statistiques fréquemment utilisées pour comparer les durées de séjour de deux groupes de patients, les durées de séjour ayant une distribution très particulière. Au terme d'une revue bibliographique, nous identifiâmes 12 méthodes statistiques fréquemment utilisés ou recommandés par les auteurs :

- Des tests statistiques paramétriques :
  - Le test de Student
  - Le test de Student après transformation logarithmique
  - Le test de Student réalisé sur les rangs des individus dans la variable
- Des tests statistiques non-paramétriques :
  - Le test de Wilcoxon non apparié, ou Mann-Whitney
  - Le test de Kruskal-Wallis
- Des méthodes basées sur des modèles de régression :
  - La régression linéaire avec lien logarithmique
  - La régression gamma avec lien logarithmique
  - La régression quantile utilisant la médiane
  - La régression de Poisson
- Des méthodes permettant l'analyse de survie (l'événement étant ici la sortie) :

- Le test du Log Rank
- Le modèle de Cox (risques proportionnels)
- La survie de paramétrique de Weibull

Nous étudiâmes tout d'abord le risque de première espèce. En utilisant ces méthodes avec un seuil de décision de 5% appliqué à la p valeur, on s'attendait à ce que le risque de première espèce fût de 5%. Afin de le vérifier empiriquement, nous utilisâmes la base nationale des séjours du PMSI comme loi de probabilité discrète des durées de séjour. A l'aide de 19 millions de simulations concernant différentes tailles d'échantillons, nous évaluâmes le risque de première espèce sous l'hypothèse nulle (les deux échantillons étant simulés à l'aide de la même loi de probabilité). Nous montrâmes ainsi que la régression gamma avec lien logarithmique, la régression médiane, la régression de Poisson, le test du Log Rank et l'analyse de survie de Weibull présentaient une élévation inacceptable du risque de première espèce (voir Figure 33).

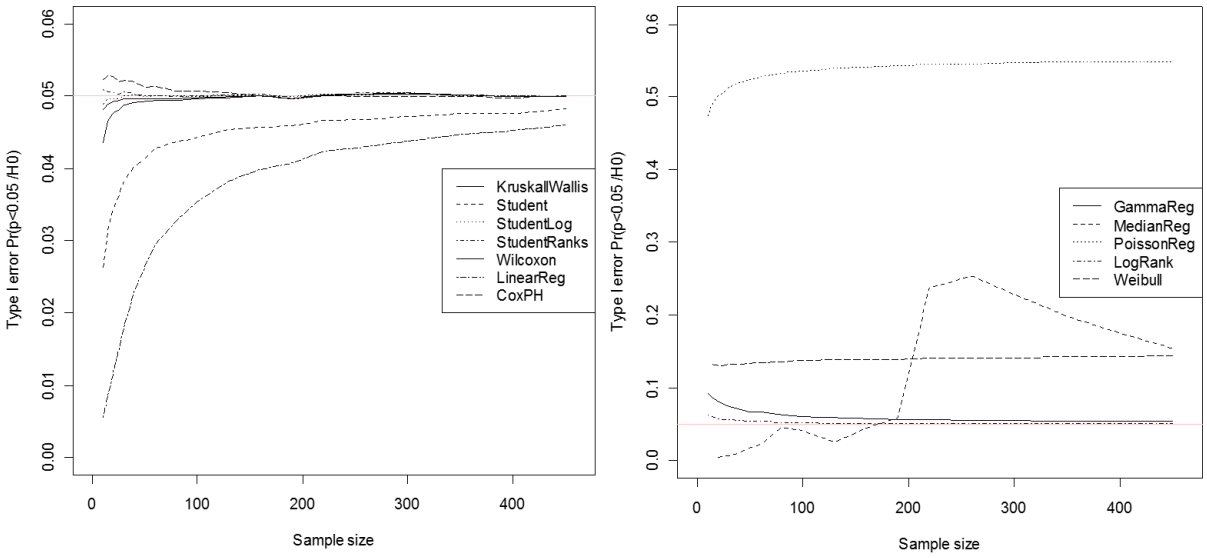


Figure 33. Risque de première espèce en fonction de la taille d'échantillon.
A gauche : 7 méthodes avec risque de première espèce acceptable (ou plus bas qu'attendu).
A droite : 5 méthodes avec élévation inacceptable du risque de première espèce.

Dans un deuxième temps, nous évaluâmes la puissance des sept méthodes dont le risque de première espèce restait acceptable, à l'aide de 3 scénarios (hypothèses alternatives) et 5,7 millions de simulations. Nous observâmes dans les 3 hypothèses alternatives testées que 4 tests statistiques avaient une puissance élevée (et très proche) : il s'agit des tests de Kruskall Wallis, de Wilcoxon-Mann-Whitney (et ce en dépit des nombreux ex-aequo) et de Student réalisé sur les rangs ou sur la variable `log(durée+1)` (voir Figure 34). En revanche, le test de Student (sur données natives), la régression linéaire avec lien logarithmique ou le modèle de Cox, pourtant fréquemment utilisés, semblèrent être peu puissants.

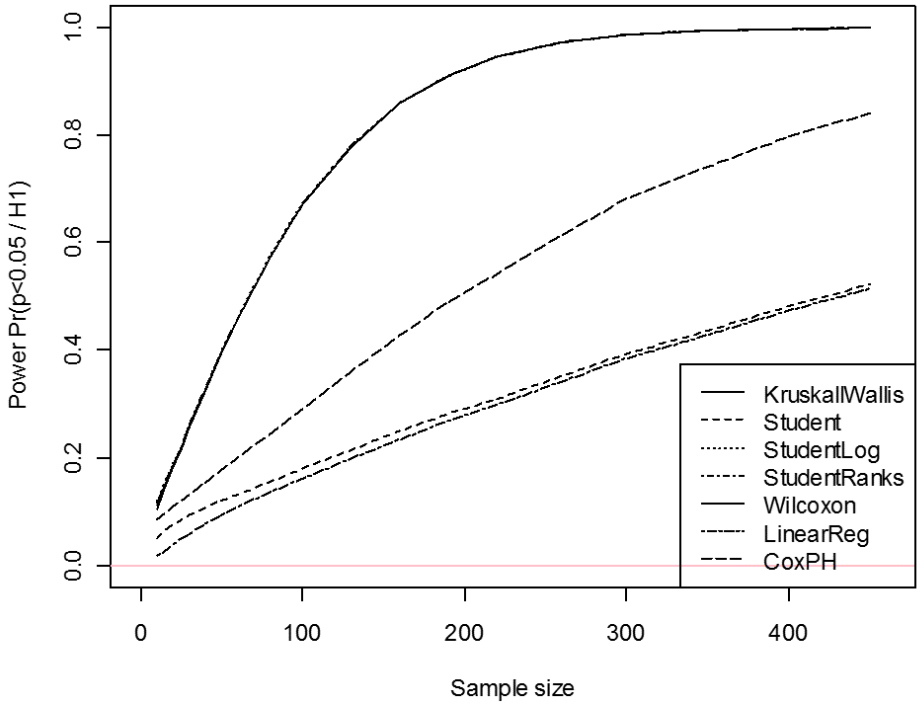
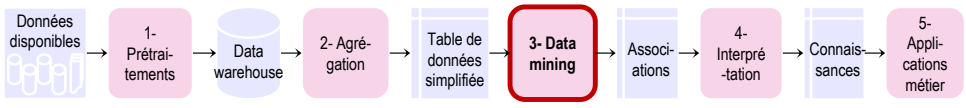


Figure 34. Puissance en fonction de la taille d'échantillon, pour les 7 tests avec risque de première espèce acceptable. Exemple d'une des trois hypothèses alternatives.

Ce travail nous permit de conseiller l'usage des tests suivants pour comparer les LOS de deux groupes de patients : le test de Wilcoxon, le test de Kruskall-Wallis ou le test de Student sur les rangs ou sur la transformée logarithmique. Nous conclûmes que les tests recommandés par certains auteurs sur le seul argument de la puissance devaient être évités car leur apparente puissance était notamment due à une inflation du risque de première espèce, ce qu'ils n'avaient pas vérifié.

2.8.2 Quelle augmentation de durée de séjours est imputable à un événement temps-dépendant ?

Position :	
Publication :	*Chazard E, Dumesnil C, Beuscart R. How much does hyperkalemia lengthen inpatient stays? About methodological issues in analyzing time-dependant events. Stud Health Technol Inform 2015;210:835–9. [texte intégral en ligne] [126]*

Nous nous demandâmes ensuite comment comparer les durées de séjour d'un groupe de patient présentant un événement avec un groupe de patient ne le présentant pas, lorsque cet événement était lui-même temps-dépendant.

De nombreux travaux, scientifiques ou administratifs, estiment le rallongement de durée de séjour imputable à la survenue d'un événement en calculant simplement la différence de durée de séjour moyenne entre les deux groupes, et en testant la nullité de cette différence en population. Or, si un événement est temps-dépendant, le simple fait qu'un patient reste hospitalisé longtemps augmente la probabilité qu'il

subisse l'événement et peut faire conclure à tort que c'est l'événement qui a rallongé le séjour.

Afin d'explorer ce problème, nous nous intéressâmes aux hyperkaliémies dans un hôpital périphérique français. Une hyperkaliémie était présente dans 2,2% des 13 741 mesures de potassium sanguin, soit 2,0% des 6 678 séjours de patients. La différence brute des durées moyennes de séjour était de 8,78 jours, ce qui aurait été interprété dans beaucoup d'études comme « les hyperkaliémies augmentent les durées de séjour de 8,78 jours en moyenne ». Nous testâmes 17 méthodes plus complexes. Nous obtînmes des résultats différents, en tout cas en faveur de l'idée que la différence réelle était nettement plus modérée : selon les méthodes employées, il était vraisemblable que l'augmentation de durée de séjour réellement causée par l'hyperkaliémie ne fût que la moitié de la différence brute observée. La Table 17 présente les résultats obtenus en fonction des 17 méthodes testées.

Table 17. Augmentation de la durée moyenne de séjour imputable à l'hyperkaliémie : résultats obtenus par 17 méthodes différentes

Méthode		Différence de LOS (jours)	Idem après exclusion des outliers (jours)
Comparaison brute		8.78	
Simulation		2.28	
Régression linéaire	Basique	6.01	4.45
	Avec interaction	4.19	4.35
	Loglinéaire	4.45	4.60
Modèle surajusté	Basique	5.74	4.48
	Avec interaction	4.58	4.61
	Loglinéaire	3.83	4.38
Analyse appariée	Basique	6.01	
	Avec interaction	5.98	
	Loglinéaire	5.43	

En résumé, alors que la différence brute de LOS observée entre les patients était de 8,78 jours, l'hypothèse la plus pessimiste ne retrouva qu'une différence de 2,28 jours réellement imputable à l'hyperkaliémie, tandis que la plupart des autres méthodes trouvaient une différence imputable à l'hyperkaliémie se situant aux alentours de 4,5 jours, soit la moitié de la différence brute observée.

Au vu de ces résultats, nous nous inquiétons de certaines conclusions couramment publiées et appuyant des décisions politiques. Ainsi par exemple, la DREES a publié les augmentations de durée de séjour imputables à certains événements indésirables liés aux soins (voir Table 18) [127]. Cette augmentation est estimée en mesurant la différence des durées de séjour, en appariant sur le GHM (autrement dit, essentiellement le motif d'entrée et l'acte chirurgical), l'âge, le sexe et le statut juridique de l'établissement. Les autres facteurs prédisant la durée de séjour et le caractère temps-dépendant de l'événement n'ont pas été pris en compte. Lorsqu'on trie ces événements par valeur décroissante de la durée supplémentaire imputable (voir Table 18), on observe que les cinq premiers sont des événements temps-dépendants, et que les trois derniers sont des événements ponctuels qui se produisent essentiellement en début de séjour, donc dont la probabilité cumulée n'augmente pas avec la durée de séjour (le PSI 15 comporte surtout des blessures

durant le temps opératoire). Forts du travail réalisé avec l'exemple de l'hyperkaliémie, qui rentre dans la catégorie « PSI 10 » de la Table 18, il nous semble que la durée imputable à ces 5 premiers événements est surestimée : c'est sans doute autant la durée de séjour élevée qui a entraîné l'événement, que l'événement qui a rallongé la durée de séjour. Ce biais méthodologique a un effet similaire sur le surcoût estimé, qui est essentiellement lié à la durée de séjour.

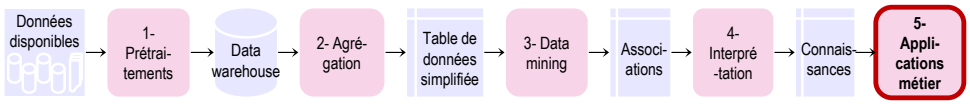
Table 18. Evénements indésirables augmentant les durées de séjour selon la DREES [127], triés par ordre décroissant (dernière colonne : note personnelle).

Evénement indésirable lié aux soins	Augmentation de LOS (j)	Surcoût (€)	Notre interprétation
PSI 13 Septicémie	19.7	23 132	Evénements temps-dépendants
PSI 7 Infection	14.7	10 942	
PSI 3 Escarre	11.2	5 769	
PSI 10 Désordre physiologique	7.3	11 948	
PSI 12 Embolie pulmonaire	5.0	4 933	
PSI 5 Oubli d'un corps étranger	2.5	4 867	Accidents ponctuels
PSI 15 Lacération ou piqûre accidentelle	1.2	2 229	
PSI 18/19 Traumatisme obstétrical	0.7	505	

Une interprétation directe de la différence brute des durées de séjour pourraient amener à mal trier les priorités d'action publique ou de prévention. Par conséquent, une intervention correctement menée sur un événement temps-dépendant aurait sans doute un effet moindre qu'escompté. Ainsi, les ressources risquent d'être allouées de manière peu efficace.

3 Principaux travaux liés aux applications

3.1 Détection automatisée des effets indésirables des médicaments

Position :	
Publication :	*Ficheur G, Chazard E, Beuscart J-B, Merlin B, Luyckx M, Beuscart R. Adverse drug events with hyperkalaemia during inpatient stays: evaluation of an automated method for retrospective detection in hospital databases. BMC Med Inform Decis Mak 2014;14:83. doi:10.1186/1472-6947-14-83. [texte intégral en ligne][128]* *Hackl WO, Ammenwerth E, Marcilly R, Chazard E, Luyckx M, Leurs P, et al. Clinical evaluation of the ADE scorecards as a decision support tool for adverse drug event analysis and medication safety management. Br J Clin Pharmacol 2013;76 Suppl 1:78–90. doi:10.1111/bcp.12185. [texte intégral en ligne][129]* *Koutkias V, Kilintzis V, Stalidis G, Lazou K, Collyda C, Chazard E, et al. Constructing Clinical Decision Support Systems for Adverse Drug Event Prevention: A Knowledge-based Approach. AMIA Annu Symp Proc 2010;2010:402–6. [texte intégral en ligne][130]* *Marcilly R, Chazard E, Beuscart-Zéphir M-C, Hackl W, Băceanu A, Kushniruk A, et al. Design of Adverse Drug Events-Scorecards. Stud Health Technol Inform 2011;164:377–81. [131]* *Ferret L, Luyckx M, Ficheur G, Chazard E, Beuscart R. Evaluation of a Computer Application for Retrospective Detection of Vitamin K Antagonist Treatment Imbalance. J Patient Saf 2016. doi:10.1097/PTS.0000000000000182. [132]* *Ferret L, Luyckx M, Merlin B, Ficheur G, Chazard E, Beuscart R. Evaluation of a computerized tool allowing retrospective detection of potential vitamin K antagonist overdoses in complex contexts. Stud Health Technol Inform 2013;192:553–6. [133]* *Koutkias VG, McNair P, Kilintzis V, Skovhus Andersen K, Niès J, Sarfati J-C, et al. From adverse drug event detection to prevention. A novel clinical decision support framework for medication safety. Methods Inf Med 2014;53:482–92. doi:10.3414/ME14-01-0027. [134]* *Perichon R, Chazard E, Beuscart R. Patients drug exchange forum corpus: toward drug safety signals detection. Stud Health Technol Inform 2015;210:1023. [135]* *Chazard E, Luyckx M, Beuscart J-B, Ferret L, Beuscart R. Routine use of the "ADE scorecards", an application for automated ADE detection in a general hospital. Stud Health Technol Inform 2013;192:308–12. [texte intégral en ligne][104]* *Chazard E, Băceanu A, Ferret L, Ficheur G. The ADE scorecards: a tool for adverse drug event detection in electronic health records. Stud Health Technol Inform 2011;166:169–79. [texte intégral en ligne][106]* *Băceanu A, Atasiei I, Chazard E, Leroy N, PSIP Consortium. The expert explorer: a tool for hospital data visualization and adverse drug event rules validation. Stud Health Technol Inform 2009;148:85–94. [texte intégral en ligne][136]* *Leroy N, Chazard E, Beuscart R, Beuscart-Zephir MC, Psip Consortium. Toward automatic detection and prevention of adverse drug events. Stud Health Technol Inform 2009;143:30–5. [texte intégral en ligne][137]*

Mon travail de recherche sur les effets indésirables liés aux médicaments (EIM) fut initialement mené dans le cadre du projet PSIP (Patient Safety through Intelligent Procedures in medication). Ce projet fut financé par la Commission Européenne dans le cadre du 7° programme cadre (FP7/2007-2013, *grant agreement* n°216130), dont le CHU de Lille était le coordonnateur. Ce projet avait pour objectifs de générer de la connaissance sur les EIM notamment par data mining, et de mettre en place des outils novateurs de prévention de ces EIM.

Selon un rapport [138], on estime que 2% des hospitalisations aux USA engendrent des effets indésirables médicamenteux, et que la moitié d'entre eux aurait pu être évitée. Cette moitié entre alors dans la définition des « erreurs médicales ». Toujours aux USA, on estime que les EIM sont responsables de 98 000 décès par an et que les coûts directs imputables à ces EIM sont de 10 Milliards de dollars. En France, la situation est comparable. On estime que 1,3 millions de patients par an souffrent d'un EIM, que 10 000 en meurent et que 35 000 survivent avec des séquelles chaque année.

La définition des EIM n'est pas évidente. En langue anglaise, les définitions ci-dessous font maintenant l'objet d'un consensus entre chercheurs. L'OMS et l'Union Européenne partagent la même définition des ADR (Adverse Drug Reactions) [139] : un ADR est une réaction à un médicament qui est nocive et inattendue, et survient à des doses habituellement utilisées chez l'homme. L'Institute Of Medicine [140,141] définit les ADE (Adverse Drug Events) comme une nuisance résultant de l'utilisation d'un médicament [142,143]. Finalement, d'après Nebeker [144], les ADE comprennent les effets des médicaments (ADR et surdosages) et les nuisances résultant de l'utilisation du médicament, incluant les diminutions de doses et l'arrêt du médicament [145], ce qui n'est pas le cas des ADR. L'Institute Of Medicine [140] précise également que les ADE sont des nuisances « plus liées au traitement qu'aux maladies du patient », intégrant ainsi la part que joue la maladie elle-même dans un effet indésirable. D'après ces définitions, les chercheurs distinguent également les *preventable ADEs* et les *non-preventable ADEs*. Les *preventable ADEs* sont assimilés à des erreurs médicales [146]. Enfin, il faut noter que les erreurs médicales sans effet clinique sur le patient ne sont pas des ADE. A cet égard, la définition de l'ADE recouvre mal le problème du *side effect*, ce dernier pouvant par exemple inclure une réaction biologique indésirable mais sans effet clinique (élévation modérée des transaminases hépatiques) ou un effet latéral bénéfique (perte d'appétence pour le tabac, perte de poids, repousse des cheveux, etc.).

Dans le texte qui suit, le terme français d'EIM sera utilisé pour désigner les ADE.

Les méthodes les plus usuelles de détection des cas d'EIM reposent sur la déclaration volontaire, or il est prouvé que cette déclaration concerne moins de 5% des EIM [112], et la revue de dossiers, efficace mais très chronophage, et ce d'autant plus que les EIM sont relativement rares. Des méthodes de *natural language processing* (NLP) ou traitement automatisé du langage (TAL) sont appliquées aux courriers de sortie, ce qui suppose néanmoins que l'EIM soit cité dans le courrier [147]. Enfin, lorsque des méthodes de data mining sont utilisées, c'est uniquement pour analyser des déclarations d'EIM [148,149], ce qui suppose que les EIM soient déclarés, ce qui est rarement le cas.

La connaissance sur les EIM est essentiellement académique. Les prescripteurs trouvent habituellement les informations nécessaires dans les résumés des caractéristiques du produit (RCP) disponibles pour chaque médicament, et

maintenus en France par l'AFSSAPS. Cependant, ces informations sont en grande quantité et submergent le lecteur. A titre d'exemple, le RCP du Previscan comporte 3 300 mots. L'information est classée par importance théorique, or les EIM surviennent avec des probabilités différentes et souvent inattendues, en fonction des connaissances des prescripteurs et de leurs pratiques de surveillance du risque [100].

Comme indiqué précédemment, nous utilisâmes des méthodes de data mining (voir section 2.5 page 54) associées à un prétraitement d'agrégation de données (voir section 2.4 page 40) et un post-traitement de filtrage, interprétation médicale et réorganisation (voir section 2.6 page 57), ce qui nous permit de générer des règles simples de détection de cas d'EIM.

Ces règles de détection des EIM furent ensuite utilisées dans les *ADE Scorecards* [106]. Cet outil innovant est capable d'analyser automatiquement des données d'hospitalisations passées, et de :

- détecter automatiquement les séjours passés suspects d'EIM
- présenter des statistiques globales sur les EIM détectés dans un service [104]
- présenter aux médecins des statistiques sur les situations dans lesquelles les EIM se sont déclenchés
- permettre aux médecins de revoir s'ils le souhaitent les cas réels détectés par le système dans leur service (voir Figure 35) [136]

La page d'accueil du logiciel présente les types d'EIM que l'outil détecte et, pour chaque type, le nombre de cas détectés mois après mois. On peut ensuite accéder à une page spécifique pour chaque EIM. Chaque page spécifique par EIM type (par exemple les hyperkaliémies) permet notamment d'accéder à chacun des cas d'EIM détectés. Un outil de visualisation, dont l'écran synthétique est présenté en Figure 35, montre les données disponibles sur ce séjour, incluant les informations du PMSI, les médicaments, les résultats de laboratoire et les courriers en texte libre. Les éléments qui selon le système expliquent l'EIM apparaissent automatiquement sur fond coloré, afin d'accélérer la revue des cas.

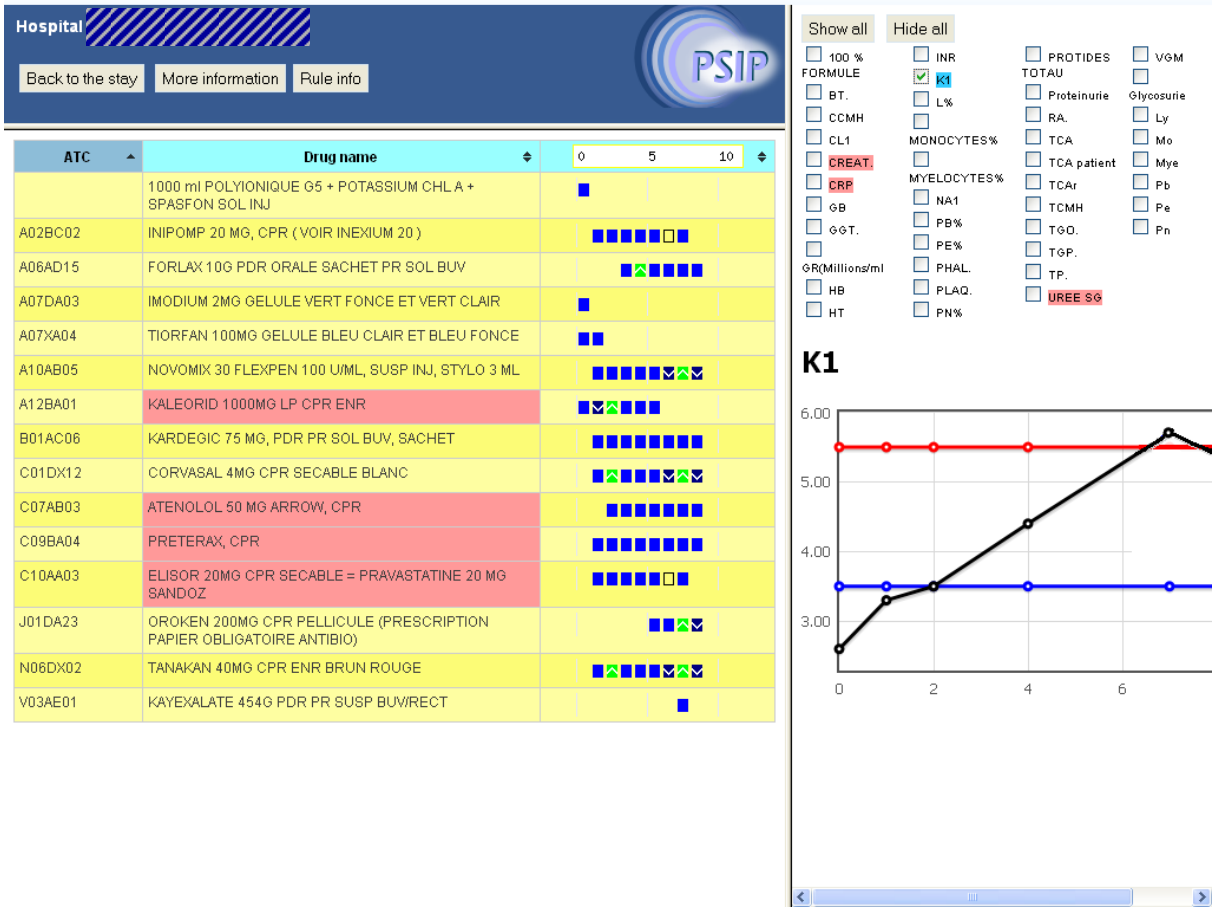


Figure 35. Revue détaillée d'un cas détecté.

Nous évaluâmes l'aptitude de ces applications à détecter des cas réels d'EIM, c'est-à-dire des cas présentant simultanément toutes les caractéristiques suivantes :

- un incident est observable (et il ne s'agit pas d'un artéfact, par exemple pas d'une fausse hyperkaliémie)
- cet incident est principalement expliqué par les médicaments, et non par la pathologie du patient
- les causes détectées par le logiciel, médicamenteuses ou non, prennent une part majoritaire dans l'explication de cet incident

Dans le cas par exemple des ADE se traduisant par une hyperkaliémie au-delà de 5,5 mmol/l, nous obtînmes les résultats suivants [15] :

- Nombre de séjours analysés : 14 747
- Nombre de séjours avec hyperkaliémie : 117 (0,793%)
- Rappel (sensibilité) : 39/41=95,1%
- Précision (valeur prédictive positive) : 39/75=52,0%
- F-mesure : 67,2%
- Nombre de cas déclarés en pharmacovigiliance : 0/41=0%
- Cas au-delà de 6 mmol/l : 11/41=26,8%
- Administration de kayexalate : 12/41=29,3%

De plus, une évaluation clinique et qualitative de l'outil ramena d'excellents résultats [129], comme rapporté en Figure 36.

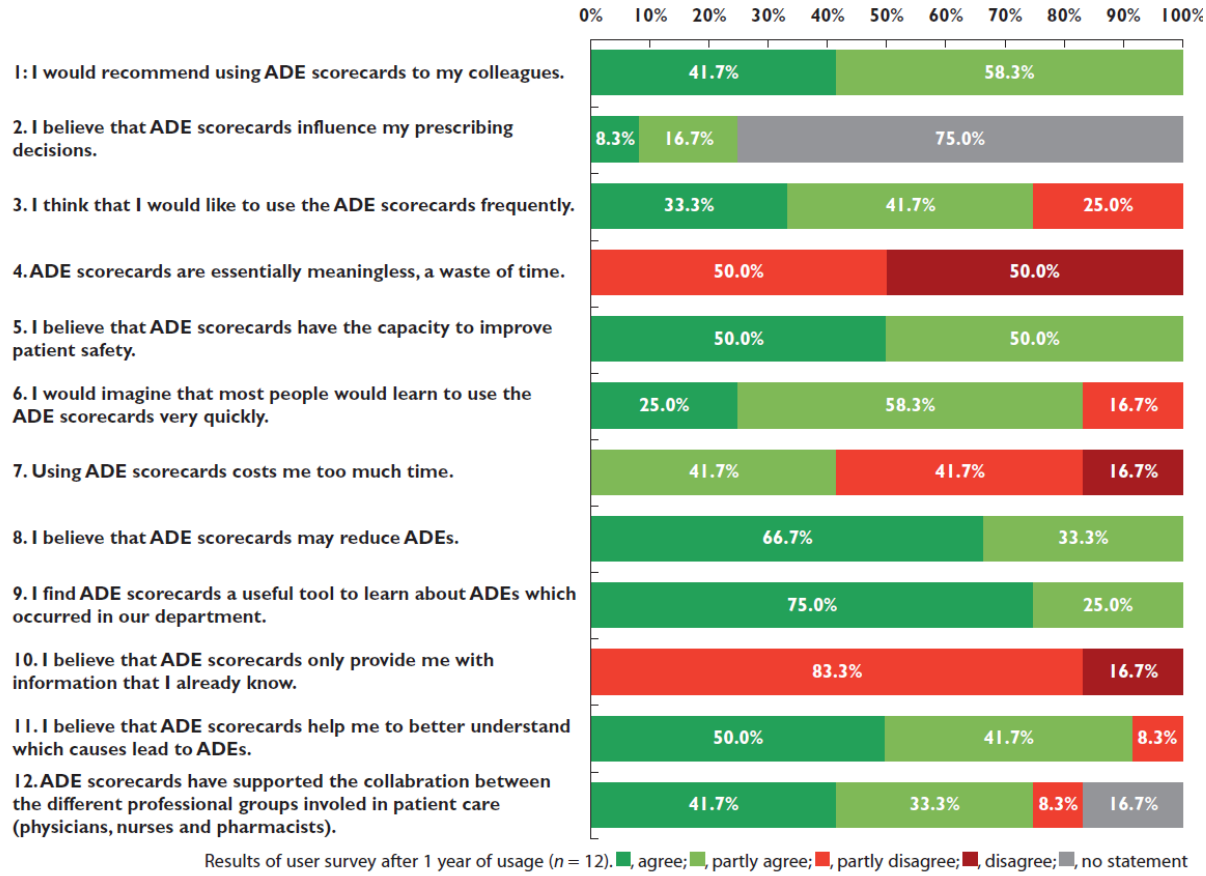
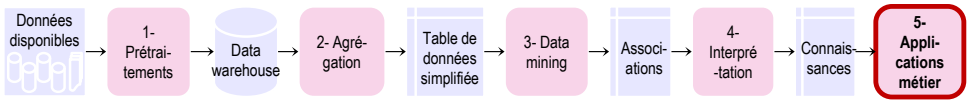


Figure 36. Evaluation qualitative des ADE Scorecards, d'après [129].

3.2 Estimation automatisée du risque d'effets indésirables des médicaments

Position :	
Publication :	*Chazard E, Beeler PE, Ficheur G, Dalleur O, Beuscart R, Bates DW. Drug-drug interactions with Vitamin K antagonists: are we focusing on the right rules? J Am Med Inform Assoc [IN PRESS] 2017.* [150]

Parmi les 236 règles de détection des EIM produites dans le cadre de la recherche présentée à la section précédente, 72% apportèrent de nouvelles connaissances, comme mentionné préalablement et notamment en Table 14. Réciproquement, nous dûmes inclure manuellement 40 règles déjà connues et qui paraissaient essentielles aux pharmacologues (par exemple aspirine & héparine → hémorragie), mais nous fûmes surpris de voir que ces règles, pourtant valides par définition, ne détectaient pas ou peu de séjours. Nous formulâmes ainsi l'hypothèse suivant :

> *Les règles de connaissance sur les EIM sont découvertes dans des contextes relativement neutres : phases de développement du médicament et phase précoce post-AMM.*
> *En vie réelle, inversement, la pratique de soins tient compte des EIM attendus, et essaie de les prévenir. Elle pourrait de ce fait faire baisser la fréquence des EIM attendus, sans affecter pour autant la fréquence des EIM peu connus ou inattendus.*

Cette hypothèse concorde avec le constat dressé par la communauté scientifique sur les effets de l'introduction des systèmes d'aide à la décision clinique (CDSS). Pour les CDSS tels qu'ils sont conçus aujourd'hui, l'*over alerting* (excès de bruit, trop faible valeur prédictive positive) représente un écueil majeur : des alertes trop nombreuses et inappropriées interrompent sans cesse le travail du clinicien. Elles peuvent induire une *alert fatigue*, c'est-à-dire un état de fatigue psychique et d'insensibilité aux messages, même pertinents [9,151,152]. Cette *alert fatigue* pourrait être responsable de la faible efficacité clinique des CDSS en termes de réduction de morbi-mortalité [152,153]. En effet, jusqu'à 96% des alertes des CDSS sont annulées par les prescripteurs de médicaments [151,154–158], essentiellement car elles ne sont pas appropriées [159,160], c'est-à-dire par exemple qu'elle décrivent un risque théorique mais dont le médecin est déjà parfaitement au courant, ou que ces alertes ne devraient pas surgir compte tenu du contexte précis du patient traité. Malheureusement, ces annulations d'alertes sont également souvent inappropriées, et certaines d'entre elles sont suivies de véritables EIM [161–164]. En effet, certaines alertes appropriées peuvent être considérées à tort comme inappropriées par les prescripteurs [151], notamment du fait même de leur nombre élevé et de l'*alert fatigue* qu'elles induisent.

Nous nous intéressâmes donc à la réduction de l'over-alerting. Pour ce faire, nous examinâmes les interactions médicamenteuses publiées par Holbrook et al. dans une revue de la littérature qui fait référence [165]. Nous les formalisâmes sous la forme *médicamentA & médicamentB → élévation_INR*, ou sous la forme *médicamentA & médicamentB → abaissement_INR*. Tirant profit des données recueillies dans le cadre du projet européen PSIP, nous évaluâmes le risque lié à chaque règle de manière empirique à l'aide d'un modèle de Cox à covariables temps-dépendantes, afin de proposer une « alerte graduée » utilisant des moyens plus ou moins interruptifs en fonction de la probabilité de survenue d'un EIM. Ceci permit, pour chaque médicament, de calculer la modification du risque d'observer un surdosage ou un sous-dosage en AVK (voir Figure 37) [166]. Nous observâmes une élévation du risque de surdosage en AVK pour le fénofibrate (*hazard ratio*=3,09), la méthylprednisolone (*hazard ratio*=3,02), et la simvastatine (*hazard ratio*=2,52). Nous observâmes un sur-risque de sous-dosage en AVK pour la dicloxacilline (*hazard ratio*=4.58) et le sucralfate (*hazard ratio*=9,0). Pour les autres médicaments, nous n'obtînmes pas de *hazard ratio* significativement différent de 1, y compris pour des médicaments très fréquemment prescrits.

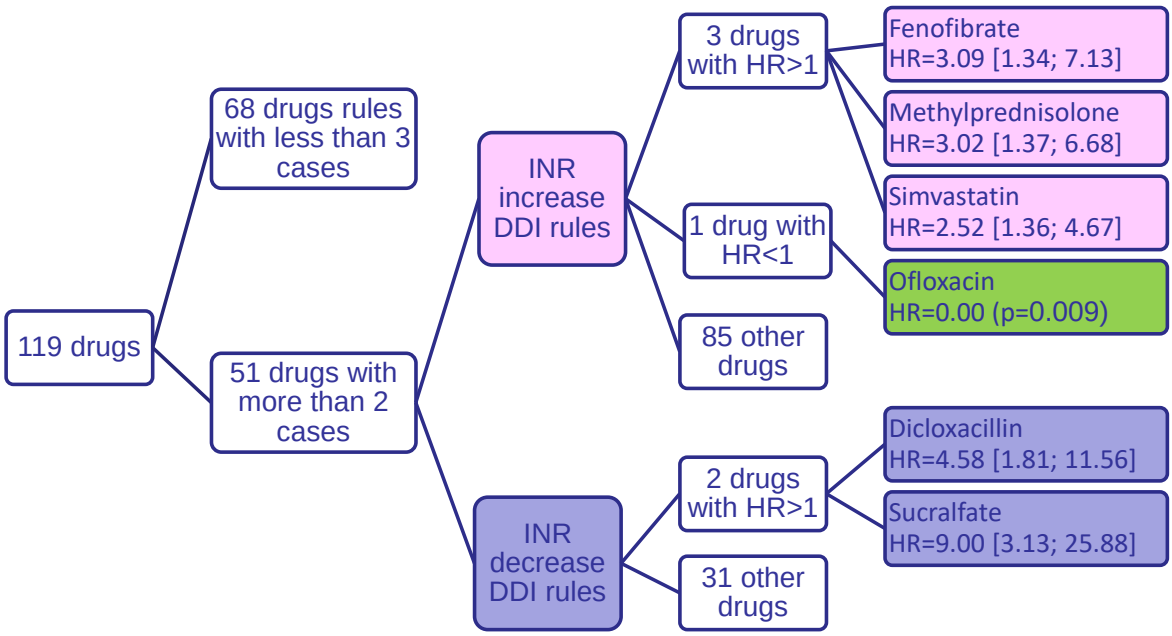


Figure 37. Evolution de l'INR en vie réelle pour 119 médicaments associés avec un AVK. INR increase : INR>5 ; INR decrease : INR<1,5 ; HR=« hazard ratio »

Les résultats de cette recherche furent assez frappants (voir Table 19) : il s'avéra que les situations identifiées comme étant à risque par les auteurs n'étaient que rarement suivies d'EIM, probablement parce que les médecins connaissaient ces situations et adaptaient leur conduite thérapeutique. Inversement, les risques semblaient plutôt liés à des interactions décrites comme étant secondaires, peut-être parce que les médecins n'avaient pas conscience du risque d'interaction médicamenteuse.

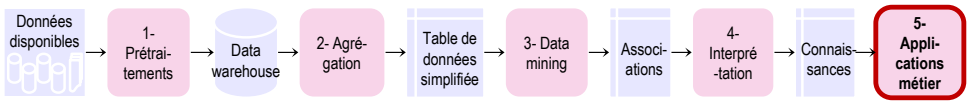
Table 19. Relation entre les résultats empiriques (colonnes 1 & 2) et la bibliographie (colonnes 3 & 4, selon Holbrook et al. [166]) HR = hazard ratio

Résultat	Médicament (avec AVK)	Causalité*	Sévérité*
INR≥5	Fénofibrate (HR>1)	Fortement probable	Modérée
	Méthylprednisolone (HR>1)	Fortement improbable	Majeure
	Simvastatine (HR>1)	Probable	Mineure
	Ofloxacine (HR<1 !!)	Possible	Majeure
	85 autres médicaments (HR pas différent de 1)	Fortement improbable : 9% Possible : 32% Probable : 30% Fortement probable : 29%	Non-clinique : 50% Mineure : 4% Modérée : 33% Majeure : 13%
INR≤1.5	Sucralfate (HR>1)	Fortement probable	Non-clinique
	Dicloxacillin (HR>1)	Probable	Modérée
	31 autres médicaments (HR pas différent de 1)	Fortement improbable : 14% Possible : 14% Probable : 29% Fortement probable : 43%	Non-clinique : 46% Mineure : 7% Modérée : 36% Majeure : 11%

En conclusion, il semble déraisonnable de remettre en question la connaissance académique sur les médicaments au seul vu de ces résultats en vie réelle. En

revanche, de tels résultats suggèrent que les alertes des systèmes d'aide à la décision devraient être calibrées non pas d'après la connaissance académique, mais bien d'après les probabilités empiriques d'EIM. Nous souhaitons défendre ce point de vue à l'avenir, afin de concevoir des systèmes d'aide à la décision contextualisés, délivrant des alertes plus pertinentes.

3.3 Adhésion aux recommandations

Position :	
Publication :	*Petit A-E, Mangeard H, Chazard E, Puisieux F. Changes in drug management of Alzheimer's disease in nursing homes: Impact of the media campaign against specific drugs for Alzheimer's disease. Encephale 2016. doi:10.1016/j.encep.2015.03.006. [167]* *Ferret L, Beuscart J-B, Ficheur G, Beuscart R, Luyckx M, Chazard E. Evaluation of compliance with recommendations of prevention of thromboembolism in atrial fibrillation in the elderly, by data reuse of electronic health records. Stud Health Technol Inform 2015;210:394–8. [texte intégral en ligne][168]* *Frély A, Chazard E, Pansu A, Beuscart J-B, Puisieux F. Impact of acute geriatric care in elderly patients according to the Screening Tool of Older Persons' Prescriptions/Screening Tool to Alert doctors to Right Treatment criteria in northern France. Geriatr Gerontol Int 2015. doi:10.1111/ggi.12474. [169]* *Ficheur G, Schaffar A, Caron A, Balcaen T, Beuscart J-B, Chazard E. Elderly Surgical Patients: Automated Computation of Healthcare Quality Indicators by Data Reuse of EHR. Stud Health Technol Inform 2016;221:92–6. [texte intégral en ligne][170]* *Chazard E, Babaousmail D, Schaffar A, Ficheur G, Beuscart R. Process assessment by automated computation of healthcare quality indicators in hospital electronic health records: a systematic review of indicators. Stud Health Technol Inform 2015;210:867–71. [texte intégral en ligne][171]*

Il nous fut également possible d'analyser les pratiques de soins, en termes de conformité aux recommandations.

3.3.1 Exemple des prescriptions inappropriées

Une première déclinaison fut la détection des prescriptions inappropriées. Elle repose simplement sur l'application d'une recommandation qui indique que tel traitement ou telle dose n'est pas recommandé dans telles circonstances. Nous obtînmes d'intéressants résultats dans le champ des anticoagulants [132,133]. Ainsi, sur 443 patients de plus de 75 ans présentant une fibrillation atriale, nous calculâmes le score HAS-BLED et vérifiâmes la conformité des prescriptions d'anticoagulants avec les recommandations de la Société Européenne de Cardiologie. Nous observâmes que les recommandations n'étaient suivies que dans 47,8% des cas. En particulier, lorsque les critères indiquaient de manière formelle une bithérapie, elle n'était appliquée que dans 3% des cas [168].

3.3.2 Principes d'une mesure automatisée de la qualité des soins

Les événements liés aux soins qui peuvent être prévenus entrent dans la problématique plus générale du **défaut de qualité des soins**. Ce défaut de qualité des soins est responsable de morbi-mortalité, qui ne peut que s'accroître du fait du vieillissement de la population, de l'accumulation des traitements préventifs et de la complexification des thérapeutiques. Le défaut de qualité des soins est notamment lié à la non-conformité des prises en charge (diagnostiques, thérapeutiques, de suivi) aux recommandations produites par la littérature blanche, les sociétés savantes et les agences publiques. Deux approches traditionnelles tentent d'augmenter cette conformité.

La première approche traditionnelle consiste à implémenter des **systèmes d'aide à la décision** (CDSS), qui émettent des alertes interruptives lorsque, pour un patient donné, une action non-conforme survient. Deux écueils majeurs existent. Le premier est l'apparition de faux positifs, et l'état de non-réceptivité des médecins qui en résulte. Ce bruit rend les CDSS inefficaces voire nuisibles [151]. Le deuxième est une altération du workflow : le déploiement des CDSS nécessite une formalisation des tâches autorisées par profil, et met ainsi fin à des pratiques pragmatiques de délégation et supervision. La charge de travail remonte généralement vers les médecins, et les autres soignants sont parfois exclus de la conduite du soin. Ces deux écueils expliquent que les CDSS aient rarement un impact clinique positif [172].

La deuxième approche traditionnelle est le suivi d'**indicateurs de processus** (pourcentages de conformité des pratiques aux recommandations, tels les IPAQSS, indicateurs pour l'amélioration de la qualité et de la sécurité des soins) [173]. Cette démarche de mesure rétrospective de l'adhésion aux recommandations permet à d'autres professionnels (ingénieurs qualité) d'investir la qualité des soins, sans interrompre le soin quotidien. Néanmoins, ces enquêtes nécessitent l'examen de dossiers patients par des experts. La charge de travail est excessive, il existe une variabilité inter-opérateur et l'audit est de fait limité à quelques dossiers et quelques pathologies. Ces indicateurs ne sont calculés que sur des petits échantillons sans mesure historique. Au fil des évolutions des indicateurs, toute comparabilité longitudinale est perdue, or les recommandations évoluent très fréquemment.

Nos travaux plus récents visent **proposer un nouveau paradigme** reprenant les avantages des deux approches traditionnelles, et **les combinant** en tirant profit de la réutilisation des données massives de santé (*data reuse, big data*).

3.3.3 Faisabilité bibliographique d'une mesure automatisée de la qualité des soins

Dans un premier temps, nous réalisâmes une revue exhaustive de la littérature afin de recenser tous les **indicateurs de processus** définis explicitement dans la littérature blanche [171]. Nous passâmes en revue 8744 articles et en conservâmes 126, qui décrivaient **440 indicateurs**. Nous cherchâmes à savoir si ces indicateurs étaient implémentables automatiquement à l'aide d'un cœur de données comprenant les données médico-administratives du PMSI, les résultats de biologie, les médicaments administrés et les courriers en texte libre. Un indicateur fut considéré comme implémentable automatiquement si, en le programmant sous forme d'algorithme informatique, il était *a priori* possible de calculer automatiquement un taux de conformité sans avoir recours à une lecture humaine du dossier patient.

A l'issue de cette revue, **343 indicateurs (77,7%)** furent classés comme **non-implémentables**, pour les raisons suivantes (total supérieur à 100%) :

- Besoin de données complémentaires sur la prise en charge ambulatoires : 209 indicateurs (61.1%)
- Besoin de données complémentaires structurées : 147 indicateurs (43.0%)
- Besoin de données complémentaires non-structurées (texte libre) : 90 indicateurs (26.3%)
- Besoin de la trace qu'une information a été donnée au patient : 29 indicateurs (8.5%)

Néanmoins, **98 indicateurs (22,3%)** furent classés comme implémentables. Ces indicateurs nécessitaient d'accéder aux données suivantes (total supérieur à 100%) :

- Diagnostics (sans date précise) : 97 indicateurs (99.0%)
- Médicaments administrés (avec date précise) : 58 indicateurs (59.2%)
- Actes médicaux, diagnostiques ou thérapeutiques (avec date précise) : 47 indicateurs (48.0%)
- Informations administratives de base (mouvements, démographie) : 29 indicateurs (29.6%)
- Résultats de biologie (avec date précise) : 20 indicateurs (20.4%)
- Comptes-rendus en texte libre (courrier de sortie, compte-rendu d'acte, etc.) analysés par simple recherche de mot clef : 19 indicateurs (19.4%),
- Capacité à chaîner les séjours antérieurs du même patient : 11 indicateurs (11.2%)
- Echelle de dépendance (activités de la vie quotidienne, AVQ) : 3 indicateurs (3.1%)

Cette revue de la littérature nous enseigna deux points importants, qui confortèrent notre approche. Tout d'abord, une part importante des indicateurs de processus publiés dans la littérature est implémentable sous forme d'indicateur automatisé en disposant d'un « cœur de données » limité (mouvements, démographie, diagnostics, actes, biologie, médicaments et comptes-rendus). Cette proportion s'élève à 22,3%, alors que nous nous attendions à trouver 5 à 15% d'indicateurs implémentables. En outre, cette revue nous enseigne que nos travaux futurs gagneraient à réaliser une jointure entre bases de données hospitalières et ambulatoires.

Nous testâmes ensuite l'implémentation de tels indicateurs sur des données réelles, afin de valider le concept.

3.3.4 Expérimentation en gériatrie

Nous réutilisâmes les données aujourd'hui recueillies en routine dans le dossier patient informatisé (données du PMSI, résultats de biologie, médicaments administrés et comptes-rendus en texte libre) et mîmes ainsi en place une cohorte historique de séjours hospitaliers [170]. Nous sélectionnâmes 9 indicateurs proposés par McGory et al. [174] dans le champ de la chirurgie des personnes âgées. Nous proposâmes une **abstraction** des indicateurs de qualité sous la forme de **fonction informatique** prenant en entrée toutes les données d'un séjour, et retournant une valeur manquante en cas de séjour exclu, la valeur 0 en cas de séjour inclus mais non-conforme, et la valeur 1 en cas de séjour inclus et conforme (voir Figure 38) [170]. Cette abstraction est compatible avec tous les indicateurs de processus

portant un jugement sur chaque séjour (ou patient), et présente de plus une particularité très appréciable : *moyenne(X) = taux de conformité*.

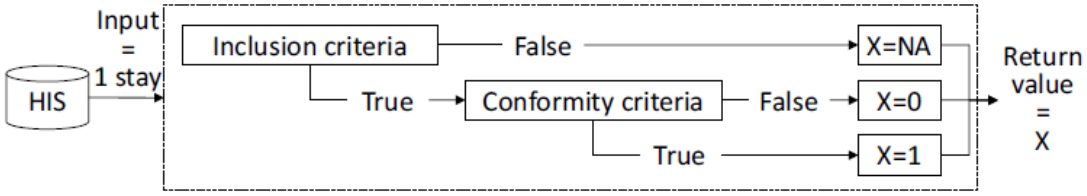


Figure 38. Abstraction d'un indicateur de conformité sous la forme de fonction informatique

Nous programmâmes ainsi chacun des 9 indicateurs, et demandâmes à un médecin expert de vérifier les résultats de chaque indicateur sur un échantillon aléatoire de séjours. En cas de mauvais résultat, l'expert dut documenter les motifs de l'échec de l'algorithme.

Cette implémentation fit appel à une connaissance de trois domaines :

- Une connaissance algorithmique suffisante
- Une bonne compréhension des données médicales
- Une très forte connaissance et expérience des données médicales informatisées, au point notamment d'intégrer quelles données étaient fiables ou non, d'imaginer des contournements acceptables, et d'anticiper à quel moment une erreur de mesure se transformait en biais systématique, et quel était le sens de ce biais

Il fallut tout d'abord interpréter ces conditions médicalement, mais en les précisant, car l'expression écrite traditionnelle repose sur une foule de sous-entendus. Ainsi, pour prendre un exemple très simple, si la prescription d'un médicament doit déclencher un examen biologique à 30 jours, il n'est jamais précisé dans une recommandation que cet examen n'est plus nécessaire si le patient est décédé entretemps. L'implémentation stricte du texte serait aussi incohérente qu'indécente.

Ensuite, il fallut transformer les données natives en information. Si par exemple une règle repose sur le concept d'insuffisance rénale, il faut la formaliser en mappant plusieurs dizaines de codes diagnostiques CIM10 (jusqu'à par exemple « Q601 Agénésie rénale, bilatérale »), une dizaine de codes CCAM d'hémodialyse, et quelques paramètres biologiques.

Enfin, il fallut développer le code informatique correspondant.

Le détail des 9 indicateurs est lisible dans l'article correspondant [170]. Nous livrons les principaux résultats en Table 20.

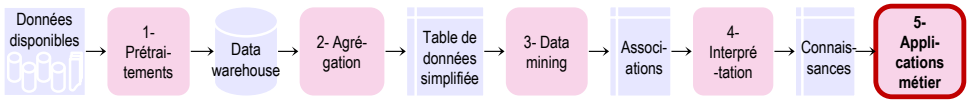
Table 20. Résultats de la mesure automatisée de 9 indicateurs de conformité en gériatrie. Précision : proportion de séjours détectés comme étant non-conformes par la machine et confirmés comme tels par l'expert

Domaine de l'indicateur	Nb séjours inclus	Taux de conformité	Précision (détection des non-conformes)	
Dosage de la clairance de la créatinine	238	86,6% [81,6 ; 90,7]	86,7% [59,5 ; 98,3]	☺
Confusion	8	75 % [34,9 ; 96,8]	0,0%	☹
Prescription de bêtabloquants	18	27,8% [9,7 ; 53,5]	46,2% [19,2 ; 74,9]	😐
Mise en place d'une antibiothérapie prophylactique IV	607	2,4% [1,5 ; 4,2]	93,3% [68,1 ; 99,8]	☺
Poursuite d'une antibiothérapie prophylactique per os	607	2,3% [1,3 ; 3,9]	93,3% [68,1 ; 99,8]	☺
Prévention des thromboses veineuses profondes	314	70,0% [65,6 ; 76,0]	53,3% [26,6 ; 78,7]	☺
Traitement de l'anémie	217	60,4% [53,5 ; 66,9]	86,7% [59,5 ; 98,3]	☺
Transfusion en cas d'anémie	19	73,7% [48,8 ; 90,9]	60,0% [14,7 ; 94,7]	😐
Fièvre postopératoire	4	75% [19,4 ; 99,4]	0,0%	☹

Il ressort de cette expérimentation que le calcul automatisé d'indicateurs de processus de mesure de la qualité des soins est **possible**, et peut permettre d'**identifier très rapidement des situations à risque** pour les patients. Ces situations correspondent au non-respect de certaines recommandations validées. Il ressort également que la mise en place d'une telle démarche nécessite de **réelles expertises**, et peut difficilement être facilitée par des ontologies ou des analyses automatisées du langage des recommandations elles-mêmes. De notre expérience, les systèmes de gestion de règles seraient rapidement pris en défaut par la complexité algorithmique. Il nous semble indispensable de programmer directement en code de bas niveau chaque recommandation. Enfin, une revue critique de cas réels est indispensable car, si certains indicateurs semblent bien fonctionner, d'autres sont en pratique inutilisables, principalement pour des raisons liées à la qualité des données. Cette qualité pouvant varier d'un hôpital à l'autre, il semble utile de réévaluer la précision des règles de détection dans chaque hôpital.

Une batterie d'indicateurs pourrait ainsi être développée pour identifier les situations à risque et alimenter les processus d'analyse qualité plus classiques, tels les revues de morbidité et mortalité.

3.4 Télécardiologie

Position :	
Publication :	*Chazard E, Marcolino MS, Dumesnil C, Caron A, Palhares DMF, Ficheur G, et al. One Million Electrocardiograms of Primary Care Patients: A Descriptive Analysis. Stud Health Technol Inform 2015;216:69–73.* [texte intégral en ligne][175] *Rosier A, Mabo P, Temal L, Van Hille P, Dameron O, Deléger L, et al. Personalized and automated remote monitoring of atrial fibrillation. Europace 2015. doi:10.1093/europace/euv234.* [176]

Nous pûmes ensuite appliquer ces méthodes à des données de télécardiologie issues d'un centre brésilien.

Le centre de télémédecine de Belo Horizonte (Minas Gerais, Brésil) a en charge l'interprétation à distance des électrocardiogrammes enregistrés par l'ensemble des médecins généralistes de l'Etat du Minas Gerais. Chaque médecin généraliste est équipé d'un électrocardiographe relié à Internet. Sitôt après sa capture, chaque ECG est envoyé en ligne au centre national de télémédecine, accompagné d'un formulaire de renseignements cliniques (symptômes, traitements, antécédents). L'ECG est alors interprété dans les 12 heures (hormis urgence) par un des 30 cardiologues d'astreinte, à distance. Lorsque ce travail fut réalisé, le centre national de Télémédecine disposait déjà d'une base de 1,2 millions d'ECG complétés de données cliniques et interprétés. Des résultats d'interprétation automatique, non utilisés à ce jour, étaient également disponibles dans le cadre d'un projet de traitement automatisé du signal en collaboration avec l'université d'Oslo. Ces données ne faisaient alors l'objet d'aucune exploitation de masse.

Nous pûmes tout d'abord analyser les données disponibles en base [175]. Nous connectâmes un interpréteur automatisé d'ECG afin d'interpréter automatiquement les millions d'ECG enregistrés dans la base de données. Cette tâche fut plus complexe qu'il n'y parait, notamment du fait de l'hétérogénéité terminologique des messages générés par l'interpréteur. Nous observâmes par exemple que près de 70% des ECG étaient dans les limites de la normalité, et que ce taux diminuait en moyenne de 7,4% tous les 10 ans d'âge du patient (voir Figure 39).

Figure 39. Proportion d'ECG non-pathologiques en fonction de l'âge.

Nous pûmes discuter les problèmes de qualité de telles bases de données, en particulier pour ce qui concerne les données cliniques codées. Ainsi par exemple, 17,3% des patients sans obésité codée avaient un BMI supérieur à 30 kg/m². Nous observâmes enfin des caractéristiques propres à la population brésilienne :

- 3% des patients étaient atteints d'une maladie de Chagas
- 95,7% des patients sous diurétiques avaient une hypertension artérielle, et 59,5% des patients avec HTA recevaient des diurétiques, conformément aux recommandations brésiliennes
- La population étudiée souffrait d'obésité : le BMI moyen était de 25.9 kg/m², 20,3% des patients avaient un BMI supérieur à 30 kg/m², et le BMI moyen augmentait de 0.15 kg/m² par année civile
- Le tabagisme semblait peu développé. Il était rapporté dans seulement 6.9% des cas.

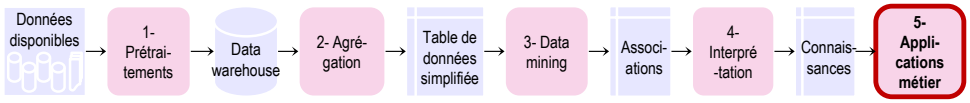
Nous étudiâmes ensuite la concordance entre l'interprétation automatique réalisée par le Glasgow Program (un programme d'interprétation automatisée reconnu [177–179]), et un gold standard issu d'une triple interprétation par des cardiologues. Nous montrâmes ainsi que les résultats de ce programme étaient corrects [180], mais nettement inférieurs à ceux annoncés par son créateur dans la littérature grise constituant notamment la notice technique du logiciel (voir Table 21). Nous tentâmes ensuite d'améliorer la détection des ECG normaux en incorporant les données cliniques, les traitements et les résultats de l'interprétation automatisée, afin d'en diminuer les faux négatifs et permettre une réorganisation du processus de revue des nouveaux ECG (en éliminant d'emblée les ECG normaux, qui représentent 60% des cas). Cette approche fut menée à l'aide de méthodes de data mining. Nous pûmes améliorer la détection des ECG de l'algorithme, néanmoins les résultats ne durent pas suffisants pour imaginer de filtrer ou réordonner automatiquement la liste d'attente d'ECG à interpréter par le centre.

Table 21. Resultats de l'évaluation du Glasgow Program

Statement	Prevalence rate with 95% conf. int.	Sensitivity	Specificity	AUC (1)
Global statements				
Non-fatal error	0.151 [0.127;0.175]	0.111	0.960	0.540
Permanent pacemaker	0.006 [0.001;0.011]	0.200	0.995	0.600
Non-sinus rhythm	0.050 [0.036;0.065]	0.511	0.962	0.740
Non-exclusive ECG abnormalities				
Left ventricular hypertrophy	0.034 [0.022;0.046]	0.600	0.954	0.790
Right ventricular hypertrophy	0.004 [0.000;0.009]	0.500	0.968	0.730
Myocardial infarction	0.031 [0.020;0.043]	0.357	0.965	0.700
Rhythm troubles				
Atrial fibrillation or flutter	0.015 [0.007;0.023]	0.923	0.987	0.960
Multifocal or ectopic atrial rhythm	0.006 [0.001;0.011]	0.600	0.988	0.790
Atrial or supraventricular extrasystole	0.022 [0.013;0.032]	0.650	0.937	0.790
Sinusal bradycardia	0.020 [0.011;0.030]	0.722	0.970	0.850
Sinusal or supraventricular tachycardia	0.020 [0.011;0.030]	0.778	0.977	0.880
Accelerated or normal junctional rhythm	0.001 [0.000;0.003]	0.000	0.997	0.500
Ventricular extrasystole	0.026 [0.015;0.036]	0.783	0.949	0.870
Conduction troubles				
First degree atrioventricular block	0.030 [0.019;0.042]	0.593	0.978	0.790
Wolff Parkinson White syndrome	0.002 [0.000;0.005]	0.500	0.999	1.000
Left bundle branch block (2)	0.094 [0.075;0.114]	0.464	0.986	0.730
Right bundle branch block (2)	0.065 [0.049;0.082]	0.603	0.995	0.800
Repolarization abnormalities				
Primary or secondary repolarization abnormality	0.339 [0.307;0.37]	0.659	0.771	0.710
Descriptive ECG abnormalities				
Bradycardia	0.020 [0.011;0.030]	0.722	0.970	0.850
Tachycardia	0.020 [0.011;0.030]	0.778	0.981	0.880
Short PR	0.001 [0.000;0.003]	1.000	0.975	0.990
QRS axis deviation	0.099 [0.079;0.119]	0.761	0.915	0.840
Long QT	0.001 [0.000;0.003]	0.000	0.997	0.510

(1): the area under the curve is computed using several points when several thresholds are available
(2): complete, incomplete or fascicular

3.5 Epidémiologie des soins

Position :	
Publication :	*Ficheur G, Ferreira Careira L, Beuscart R, Chazard E. EpiHosp: A web-based visualization tool enabling the exploratory analysis of complications of implantable medical devices from a nationwide hospital database. Stud Health Technol Inform 2015;210:409–13.* [texte intégral en ligne][181] *Ficheur G, Caron A, Beuscart J-B, Ferret L, Putman S, Beuscart R, et al. The risks of pulmonary embolism and upper gastrointestinal bleeding beyond 35days after total hip replacement for coxarthrosis among middle-aged patients: A cross-over cohort. Prev Med 2016. doi:10.1016/j.ypmed.2016.09.010.* [103]

Enfin, nous nous intéressâmes à l'application de ces techniques pour mettre en évidence de nouvelles connaissances épidémiologiques, non pas sur les pathologies mais sur leur prise en charge (épidémiologie des soins).

Il est couramment admis que l'intérêt de la réutilisation de données en épidémiologie (strictement, par opposition à l'épidémiologie des soins, définie plus bas) est limité notamment par le biais de recrutement et d'information : la population hospitalisée ne représente pas la population totale. Parmi les patients absents, on trouve indifféremment des patients dont l'état pathologique est moins sévère, d'autres patients dont l'état pathologique est similaire à ceux hospitalisés, et enfin des patients dont l'état est tellement sévère qu'ils sont décédés en-dehors de structure hospitalière.

En revanche, l'**épidémiologie des soins** pose moins de problème. Elle consiste notamment à décrire les prises en charge hospitalières en soi, et non les pathologies. Par définition, il n'y a pas de biais de recrutement dans ce cas précis.

Nous nous intéressâmes tout d'abord aux poses de dispositifs médicaux implantables dans les hôpitaux publics ou privés non-lucratifs (ex-DGF) d'après la base nationale du PMSI. Les établissements privés lucratifs (ex-OQN) renseignant ces variables dans les fichiers de RSF, elles sont malheureusement absentes des bases nationales du PMSI. Nous développâmes un outil simple de requête qui permet à l'utilisateur de choisir un dispositif par son code LPP (qui est en fait automatique renseigné depuis un arbre de choix intuitif et agréable à utiliser) (voir Figure 40).

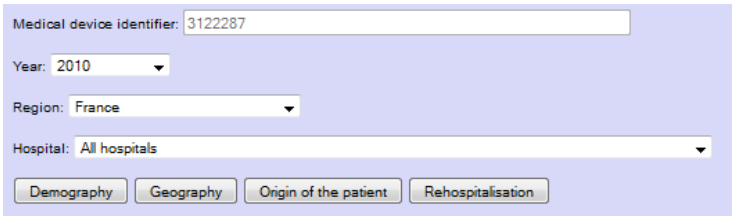


Figure 40. Ecran de choix d'un ou plusieurs codes LPP

L'utilisateur peut ensuite visualiser diverses informations, comme par exemple l'origine géographique des patients (voir Figure 41) et, sur la droite de cette même figure, les établissements de prise en charge des patients vivant dans l'aire géographique sélectionnée avec la souris.

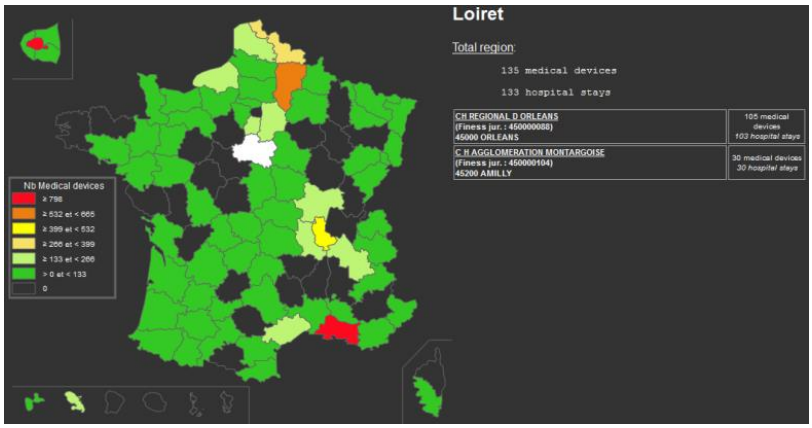


Figure 41. Origine géographique des patients.

La fonctionnalité la plus intéressante est sans doute l'analyse automatique des ré-hospitalisations. Un tableau s'affiche automatiquement (voir Figure 42 sur la gauche) et, lorsqu'on sélectionne un des motifs, une courbe de survie s'affiche également (voir Figure 42 sur la droite). Toutes ces requêtes étant réalisées à la volée sur la base nationale en quelques secondes, l'enjeu du développement fut de préparer les requêtes à l'aide de volumineuses tables de données pré-agrégées.

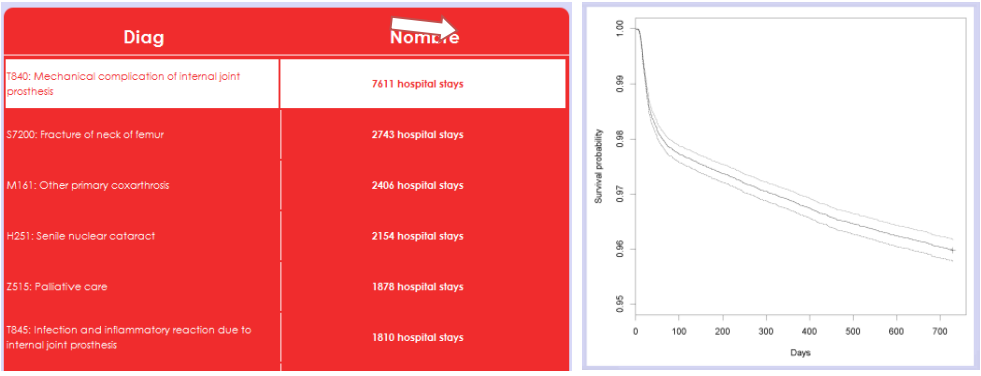


Figure 42. Motifs de réadmission des patients et courbe de survie correspondant au motif sélectionné

Sur cette même base nationale du PMSI, nous poursuivons actuellement des travaux descriptifs et inférentiels que nous jugeons très prometteurs. A ce jour, aucun de ces travaux n'a été définitivement accepté par un journal, du fait de leur caractère très récent.

4 Conclusion de l'exposé des travaux de recherche

Nous avons montré par nos travaux que la réutilisation des données produites en routine au cours du soin était possible, et qu'elle pouvait donner des résultats utiles à la collectivité dans différents domaines (effets indésirables du médicament, qualité des soins, télécardiologie, épidémiologie des soins, etc.).

Nous avons également montré que cette réutilisation posait différents problèmes techniques (par exemple anonymisation, interopérabilité) et méthodologiques (notamment en ce qui concerne la gestion de la temporalité). Nous avons rencontré plusieurs de ces problèmes, et avons tenté chaque fois que possible de trouver, publier et évaluer des solutions utilisables par les chercheurs.

Ces travaux ont également mis en avant l'importance de l'agrégation de données, phase préalable à l'analyse statistique.

Enfin, il semble que la réutilisation des données ouvre des perspectives importantes. Pour autant, il faut bien garder à l'esprit que de telles approches resteront d'un niveau de preuve inférieur à celui par exemple des essais randomisés contrôlés. Nous défendons néanmoins le point de vue que, à terme, des études de *data reuse* pourraient devenir des préalables utiles puis nécessaires à l'obtention de financements pour des études plus traditionnelles et plus coûteuses.

Présentation des activités d'encadrement

Cette section présente uniquement les encadrements académiques formels achevés et soutenus avant juin 2017. Les encadrements amorcés mais non soutenus sont dénombrés mais non détaillés.

1 Thèses d'université (2 achevées, 1 en cours)

J'ai co-encadré deux thèses d'université, toutes deux placées sous la direction du Professeur Régis Beuscart, au sein du CERIM :

- Réutilisation de données hospitalières pour la recherche d'effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical implantable. Ficheur, Grégoire. Ph.D. Thesis, Université du Droit et de la Santé - Lille II, Lille, France, November 2015
- Anticoagulants oraux : réutilisation de données hospitalières informatisées dans une démarche de soutient à la qualité des soins. Ferret, Laurie. Ph.D. Thesis, Université du Droit et de la Santé - Lille II, Lille, France, December 2015

Ces deux thèses donnent encore lieu à publications actuellement. Les publications déjà référencées dans Medline sont présentées ci-dessous (voir Table 22).

Table 22. Thèses d'université co-encadrées ayant donné lieu à des publications déjà référencées

Thèse d'université	Publication
Réutilisation de données hospitalières pour la recherche d'effets indésirables liés à la prise d'un médicament ou à la pose d'un dispositif médical implantable. Ficheur, G. Ph.D. Thesis, Université du Droit et de la Santé - Lille II, Lille, France, November 2015	Ficheur G, Ferreira Careira L, Beuscart R, Chazard E. EpiHosp: A web-based visualization tool enabling the exploratory analysis of complications of implantable medical devices from a nationwide hospital database. Stud Health Technol Inform 2015;210:409–13.
	Ficheur G, Caron A, Beuscart J-B, Ferret L, Putman S, Beuscart R, Chazard E. The risks of pulmonary embolism and upper gastrointestinal bleeding beyond 35days after total hip replacement for coxarthrosis among middle-aged patients: A cross-over cohort. Prev Med 2016. doi:10.1016/j.ypmed.2016.09.010.
Anticoagulants oraux : réutilisation de données hospitalières informatisées dans une démarche de soutient à la qualité des soins. Ferret, L. Ph.D. Thesis, Université	Ferret L, Beuscart J-B, Ficheur G, Beuscart R, Luyckx M, Chazard E. Evaluation of compliance with recommendations of prevention of thromboembolism in atrial fibrillation in the elderly, by data reuse of electronic health records. Stud Health Technol Inform 2015;210:394–8.

Thèse d'université	Publication
du Droit et de la Santé - Lille II, Lille, France, December 2015	Ferret L, Luyckx M, Ficheur G, Chazard E, Beuscart R. Evaluation of a Computer Application for Retrospective Detection of Vitamin K Antagonist Treatment Imbalance. J Patient Saf 2016. doi:10.1097/PTS.0000000000000182.

2 Thèses de médecine (15 achevées, 6 en cours)

J'ai dirigé 15 thèses de médecine achevées à ce jour :

- Dugast JB. Implantation de dispositifs d'assistance ventriculaire implantables en France de 2008 à 2014 : étude de la base nationale du PMSI. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2017.
- Bonte C. Le taux de césariennes reste stable de 2008 à 2014. Une étude de la base nationale du PMSI. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2017.
- Martincic C. Chirurgie bariatrique en France de 2008 to 2014: techniques, complications et préférences des chirurgiens. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2017.
- George A. Mise au point d'un framework d'analyses statistiques pour la réutilisation des bases de données médico-administratives. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2017.
- Delrot C. Les praticiens ont-ils confiance en l'interprétation automatisée des électrocardiogrammes ? Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2017.
- Melot E. Qu'est-ce qu'une maladie chronique ? Apport de l'analyse des données de la base nationale du PMSI. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2017.
- Girier N. Epidémiologie de la pose de prothèses de hanche en France : analyse de la base nationale du PMSI de 2008 à 2014. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2017.
- Degoul S. Comment mesurer la performance d'un test diagnostique ? Présentation et comparaison d'indicateurs. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2016.
- Baro E. Vers une définition des big data en santé basée sur la littérature. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2015.
- Dumesnil C. Quelle est l'augmentation des durées de séjour liée aux hyperkaliémies : comparaison de différentes méthodes statistiques. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2014.
- Schaffar A. Etude de la faisabilité de l'implémentation d'indicateurs automatisés de la qualité des soins. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2013.
- Mouret-Kubiak C. Dé-identification automatisée de courriers médicaux : proposition et évaluation de la méthode FASDIM. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2012.

- Miquel P-H. Conception et déploiement d'indicateurs médico-économiques dans un groupe d'hospitalisation privé. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2011.
- Genty M. Détection automatisée des effets indésirables médicamenteux : revue et validation des cas détectés. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2010.
- Ficheur G. Détection automatisée d'effets indésirables médicamenteux dans des contextes hospitaliers variables. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2010.

Parmi celles-là, 5 ont donné lieu à publication référencées Medline (voir Table 23). La place de premier auteur revenait à l'étudiant chaque fois qu'il acceptait de rédiger l'article.

Table 23. Thèses de médecine dirigées ayant donné lieu à publication

Thèse de médecine	Publication
Baro E. Vers une définition des big data en santé basée sur la littérature. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2015.	Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. Biomed Res Int 2015;2015. doi:10.1155/2015/639021.
Dumesnil C. Quelle est l'augmentation des durées de séjour liée aux hyperkaliémies : comparaison de différentes méthodes statistiques. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2014.	Chazard E, Dumesnil C, Beuscart R. How much does hyperkalemia lengthen inpatient stays? About methodological issues in analyzing time-dependant events. Stud Health Technol Inform 2015;210:835–9.
Schaffar A. Etude de la faisabilité de l'implémentation d'indicateurs automatisés de la qualité des soins. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2013.	Ficheur G, Schaffar A, Caron A, Balcaen T, Beuscart J-B, Chazard E. Elderly Surgical Patients: Automated Computation of Healthcare Quality Indicators by Data Reuse of EHR. Stud Health Technol Inform 2016;221:92–6.
Mouret-Kubiak C. Dé-identification automatisée de courriers médicaux : proposition et évaluation de la méthode FASDIM. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2012.	Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inform 2014;83:303–12. doi:10.1016/j.ijmedinf.2013.11.005.
Ficheur G. Détection automatisée d'effets indésirables médicamenteux dans des contextes hospitaliers variables. Doctorat de médecine. Université du Droit et de la Santé - Lille II, 2010.	Chazard E, Ficheur G, Merlin B, Serrot E, PSIP Consortium, Beuscart R. Adverse drug events prevention rules: multi-site evaluation of rules from various sources. Stud Health Technol Inform 2009;148:102–11.

3 Mémoires de master 2 (4 achevés)

J'ai dirigé 4 mémoires de master 2 :

- Dumesnil C. Quelle est l'augmentation des durées de séjour liée aux hyperkaliémies ? Comparaison de différentes méthodes statistiques. Mémoire de Master 2. Université Paris Sud, Paris XI, 2013.
- Babaousmail D. Qualité des soins : indicateurs de conformité aux recommandations. Mémoire de Master 2. Université du Droit et de la Santé - Lille II, 2013.
- Lepert H. A la conquête d'un marché de niche : Cas du lancement d'une formation initiale supérieure en santé, Discipline de l'information médicale. Mémoire de Master 2. Université des Sciences et Technologies, Lille 1, 2012.
- Génin M. Application du datamining dans le cadre de la prévention des effets indésirables liés aux médicaments. Mémoire de Master 2. Université des Sciences et Technologies, Lille 1, 2009.

Les trois mémoires de master recherche ont donné lieu à publications référencées Medline. Elles sont présentées ci-dessous (voir Table 24).

Table 24. Mémoires de master recherche dirigés ayant donné lieu à des publications déjà référencées

Mémoire de master	Publication
Dumesnil C. Quelle est l'augmentation des durées de séjour liée aux hyperkaliémies ? Comparaison de différentes méthodes statistiques. Master. Thesis, Université Paris Sud, Paris XI, Lille, France, October 2014.	Chazard E, Dumesnil C, Beuscart R. How much does hyperkalemia lengthen inpatient stays? About methodological issues in analyzing time-dependant events. Stud Health Technol Inform 2015;210:835–9.
Babaousmail D. Qualité des soins : indicateurs de conformité aux recommandations. Mémoire de Master 2. Université du Droit et de la Santé - Lille II, 2013.	Chazard E, Babaousmail D, Schaffar A, Ficheur G, Beuscart R. Process assessment by automated computation of healthcare quality indicators in hospital electronic health records: a systematic review of indicators. Stud Health Technol Inform 2015;210:867–71.
Génin M. Application du datamining dans le cadre de la prévention des effets indésirables liés aux médicaments. Mémoire de Master 2. Université des Sciences et Technologies, Lille 1, 2009.	Chazard E, Ficheur G, Merlin B, Genin M, Preda C, PSIP consortium, et al. Detection of adverse drug events detection: data aggregation and data mining. Stud Health Technol Inform 2009;148:75–84.

Prospective de recherche

Je poursuivrai mes travaux de recherche dans la réutilisation et fouille de données massives de santé produites en routine au cours du soin, et ce notamment dans le champ particulier de la **prévention de la perte d'autonomie du sujet âgé**.

Nous présenterons notre prospective de recherche à travers trois axes :

- Un axe thématique et applications médicales
- Un axe sources de données
- Un axe méthodologies

Nous évoquerons les collaborations à l'issue de cette présentation.

1 Axe thématique et applications médicales

La perte d'autonomie du sujet âgé est un processus multifactoriel.

Cet aspect multifactoriel peut s'entendre transversalement, au sens où, à un instant donné, plusieurs facteurs concourent à la survenue d'un événement péjoratif. Ainsi par exemple, les antihistaminiques de première génération peuvent concourir chez le sujet âgé à la survenue d'une confusion ou d'une insuffisance rénale, en particulier en combinaison avec d'autres psychotropes ou une déshydratation [182].

Il peut également s'entendre longitudinalement, au sens où plusieurs facteurs peuvent s'enchaîner et être associés dans le temps à la survenue d'un événement péjoratif. Ainsi par exemple, une fracture de hanche chez une personne âgée est associée à une mortalité à un an de 25 à 50% selon les études [183].

Ces facteurs sont divers, et on pourra ainsi citer plusieurs facteurs de risque ou facteurs prédictifs de la perte d'autonomie :

- Les effets indésirables des médicaments et l'iatrogénie médicamenteuse
- Le défaut de qualité des soins en hospitalisation
- Les maladies neurodégénératives
- La séquence des soins (séquence d'épisodes ambulatoires et hospitaliers)

Nous présenterons ci-après les travaux que nous comptons mener dans ces domaines, et comment nous apporterons notre soutien méthodologique aux travaux menés par d'autres chercheurs.

1.1 Le programme Paerpa

L'Etat a mis en place en 2014 le programme Paerpa [184] sur des territoires pilotes. Ce programme s'adresse aux personnes âgées de 75 ans et plus, dont l'autonomie est susceptible de se dégrader pour des raisons d'ordre médical ou social. L'objectif

est de préserver leur autonomie, en leur permettant de recevoir « les bons soins par les bons professionnels, dans les bonnes structures au bon moment, le tout au meilleur coût » [184]. A cet effet, il s'agit de repérer les principaux facteurs d'hospitalisation évitables : dépression, chute et problèmes liés aux médicaments. Le dispositif est articulé autour de cinq champs d'action :

- Renforcer le maintien à domicile
- Améliorer la coordination des intervenants et des interventions
- Sécuriser la sortie d'hôpital
- Éviter les hospitalisations inutiles
- Mieux utiliser les médicaments

Sous l'impulsion du Dr Jean-Baptiste Beuscart, PHU dans notre équipe de recherche, notre équipe a obtenu de participer à ce programme. Actuellement, nous collectons des données chaînées et nominatives concernant tous les patients âgés de plus de 75 ans du Valenciennois-Quercitain (plus de 30 000 patients). Ces données concernent tant les prises en charge hospitalières qu'ambulatoires.

1.2 Médicament et perte d'autonomie

1.2.1 Détection automatisée des effets indésirables du médicament

Les effets indésirables des médicaments touchent en particulier les personnes âgées, et sont un facteur de risque de perte d'autonomie. Dans le cadre du projet européen PSIP, nous avons mis en œuvre des techniques de data mining pour détecter les effets indésirables des médicaments [15], les présenter rétrospectivement auprès des cliniciens et les analyser statistiquement [104], et tenter de les prévenir à l'aide de systèmes d'aide à la décision [134].

Au contact de David W. Bates à Boston, nous avons observé combien l'approche du projet PSIP était en avance à l'époque. Les idées développées dans PSIP commencent à ce jour à trouver un écho dans la communauté scientifique. Nous souhaitons poursuivre les travaux débutés à Boston et maintenir une collaboration régulière avec cette équipe, afin de mieux valoriser notre savoir-faire.

Parallèlement à cela, nous discutons actuellement avec des centres hospitaliers publics et privés de la région Hauts-de-France (anciennement Nord-Pas de Calais-Picardie) afin d'implémenter les méthodes et logiciels développés dans le cadre du projet PSIP. Enfin, nous espérons tirer profit de l'informatisation de la prescription au CHU de Lille, ainsi que de la mise en place d'un entrepôt de données cliniques pour poursuivre ces recherches, voire implémenter ces méthodes et logiciels à titre expérimental au CHU de Lille.

1.2.2 Prévention de la iatrogénie médicamenteuse au CHU de Lille

Au CHU de Lille, en mai 2017, 1400 lits sont déjà concernés par la prescription connectée (médecine, chirurgie, soins de suite et réadaptation). Sur ces lits, les prescriptions médicamenteuses sont contrôlées par 20 pharmaciens cliniciens. Il peut s'agir tantôt de revue à distance sur le logiciel de dossier patient, tantôt de présence physique du pharmacien dans le service clinique. Les pharmaciens du CHU de Lille réalisent ainsi actuellement près de 10 000 analyses pharmaceutiques

par mois, et 79% des modifications proposées par les pharmaciens sont acceptées par les médecins.

A la demande de nos collègues pharmaciens, nous souhaitons les aider à mettre en place deux fonctionnalités attendues, en parallèle avec l'acquisition d'un logiciel de support à l'analyse d'ordonnances destiné aux pharmaciens.

Tout d'abord, le temps de pharmacien consacré à la vérification des ordonnances est limité, tandis que, du fait du déploiement de la prescription connectée, le volume d'ordonnances à traiter augmente progressivement. Nous souhaitons mettre en place un système d'alertes pondérées, et de plus que cette pondération soit valable à l'échelle de l'ordonnance et non plus de la ligne de prescription, de manière à ce que le pharmacien, dans une période de temps limitée, puisse revoir en priorité les ordonnances les plus porteuses de risque. A l'inverse du projet de recherche désigné dans la section précédente, il s'agirait ici d'utiliser les connaissances acquises lors des recherches afin d'améliorer des logiciels existants, de manière plus opérationnelle, et non de mettre au point de nouveaux processus.

Ensuite, il n'est actuellement pas possible de tirer profit des informations saisies par le pharmacien à l'occasion de la revue d'ordonnances. Ces informations sont bien utilisées pour le soin individuel du patient, mais il n'est pas possible de les réutiliser à des fins d'études de groupes, et encore moins de les recouper avec les informations du dossier patient. Nous souhaitons rendre ces informations (revue d'ordonnance et informations sur le patient) disponibles pour l'évaluation et la recherche, dans un entrepôt de données pharmaceutiques, et aider les pharmaciens pour l'exploitation de ces données.

1.2.3 Pharmacovigilance nationale

Nous avons été sollicités récemment par la cellule de pharmacovigilance du CHU de Lille pour confirmer ou infirmer des signaux de pharmacovigilance. Il s'agit de médicaments potentiellement impliqués dans l'apparition d'une sclérose latérale amyotrophique et d'épisodes délirants aigus. Ces médicaments appartenant à la liste des molécules onéreuses, ils sont tracés dans les bases nationales du PMSI. La mise en œuvre de méthodes de survie à covariables temps-dépendantes devrait nous permettre de confirmer ou infirmer ces signaux en calculant des *hazard ratios* ajustés dans le cadre de véritables cohortes historiques en France entière.

Cet exemple illustre les possibilités données par le *data reuse* de bases nationales, qu'il s'agisse des bases du PMSI ou du SNIIRAM. En cas de besoin d'informations plus détaillées mais concernant des événements moins rares, il sera possible d'utiliser des bases de données de quelques établissements hospitaliers.

1.3 Défaut de qualité des soins en hospitalisation et perte d'autonomie

Nous avons vu que le **défaut de qualité des soins**, au sens de la non-adhésion aux recommandations, était responsable de morbi-mortalité. Alors que les approches traditionnelles consistent en des alertes interruptives (systèmes d'aide à la décision, CDSS), ou aux indicateurs de processus (pourcentages de conformité) mesurés par des experts, nous avons proposé un **nouveau paradigme** reprenant les avantages

des deux approches traditionnelles, et **les combinant** en tirant profit de la réutilisation des données massives de santé (*data reuse, big data*). Ce point a été détaillé en section 3.3, page 75.

Nous avons montré, à l'aide d'une revue de la littérature [171], que 22,3% des indicateurs pouvaient à dire d'expert être implémentés de manière à être calculés automatiquement par data reuse (soit **98 indicateurs** au moment de cette revue de la littérature).

Nous avons ensuite réalisé une expérimentation sur données réelles, en utilisant arbitrairement 9 indicateurs proposés en **chirurgie de la personne âgée** [170]. Nous avons montré que pour 5 d'entre eux, la valeur prédictive positive était supérieure à 85%, ce qui permettait leur utilisation en vie réelle.

Sur la base de cette seule revue de la littérature, **une cinquantaine d'indicateurs** pourrait ainsi être implémentée pour identifier les situations à risque et alimenter les processus d'analyse qualité plus classiques, tels les revues de morbidité et mortalité. Nous envisageons d'étendre minutieusement cette démarche, afin de développer à terme un **kit de screening rapide des défauts potentiels de qualité des soins**. Ce kit, tel les kits de screening de toxiques, permettrait d'identifier très rapidement des défauts potentiels de qualité. Il serait ensuite possible de les investiguer sans perdre de temps dans les domaines où la qualité ne fait pas défaut, à l'aide de méthodes plus traditionnelles. Ce kit permettrait également, après validation métrologique dans l'établissement, de suivre au long cours l'évolution des problèmes de qualité des soins.

1.4 Identification de facteurs de risque de maladies neurodégénératives

En collaboration avec l'UMR 1171 « Troubles cognitifs dégénératifs et vasculaires » dirigée par le Pr Régis Bordet, je co-encadre la thèse d'Université de Michaël Rochoy depuis décembre 2015. Cette thèse porte sur l'analyse de la base nationale du PMSI, afin de découvrir des facteurs de risque de démences, par data mining. L'approche se veut relativement neutre et systématique, afin d'identifier des facteurs de risque potentiels qui devront naturellement être confirmés par des approches plus traditionnelles. Ces facteurs de risque pourraient être de toutes natures : autres pathologies chroniques ou aiguës, actes thérapeutiques, dispositifs médicaux, médicaments (limités aux médicaments tracés dans le PMSI), ou même éléments de parcours du patient à travers les structures de soins.

Si cette approche est valide, elle pourra être utilisée de manière similaire pour identifier (ou valider) des facteurs de risque aux maladies chroniques multifactorielles ou d'étiologie inconnue. On peut citer par exemple la sclérose latérale amyotrophique.

1.5 Séquence de soins

1.5.1 Epidémiologie des soins nationale et parcours du patient

L'objectif est de décrire différentes prises en charge en termes de volume et répartition d'activité, puis d'identifier le devenir des patients, les complications et les facteurs de risque de ces complications. En collaboration avec les cliniciens du CHU de Lille, nous avons ainsi exploré les thèmes suivants, qui sont en cours de publication :

- Chirurgies de l'obésité et complications
- Césariennes
- Pose de dispositifs d'assistance ventriculaire implantables
- Pose de prothèses de hanche

Nous souhaitons également étendre ces travaux à la description sur données nationales de l'évolution de différentes pathologies chroniques, faisant alors référence à la notion de parcours à travers des stades pathologiques, comme par exemple : intolérance au glucose, diabète non-compliqué, diabète compliqué, etc.

1.5.2 Soins primaires

Tandis que de nombreux politiques et administrations appellent de leurs vœux la réorganisation du système de soins autour des soins primaires, les études scientifiques réalisées en soins primaires sont à ce jour relativement peu nombreuses, en particulier dans le champ du *data reuse*. Nous souhaitons développer des recherches dans ce domaine. En guise de preuve de concept, nous avons déjà entamé des recherches sur les thématiques suivantes :

- Incidence, prise en charge et facteurs de risque de gale et de récidive de gale
- Evolution au long cours du poids sous traitements contraceptifs et antiépileptiques
- Impact de la vaccination contre la grippe ou le pneumocoque
- Utilisation de l'électrocardiographe et de l'interprétation automatisée en soins primaires

Ces premières études permettent d'appréhender la faisabilité du *data reuse* sur des thématiques relativement simples. L'étape suivant est de développer une approche similaire dans la thématique de la perte d'autonomie.

1.5.3 Prévention de la chute en maison de retraite

Nous participons actuellement au projet ANR Prudence, dont la finalité est de détecter la chute des sujets âgés en maison de retraite, à l'aide de capteurs thermiques à bas coût. La mise en place d'une telle technologie permettrait à terme de secourir plus rapidement les personnes chutant en institution, limitant ainsi le risque d'hématome compressif, d'escarre, de rhabdomyolyse, de déshydratation, etc.

1.6 Support à la recherche : preuve de concept avant demande de financement

Tandis que jusque-là nous avons évoqué nos propres travaux de recherche, nous mettrons également ces données et méthodologiques au service des travaux des autres chercheurs, portant notamment sur la thématique de la perte d'autonomie du sujet âgé.

Nos propres travaux nous ont permis de mieux appréhender les bases de données utilisables, et d'automatiser ou systématiser certaines tâches intermédiaires très consommatrices de temps. Nous souhaitons par la suite proposer la réalisation de ce type de travaux sous la forme d'un service offert aux investigateurs qui souhaitent déposer une demande de financement, par exemple dans le cadre d'appels à projets. Ainsi, dans certains cas, il sera possible de réaliser une preuve de concept par *data reuse* avant même le dépôt d'une demande de financement pour un projet de recherche traditionnel (prospectif ou rétrospectif hors *data reuse*). L'approche par data reuse pourrait permettre par exemple de mieux estimer le nombre de sujets nécessaires, apporter des arguments de faisabilité, voire parfois d'apporter une preuve de concept, maximisant ainsi les chances que le projet soit financé.

2 Axe données

Les projets énumérés ci-dessus feront appel à différentes sources de données. Nous les présentons ci-dessous.

2.1 Bases nationales de données de santé (PMSI, SNIIRAM)

Nous poursuivons une partie de nos recherches sur les bases nationales du PMSI, comprenant actuellement approximativement 27 millions d'épisodes de court séjour par an, mais également des épisodes d'hospitalisation en psychiatrie, moyen séjour et hospitalisation à domicile. Après plusieurs années de gestion des données et analyses informatiques et statistiques, nous avons à présent atteint un niveau de maturité nous permettant de réaliser de nombreux types d'études sur cette base de données. Nous détenons les données de 2008 à 2014 (soit près de 250 millions de séjours), selon l'ancien cadre de mise à disposition des données par l'ATIH, et avec une autorisation CNIL suffisamment englobante. Cet accès direct aux données permet de réaliser des traitements avec les outils les plus adaptés, sans limitation autre que l'autorisation CNIL obtenue.

Le mode d'accès à ces données a désormais changé, avec la mise en place du SNDS (système national des données de santé), instauré par la loi du 26 janvier 2016 [185,186]. Ce nouveau mode d'accès implique de lancer les traitements à distance, sur un serveur contrôlé par le SNDS, doté d'outils limités. Les données postérieures à 2014 ne pourront jamais être accessibles autrement que via le SNDS.

Il nous faudra désormais entrer dans ce nouveau cadre, afin de poursuivre nos recherches. Les démarches règlementaires ont déjà été amorcées.

Parallèlement, nous souhaitons également être habilités à accéder à la base de données du SNIIRAM (système national d'identification inter-régimes de l'assurance maladie). Cette base de données inclut les données du PSMI, mais également les données de décès du CEPI-DC et les données de consommation ambulatoire de soins. Ces trois sources de données sont chaînées selon un identifiant unique de l'assuré. Cette base de données présente ainsi l'immense avantage de permettre d'étudier le parcours complet du patient, y compris la consommation de médicaments ou soins ambulatoires, et le décès extrahospitalier.

2.2 Entrepôt de données hospitalières

2.2.1 Au CHU de Lille

Comme nous l'avons développé dans ce mémoire, les dossiers médicaux représentent une source d'information très importante pour la recherche scientifique. Ces données sont actuellement peu valorisées dans notre établissement. Nous avons pour projet de mettre en place un entrepôt de données cliniques permettant d'exploiter ces données à des fins de recherche mais également d'amélioration et d'étude des processus au CHU de Lille. Nous avons mis en place et animons un **groupe de travail** « accès aux données du dossier patient à des fins d'analyse », composé de **12 membres** dont 4 PU-PH. Nous avons également mené une série de **27 entretiens**, notamment avec chaque chef de pôle du CHU de Lille.

Une **charte de réutilisation des données** a ainsi été élaborée (actuellement 35 pages). Au fil des entretiens, cette charte a évolué et semble actuellement recueillir un **véritable consensus** auprès des Chefs de Pôles, sur les principes proposés. Ces principes garantissent en même temps la souplesse d'utilisation, et le contrôle complet des pôles sur l'utilisation qui est faite de leurs données.

Durant le même temps, sous l'impulsion du pôle d'anesthésie réanimation et du CIC-IT du CHU de Lille, un entrepôt de données relatives à l'anesthésie, basé en particulier sur le logiciel d'anesthésie Diane, a été mis en place et est actuellement géré dans notre service par Antoine Lamer, ingénieur de recherche. Nous avons pour projet d'exploiter cette base de données en collaboration avec le pôle d'anesthésie réanimation.

Les différents entretiens menés et l'expérience que nous avons déjà acquise sur toutes les étapes du traitement indiquent qu'il n'existe plus d'obstacle technique insurmontable à la mise en place d'un entrepôt incluant tous les pôles, et concernant les données du PMSI, les résultats de biologie, les médicaments prescrits et les courriers et comptes-rendus.

2.2.2 Réseau de données d'établissements de la Région

Nous avons dans le cadre du projet PSIP mis en œuvre un modèle de données incluant les données du PMSI, les résultats de biologie, les médicaments prescrits et les courriers et comptes-rendus. Nous avons pu mettre en œuvre également toutes les étapes préalables à l'analyse de données, en particulier :

- Les étapes d'ETL (*export tranform and load*)
- Les étapes de contrôle qualité et correction de données
- Les étapes d'alignement terminologique (*mapping* et conversion d'unités des données de biologie [91], rajout de codes ATC aux médicaments sans code telles les perfusions, etc.)
- L'anonymisation des données structurées et des courriers en texte libre [97]

Ces opérations ont été réalisées par nos soins au CH de Denain, au CH de Valenciennes, et sous notre supervision dans deux hôpitaux danois, un hôpital bulgare et au CHU de Rouen. Ces expériences très diverses et toutes réussies nous ont permis de généraliser le processus complet, au point d'envisager sa portabilité. Par ailleurs, nous avons développé de bonnes relations avec de nombreux hôpitaux de la région Hauts de France.

Nous souhaitons proposer à une dizaine d'hôpitaux de la région Hauts de France d'installer chez eux un entrepôt de données selon un modèle de données uniformisé. Nous pourrions ainsi développer des programmes de recherche sur ces données. Les scripts d'analyse (programmation en statistiques), une fois développés sur des données test, pourraient être exécutés sur chaque site sans effort marginal. Les établissements, s'ils l'acceptent, pourraient ainsi partager les résultats d'analyse et participer aux projets de recherche, tout en contrôlant leurs données et leur utilisation. Les composants développés par l'un ou l'autre pourraient immédiatement profiter aux autres établissements.

2.3 Données ambulatoires

2.3.1 Cabinets de médecine générale

Le partenariat mis en place avec un cabinet de groupe de médecine générale de la métropole lilloise nous donne l'opportunité d'analyser depuis quelques mois les données relatives au suivi de 5 500 patients différents, soit près de 100 000 consultations. Nous avons pour l'instant initié des recherches portant sur des sujets exclusivement centrés sur la prise en charge ambulatoire, sans lier ces données aux séjours hospitaliers des patients.

Ces données comprennent notamment :

- Des données sur le patient (démographie, couverture assurancielle)
- Des données sur les contacts (dates, motifs, diagnostics)
- Des données sur les médicaments prescrits (produit, dose, indication thérapeutique)
- Les résultats de biologie transmis par les laboratoires de biologie médicale et automatiquement importés
- Des données de vaccination

2.3.2 Réseau de données ambulatoires

Le marché du logiciel de gestion des patients des médecins généralistes est relativement concentré. Ainsi, alors que près de 90% des médecins généralistes seraient informatisés, quatre éditeurs se partagent 70% du marché [187]. Cela signifie qu'un nombre important de médecins généralistes disposent des mêmes

modules d'extraction de données. Un exemple de données disponibles est présenté en section 2.3.1 page 97.

Nous souhaitons constituer et animer un réseau de médecins généralistes, et leur proposer de constituer chez eux un entrepôt de données selon un modèle uniformisé. De la même manière que pour les hôpitaux, nous développerions des programmes de recherche sur ces données. Les scripts d'analyse correspondants seraient exécutés chez les médecins. Ils pourraient ainsi partager les résultats d'analyse et participer aux projets de recherche, tout en contrôlant leurs données et leur utilisation. Des bénéfices secondaires pourraient leur être proposés, comme des modules de rappel d'actes à proposer (vaccination ou examens périodiques selon le profil pathologiquee, etc.).

2.3.3 Chaînage entre données hospitalières et ambulatoires

L'analyse de données ambulatoires seules peut souvent s'avérer insuffisante. En voici une illustration simple. On peut se demander si les vaccinations antipneumococciques diminuent la morbidité chez les patients atteints de BPCO. Les données ambulatoires sont indispensables car la vaccination est réalisée en ville, et que les consultations sont des événements suffisamment fréquents pour permettre une détection sensible des exacerbations. Néanmoins, il n'aurait aucun sens d'évaluer la variation du nombre de consultations, si dans le même temps les patients hospitalisés pour exacerbation de BPCO ne sont pas détectés.

Notre objectif est donc de pouvoir mettre en place des entrepôts de données croisant des données hospitalières et ambulatoires des mêmes patients. Sous l'impulsion du Dr Jean-Baptiste Beuscart, PHU dans notre équipe de recherche, et dans le cadre du projet Paerpa [184], nous avons obtenu l'autorisation de réunir dans un même entrepôt les données hospitalières du CH de Denain, du CH de Valenciennes, et les données ambulatoires du territoire correspondant. Les données pourront être chaînées. L'analyse de ces données est très prometteuse.

2.4 Dispositifs médicaux

En sus des données décrites précédemment (PMSI, biologie, médicaments, courriers), nous anticipons que les données analysables intègreront de plus en plus de **données massives produites en routine par les dispositifs médicaux**. Cela posera de nouveaux problèmes techniques liés à la dimension de ces données, mais surtout au besoin de qualifier médicalement des données de signal, qui sont fondamentalement « neutres ». Le croisement de ces données avec les autres données disponibles (notamment pathologies et traitements) sera riche d'enseignements. Afin de préparer cette évolution de la thématique, je collabore de plus en plus avec les **CIC-IT** de Lille et de Rennes, notamment à travers le **projet ANR Prudence** accepté en 2016.

3 Axe méthodologies

Les projets énumérés ci-dessus nécessiteront pour certains d'entre eux de mettre au point et évaluer de nouvelles méthodes d'analyse. Nous les présentons ci-dessous.

3.1 Visualisation du parcours du patient

Nous avons par le passé conçu et développé une application de représentation graphique du parcours des patients à travers les unités médicales d'un établissement hospitalier [18]. Au fil du projet **ANR CLINMINE**, le besoin de composants graphiques plus génériques permettant visualiser les parcours s'est fait ressentir. L'ensemble des chercheurs du projet s'est accordé sur plusieurs nécessités :

- Abstraire la notion de parcours à travers des états discrets (parcours à travers des établissements, des unités médicales, des modalités de prise en charge, des stades d'une pathologie, etc.)
- Identifier les composants graphiques existants pouvant être réutilisés
- Proposer et développer de nouveaux composants graphiques
- De la même manière, identifier les composants statistiques existant, dont ceux produits à l'occasion du projet CLINMINE, et en proposer de nouveaux
- Rendre disponibles ces anciens et nouveaux composants graphiques et statistiques à travers une interface unique sous la forme d'une librairie développée pour R [188]
- Développer un langage de requête ou étendre un langage de requête existant, afin de permettre des requêtes gérant à la fois la notion de lieu et de temps.

Ces idées devraient constituer les fondements d'un **dépôt de projet ANR, CLINMINE 2**, prévu pour 2018.

3.2 Visualisation des données individuelles du patient

Les projets de réutilisation de données que nous avons menés nous ont fréquemment conduits à visualiser les données individuelles du patient, à des fins variées :

- contrôle qualité des données
- contrôle qualité des processus (validation des algorithmes de visu)
- validation experte des signaux produits de manière automatisée (sensibilité, spécificité, etc.)

Nous avons pour ce faire développé une application de visualisation des séjours au décours du projet PSIP [136]. Les approches de visualisation actuelles, qu'elles relèvent des logiciels actuels de dossier patient ou d'applications de recherche telle la nôtre, nous semblent relativement pauvres et trop dépendantes du type de données visualisées.

Nous souhaitons mener des recherches pour améliorer les interfaces graphiques de visualisation de données individuelles (par exemple le séjour d'un patient, ou la période de suivi d'un patient), intégrant les données que nous rencontrons habituellement (par exemple séquences et mouvements du patient, diagnostics, actes, résultats de biologie, médicaments, courriers, etc.) en gérant à la fois :

- Un axe temporel unique (*timeline*)
- La notion de hiérarchie des terminologies représentées
- La possibilité de réduire ou développer l'information
- La possibilité de faire apparaître à un moment donné toutes les informations relatives à un même concept mais de types différents (exemple pour l'hémostase : diagnostics de troubles de l'hémostase, actes d'hémostase, INR, TP, plaquettes, facteurs de coagulation, AVK, vitamine K, etc.)

Si cette démarche aboutit, elle pourrait à la fois faciliter les revues de cas dans le cadre des recherches cliniques, mais également avoir un impact favorable sur les interfaces de logiciels de dossier patient.

3.3 Analyse de données spatiales

Nous souhaitons développer et « packager » des méthodes d'analyse de données spatiales, afin d'améliorer la détection de concentrations géographiques des cas, et de permettre la réalisation d'études écologiques, croisant les données analysées avec des indicateurs écologiques agrégés produits par exemple par l'INSEE (données de pollution, de défaveur sociale, etc.). Cette approche exige à la fois de maîtriser les statistiques de *scan*, les problèmes plus prosaïques d'alignement terminologique géographique, etc.

Ces approches sont actuellement développées dans notre équipe de recherche par Michaël Génin, MCU de Statistiques. Nous souhaitons contribuer à son développement, et proposer des applications sur les sources de données citées précédemment.

3.4 Analyse de données temporelles

Nous souhaitons développer et « packager » des méthodes d'analyse de données temporelles, qu'il s'agisse par exemple de détection de séquences d'événements, de transitions entre plusieurs états de pathologie ou de prise en charge, de risques concurrents, etc.

Ces approches sont actuellement développées dans notre équipe de recherche par Evgeniya Babykina (MCU en statistiques) et Jean-Baptiste Beuscart (PHU en gériatrie). Nous souhaitons contribuer à son développement, et proposer des applications sur les sources de données citées précédemment.

4 Collaborations

Au fil des trois axes développés dans notre prospective de recherche (thématique, données, méthodologies), nous avons ainsi évoqué plusieurs collaborations déjà amorcées :

- Des collaborations au CHU de Lille :
  - avec le service de pharmacologie et l'UMR 1171
  - avec le service de pharmacie
  - avec le pôle de Gériatrie
  - avec le CIC-IT, incluant notamment le laboratoire Evalab
  - avec les différentes spécialités cliniques représentées au CHU
  - avec le DIM et la DRN (direction informatique)
- Des collaborations régionales avec des établissements de soins :
  - Avec le CH de Valenciennes et le CH de Denain puis, nous l'espérons, avec une dizaine de CH de la région Hauts de France
  - Avec un cabinet de groupe de la métropole lilloise puis, nous l'espérons, avec un réseau étendu de médecins généralistes
- Des collaborations scientifiques régionales ou nationales :
  - Avec Inria (équipe Modal à Lille) et le CNRS (équipe Cristal à Lille)
  - Avec l'Inserm (LTSI UMR 1099 à Rennes)
- Des collaborations scientifiques internationales
  - En particulier avec le service du Dr David w. Bates, au BWH de Boston et à Harvard University

Je souhaite rester rattaché à l'EA2694 « Santé Publique, épidémiologie et qualité des soins », et participer aux évolutions ou recompositions de cette équipe à partir de l'année 2019.

Liste des publications

1 Articles publiés dans des revues internationales à comité de lecture (n=46)

Ces 46 articles se répartissent ainsi : 16 en premier auteur, 10 en deuxième auteur, 7 en dernier auteur et 11 en position autre. Ces articles sont positionnés sur notre schéma de *data mining* et *data reuse* en annexe. La liste complète est présentée ci-dessous :

1. Ficheur G, Caron A, Beuscart J-B, Ferret L, Jung Y-J, Garabedian C, et al. Case-crossover study to examine the change in postpartum risk of pulmonary embolism over time. BMC Pregnancy Childbirth 2017;17:119. doi:10.1186/s12884-017-1283-y.
2. Chazard E, Ficheur G, Beuscart J-B, Preda C. How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests. Value in Health 2017. doi:10.1016/j.jval.2017.02.009.
3. Chazard E, Beeler PE, Ficheur G, Dalleur O, Beuscart R, Bates DW. Drug-drug interactions with Vitamin K antagonists: are we focusing on the right rules? [IN PRESS] 2017.
4. Rosier A, Mabo P, Temal L, Van Hille P, Dameron O, Deleger L, et al. Remote Monitoring of Cardiac Implantable Devices: Ontology Driven Classification of the Alerts. Stud Health Technol Inform 2016;221:59–63.
5. Petit A-E, Mangeard H, Chazard E, Puisieux F. [Changes in drug management of Alzheimer's disease in nursing homes: Impact of the media campaign against specific drugs for Alzheimer's disease]. Encephale 2016. doi:10.1016/j.encep.2015.03.006.
6. Ficheur G, Schaffar A, Caron A, Balcaen T, Beuscart J-B, Chazard E. Elderly Surgical Patients: Automated Computation of Healthcare Quality Indicators by Data Reuse of EHR. Stud Health Technol Inform 2016;221:92–6.
7. Ficheur G, Caron A, Beuscart J-B, Ferret L, Putman S, Beuscart R, et al. The risks of pulmonary embolism and upper gastrointestinal bleeding beyond 35days after total hip replacement for coxarthrosis among middle-aged patients: A cross-over cohort. Prev Med 2016. doi:10.1016/j.ypmed.2016.09.010.
8. Ferret L, Luyckx M, Ficheur G, Chazard E, Beuscart R. Evaluation of a Computer Application for Retrospective Detection of Vitamin K Antagonist Treatment Imbalance. J Patient Saf 2016. doi:10.1097/PTS.0000000000000182.
9. Caron A, Chazard E, Muller J, Perichon R, Ferret L, Koutkias V, et al. IT-CARES: an interactive tool for case-crossover analyses of electronic medical records for patient safety. J Am Med Inform Assoc 2016. doi:10.1093/jamia/ocw132.
10. Rosier A, Mabo P, Temal L, Van Hille P, Dameron O, Deléger L, et al. Personalized and automated remote monitoring of atrial fibrillation. Europace 2015. doi:10.1093/europace/euv234.
11. Perichon R, Chazard E, Beuscart R. Patients drug exchange forum corpus: toward drug safety signals detection. Stud Health Technol Inform 2015;210:1023.

12. Frély A, Chazard E, Pansu A, Beuscart J-B, Puisieux F. Impact of acute geriatric care in elderly patients according to the Screening Tool of Older Persons' Prescriptions/Screening Tool to Alert doctors to Right Treatment criteria in northern France. Geriatr Gerontol Int 2015. doi:10.1111/ggi.12474.

13. Ficheur G, Ferreira Careira L, Beuscart R, Chazard E. EpiHosp: A web-based visualization tool enabling the exploratory analysis of complications of implantable medical devices from a nationwide hospital database. Stud Health Technol Inform 2015;210:409–13.

14. Ferret L, Beuscart J-B, Ficheur G, Beuscart R, Luyckx M, Chazard E. Evaluation of compliance with recommendations of prevention of thromboembolism in atrial fibrillation in the elderly, by data reuse of electronic health records. Stud Health Technol Inform 2015;210:394–8.

15. Djennaoui M, Ficheur G, Beuscart R, Chazard E. Improvement of the quality of medical databases: data-mining-based prediction of diagnostic codes from previous patient codes. Stud Health Technol Inform 2015;210:419–23.

16. Chazard E, Marcolino MS, Dumesnil C, Caron A, Palhares DMF, Ficheur G, et al. One Million Electrocardiograms of Primary Care Patients: A Descriptive Analysis. Stud Health Technol Inform 2015;216:69–73.

17. Chazard E, Dumesnil C, Beuscart R. How much does hyperkalemia lengthen inpatient stays? About methodological issues in analyzing time-dependant events. Stud Health Technol Inform 2015;210:835–9.

18. Chazard E, Babaousmail D, Schaffar A, Ficheur G, Beuscart R. Process assessment by automated computation of healthcare quality indicators in hospital electronic health records: a systematic review of indicators. Stud Health Technol Inform 2015;210:867–71.

19. Caron A, Clement G, Heyman C, Aernout E, Chazard E, Le Tertre A. An original imputation technique of missing data for assessing exposure of newborns to perchlorate in drinking water. Stud Health Technol Inform 2015;210:860–4.

20. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. Biomed Res Int 2015;2015. doi:10.1155/2015/639021.

21. Koutkias VG, McNair P, Kilintzis V, Skovhus Andersen K, Niès J, Sarfati J-C, et al. From adverse drug event detection to prevention. A novel clinical decision support framework for medication safety. Methods Inf Med 2014;53:482–92. doi:10.3414/ME14-01-0027.

22. Ficheur G, Chazard E, Beuscart J-B, Merlin B, Luyckx M, Beuscart R. Adverse drug events with hyperkalaemia during inpatient stays: evaluation of an automated method for retrospective detection in hospital databases. BMC Med Inform Decis Mak 2014;14:83. doi:10.1186/1472-6947-14-83.

23. Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inform 2014;83:303–12. doi:10.1016/j.ijmedinf.2013.11.005.

24. Pruvost I, Dubos F, Chazard E, Hue V, Duhamel A, Martinot A. The value of body weight measurement to assess dehydration in children. PLoS ONE 2013;8:e55063. doi:10.1371/journal.pone.0055063.

25. Hackl WO, Ammenwerth E, Marcilly R, Chazard E, Luyckx M, Leurs P, et al. Clinical evaluation of the ADE scorecards as a decision support tool for adverse drug event analysis and medication safety management. Br J Clin Pharmacol 2013;76 Suppl 1:78–90. doi:10.1111/bcp.12185.

26. Ficheur G, Chazard E, Merlin B, Ferret L, Luyckx M, Beuscart R. Supervised analysis of drug prescription sequences. Stud Health Technol Inform 2013;192:293–7.

27. Ferret L, Luyckx M, Merlin B, Ficheur G, Chazard E, Beuscart R. Evaluation of a computerized tool allowing retrospective detection of potential vitamin K antagonist overdoses in complex contexts. Stud Health Technol Inform 2013;192:553–6.

28. Chazard E, Luyckx M, Beuscart J-B, Ferret L, Beuscart R. Routine use of the “ADE scorecards”, an application for automated ADE detection in a general hospital. Stud Health Technol Inform 2013;192:308–12.

29. Franco Contreras J, Coatrieux G, Chazard E, Cuppens F, Cuppens-Boulahia N, Roux C. Robust lossless watermarking based on circular interpretation of bijective transformations for the protection of medical databases. Conf Proc IEEE Eng Med Biol Soc 2012;2012:5875–8. doi:10.1109/EMBC.2012.6347330.

30. Chazard E, Bernonville S, Ficheur G, Beuscart R. A statistics-based approach of contextualization for adverse drug events detection and prevention. Stud Health Technol Inform 2012;180:766–70.

31. Marcilly R, Chazard E, Beuscart-Zéphir M-C, Hackl W, Baceanu A, Kushniruk A, et al. Design of Adverse Drug Events-Scorecards. Stud Health Technol Inform 2011;164:377–81.

32. Ficheur G, Chazard E, Schaffar A, Genty M, Beuscart R. Interoperability of medical databases: construction of mapping between hospitals laboratory results assisted by automated comparison of their distributions. AMIA Annu Symp Proc 2011;2011:392–401.

33. Coatrieux G, Chazard E, Beuscart R, Roux C. Lossless watermarking of categorical attributes for verifying medical data base integrity. Conf Proc IEEE Eng Med Biol Soc 2011;2011:8195–8. doi:10.1109/IEMBS.2011.6092021.

34. Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed 2011;15:823–30. doi:10.1109/TITB.2011.2165727.

35. Chazard E, Baceanu A, Ferret L, Ficheur G. The ADE scorecards: a tool for adverse drug event detection in electronic health records. Stud Health Technol Inform 2011;166:169–79.

36. Merlin B, Chazard E, Pereira S, Serrot E, Sakji S, Beuscart R, et al. Can F-MTI semantic-mined drug codes be used for adverse drug events detection when no CPOE is available? Stud Health Technol Inform 2010;160:1025–9.

37. Koutkias V, Kilintzis V, Stalidis G, Lazou K, Collyda C, Chazard E, et al. Constructing Clinical Decision Support Systems for Adverse Drug Event Prevention: A Knowledge-based Approach. AMIA Annu Symp Proc 2010;2010:402–6.

38. Leroy N, Chazard E, Beuscart R, Beuscart-Zephir MC, Psip Consortium. Toward automatic detection and prevention of adverse drug events. Stud Health Technol Inform 2009;143:30–5.

39. Chazard E, Preda C, Merlin B, Ficheur G, PSIP consortium, Beuscart R. Data-mining-based detection of adverse drug events. Stud Health Technol Inform 2009;150:552–6.

40. Chazard E, Merlin B, Ficheur G, Sarfati J-C, PSIP Consortium, Beuscart R. Detection of adverse drug events: proposal of a data model. Stud Health Technol Inform 2009;148:63–74.

41. Chazard E, Ficheur G, Merlin B, Serrot E, PSIP Consortium, Beuscart R. Adverse drug events prevention rules: multi-site evaluation of rules from various sources. Stud Health Technol Inform 2009;148:102–11.

42. Chazard E, Ficheur G, Merlin B, Genin M, Preda C, PSIP consortium, et al. Detection of adverse drug events detection: data aggregation and data mining. Stud Health Technol Inform 2009;148:75–84.

43. Baceanu A, Atasiei I, Chazard E, Leroy N, PSIP Consortium. The expert explorer: a tool for hospital data visualization and adverse drug event rules validation. Stud Health Technol Inform 2009;148:85–94.

44. Puech P, Chazard E, Lemaitre L, Beuscart R. DicomWorks Teleradiology: secure transmission of medical images over the internet at low cost. Conf Proc IEEE Eng Med Biol Soc 2007;2007:6706–9. doi:10.1109/IEMBS.2007.4353899.

45. Chazard E, Beuscart R. Graphical representation of the comprehensive patient flow through the hospital. AMIA Annu Symp Proc 2007:110–4.

46. Chazard E, Puech P, Gregoire M, Beuscart R. Using Treemaps to represent medical data. Stud Health Technol Inform 2006;124:522–7.

2 Articles dans des revues nationales à comité de lecture (n=5)

La liste complète est présentée ci-dessous :

1. Beuscart J-B, Ficheur G, Miqueu M, Luyckx M, Perichon R, Puisieux F, et al. Co-prescriptions of psychotropic drugs to older patients in a general hospital. European Geriatric Medicine 2017;8:84–9. doi:10.1016/j.eurger.2016.11.012.

2. Contreras JF, Coatrieux G, Cuppens F, Cuppens-Boulahia N, Chazard E, Roux C. Tatouage Robuste et Réversible pour la Traçabilité de Bases de Données Relationnelles en Santé. ResearchGate, 2013.

3. Chazard E, Baceanu A, Ficheur G, Marcilly R, Beuscart R. Les «ADE Scorecards»: Un outil de détection par data mining des effets indésirables liés aux médicaments dans les dossiers médicaux (projet PSIP). In: Staccini PPM, Harmel DA, Darmoni PSJ, Gouider PR, editors. Systèmes d'information pour l'amélioration de la qualité en santé, Springer Paris; 2011, p. 177–88. doi:10.1007/978-2-8178-0285-5_16.

4. Beuscart R, Chazard E, Souf N. De l'innovation au remboursement. IRBM 2010;31:26–9. doi:10.1016/j.irbm.2009.11.004.

5. Chazard E, Preda C, Merlin B, Ficheur G, Beuscart R. Détection et prévention des effets indésirables liés aux médicaments par data-mining. IRBM 2009;30:192–6. doi:10.1016/j.irbm.2009.05.008.

3 Communications avec actes (n=23)

La liste complète est présentée ci-dessous :

1. Martincic C, Balcaen T, Georges A, Baro E, Ficheur G, Chazard E. La chirurgie bariatrique en France de 2008 à 2014 : triplement de l'activité et fort recul de l'anneau gastrique. Revue d'Épidémiologie et de Santé Publique 2017;65, Supplement 1:S20. doi:10.1016/j.respe.2017.01.044.

2. Balcaen T, Chazard E, Ganry O, Caillet P. Validité de la mesure de l'incidence des cancers en France à partir de la base de données du Programme de médicalisation des systèmes d'information : revue systématique de la littérature de 2001 à 2015. Revue d'Épidémiologie et de Santé Publique 2017;65, Supplement 1:S28. doi:10.1016/j.respe.2017.01.066.

3. Houyengah F, Kyndt X, Durand-Joly I, Blanckaert K, Janssoone N, Chazard E. Les départements de l'information médicale, alerteurs dans le parcours des patients à risque à l'hôpital. Expérience des CH de Valenciennes, Dunkerque et Romorantin. Revue d'Épidémiologie et de Santé Publique 2016;64, Supplement 1:S22. doi:10.1016/j.respe.2016.01.070.

4. Chazard E, Preda C, Beuscart R. Comparer la durée de séjour de deux groupes de patients : quel test choisir ? Comparaison des risques alpha et bêta de douze tests statistiques. Revue

d'Épidémiologie et de Santé Publique 2016;64, Supplement 1:S32–3. doi:10.1016/j.respe.2016.01.013.

5. Chazard E, Ficheur G, Beuscart R. Risque hémorragique sous anti-vitamines K : quelles sont réellement les interactions prioritaires ? Revue d'Épidémiologie et de Santé Publique 2016;64, Supplement 1:S11–2. doi:10.1016/j.respe.2016.01.041.

6. Muller J, Ficheur G, Ferreira Carreira L, Chazard E, Beuscart R. Réutilisation du fichier FichComp de la base nationale du Programme de médicalisation de systèmes d'information pour explorer les complications mécaniques des prothèses totales de hanche. Revue d'Épidémiologie et de Santé Publique 2015;63, Supplement 1:S14–5. doi:10.1016/j.respe.2015.01.029.

7. Ficheur G, Ferreira Carreira L, Chazard E, Beuscart R. EpiHosp : un outil web permettant l'exploration des poses de dispositifs médicaux implantables et de leurs complications par la réutilisation de la base nationale du PMSI. Revue d'Épidémiologie et de Santé Publique 2015;63, Supplement 1:S10. doi:10.1016/j.respe.2015.01.018.

8. Djennaoui M, Ficheur G, Aernout E, Beuscart R, Chazard E. Construction et évaluation de règles de prédiction de diagnostics à partir des bases de données hospitalières : application au contrôle qualité des données médico-administratives. Revue d'Épidémiologie et de Santé Publique 2015;63, Supplement 1:S11. doi:10.1016/j.respe.2015.01.020.

9. Caron A, Clément G, Heyman C, Aernout E, Chazard E, Le Tertre A. Détermination de l'exposition de 394 979 nouveau-nés par imputation multiple de données manquantes dans une étude épidémiologique. Revue d'Épidémiologie et de Santé Publique 2015;63, Supplement 1:S9. doi:10.1016/j.respe.2015.01.016.

10. Schaffar A, Babaousmail D, Ficheur G, Beuscart R, Chazard E. Étude de la faisabilité de l'implémentation d'indicateurs automatisés de la qualité des soins en France. Revue d'Épidémiologie et de Santé Publique 2014;62, Supplement 3:S81. doi:10.1016/j.respe.2014.01.033.

11. Dumesnil C, Beuscart R, Chazard E. Comparer les durées de séjour selon qu'un événement indésirable temps-dépendant survient : évaluation et correction du risque de première espèce. Revue d'Épidémiologie et de Santé Publique 2014;62, Supplement 3:S99. doi:10.1016/j.respe.2014.01.090.

12. Chazard E, Dumesnil C, Marcolino MS, Caron A, Alkmim MB, Pinho-Ribeiro AL. Exploitation automatisée des données électrocardiographiques pour le codage : mise en place et évaluation. Revue d'Épidémiologie et de Santé Publique 2014;62, Supplement 3:S76. doi:10.1016/j.respe.2014.01.017.

13. Aernout E, Ficheur G, Djennaoui M, Chazard E, Beuscart R. Codage automatisé à partir des comptes-rendus d'actes : construction et évaluation de règles de prédiction par une méthode mixte associant fouille de texte et validation experte. Revue d'Épidémiologie et de Santé Publique 2014;62, Supplement 3:S93. doi:10.1016/j.respe.2014.01.070.

14. Ficheur G, Genty M, Chazard E, Flament C, Beuscart R. Proposition d'une méthode automatisée calculant la valeur moyenne d'un diagnostic associé significatif. Revue d'Épidémiologie et de Santé Publique 2013;61, Supplement 1:S18–9. doi:10.1016/j.respe.2013.01.047.

15. Chazard E, Miquel P-H, Genty M, Beuscart R. « Planifadmission », outil open-source d'aide à la planification des admissions programmées basé sur une prédiction statistique des durées de séjour. Revue d'Épidémiologie et de Santé Publique 2013;61, Supplement 1:S5–6. doi:10.1016/j.respe.2013.01.005.

16. Ficheur G, Beuscart J-B, Schaffar A, Chazard E. Interopérabilité des bases de données médicales : proposition d'une méthode de mise en correspondance des bases de biologie optimisant leur exploitation. Revue d'Épidémiologie et de Santé Publique 2012;60, Supplement 1:S19. doi:10.1016/j.respe.2011.12.114.

17. Chazard E, Mouret-Kubiak C, Ficheur G, Beuscart R. Déidentification automatisée de courriers médicaux : la méthode FASDIM. Revue d'Épidémiologie et de Santé Publique 2012;60, Supplement 1:S18. doi:10.1016/j.respe.2011.12.112.

18. Chazard E, Ficheur G, Bernonville S, Beuscart J-B, Beuscart R. Contextualisation pour la détection et la prévention des effets indésirables médicamenteux. Revue d'Épidémiologie et de Santé Publique 2012;60, Supplement 2:S143–4. doi:10.1016/j.respe.2012.06.376.

19. Ficheur G, Chazard E, Messai R, Beuscart R. Codage automatisé : proposition d'une méthode utilisant une ontologie médicale construite par fouille de textes. Revue d'Épidémiologie et de Santé Publique 2011;59, Supplement 2:S51. doi:10.1016/j.respe.2011.03.026.

20. Chazard E, Ficheur G, Baceanu A, Marcilly R, Beuscart R. Les « ADE Scorecards », outil de détection et visualisation des effets indésirables médicamenteux. Revue d'Épidémiologie et de Santé Publique 2011;59, Supplement 2:S54. doi:10.1016/j.respe.2011.03.036.

21. Genty M, Chazard E, Legrand B, Beuscart R. Réduction du temps d'attente des patients et des médecins libéraux par mutualisation des patientèles en cabinet de groupe. Revue d'Épidémiologie et de Santé Publique 2010;58, Supplement 1:S21. doi:10.1016/j.respe.2010.02.053.

22. Chazard E, Salleron J, Génin M, Ficheur G, Duhamel A. Détection et prévention des effets indésirables médicamenteux par fouille automatisée des dossiers patients électroniques. Revue d'Épidémiologie et de Santé Publique 2010;58, Supplement 1:S8. doi:10.1016/j.respe.2010.02.013.

23. Grenier JL, d'Escrivan N, Vincent D, Chazard E, Lepeut M. P107 Évaluation rétrospective de 20 ans d'activité d'une unité de diabétologie centrée sur une éducation thérapeutique du patient – Étude préliminaire. Diabetes & Metabolism 2008;34, Supplement 3:H74. doi:10.1016/S1262-3636(08)73019-2.

4 Ouvrages et chapitres pédagogiques (n=2)

J'ai contribué à un chapitre de deux ouvrages ci-dessous, portés par le CIMES (Collège des enseignants d'Informatique médicale, bioMathématiques, méthodes en Epidémiologie et Statistiques) :

1. Venot, Alain, Burgun, Anita, Quantin, Catherine. Medical Informatics, e-Health - Fundamentals and Applications. Paris, France: Springer; 2013.

2. Venot, Alain, Burgun, Anita, Quantin, Catherine. Informatique Médicale, e-Santé – Fondements et applications. Paris, France: Springer; 2013.

5 Rapports (n=5)

Les rapports ci-dessous ont été produits dans le cadre de projets européens FP7 ou de projets ANR :

1. Livrable D1.1 : données médicales, confidentialité et secret médical. Projet ANR Clinmine. Chazard, E.; Ficheur, G.; and Perichon, R. Technical Report Agence Nationale de la Recherche, Lille, France, February 2014.
2. Final set of CDSS modules. PSIP deliverable 5.2. Andersen, K. S.; Koutkias, V.; Frandsen, J. R.; Jensen, S.; McNair, P.; Baceanu, A.; Chazard, E.; Niès, J.; and Pereira, S. Technical Report European Research Council, July 2010.
3. Results of data & semantic mining. PSIP deliverable 2.3. Chazard, E.; Preda, C.; Bernonille, S.; Băceanu, A.; Ficheur, G.; Genty, M.; Darmoni, S.; Sakji, S.; Pereira, S.; Tessier, S.; Saur, F.; Serrot, E.; Kergourlay, I.; Beuscart, R.; and Cacciabue, C. Technical Report European Research Council, August 2010.
4. First results of data mining. PSIP deliverable 2.1. Chazard, E.; Preda, C.; Beuscart, R.; Băceanu, A.; and Niculescu, C Technical Report European Research Council, August 2008.
5. Structures and Data Models of the Data repositories available in the PSIP project. PSIP deliverable 1. Bernard, O.; Koncar, M; Sarfati, J.; Chazard, E.; and Niès, J. Technical Report European Research Council, April 2008.

6 Mémoires académiques (n=5)

Au cours de mes études, j'ai produit les mémoires et thèses suivants :

1. Automated detection of adverse drug events by data mining of electronic health records. Chazard, E. Ph.D. Thesis, Université du Droit et de la Santé - Lille II, Lille, France, February 2011.
2. Data Mining et prévention des effets indésirables liés aux médicaments. Chazard, E. Master Thesis, Université Paris Sud, Paris XI, Paris, France, 2008.
3. Etude rétrospective de 17 ans d'activité du Centre d'Education pour le TRAitement du DIabète et des Maladies de la Nutrition du CHG de Roubaix : traitement de base de données, analyse, indicateurs d'utilité. Chazard, E. Medicine Thesis, Université du Droit et de la Santé - Lille II, Lille, France, January 2006.
4. Les représentations graphiques, support de la décision de gestion hospitalière. Chazard, E. Master Thesis, Université des Sciences et Technologies, Lille 1, Lille, France, 2006.
5. DIND PMSI : Dind Is Not Datim - Système de détection d'atypies dans les fichiers de RSS. Chazard, E. Master Thesis, Université du Droit et de la Santé, Lille 2, Lille, France, 2005.

Tirés à part de publications significatives

Les publications ci-après sont des tirés à part. Les numérotations des pages et les références bibliographiques sont donc autonomes.

VALUE IN HEALTH ▮ (2017) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests

Emmanuel Chazard, MD, PhD[1,2,*], *Grégoire Ficheur, MD, PhD*[1,2], *Jean-Baptiste Beuscart, MD, PhD*[1,3], *Cristian Preda, PhD*[4,5]

[1]*EA2694 Santé publique: épidémiologie et qualité des soins, Université Lille, Lille, France;* [2]*Public Health Department, CHU Lille, Lille, France;* [3]*Geriatrics Department, CHU Lille, Lille, France;* [4]*Laboratory of Mathematics Paul Painlevé, Université Lille, Lille, France;* [5]*Inria Lille Nord Europe, MODAL, Villeneuve-d'Ascq, France*

ABSTRACT

Background: Although many researchers in the field of health economics and quality of care compare the length of stay (LOS) in two inpatient samples, they often fail to check whether the sample meets the assumptions made by their chosen statistical test. In fact, LOS data show a highly right-skewed, discrete distribution in which most of the observations are tied; this violates the assumptions of most statistical tests. **Objectives:** To estimate the type I and type II errors associated with the application of 12 different statistical tests to a series of LOS samples. **Methods:** The LOS distribution was extracted from an exhaustive French national database of inpatient stays. The type I error was estimated using 19 sample sizes and 1,000,000 simulations per sample. The type II error was estimated in three alternative scenarios. For each test, the type I and type II errors were plotted as a function of the sample size. **Results:** Gamma regression with log link, the log rank test, median regression, Poisson regression, and Weibull survival analysis presented an unacceptably high type I error. In contrast, the Student standard t test, linear regression with log link, and the Cox models had an acceptable type I error but low power. **Conclusions:** When comparing the LOS for two balanced inpatient samples, the Student t test with logarithmic or rank transformation, the Wilcoxon test, and the Kruskal-Wallis test are the only methods with an acceptable type I error and high power.

Keywords: length of stay, methodology, outcome measurement, statistics.

Introduction

The Inpatient Length of Stay

The *inpatient length of stay* (LOS) is defined as the period during which a patient is confined to a hospital or any other health care establishment [1]. The LOS is often studied by clinical researchers as a guide to the putative benefit of a treatment of interest. A shorter LOS (relative to a reference treatment or standard of care) may indicate clinical benefit, whereas a longer LOS may indicate the greater occurrence of treatment-related adverse events [2,3]. Conversely, the LOS is also an important risk factor for adverse events [4–6]. Furthermore, the LOS is frequently used as a key indicator of operational efficiency and sometimes as a proxy for quality-of-care processes [7]. Health economists also use the LOS to estimate health expenditure because health care establishments mainly have fixed costs (such as salaries). Consequently, more than 2200 articles a year refer to the LOS in their abstract or title (according to the PubMed database; Table 1). Furthermore, researchers use various statistical methods to compare the LOS in two patient samples.

Modeling the LOS

Fitting the LOS distribution using various statistical models has been extensively studied [8–20]. Although the LOS is always considered as the dependent variable (Y), the independent variables (X_i) considered to be statistically significant vary as a function of the selected model [8,9,17,19]. Furthermore, when two methods generate different results after application to real data, researchers are unable to determine which result is true. One can therefore hypothesize that if the goal is to identify statistically significant explanatory variables, the type I and type II errors will differ from one method to the other. The scientific literature, however, does not provide any guidance on choosing the most appropriate method.

Conflicts of interest: The authors declare that there are no conflicts of interest.

* *Address correspondence to:* Emmanuel Chazard, Public Health Department, Faculté de Médecine de Lille, CERIM, Lille Cedex F-59045, France.

E-mail: emmanuel.chazard@univ-lille2.fr.

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2017.02.009

Table 1 – Number of articles referenced in PubMed during the last 10 y (January 2006 to December 2015).

Concept	Cited in the title	Cited in the title or the abstract
Length of stay (LOS)*	1,704	22,548
Type I error†	100	2,180
Type II error‡	223	4,925
Type I error† and LOS*	0	4
Type II error‡ and LOS*	0	12

* Keyword: "length of stay."
† Keywords: "type 1 error," "type I error," "first type error," or "alpha risk."
‡ Keywords: "type 2 error," "type II error," "second type error," "beta risk," or "statistical power."

Comparing the LOS Distributions in Two Independent Samples

Researchers often want to compare the mean LOS of two independent samples. This can be done with three families of methods [13,21]: 1) bivariate statistical tests (i.e., parametric tests such as the Student t test or nonparametric tests such as the Wilcoxon test); 2) regression models (such as gamma regression [22]), in which the LOS is the dependent variable Y and the samples are labeled by a binary independent variable X; and 3) survival analyses (e.g., the log rank test or the Cox models) in which the discharge is the observed outcome and the LOS is the time to the outcome. Although there are no censored values (because the patient is always discharged), survival analyses can be used as "less parametric" alternatives to traditional statistical tests [9,13]. We will now provide a more precise description of 12 of these methods.

Type I and Type II Errors in the Comparison of Two Samples

In the present case, the type I error (the alpha risk) corresponds to the probability with which a test will detect a significant difference in the mean LOS between two samples—even though the samples have been drawn from the same population. In most studies, the null hypothesis is rejected when the P value is less than 0.05, which leads researchers to assume that the type I error is 5%. In the field of medical research, increasing the type I error beyond 5% is considered to be unacceptable, because it could generate erroneous knowledge and prompt physicians to make inappropriate diagnostic and therapeutic decisions [23]. The type II error (the beta risk) corresponds to the probability with which a statistical test will not detect a difference in the mean LOS between two samples, even though the samples have been drawn from populations whose means were different. The power is defined as 1 − type II error.

Type I and type II errors have been extensively studied in the literature on variables not related to the LOS (Table 1). The literature contains many general assertions about type I and type II errors in statistical tests [21]. These assertions are not always evidence-based and cannot be generalized without considering the distribution of the variable under investigation. For example, Skovlund and Fenstad [24] developed an algorithm for determining the best way of comparing two samples. The appropriate choice depended on the equality of variance, the sample imbalance, and (most importantly) the skewness.

The Distribution of LOS Data

LOS data have a very particular distribution: a highly right-skewed, discrete, positive distribution with many tied observations, with values concentrated around the median [8,14,20,25,26]. In some health care establishments, the LOS distribution might be multimodal and depend on how care is organized. Last, LOS samples may contain a few outliers with extremely high values. Expert opinion suggests that these outliers should be excluded from analysis [25–27]; unfortunately, automated approaches have yet to be developed.

A Lack of Specific Research on Comparing LOS

Although mean LOS values are often compared using statistical tests, there is a lack of knowledge in this field. In particular, the type I and type II errors have not been empirically evaluated for this specific distribution (see Table 1). To our knowledge, all the studies in this field to date have assumed that the type I error is always controlled: researchers have assumed that under the null hypothesis, there is a 5% probability that the P value is less than 0.05. We believe that this assumption should be questioned. Furthermore, most investigators do not check the validity of the assumptions of the chosen statistical tests [21]. Consequently, it is important to empirically check the type I and type II errors even when statistical tests are inappropriately used (e.g., a parametric test applied to a small sample or a rank test applied to tied observations).

Hence, the objective of the present study was to empirically evaluate statistical tests that are frequently used to compare the mean LOS of two independent samples, with regard to the type I error under the null hypothesis and the type II error under three realistic, alternative hypotheses. We evaluated the tests even when their assumptions were not met so as to determine and consider the consequences of inappropriate application.

Methods

Estimation of the LOS Distribution Function

We first queried the French nationwide hospital discharge database programme de médicalisation des systèmes d'information (PMSI) to obtain the LOS empirical distribution function. This database is based on compulsory, standardized discharge reports on all patients admitted to nonprofit acute-care hospitals in France and is used to calculate a significant proportion of a hospital's public funding. Each discharge report describes the patient's administrative and demographic data, diagnoses, and medical procedures. The database query included all inpatient stays for 2012 and excluded outpatients and iterative ambulatory treatments (dialysis etc.). The total number of included stays was 9,895,673. For each inpatient stay, the LOS is defined as an integral number of calendar days (see Equation 1).

$$\mathrm{LOS} = \mathrm{Discharge\_date} - \mathrm{Admission\_date} + 1. \quad (1)$$

The probability density function of LOS was estimated from the result of the database query. Univariate statistics were calculated.

The Statistical Tests

This section introduces the statistical tests used to assess the difference between the LOS X of two independent samples of patients. Nevertheless, actual examples of use will be given later in the article. Here, X is the random variable representing the LOS, μ is the mean of X, and G is the binary variable "group name," with possible values of $\{G_1; G_2\}$.

Bivariate methods are used to test whether $\mu_{X_{G1}}$ and $\mu_{X_{G2}}$ are different. For each test, the P value is considered to be the main output. Methods based on regression models are used to test whether X can be explained by G. In that case, a coefficient β_G is calculated for G and is tested against the null hypothesis $\beta_G = 0$. The P value of this test is again considered to be the main output. For survival-based methods, all subjects are assumed to have an

event (i.e., discharge), and X is used as the time to event. The tests are then performed the same way as described earlier.

All simulations and statistical tests were performed using R software and the *MASS*, *survival*, and *quantreg* libraries (R Foundation for Statistical Computing, Vienna, Austria) [28–30]. The statistical tests and their corresponding R codes are given in Table 2. Hereafter, the tests are referred to by the abbreviations given in Table 2 (e.g., *Student* for Student *t* test).

Table 2 – List of the statistical tests and the corresponding R code*,†.

Statistical method (and abbreviation)	R code to get the P value
Statistical tests	
Kruskal-Wallis test (KruskalWallis)	obj <-kruskal.test(x=X, g=G) pval<-obj$p.value
Student *t* test (Student)	obj <- t.test(X1,X2, var.equal=FALSE) pval<-as.numeric(obj$p.value)
Student *t* test with log transformation (StudentLog)	X1 <- log(X1) X2 <- log(X2) # then the same code as Student
Student *t* test on the ranks (StudentRanks)	X1 <-rank(X)[G==0] X2 <-rank(X)[G==1] # then the same code as Student
Wilcoxon or Mann-Whitney test (Wilcoxon)	obj <-wilcox.test(X1,X2) pval<-obj$p.value
Methods based on regression models	
Gamma regression with a log link (GammaReg)	fam <-Gamma(link="log") obj <-glm(X~G, family=fam) pval<-summary(obj)$coefficients[2,4]
Linear regression with a log link (LinearReg)	fam <-gaussian(link="log") obj <-glm(X~G, family=fam) pval<-summary(obj)$coefficients[2,4]
Median regression (MedianReg)	library(quantreg) obj <-rq(formula=X~G, tau=0.5) pval<-summary(obj,se="ker")$coefficients[2,4]
Poisson regression (PoissonReg)	fam<-poisson obj <-glm(total ~ groups, family=fam) pval<-summary(obj)$coefficients[2,4]
Survival methods	
Cox proportional hazard model (CoxPH)	library(survival) obj <-coxph(formula=Surv(X)~G) pval<-summary(obj)$coefficients[5];
Log rank test (LogRank)	library(survival) obj <-survdiff(formula=Surv(X)~G) pval<-1-pchisq(obj$chisq, df=1)
Weibull survival (Weibull)	library(survival) obj <-survreg(formula=Surv(X)~G, dist="weibull") pval <-1-pchisq(2*diff(obj$loglik), sum(obj$df)-obj$idf)

LOS, length of stay.

* An example of implementation is available in Appendix B in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

† In the code, X1 and X2 are two samples of LOS. X results from pooling X1 and X2, then G is the binary variable corresponding to the groups. pval is the P value computed for each test.

Estimation of the Type I Error

When a P value of less than 0.05 is the threshold for statistical significance, the statistical tests' type I error is expected to be 5%. We empirically estimated this risk by simulation. For each fixed sample size n ∈ {10; 15; 20; 25; 30; 40; 50; 60; 80; 100; 130; 160; 190; 220; 260; 300; 350; 400; 450}, two samples were randomly drawn with replacement from the LOS population described earlier. The statistical tests were performed under the null hypothesis, and the P value of each statistical test was recorded. This process was iterated 1,000,000 times.

Next, the empirical cumulative distribution function (ECDF) of the P values, $F(P)$, from each test was estimated under the null hypothesis. In theory, each ECDF should be a straight line corresponding to Equation 2, given that Equation 3 is true. To verify this hypothesis, Equation 4 was then used to estimate the empirical type I error for each test with a 5% threshold.

$$\forall X \in [0;1],\ Y = X, \tag{2}$$

$$\forall P_0,\ P_0 = P(P < P_0), \tag{3}$$

$$\text{Empirical type I error} = P(P < 0.05/H_0). \tag{4}$$

Estimation of the Type II Error, Power, and Relative Efficiency

The type II error can be estimated only for a specific alternative hypothesis and should generally be as high as possible—although the value is constrained by the type I error. In the present study, we simulated three simple hypotheses. According to the alternative hypothesis H_{1a}, the first sample was drawn from the original LOS distribution and the second sample was drawn from a distribution that is shifted to the right by 1 day. In the alternative hypothesis H_{1b}, each individual in the second sample had its LOS value increased randomly by 0, 1, or 2 days, with a probability of 1/3 for each possible value. Last, the alternative hypothesis H_{1c} is a real-life hypothesis: the first sample was composed of inpatient stays without surgery, and the second sample was composed of inpatient stays with surgery.

For each fixed sample size n ∈ {10; 15; 20; 25; 30; 40; 50; 60; 80; 100; 130; 160; 190; 220; 260; 300; 350; 400; 450}, two samples were drawn with replacement. For the H_{1a} and H_{1b} hypotheses, both samples were drawn from the original LOS distribution, and the second sample was then transformed (depending on the hypothesis). For the H_{1c} hypothesis, the nationwide LOS data were first separated into inpatient stays without surgery and inpatient stays with surgery, and one sample was drawn from each of the data sets. The statistical tests were performed, and the P value of each statistical test was recorded. This process was iterated 100,000 times. Next, the ECDF of the P value from each test was estimated. The empirical power was then estimated for each test and for each hypothesis with a 5% threshold, using Equation 5.

$$\text{Power} = 1\text{-Type II error} = P(P < 0.05/H_1). \tag{5}$$

For each hypothesis and for each statistical test, we calculated the efficiency relative to *Student* (because it is by far the most frequently used test in the literature). Here, efficiency was defined as the sample size that is required for a method to achieve a predefined power (0.5 in the present study). Hence, the efficiency of a test t_1 relative to *Student* is the efficiency of *Student* divided by the efficiency of t_1. Accordingly, a test t_1 was considered to be more efficient than *Student* if the relative efficiency was greater than 1, and a test t_2 was considered more efficient than a test t_1 if its relative efficiency was greater than the value for t_1.

If a test did not achieve a power of 0.5 with the largest sample size (300), we recorded the highest power achieved. Because only

19 sample sizes were tested, the relative efficiency at a power of 0.5 was estimated by linear interpolation.

Workflow and Decision Thresholds

The type I error was estimated for the 12 statistical methods and the 19 sample sizes. If one or more of the type I errors exceeded 5.5%, the method in question was excluded from the power analysis. If one or more of the type I errors were between 5.1% and 5.5%, the method was classified as being subject to caution. All other methods were classified as being appropriate for use.

Next, each statistical test's power was estimated for each of the alternative hypotheses H_{1a}, H_{1b}, and H_{1c} and each value of n. The efficiency was estimated relative to *Student*. The tests were then classified into three efficiency groups: low, moderate, and high. The overall workflow is depicted in Appendix Figure C3 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

Ethics

The present study used anonymized secondary data that had been collected in the course of routine care. In line with the French legislation on data set analyses, approval by an investigational review board was neither required nor sought. The work complied with the tenets of the Declaration of Helsinki. The collection and analysis of data was authorized by the French national data protection commission (*Commission Nationale de L'informatique et des Libertés*, Paris, France).

Results

LOS Distribution with Real Hospital Data

The LOS distribution was derived from the total number of inpatient stays in France in 2012. The data distribution was discrete, as shown in Figure 1. The LOS values ranged from 1 day to 1247 days (3.41 years). The mode was 1 day and was associated with a probability of 28.1%. The mean LOS was 5.44 ± 7.80 days. The median LOS was 3 days, and 33.4% of all values were concentrated around the median ± 1 day. The distribution was strongly skewed to the right, with a skewness of 9.26 and a kurtosis of 353.9. Detailed results are provided in Appendix A in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009. A large minority of the patients (26.8%) were hospitalized several times, and 2.8% of patients were hospitalized 5 times or more. Removal of duplicates modified the distribution slightly: the mean LOS was 5.28 ± 7.65 days, the median was still 3 days, 34.7% of the LOS values were between 2 and 4 days, the skewness was 11.7, and the kurtosis was 340.7. These metrics are provided for information purposes only; hereafter, only the raw distribution will be described.

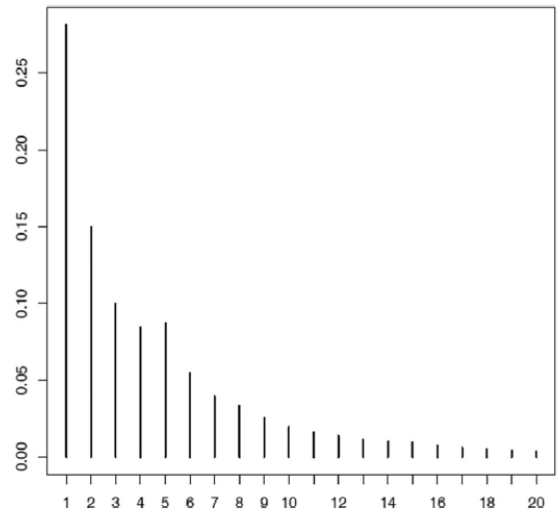


Fig. 1 – Estimated probability density (*y*-axis) function of the length of stay (*x*-axis, truncated to 20 days).

Estimation of the Type I Error

Under the null hypothesis, the ECDF of the *P* values from each of the 12 statistical tests was estimated for each sample size. The 228 ECDF curves (12 tests × 19 sample sizes) were used to estimate the type I error with *P* less than 0.05. Three groups of tests were identified (Fig. 2). The *KruskallWallis*, *LinearReg*, *Student*, *StudentLog*, *StudentRanks*, and *Wilcoxon* tests were found to be appropriate; it is particularly noteworthy that the type I errors for *Student* and *LinearReg* were well less than 5%. Caution should be exercised with regard to *CoxPH*; this method was associated with a moderately high type I error for sample sizes smaller than 80. The *GammaReg*, *LogRank*, *MedianReg*, *PoissonReg*, and *Weibull* tests were excluded because of very high type I errors, and so should not be used for LOS comparisons.

Detailed methods and results are provided in Appendix D in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

Estimation of the Type II Error

The type II error and the power were estimated for the three alternative hypotheses, the 19 sample sizes, and the seven tests with an acceptable type I error. The 399 resulting ECDF curves were used to estimate the power with *P* less than 0.05.

Under the H_{1a} hypothesis, the power was high for *KruskallWallis*, *StudentLog*, *StudentRanks*, and *Wilcoxon* (Fig. 3), moderate for *CoxPH*, and low for *Student* and *LinearReg*. The results under the H_{1b} hypothesis were very similar. Under the H_{1c} hypothesis, the power was low for every test and increased linearly with the sample size. Table 3 presents the relative efficiency with respect to *Student*. In the high-power group, the relative efficiencies ranged from 7.9 to 8.5 under the H_{1a} hypothesis and were around 6.3 under the H_{1b} hypothesis. Within this group of tests, *StudentLog* was the most efficient. Under the H_{1c} hypothesis, the highest relative efficiency was observed for *CoxPH* (1.172). Additional methods and results are provided in Appendix C in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.02.009.

The Tests' Assumptions

When we looked at whether the statistical tests' underlying assumptions were violated, we noticed that the LOS, its logarithm, and its ranks did not follow a normal distribution. Consequently (given the central limit theorem), *Student*, *StudentLog*, and *StudentRanks* were not valid for an n value of less than 30 but were valid for an n value of 30 or more. The assumptions were always met for *KruskalWallis* and *Wilcoxon*, despite a high proportion of tied data (e.g., ties accounted for 37% of the sample for n = 10, 50% for n = 30, and 68% for n = 300). When we analyzed the residuals of 76 random experiments (4 per sample size), we observed that the residuals of the regressions were never normally distributed for *LinearReg*, *MedianReg*, *GammaReg*, and *PoissonReg*. We next examined survival methods. The proportional hazard assumption of the *CoxPH* was violated in only 4 of the 76 random experiments. When we drew quantile-quantile plots of the event times, we observed that they fitted a *Weibull*

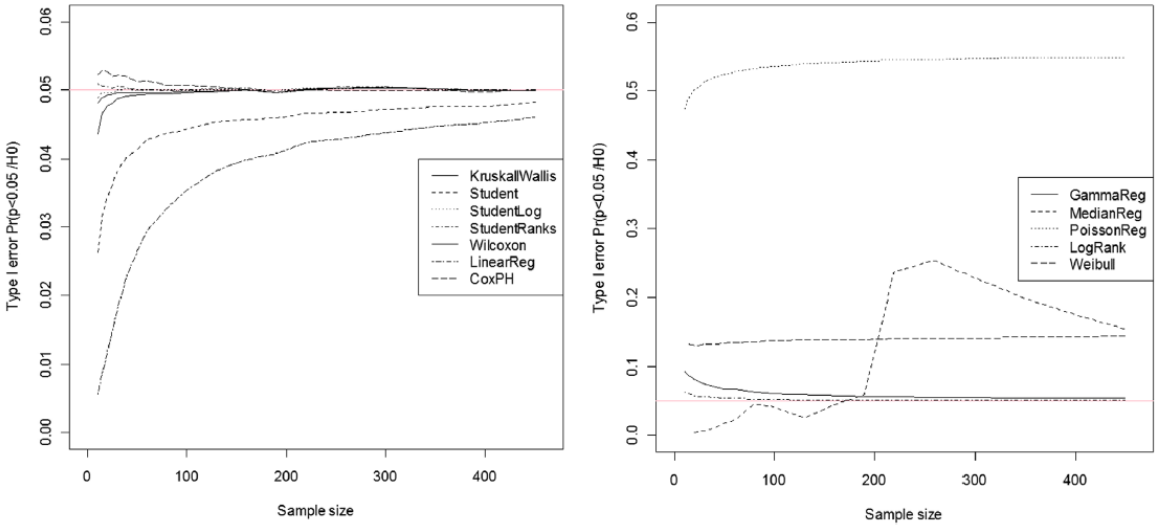


Fig. 2 – Type I error as a function of the sample size (left: seven nonexcluded methods; right: five methods with unacceptable inflation of type I error).

distribution in 73 of the 76 random experiments. Last, the assumptions of *LogRank* were always met, because there was no censoring in our simulations.

Discussion

In the present study, we evaluated 12 statistical tests that are frequently used to compare the LOS of 2 balanced samples. Given the very particular distribution of LOS data in administrative databases (a discrete, highly right-skewed distribution with tied observations and many outliers), the tests' validity cannot be guaranteed and, in practice, is not checked often enough [21]. We had three main findings. First, some methods should not be used because they have a type I error more than 5% if the null hypothesis is rejected at a *P* value of less than 0.05 (*GammaReg*, *LogRank*, *MedianReg*, *PoissonReg*, and *Weibull*). Second, some of the methods with an acceptable type I error are also similarly and highly efficient, relative to *Student* (*KruskallWallis*, *StudentLog*, *StudentRanks*, and *Wilcoxon*). Third, the differences in the tests' performance under three alternative hypotheses suggest that the relative efficiency depends on the characteristics of the samples being compared.

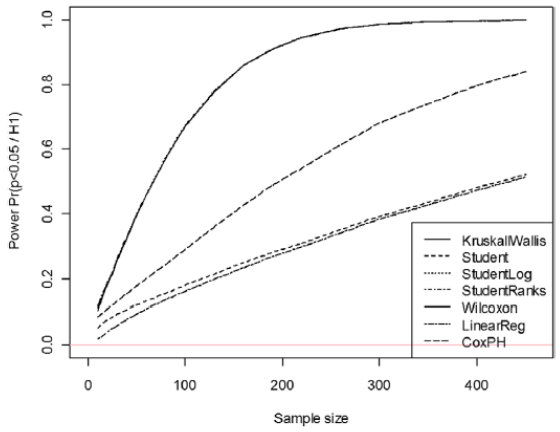


Fig. 3 – Power as a function of the sample size for the seven methods that are not excluded, for the alternative hypothesis H_{1b} (all the LOS values are increased at random by 0, 1, or 2 d). *LOS*, length of stay.

We focused on the 12 tests used most frequently in the literature. Some other tests could have been considered, and we did not evaluate approaches such as data truncation. We did not test the effects of outlier removal because a standardized consensus approach is not available, and so the number of solutions would have been infinite [25–27].

The present work was based on the French national inpatient stay database. The data distribution was exhaustively measured and not estimated. Furthermore, we performed simulations strictly, and the type I and type II errors were estimated in the absence of previous hypotheses. The validity of the statistical methods' assumptions was not checked during simulations because the objective was to evaluate the consequences of this violation, just as many researchers fail to check those assumptions in their peer-reviewed articles [21]. Our results demonstrated that the type I and type II errors obtained are not necessarily related to the validity of the tests' assumptions. For instance, *StudentLog* and *StudentRanks* produced very good results even when the sample size was too small. In contrast, *LogRank*

Table 3 – Efficiency of the nonexcluded tests relative to the Student *t* test (a high value denotes a high power).

Test	Hypothesis H_{1a}	Hypothesis H_{1b}	Hypothesis H_{1c}
CoxPH	2.311	2.164	1.172
KruskallWallis	7.968	6.315	1.005
LinearReg	0.981	0.980	0.971
Student	1	1	1
StudentLog	8.534	6.360	1.040
StudentRanks	7.992	6.331	1.006
Wilcoxon	7.939	6.301	1.005

produced a high type I error and *CoxPH* produced an intermediate power level when used under valid conditions.

The present study had a number of limitations. First, we did not take account of the fact that patients made more than one stay. Second, we chose to use balanced samples ($n_1 = n_2$), and the three alternative hypotheses used to estimate the type II error could not cover all possible configurations. We have focused on mean differences but did not evaluate the effect of differences in variance (heteroskedasticity) or skewness, which might even impair the application of nonparametric tests [24]. A simulation study should be based on a hypothesis that corresponds precisely to the situation to be tested.

Nevertheless, our results also suggest that whichever statistical test is applied, its *P* value should be corrected by using bootstrap techniques [31]. In short, bootstrap techniques enable one to replace the *P* value by an *F*(*P*) value (where *F* is the ECDF of the *P* values under the null hypothesis) for each performance of the test. Even when *P* values are corrected by bootstrapping, differences between statistical tests may nevertheless still exist.

To compare the LOS of two inpatient samples, we recommend using the Student *t* test with logarithmic or rank transformation, the Kruskal-Wallis test, or the Wilcoxon (Mann-Whitney) test. Our results agree with those of a previous study [21] in which the Wilcoxon and Kruskal-Wallis tests had a higher power than the Student *t* test (without transformation) for the analysis of LOS in an emergency department (as a continuous variable). It is often stated that the Wilcoxon and Kruskal-Wallis tests' validity can be questioned for tied data. Nevertheless, our simulation shows that ties are not problematic in this context. The tiny difference between the Wilcoxon and Kruskal-Wallis tests was simply because of the different way in which they handle tied observations, which are prevalent. It is noteworthy that the Student *t* test with logarithmic or rank transformation is just as efficient as the Wilcoxon and Kruskal-Wallis tests, and it may be more accessible for nonstatisticians. Data transformation, however, has some important drawbacks. Transformation can easily be applied if the purpose of the test is solely to compute a *P* value. Nevertheless, such a transformation can be likened to a "black box," which makes it more difficult to interpret the effect size and its confidence interval.

The Student *t* test is by far the most frequently applied test in this context. Our present results showed that this test was very conservative (i.e., with a type I error well less than 5%) but had low power. This observation is very reassuring because in real life, many processes tend to increase the false-discovery rate. Many statistical studies are data-driven (in which tests are performed *because* something is visible in the descriptive analyses); repeated statistical tests are often performed without Bonferroni or Šidák correction, underpublication bias is produced because negative results are not to be submitted or not accepted by journals, and public opinion and decision makers often focus on studies that reject the null hypothesis (even when many other published studies do not).

With regard to methods based on regression models, our results suggest that gamma, median, and Poisson regressions should be avoided because of their type I error, and that linear regression should be avoided because of its low power (the Cox model is preferable). Some researchers have reported the superiority of gamma regression [10,12] and the inferiority of the Cox model [13], although these studies sought to fit the LOS by decreasing the residuals of the prediction and were not designed to evaluate type I and type II errors.

Conclusions

To compare the LOS of two balanced samples of inpatients, we recommend the Student *t* test with logarithmic or rank transformation, the Kruskal-Wallis test, or the Wilcoxon (Mann-Whitney) test. If the LOS distribution differs greatly from that used in the present simulation study, we recommend using bootstrap techniques to recalibrate the chosen statistical method.

Source of financial support: This study did not receive any funding.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2017.02.009 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Medical Subject Headings. Length stay. Available from: http://www.ncbi.nlm.nih.gov/mesh?term=length%20of%20stay. [Accessed May 18, 2016].
[2] Classen DC, Pestotnik SL, Evans RS, et al. Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. JAMA 1997;277:301–6.
[3] Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA 2003;290:1868–74.
[4] Bates DW, Miller EB, Cullen DJ, et al. Patient risk factors for adverse drug events in hospitalized patients. ADE Prevention Study Group. Arch Intern Med 1999;159:2553–60.
[5] Morimoto T, Sakuma M, Matsui K, et al. Incidence of adverse drug events and medication errors in Japan: the JADE study. J Gen Intern Med 2011;26:148–53.
[6] Hauck K, Zhao X. How dangerous is a day in hospital? A model of adverse events and length of stay for medical inpatients. Med Care 2011;49:1068–75.
[7] Diercks DB, Roe MT, Chen AY, et al. Prolonged emergency department stays of non-ST-segment-elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events. Ann Emerg Med 2007;50:489–96.
[8] Austin PC, Rothwell DM, Tu JV. A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. Health Serv Outcomes Res Methodol 2002;3:107–33.
[9] Dudley RA, Harrell FE Jr, Smith LR, et al. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. J Clin Epidemiol 1993;46:261–71.
[10] Marazzi A, Paccaud F, Ruffieux C, Beguin C. Fitting the distributions of length of stay by parametric models. Med Care 1998;36:915–27.
[11] Lee AH, Fung WK, Fu B. Analyzing hospital length of stay: mean or median regression? Med Care 2003;41:681–6.
[12] Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ 2005;24:465–88.
[13] Basu A, Manning WG, Mullahy J. Comparing alternative models: log vs Cox proportional hazard? Health Econ 2004;13:749–65.
[14] Lee AH, Gracey M, Wang K, Yau KKW. A robustified modeling approach to analyze pediatric length of stay. Ann Epidemiol 2005;15:673–7.
[15] Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ 2001;20:461–94.
[16] Samore MH, Shen S, Greene T, et al. A simulation-based evaluation of methods to estimate the impact of an adverse event on hospital length of stay. Med Care 2007;45:S108–15.
[17] Faddy M, Graves N, Pettitt A. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. Value Health 2009;12:309–14.
[18] Singh CH, Ladusingh L. Inpatient length of stay: a finite mixture modeling analysis. Eur J Health Econ 2010;11:119–26.
[19] Ravangard R, Arab M, Rashidian A, et al. Comparison of the results of Cox proportional hazards model and parametric models in the study of length of stay in a tertiary teaching hospital in Tehran, Iran. Acta Med Iran 2011;49:650–8.
[20] Moran JL, Solomon PJ. A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and New Zealand intensive care adult patient data-base, 2008–2009. BMC Med Res Methodol 2012;12:68.
[21] Qualls M, Pallin DJ, Schuur JD. Parametric versus nonparametric statistical tests: the length of stay example. Acad Emerg Med 2010;17:1113–21.

[22] Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. J Health Econ 1999;18:153–71.
[23] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference, part I. Biometrika 1928;20A:175–240.
[24] Skovlund E, Fenstad GU. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? J Clin Epidemiol 2001;54:86–92.
[25] Ruffieux C, Marazzi A, Paccaud F. Exploring models for the length of stay distribution. Soz Praventivmed 1993;38:77–82.
[26] Weissman C. Analyzing intensive care unit length of stay data: problems and possible solutions. Crit Care Med 1997;25:1594–600.
[27] Lee AH, Xiao J, Vemuri SR, Zhao Y. A discordancy test approach to identify outliers of length of hospital stay. Stat Med 1998;17: 2199–206.
[28] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011.
[29] Therneau T, Lumley T. Survival: survival analysis, including penalised likelihood. R package version 2.36-5. Available from: http://CRAN.R-project.org/package=survival. [Accessed January 8, 2014].
[30] Koenker R. quantreg: Quantile Regression. 2015.
[31] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press; 1994.

INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS XXX (2013) XXX–XXX

journal homepage: www.ijmijournal.com

Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records

Emmanuel Chazard[a,*], *Capucine Mouret*[b], *Grégoire Ficheur*[a], *Aurélien Schaffar*[a], *Jean-Baptiste Beuscart*[c], *Régis Beuscart*[a]

[a] *Department of Public Health, CHU Lille, UDSL EA 2694, Univ Lille Nord de France, F-59000 Lille, France*
[b] *Department of Occupational Medicine, CHU Lille, F-59000 Lille, France*
[c] *Department of Geriatrics, CHU Lille, UDSL EA 2694, Univ Lille Nord de France, F-59000 Lille, France*

ARTICLE INFO

Article history:
Received 26 March 2012
Received in revised form
28 November 2013
Accepted 28 November 2013

Keywords:
Anonymization
De-identification
Confidentiality
Free text
Natural language processing

ABSTRACT

Purpose: Medical free-text records enable to get rich information about the patients, but often need to be de-identified by removing the Protected Health Information (PHI), each time the identification of the patient is not mandatory. Pattern matching techniques require pre-defined dictionaries, and machine learning techniques require an extensive training set. Methods exist in French, but either bring weak results or are not freely available. The objective is to define and evaluate FASDIM, a Fast And Simple De-Identification Method for French medical free-text records.

Methods: FASDIM consists in removing all the words that are not present in the authorized word list, and in removing all the numbers except those that match a list of protection patterns. The corresponding lists are incremented in the course of the iterations of the method.

For the evaluation, the workload is estimated in the course of records de-identification. The efficiency of the de-identification is assessed by independent medical experts on 508 discharge letters that are randomly selected and de-identified by FASDIM. Finally, the letters are encoded after and before de-identification according to 3 terminologies (ATC, ICD10, CCAM) and the codes are compared.

Results: The construction of the list of authorized words is progressive: 12 h for the first 7000 letters, 16 additional hours for 20,000 additional letters. The Recall (proportion of removed Protected Health Information, PHI) is 98.1%, the Precision (proportion of PHI within the removed token) is 79.6% and the F-measure (harmonic mean) is 87.9%. In average 30.6 terminology codes are encoded per letter, and 99.02% of those codes are preserved despite the de-identification.

Conclusion: FASDIM gets good results in French and is freely available. It is easy to implement and does not require any predefined dictionary.

* *Corresponding author at:* Pôle de santé publique, 154 rue Yersin, CHRU de Lille, 59037 Lille Cedex, France. Tel.: +33 3 20 44 60 38.
E-mail addresses: emmanuel.chazard@univ-lille2.fr, emmanuelchazard@yahoo.fr (E. Chazard).
1386-5056/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijmedinf.2013.11.005

1. Introduction

1.1. A need for de-identifying discharge letters

Computerized free-text medical records are important information sources for research. In most countries, each time a patient is discharged from a healthcare facility, a discharge letter has to be written: it summarizes all the pertinent information from the reason for admission to the discharge drug treatment. Those letters are routinely produced and provide the researchers with a big amount of medical information. On the other hand, the confidentiality must imperatively be respected: as soon as a discharge letter is not used with direct benefit to the patient and if the patient does not need to be identified, the letter must be de-identified. The anonymization consists in removing the patients' names from the records: unfortunately, other pieces of information enable to identify the patients. The de-identification is a more exhaustive removal of the entire Protected Health Information (PHI), so that the patients cannot be identified, directly nor indirectly. In the US, privacy rules have been enacted by the Department of Health and Human Services further to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [1]. In order to de-identify a high number of records, it is necessary to use automated methods, as manual methods require too high workload [2].

1.2. State of the art

Several methods exist for automated de-identification of free-text records [3], including procedures reports and discharge letters.

Pattern matching methods [4–16] consist in applying rules that enable to keep or remove some words that belong to dictionaries that have been predefined by experts or institutions. For instance, it is possible to remove all the words that belong to a list of town names, or to preserve all the words that belong to a list of medical terms (such as the Unified Medical Language System [17]). Additional rules may be used to take into account words declension and verbs conjugation. This approach requires that such lists are available. When they exist, those lists are language-dependent, and are suitable for a specific context only (e.g. town names or current family names are useless in another country).

Machine learning methods [14,18–26] are derived from artificial intelligence. A learning phase requires that a corpus of records is previously de-identified manually by experts. Those methods are often very efficient, depending on the quality and the completeness of the learning corpus.

Whatever the method used, the de-identification is evaluated by computing three rates:

- The recall (or sensitivity or completeness, Eq. (1)), which is the proportion of removed token within the PHI. A high recall enables to preserve the confidentiality.
- The precision (or positive predictive value or correctness, Eq. (2)), which is the proportion of PHI within the removed token. A high precision enables to preserve the readability of the text.
- The *F*-Measure, which is the harmonic mean of the recall and the precision (Eq. (3)).

$$\text{recall} = R = \frac{TP}{\#\text{identifiers}} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{precision} = P = \frac{TP}{\#\text{removed}} = \frac{TP}{TP + FP} \quad (2)$$

$$F\text{-measure} = F = \left(\frac{R^{-1} + P^{-1}}{2}\right)^{-1} \quad (3)$$

Table 1 presents the main results obtained in the literature by the authors for medical free-text de-identification. Most of methods are developed for English language and can hardly be used for other languages. Some methods have been developed in French, but either their results are disappointing, or they are not freely available.

1.3. Unsolved situations

Despite the good results obtained by many methods, text de-identification is still not obvious and some situations may not be addressed with current tools. We shall illustrate it through 4 use cases.

Case 1: a team has to de-identify English free-text records using pattern-matching. Some tools are freely available. However, it cannot be guaranteed that those tools could be applied in a different context without any adaptation. Indeed, pattern matching techniques rely on lists of words that are context-dependent: for instance "lime tree" should be removed in most reports as it is often part of a street name, but should not be removed in an allergy-related report. Lists of town names or family names also depend on the country. Finally, misspellings are most often not taken into account by existing methods.

Case 2: a team has to de-identify English free-text records using machine learning. Here again, some tools are freely available but, in a like manner, machine learning techniques require a pre-existing corpus of de-identified records. Such corpuses are available in English [11,36,37], but they may be used only if the type of document to de-identify is the same as the documents of the training corpus.

Case 3: a team has to de-identify French free-text records (the problem is the same with most of non-English languages): no free and efficient method, no list of words, and no training corpus are available. Everything has to be built.

Case 4: a team has only little time (e.g. 1 man-week) to de-identify a few records (e.g. 25,000 records). Whatever the language, the context and the technique, it will probably take more time to understand, adapt, implement and execute an existing tool.

The conception of FASDIM relies on the idea that a simple de-identification technique could enable to de-identify French discharge letters with an acceptable workload, particularly when the number of records is low. The main idea is to supply the workload in the course of the method, and not before the first document can be de-identified.

Table 1 – Results of authors for medical free-text records de-identification.

Author	Method	Language	Precision	Recall	F-measure
Aberdeen et al. (2010) [18]	Machine learning	EN	94.3%	97.8%	96%
Aramaki and Miyo (2006) [19]	Machine learning	EN	–	–	–
Beckwith et al. (2006) [4]	Pattern matching	EN	98.3%	–	–
Deleger et al.	Machine learning	EN	92.8%	92.8%	92.8%
(2013) [25]	Machine learning	EN	95.1%	91.9%	93.5%
	Manual	EN	93.9%	92.1%	93%
	Manual	EN	88.5%	94.6%	91.4%
Dorr et al. (2006) [2]	Manual	EN	95.9%	99.6%	–
Ferrández et al.	Machine learning	EN	96%	70%	74%
(2012) [26]	Machine learning	EN	95%	76%	79%
Friedlin and McDonald (2008) [7]	Pattern matching	EN	–	99.47%	–
Grouin et al. (2009)	Pattern matching	FR	92%	83%	–
[9]		EN	23%	65%	–
Neamatullah et al. (2008) [11]	Pattern matching	EN	75%	94%	–
Ruch et al. (2000) [12]	Pattern matching	FR	–	99%	–
Sweeney (1996) [13]	Pattern matching	EN	–	–	–
Szarvas et al. (2007) [22]	Machine learning	EN	–	–	96%
Taira et al. (2002) [14]	Pattern matching and machine learning	EN	99%	94%	–
Thomas et al. (2002) [15]	Pattern matching	EN	–	98.7%	–
Tu et al. (2010) [28]	Pattern matching	EN	91.3%	88.3%	90%
Uzuner et al. (2007) [23]	Machine learning	EN	99%	97%	98%
Velupillai et al. (2009) [16]	Pattern matching	SW	3–9%	56–76%	4–16%
Wellner et al. (2007) [24]	Machine learning	EN	98%	96%	96%

1.4. *Objectives*

The first general objective of this work is to design and implement FASDIM, a Fast And Simple De-Identification Method for clinical free-text records. The second general objective is to evaluate the method.

To reach the first general objective, operational objectives are (1) to design a method that reaches good results in French using completely unstructured free-text records, but (2) is as independent as possible from the language structure (i.e. for instance does not consider the declension of words and the conjugation of verbs) and (3) does not rely on any pre-existing material (list of words or corpus of de-identified documents), in order to (4) be easily and fast reproducible from scratch by any hospital or research team.

To reach the second general objective, operational objectives are (5) to objectively compute traditional evaluation metrics but also (6) to evaluate the preservation of medical information and (7) to evaluate the workload required to implement the method.

The method is implemented and evaluated in French, but the examples that are given in this paper here are translated into English.

2. Definition of FASDIM

FASDIM stands for Fast And Simple De-Identification Method. This method is composed of 3 steps (Fig. 1).

2.1. *Step 1 (automated): preparation of the records and patterns extraction*

The first step of the algorithm consists of an automated treatment of the free-text records (Fig. 1). The records are loaded as simple text, and the typography is simplified: the text is lowercased, the accents are removed from the letters, and special characters are replaced by simpler characters (Fig. 2).

Optionally, the first name and the last name of the patient can be available in the Hospital Information System (HIS), with a link to his or her letter(s). In that case, the first name and last name of the patient are extracted from the HIS and are removed from the corresponding letters. For that purpose, the first name and the last name are split on spaces and punctuation marks, and each token is removed from the text. This step is optional. Finally, the titles (Mr, Mrs, Dr, etc.) and the 1 or 2 following words are removed by means of 48 regular expressions. An example is displayed in Fig. 3.

This process enables to get "prepared records". Finally, regular expressions are applied to those records in order to extract (1) a list of all the available words, as well as their frequencies, and (2) a list of all the different patterns that involve numbers (each pattern include the word before and the word after a number), as well as their frequencies.

The results of the first step are:

- a set of prepared records, that are ready to undergo the third step;
- a list of words, that will be filtered in the second step;
- a list of patterns including numbers, that will be filtered in the second step.

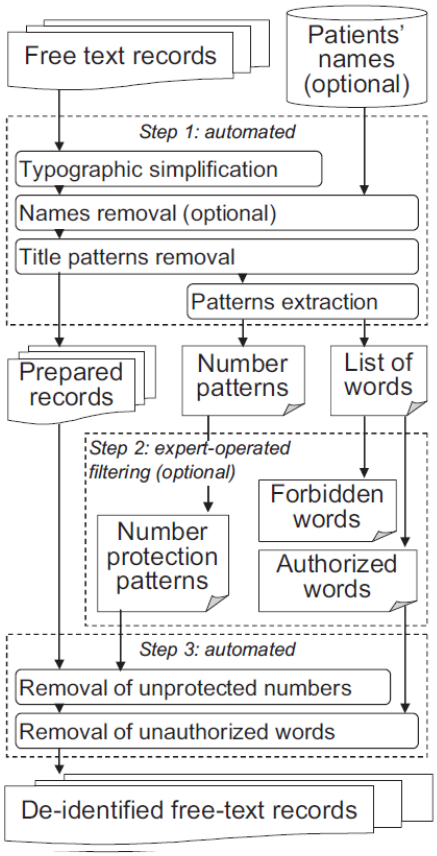


Fig. 1 – The algorithm of FASDIM consists of 3 steps. Steps 1 and 3 are fully automated. Step 2 is expert-operated but does not necessarily require to be performed for each iteration of FASDIM.

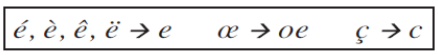


Fig. 2 – Examples of typographic simplification.

2.2. *Step 2 (expert-operated, sporadic): filtering of the lists*

The second step consists of a review and a manual filtering of the lists generated in the first step (Fig. 1). It is important to understand that this step can even be performed very fast on a little set of letters, depending of the needs. In addition, it can also be performed incrementally e.g. if there is a certain number of records to de-identify each month.

Experts are asked to review the list of patterns including numbers that are discovered in the first step. The experts can review the corresponding letters stored in the database if necessary. From the list, they define a list of "number protection patterns": all the numbers that match the patterns can be kept without confidentiality threat. This is illustrated in Fig. 4. Those number protection patterns mostly include prepositions, measure units, clinical parameters, galenic forms and drug names.

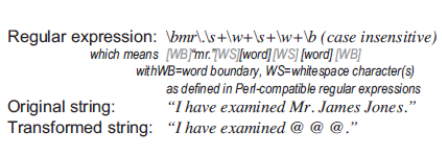


Fig. 3 – Example of title pattern removal.

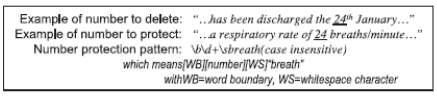


Fig. 4 – Example of number consideration.

The experts are also asked to review the list of different words discovered in the first step. They can review the corresponding letters stored in the database if necessary. They filter that list in order to get a list of "authorized words": all the words of this list can be kept without confidentiality threat. The other words are put into a list of "forbidden words". That second list is not useful for the third step, but prevents from reviewing those words again during next iterations of the second step. This second step is crucial and the list of authorized words is not simply a list of common words:

- this list should include some words that are not common words, such as:
  - some misspellings, e.g. "Ferosemide" instead of "Furosemide",
  - some medical proper names, e.g. "Prader-Willi",
  - some medical abbreviations or acronyms, e.g. "HTN" for "high blood pressure".
- this list should exclude some words despite they are common words, such as:
  - words that refer to dates, e.g. "tomorrow", "Monday", "January",
  - words that refer to places, e.g. "street", "cardiology", "hospital", "town",
  - words that are frequently present in street names, town names or names of healthcare facilities, e.g. "liberty", "square", "street" or "forest".

Many choices at this step are not obvious, and those choices are probably impacted by the context of the de-identification. However the experts are asked to value confidentiality over legibility of the de-identified text.

2.3. *Step 3 (automated): de-identification*

Finally, the numbers that match the number protection patterns defined in step 2 are protected. All the numbers that are not protected are removed from the text. All the words that do not belong to the authorized list are removed from the text. At the end of the process, the text is de-identified (Fig. 1).

2.4. *Practical use of the 3 steps*

Steps 1 and 3 are fully automated. Step 2 is an expert-operated filtering of lists and patterns. Contrary to classical pattern-matching techniques, the lists do not have to be written before

the de-identification process: they are filtered in the course of the use of the method. If a small number of records are de-identified, the list of words and patterns is short and then can be filtered very fast. FASDIM could typically be used as follows. During the first iterations, the 3 steps are performed. During next iterations, the 2nd step can possibly be discarded. The absence of 2nd step does not threat the confidentiality: the only risk is to over-scrub, i.e. to remove too many tokens from the text. Indeed, lists of words and patterns are only used to protect some tokens of the records, and never to identify the words that should be deleted, contrary to some other pattern matching methods. However the 2nd step should be performed from time to time. Another way is to directly get the list of words and patterns from another user of the method [27].

3. Material and method of the evaluation

The FASDIM method has been first developed to meet the needs of a research project, with imposed deadlines: that explains why the numbers of records at each step are not regular. Seven successive sets of unstructured discharge letters are extracted from the HIS of a general French hospital:

- A first set of 20 records used to develop and test the method.
- Successive cumulative sets of records: 7012 then 9503 then 16,009 then 17,812 then 23,493 letters.
- Finally, from the last cumulated extraction of 28,540 records, 1000 records that do not belong to the 23,493 first records are randomly selected to build an evaluation set, and are excluded from the learning set.

This way we obtain 6 cumulated learning sets (the latest one contains 27,540 letters) and 1 evaluation set of 1000 records (due to time restrictions, only the first 508 of them are annotated by the experts and used for the evaluation). The names of the corresponding patients are simultaneously extracted from the HIS, with an identifier that enables to link each patient name to the corresponding letters.

A list of the categories of PHI to remove is obtained from the HIPAA [1]. That list is complemented using the names and addresses of healthcare providers as, according to several authors, they could enable to identify the patients [4,7,11,12,22,24,28,29]. The list of PHI categories is presented in Table 2.

The evaluation consists of 3 phases. For phases 1 and 2, the 508 unstructured discharge letters are de-identified by FASDIM, using the patients' names (Fig. 5). The evaluation phases 1 and 2 are performed by 3 independent experts who are physicians and are aware of confidentiality rules and health terminologies. The third evaluation phase is performed by the developer of the tool. Ninety-five percent confidence intervals are provided when appropriate.

3.1. *First evaluation phase*

The 508 original discharge letters and the 508 de-identified letters are reviewed by an independent expert (Fig. 5, middle). The expert is in charge of counting:

Table 2 – The list of Protected Health Information categories is obtained from the HIPAA. The items marked with (*) are frequently added by authors and will be used in this work.

	Protected Health Information categories
1.	Names
2.	Geographic subdivisions smaller than a State
3.	All elements of dates (except year), ages over 89
4.	Telephone numbers
5.	Fax numbers
6.	Electronic mail addresses
7.	Social security numbers
8.	Medical record numbers
9.	Health plan beneficiary numbers
10.	Account numbers
11.	Certificate/license numbers
12.	Vehicle identifiers and serial numbers
13.	Device identifiers and serial numbers
14.	Web Universal Resource Locators (URLs)
15.	Internet Protocol (IP) address numbers
16.	Biometric identifiers, including finger and voice prints
17.	Full face photographic images
18.	Any other unique identifying number, characteristic, or code
19.	(*) Names of health personnel, or health facilities
20.	(*) Geographic location of health facilities

- the PHI tokens that are removed by FASDIM (true positives, TP),
- the PHI tokens that are not removed by FASDIM (false negatives, FN),
- the tokens that are removed by FASDIM but are not PHI (false positives, FP).

The counting process is strict. For instance, if an error appears several times in the same letter, it is counted several times also, which is not systematic in the literature [30]. The numbers above enable to compute the precision, the recall and the *F*-measure as defined in Section 1 (Eqs. (1)–(3)).

3.2. *Second evaluation phase*

If a token is falsely removed (i.e. false positive), it might unequally alter the readability of the document, and in particular the medical information: indeed, the removed token could be either an insignificant word or a medical term. The second evaluation phase deals with that issue by evaluating the preservation of medical information. For that purpose (Fig. 5, right), experts are asked to encode the anonymized discharge letters using several terminologies. Then, the same experts encode the original discharge letter using the same terminologies. It is chosen to encode exhaustively all the concepts (e.g. a disease and all the related symptoms that are described in the letter). This enables to compare the codes that are chosen after and before the anonymization process and thus to compute the preservation rate of the medical information (number of codes after/number of codes before). The terminologies are:

- The ATC for the drug names [31]
- The ICD10 for three categories of information [32]:
  - Diseases, symptoms and other factors (most of the codes)

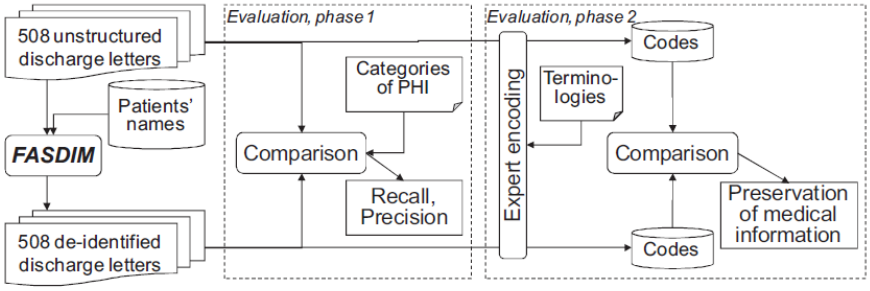


Fig. 5 – The first evaluation phase measures the recall and precision of the de-identification. The second evaluation phase measures the preservation of medical information according to three terminologies.

- Some abnormal laboratory results (e.g. R73.9 – hyperglycemia)
- Some medical procedures (e.g. Z49.1 – extracorporeal dialysis)

- The CCAM, a French classification, for the medical diagnostic or therapeutic procedures [33]

3.3. *Third evaluation phase*

The aim of the third phase is to measure the cumulated time required to implement the method and to perform the second step of the method, which consists in filtering the number patterns and the list of authorized words, this second task being from far the most important one.

3.4. *Ethics*

The persons who had access to real free-text medical letters are all physicians (doctors or students) and are bound by professional secrecy. The medical letters have been handled by respecting confidentiality rules. The study is covered by the general authorization of the hospital about observational studies. All the patients of the hospital are informed that their medical data can be used for observational studies.

4. Results of the evaluation

4.1. *First evaluation phase*

The accuracy of the de-identification is evaluated using 508 discharge letters (Table 3). The recall is 98.1% [97.8%; 98.4%], the precision is 79.6% [78.9%; 80.3%] and the F-measure is 87.9%. Many auxiliary verbs are over-scrubbed because they stand nearby family names, but it does not alter the legibility of the text. If the suppression of auxiliary verbs is ignored, the precision reaches 89.2% and the F-Measure reaches 93.4%.

Table 3 – Results of the first evaluation phase.

Measures	Values
Number of discharge letters	508
Total number of PHI	9914
Mean number of token per letter	510
Mean number of PHI per letter	19.5
False positives (FP)	2537
False negatives (FN)	183
Mean number of FN per letter	0.36
Recall (R)	98.1%
Precision (P)	79.6%
F-measure (F)	87.9%

In average, 0.36 PHI token are inappropriately preserved per letter [0.318; 0.402]. Those PHI token can be categorized as in Table 4. None of those PHIs is a patient name or a complete date.

Table 4 – Description of the false negatives (PHI inappropriately preserved).

Categories of forgotten PHI token	Proportion
Partial information about a place	63.7%
Healthcare professional's name	23%
Patient's weight	5.5%
Part of date or patient's age	4.4%
Health facility name	3.3%

4.2. *Second evaluation phase*

The preservation of medical information is evaluated through a double encoding process using 3 terminologies, before and after the de-identification. Each letter contains in average 30.6 codes from those terminologies (15,563 codes in 508 letters). Despite the de-identification, 99.02% of the codes are preserved. This rate is detailed per terminology in Table 5.

4.3. *Third evaluation phase*

The time required to develop FASDIM is displayed in Fig. 6. An incompressible time has been necessary to develop a simple and functional version of the program (12 h). Then, additional time is required for each iteration (respectively 7012, 9503, 16,009, 17,812, 23,493 then 27,540 letters) to update the number

Table 5 – Preservation rate of the medical information.

Terminology	Preservation rate
CCAM – medical procedures	99.66%
ICD10 – diagnoses, symptoms and others	99.49%
ICD10 – procedures	98.92%
ICD10 – abnormal laboratory results	96.99%
ATC – drugs	98.84%
All terminologies	99.02%

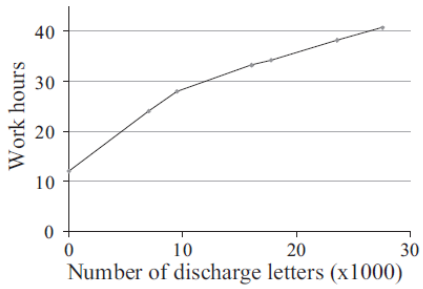


Fig. 6 – Workload as a function of the number of discharge letters to de-identify (including the development of the software).

protection patterns and mostly the list of authorized words. In summary, 28 h are necessary to de-identify 7000 letters or 40 h for 27,000 letters when no pre-existing material or piece of software is available.

Indeed, additional letters bring new unlisted words (Fig. 7), with an increasing proportion of misspellings [28], but the de-identification process has to take them into account. After de-identification of 27,540 discharge letters, the lists contain about 17,600 authorized words and 512 number protection patterns. Those lists are freely available on the Web [27] and will be updated so that it should require less time for another team to use the FASDIM method on French discharge letters.

5. Discussion

In this work, FASDIM, a Fast And Simple De-Identification Method for clinical unstructured free-text records, has been defined and evaluated in French language. The operational objectives defined in Section 1 have been reached (Table 6).

Objectives 1, 5 and 6: the method reaches very good results in French. The recall is 98.1%, the precision is 79.6% and the F-measure is 87.9%. A less strict evaluation gives a precision of 89.2% and an F-Measure of 93.4%. Moreover, 99.0% of the medical information is preserved after the identification.

Objective 2: the method does not rely on a strong knowledge of the language. Declensions, conjugations, syntax of the sentences and semantic links within sentences are ignored. That enables to get a simple and fast to develop method; however the results remain good enough.

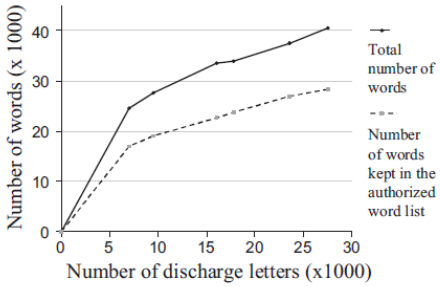


Fig. 7 – Number of different words as a function of the number of discharge letters.

Table 6 – Summary of the operational objectives defined in Section 1.

#	To design a method that:
1	reaches good results in French
2	is as independent as possible from the language structure
3	does not rely on any pre-defined corpus of words
4	can be easily reproducible by any hospital or research team
#	To evaluate the method by:
5	computing traditional evaluation metrics
6	evaluating the preservation of medical information
7	evaluating the workload required

Objective 3: the method does not rely on any pre-defined corpus of words: such a corpus is defined iteratively during the second step of the method (Fig. 1). That enables to get de-identified letters very early, and to require only little time when there are few letters to de-identify.

Objectives 4 and 7: the method can easily be reproduced with any material by any hospital or research team and a little time is required. Indeed, objectives 2 and 3 make the method very simple to implement. As a consequence, first results can be obtained fast, as shown in the workload evaluation. In summary, it seems possible to spend 12 h to develop the software, and then in average 1 h of additional work for each new set of 1000 letters. On the contrary, in traditional pattern matching methods, the lists of words and patterns must be defined before de-identifying the first letter, That evaluation enables a team who would like to use or reproduce the method to predict the required workload. It also suggests that other methods should be preferred to de-identify very large amount of letters (more than 200,000), this is probably due to the fact that FASDIM does not support language-dependent advanced features, such as declensions and conjugations.

The method has several advantages. In contrast to traditional pattern matching, no predefined list of words is necessary. This is an advantage in particular for countries and languages where such lists are not available, or when misspellings are frequent in the letters. In contrast to machine learning, no learning corpus of de-identified free-text records is required, and the method is simple to implement from scratch. Despite a strict evaluation process, the method provides results that are comparable to the best methods in English and French.

The main drawback of FASDIM is to require an additional work to filter the new words to increment the authorized word list. This task is a tedious work that requires an implicit knowledge about language, care and medicine, and medical context. In our study it has been performed by physicians who were allowed to read the letters. However, this task requires much less time than constructing or adapting exhaustive lists. The lists that have been used can be downloaded from the Web [27] or in the additional material. However, they should probably be adapted function of the context of use. Indeed the performance of de-identification methods depend on the nature of the documents [18], and de-identification may also be applied to non-medical records such as nursing documents [35]. For instance, we have considered “acacia” as being frequently associated with street names, but such a word should not be removed from records of occupational medicine or allergy.

Another example: free text written in short fields such as in electronic health records may contain more abbreviations or typos. In both examples, using the method without updating the lists of words could lead to over-scrub the text. It would alter the legibility of the documents but should not threaten patient confidentiality.

FASDIM has another drawback: it is able to delete PHIs, but it is not able to tag them or to determine their type (e.g. name, address, date, etc.). This is due to the fact that the method only identifies the token that should be preserved, not the token that should be removed, contrary to other pattern-matching methods, such as de-id [11] that uses for instance known PHI, potential PHI and PHI indicator look-up tables. The approach of FASDIM is simpler, which enables to always value confidentiality over legibility of the de-identified text. As a consequence, FASDIM cannot be used as a pseudonymization tool, contrary to other methods.

FASDIM little relies on regular expressions, contrary to other pattern matching methods such as de-id [11]. We use only 48 regular expressions to remove titles, and then the experts are able to list simple number protection patterns. However, as FASDIM is mainly based on the removal of all the token that are not protected, and not the removal of specific patterns, more complex regular expression (such as address detection) are not necessary, which makes the software easier to maintain. In the evaluation study, no letter contains a complete address after de-identification, but 63.7% of the remaining PHIs are partial information about a place. Perhaps a more complete approach based on regular expression would enable to improve that point.

During the first step of FASDIM, the patients' names are optionally extracted from the HIS and removed from the text. This operation is a useful precaution, but is not sufficient, as there are frequently misspellings of names, and not indispensable, as the title patterns removal and the removal of unauthorized words may delete the names. The title patterns removal works well for formal discharge letters where titles (Mr, Mrs, Dr, etc.) are commonly used before person names, but might be less efficient for clinical notes that are less formal. On the other hand, when the family name is a common word (e.g. "Little"), the inappropriate disappearance of such a word may enable to guess the family name. This drawback can be sidestepped with pseudonymization, which for instance consists in replacing the family names by other family names: indeed it can be decided not to replace family names that are also common names, but then the reader cannot guess whether it is the original name or a pseudonym [34]. However, the names of the patients are more easily concealable in structured text.

As in most of the scientific papers dealing with de-identification evaluation, we have considered the patients' weights as "biometric identifiers" and therefore as PHIs (and 5.5% of our false negatives are patients' weights). This strict implementation may not be appropriate in medicine, as the patients' weights are important information, e.g. for obese or malnourished patients, or drug dose calculation. However, for a practical use of FASDIM, patients' weights could easily be conserved through the use of an appropriate number protection pattern as illustrated in Fig. 4.

This work also introduces a new way to evaluate a de-identification method: the second evaluation phase estimates the proportion of medical information that has been preserved by the method. This point is important as the over-scrubbing of words has a variable importance depending on the word that is inappropriately removed. It demonstrates that the use of FASDIM would not alter the usability of the de-identified discharge letters for medical research or for activity-based payment systems.

Finally, as the FASDIM method is thought to be as language-independent as possible (cf. Objective 2), the same approach could probably be tested in other languages, although we cannot be sure it would be appropriate. However, such extension would be interesting as most of methods are designed only for English language.

6. Conclusion

FASDIM is a fast and simple algorithm that enables to de-identify French free-text discharge letters. It preserves the patient confidentiality without threatening medical information. Is seems to be suitable especially when a medium corpus of letters has to be de-identified in a limited amount of time. Examples of source code and lists of words are freely available on the web [27]. The same method should be experimented and evaluated on other types of texts, including less formal texts (such as clinical notes). Its ability to work in other languages should also be evaluated.

Role of the funding source

The study sponsor had no role on the study or the writing of the manuscript.

Author contributions

Emmanuel Chazard and Grégoire Ficheur have designed and implemented FASDIM. Capucine Mouret has performed the bibliographic analysis. Emmanuel Chazard and Capucine Mouret have designed the evaluation methodology. Capucine Mouret, Aurélien Schaffar and Jean-Baptiste Beuscart have performed the evaluation of FASDIM. The article has been written by Emmanuel Chazard and Capucine Mouret, and has been reviewed by all the authors, especially Régis Beuscart who is the department chair.

Conflicts of interest

The authors are employed by a public organization. There is no business exploitation or patent of the method presented and evaluated in this paper.

The evaluators and the developers of the system tested are not the same persons, do not belong to the same department, but are employed by the same hospital and frequently work together.

Summary points
What was already known:

- Several methods exist for free-text de-identification.
- Pattern matching methods require that dictionaries are already available.
- Machine learning require that a corpus of manually de-identified free-text is available.
- There is no freely available method for French language.

What this study added to our knowledge:

- FASDIM is a new method related to pattern matching. It brings good results in French (recall 98.1%, precision 79.6%, and F-measure 87.9%).
- The effect of de-identification can be evaluated by measuring how much of the medical content is retained after de-identification, by means of coding (e.g. ICD10, ATC, CCAM).
- FASDIM preserves 99.02% of the codes through the de-identification.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007–2013) under Grant Agreement no. 216130 – the PSIP project.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmedinf.2013.11.005.

REFERENCES

[1] Summary of the HIPAA Privacy Rule, 2013, http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary (last visited 27.11.13).
[2] D.A. Dorr, W.F. Phillips, S. Phansalkar, S.A. Sims, J.F. Hurdle, Assessing the difficulty and time cost of de-identification in clinical narratives, Methods Inf. Med. 45 (3) (2006) 246–252.
[3] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Med. Res. Methodol. 10 (2010) 70.
[4] B.A. Beckwith, R. Mahaadevan, U.J. Balis, F. Kuo, Development and evaluation of an open source software tool for deidentification of pathology reports, BMC Med. Inform. Decis. Mak. 6 (2006) 12.
[5] J.J. Berman, Concept-match medical data scrubbing. How pathology text can be used in research, Arch. Pathol. Lab. Med. 127 (6) (2003) 680–686.
[6] E.M. Fielstein, S.H. Brown, T. Speroff, Algorithmic De-identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings; Medinfo, 2004.
[7] F.J. Friedlin, C.J. McDonald, A software tool for removing patient identifying information from clinical documents, J. Am. Med. Inform. Assoc. 15 (5) (2008) 601–610.
[8] D. Gupta, M. Saul, J. Gilbertson, Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research, Am. J. Clin. Pathol. 121 (2) (2004) 176–186.
[9] C. Grouin, A. Rosier, O. Dameron, P. Zweigenbaum, Testing tactics to localize de-identification, Stud. Health Technol. Inform. 150 (2009) 735–739.
[10] F.P. Morrison, et al., Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? J. Am. Med. Inform. Assoc. 16 (1) (2009) 37–39.
[11] I. Neamatullah, M.M. Douglass, L.H. Lehman, et al., Automated de-identification of free-text medical records, BMC Med. Inform. Decis. Mak. 8 (2008) 32.
[12] P. Ruch, R.H. Baud, A.M. Rassinoux, P. Bouillon, G. Robert, Medical document anonymization with a semantic lexicon, Proc. AMIA Symp. (2000) 729–733.
[13] L. Sweeney, Replacing personally-identifying information in medical records, the Scrub system, Proc. AMIA Annu. Fall Symp. (1996) 333–337.
[14] R. Taira, A. Bui, H. Kangarloo, Identification of patient name references within medical documents using semantic selectional restrictions, Proc. AMIA Symp. (2002) 757–761.
[15] S.M. Thomas, et al., A successful technique for removing names in pathology reports using an augmented search an replace method, Proc. AMIA Symp. (2002) 777–781.
[16] S. Velupillai, H. Dalianis, M. Hassel, G.H. Nilsson, Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial, Int. J. Med. Inform. 78 (December (12)) (2009) e19–e26.
[17] UMLS, 2013, http://www.nlm.nih.gov/research/umls (last visited 27.11.13).
[18] J. Aberdeen, S. Bayer, R. Yeniterzi, et al., The MITRE Identification Scrubber Toolkit: design, training, and assessment, Int. J. Med. Inform. 79 (December (12)) (2010) 849–859.
[19] E. Aramaki, K. Miyo, Automatic deidentification by using sentence features and label consistency, in: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[20] J. Gardner, L. Xiong, HIDE: an integrated system for health information de-identification, in: Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems, 2008, pp. 254–259.
[21] K. Hara, Applying a SVM based chunker and a text classifier to the de-id challenge, in: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[22] G. Szarvas, R. Farkas, R. Busa-Fekete, State-of-the-art anonymization of medical records using an iterative machine learning framework, J. Am. Med. Inform. Assoc. 14 (October (5)) (2007) 574–580.
[23] Ö. Uzuner, T.C. Sibanda, Y. Luo, P. Szolovits, A de-identifier for medical discharge summaries, Artif. Intell. Med. 42 (January (1)) (2008) 13–35.
[24] B. Wellner, M. Huyck, S. Mardis, et al., Rapidly retargetable approaches to de-identification in medical records, J. Am. Med. Inform. Assoc. 14 (5) (2007) 564–573.
[25] L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser, L. Stoutenborough, I. Solti, Large-scale evaluation of automated clinical note de-identification and its impact on information extraction, J. Am. Med. Inform. Assoc. 20 (January (1)) (2013) 84–94, http://dx.doi.org/10.1136/amiajnl-2012-001012 (Epub August 2, 2012).

[26] O. Ferrández, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, Evaluating current automatic de-identification methods with Veteran's health administration clinical documents, BMC Med. Res. Methodol. 12 (July) (2012) 109.
[27] The FASDIM Web Page, 2013, http://www.fasdim.org (last visited 27.11.13).
[28] K. Tu, J. Klein-Geltink, T.F. Mitiku, C. Mihai, J. Martin, De-identification of primary care electronic medical records free-text data in Ontario, Canada, BMC Med. Inform. Decis. Mak. 10 (2010) 35.
[29] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (5) (2007) 550–563.
[30] M.M. Douglass, G.D. Clifford, A. Reisner, W.J. Long, G.B. Moody, R.G. Mark, Computer-Assisted De-Identification of Free-text Nursing Notes, vol. 32, 2005, pp. 331–334.
[31] Anatomical Therapeutic Chemical Classification System, 2013, http://www.whocc.no/atc_ddd_index/ (last visited 27.11.13).
[32] International Classification of Diseases, 2013, http://www.who.int/whosis/icd10 (last visited 27.11.13).
[33] French Common Classification of Medical Procedures, 2013, http://ccam.ameli.fr (last visited 27.11.13).
[34] T. Neubauer, J. Heurix, A methodology for the pseudonymization of medical data, Int. J. Med. Inform. 80 (March (3)) (2011) 190–204.
[35] H. Suominen, T. Lehtikunnas, B. Back, H. Karsten, T. Salakoski, S. Salanterä, Applying language technology to nursing documents: pros and cons with a focus on ethics, Int. J. Med. Inform. October (76) (Suppl. 2) (2007) S293–S301.
[36] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (September–October (5)) (2007) 550–563 (Epub June 28, 2007).
[37] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (June (23)) (2000) E215–E220.

Data Mining to Generate Adverse Drug Events Detection Rules

Emmanuel Chazard, Grégoire Ficheur, Stéphanie Bernonville, Michel Luyckx, and Régis Beuscart

***Abstract*—Adverse drug events (ADEs) are a public health issue. Their detection usually relies on voluntary reporting or medical chart reviews. The objective of this paper is to automatically detect cases of ADEs by data mining. 115 447 complete past hospital stays are extracted from six French, Danish, and Bulgarian hospitals using a common data model including diagnoses, drug administrations, laboratory results, and free-text records. Different kinds of outcomes are traced, and supervised rule induction methods (decision trees and association rules) are used to discover ADE detection rules, with respect to time constraints. The rules are then filtered, validated, and reorganized by a committee of experts. The rules are described in a rule repository, and several statistics are automatically computed in every medical department, such as the confidence, relative risk, and median delay of outcome appearance. 236 validated ADE-detection rules are discovered; they enable to detect 27 different kinds of outcomes. The rules use a various number of conditions related to laboratory results, diseases, drug administration, and demographics. Some rules involve innovative conditions, such as drug discontinuations.**

***Index Terms*—Adverse drug events (ADEs), data mining, decision trees, electronic health records, patient safety.**

I. Introduction

ADVERSE drug events (ADEs) endanger patients as they are the most common type of iatrogenic injury [1]. They can be defined as "injuries due to medication management rather than the underlying condition of the patient" [2]. ADEs can be split into two categories: preventable ADEs that are medication errors leading to patient harm, and nonpreventable ADEs that are called adverse drug reactions [3].

Different methods are used to identify ADEs [4]–[6], the most prominent ones being chart reviews and reporting systems. Retrospective medical chart reviews constitute the main source of reliable epidemiological knowledge on ADEs, but the method is extremely time and resource consuming. Reporting systems are the most ancient methods: they are useful for the analysis of contributing factors of ADEs, but all reporting systems suffer from important under-reporting biases [5], [7]. Another way to detect ADEs is to mine free-text reports by means of natural language processing [8]–[11], assuming that the ADEs are described in the reports, which is not frequent. Data mining is sometimes used in the field of ADE detection. But it was mainly used to analyze voluntary ADE reports [12]–[17] by means of supervised rule induction methods such as decision trees, association rules, or Bayesian neural networks, and not to analyze hospitalization records. As a consequence, the results can only be used to analyze other voluntary ADE reports.

In the literature, the automated detection of ADE cases in hospital records always relies on ADE detection rules. Whatever their origin, the ADE detection rules always consist of one or two conditions that lead to an outcome. Those conditions are simple: two drugs [18]–[25], a drug and a laboratory result [5], [18], [19], [26]–[28], [30], a drug alone [5], [18], [21], [22], a drug and one patient's characteristic [5], [18], [24], or a drug and a drug allergy [18], [24], [27]. Those works are not able to mix more complex patterns of conditions, and the effects of drug discontinuation are ignored. Those rules usually lead to overalerting, as they detect many potential cases that are not real ADE cases [29]. This is notably due to the absence of contextualization of the knowledge (the same rules are applied in every medical department) and the absence of segmentation of the population (the rules do not involve additional conditions that could increase the probability of the outcome).

II. Objectives

In order to improve the patient safety and avoid the under-reporting bias, this paper aims at automatically discovering ADEs that occurred in inpatients. This will be done by identifying situations at risk of ADE by data mining of routinely collected data of past hospitalizations. In those data, the ADEs are not explicitly flagged as no preliminary review is performed. Outpatients' ADEs leading to hospitalization will not be studied.

A list of outcomes will first be defined, and the link between those outcomes and prior drug administrations or discontinuations will be studied by means of supervised rule induction techniques applied on a training set. Rules will be obtained, in which an outcome is explained by a set of drugs in combination with a clinical background, in the form of ADE detection rules (e.g., *drug_A* & *background_B* → *outcome_C*). Then those rules will be applied onto past hospital stays of an evaluation set to get contextualized statistics such as the confidence (e.g., probability of *outcome_C* when *drug_A* and *background_B* are present).

Regarding data mining techniques, two issues have to be solved: 1) the temporal constraints have to be taken into account; and 2) we have to use supervised rule-induction methods, although the ADEs are not explicitly flagged in the routinely

Manuscript received December 25, 2010; revised May 14, 2011 and July 28, 2011; accepted July 29, 2011. Date of publication; date current version; This work was supported by the European Community's Seventh Framework Program (FP7/2007-2013) under Grant Agreement n° 216130: the PSIP Project.

The authors are with the Université Lille Nord de France, F-59000 Lille, France (e-mail: emmanuel.chazard@univ-lille2.fr; gregoire.ficheur@gmail.com; stephanie.bernonville@univ-lille2.fr; mluyckx@ch-denain.fr; regis.beuscart@univ-lille2.fr).

Digital Object Identifier 10.1109/TITB.2011.2165727

TABLE I
DESCRIPTION OF THE HOSPITALS AND STAYS USED

Hospital	Number of stays included	Age in years *mean (sd)*	Men proportion	Duration in days *mean (sd)*	Wards
French #1	50,072	52.8 (21.6)	29.2%	5.48 (6.10)	Medicine surgery obstetrics
French #2	1,367	71.4 (18.4)	42.1%	11.4 (15.1)	Geriatrics
French #3	7,846	45.4 (27.5)	51.6%	10.7 (15.3)	Geriatrics and Cardiology
Danish #1	26,245	55.6 (25.9)	40.4%	4.56 (11.8)	Medicine surgery obstetrics
Danish #2	23,067	53.1 (22.6)	44.8%	4.51 (8.41)	Medicine surgery obstetrics
Bulgarian	6,880	49.4 (16.1)	26.4%	6.96 (2.54)	Endocrinology

collected data, which are usually required in the classical rule induction method.

III. MATERIAL

A. Electronic Records of Past Hospital Stays

In order to analyze past hospital stays, data are extracted from several hospitals' electronic health records (EHRs) to feed a common repository with past fully anonymized hospital stays. The repository fits a common data model that has been designed within the PSIP Project (patient safety thought intelligent procedures in medication), a European project that aims at facilitating the development of knowledge on ADE, and improving the medication cycle in hospital environments [30], [31]. Only routinely collected data are used: no data have to be specifically recorded for the project. For each hospital stay, those data include the following.

1) Medical and administrative information (e.g., age, gender, admission date, medical department, etc.).
2) Diagnoses encoded using the International Classification of Diseases, tenth version (ICD10).
3) Medical procedures encoded using national classifications, including therapeutic and diagnostic procedures.
4) Drugs administered to the patient, encoded using the Anatomical Therapeutic Chemical classification (ATC).
5) Laboratory results encoded using the International Union of Pure and Applied Chemistry classification.
6) Anonymized free-text records, such as the discharge letter.

The data from EHRs are provided by six hospitals that are part of the PSIP Project. This study is performed using 115 447 records from six hospitals (see Table I). They allow for a four-year follow up (from 2007 to 2010).

B. ADE Detection Rules From the Summaries of Product Characteristics

In many countries, the summaries of product characteristics (SPCs) describe the official and exhaustive information about ADEs. They can be used to support the drug prescription process. They are available for healthcare professionals through websites and various supports, and for patients through patient information leaflets. In this paper, those SPCs are necessary 1) to get an exhaustive list of the possible outcomes that can be observed due to ADEs, irrespective of the causes and 2) to get a reliable set of rules to validate the results of data mining.

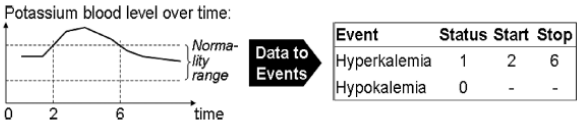


Fig. 1. Example of aggregation of laboratory results: several values of blood potassium over time enable to search for simple events.

In France, the SPCs are managed by the French drug agency (AFSSAPS, French agency for sanitary security of health products). A structured version of the SPCs is provided by the Vidal Company in the form of rules where one or two conditions lead to an outcome (e.g., Furosemide → hypokalemia). In those rules, the kind of outcome is described using free text. More than 500 000 rules are available, those rules lead to 228 different kinds of clinical or paraclinical outcomes. The paraclinical outcomes are mainly laboratory or electrocardiographic abnormal results.

IV. METHODS

A. Aggregation of the Complex Data of the Stays Into Simple Events

1) General Principles: The data described in the data repository are characterized by a complex data scheme, very numerous classes (about 17 000 codes for ICD10, about 5400 codes for the ATC, etc.) and repeated measurements throughout the hospitalization (e.g., laboratory parameters and drug administrations). Those characteristics make those data too complex to be mined using statistical methods. The aim of the data-to-event aggregation process is to automatically get a simpler representation of data for data mining purposes.

Aggregation engines are developed in order to transform the available data into information described as sets of events. For each kind of data (administrative information, diagnoses, drugs, and laboratory results), a specific aggregation engine is developed and fed with a mapping. Each mapping is described by means of extensible markup language (XML) files outside the engine. The aggregation engines enable to describe the events in terms of binary variables complemented by start and stop dates. Those engines are not static and can be adapted with respect to the context.

2) Example of Aggregation of Laboratory Results: In the example displayed in Fig. 1, for a given stay, several measures of potassium are available. Potassium is an electrolyte; its level in the blood should not reach too low or too high values; otherwise, it could lead to lethal heart arrhythmias. The aim of the aggregation process is to get simple information from those repeated measures. In the case displayed in Fig. 1, there is a hyperkalemia (too high potassium value) from day 2 to day 6 and no hypokalemia. Finally, the various measures can be summarized into two binary variables: hypokalemia = 0 and hyperkalemia = 1. In that case, a start and stop date can be added. Such variables are easier to mine using statistical methods.

3) New Variables Made Available by the Aggregation: The aggregation engines transform the data into several binary variables that can easily be mined.

1) Fifteen variables related to demographic and administrative information.
2) Forty-eight variables related to chronic diseases.
3) Five hundred variables related to drug administration or drug discontinuations. The classification considers pharmacodynamics and pharmacokinetics, although most of the existing drug classifications are based on indications.
4) Thirty-five variables related to laboratory value abnormalities.

B. Identification of the Outcomes in Relation With ADEs

As described in Section III, a list of outcomes is extracted from the summaries of product characteristics. The outcomes are traced in the data essentially by screening the laboratory results and administered drugs; this is possible through different ways depending on the category of outcome. For instance, the occurrence of a hyperkalemia (laboratory-related outcome) is directly traced using the potassium level in the blood. The occurrence of a hemorrhage under vitamin K antagonists (VKA) can be traced through different ways: 1) an increase of the international normalized ratio (INR), a laboratory parameter that rises up in case of VKA overdose; and 2) the vitamin K administration, an antidote which is prescribed in case of hemorrhage under VKA.

The structured SPC database describes 228 different kinds of outcomes. 83 (37%) of those outcomes are traceable in this paper, due to the available data. Duplicate entries are then removed; for instance, in the initial list, "hyperbilirubinemia" is also described using two synonyms, "bilirubinemia higher than twice the normal upper bound" and "jaundice." As a consequence, those 83 outcomes are traced through 56 different variables. Those outcomes correspond to life-threatening ADEs, such as hyperkalemia of hemorrhage hazard. Unfortunately, some outcomes cannot be traced in the data. This is the case especially for minor clinical incidents such as nausea or gastric pain cannot be traced. Those outcomes could correspond to ICD10 codes but in most hospitals, such codes are not flagged with a date.

C. Data-Mining-Based Induction of ADE Detection Rules

The knowledge about ADEs can be expressed using rules where some conditions lead to an outcome. Some of the variables computed by the aggregation process can be used as outcome (e.g., death) and some other ones can be used as conditions (e.g., chronic renal failure). In total, 588 variables can be used as conditions to explain 56 different outcomes. The objective here is to automatically link conditions with outcomes and then to discover ADE detection rules using data mining techniques.

After a complete review of the available data mining supervised and unsupervised techniques, and after several experiments, it was decided to use decision trees (with the CART method: Classification and Regression Trees) [32]–[39] and association rules [40]. Both methods enable to identify several decision rules containing 1 to K conditions such as

Rule A: drug_X & age≥70 → renal failure (probability= 15%)
Rule B: drug_X & age<70 → renal failure (probability= 3%)

Fig. 2. Example of two rules. A "segmentation" condition is underlined: it does not explain why the outcome occurs but deeply changes its probability.

IF(condition_1 &. . .& condition_K) THEN outcome.

The dataset (92 486 stays) is split into a learning set that is used for the rule induction (31 579 stays of the year 2007) and an evaluation set (60 907 stays of years 2008–2010). Decision trees and association rules are automatically launched for each outcome in each hospital and each medical department. The datasets are managed during the rule induction so that temporal constraints are taken into account. For a given outcome, only conditions that are compatible regarding time are tested: each condition must be an event that occurs before the outcome and is still active or has ended less than a fixed delay before the outcome occurs.

Both methods produce thousands of rules that must be filtered. Most of the outcomes are due to the patient's medical background rather than the drugs administered to him. For that reason, the rules are automatically filtered in order to keep the ones in which a drug is involved. Only the rules that increase the probability of the outcome and that have at least one of the following condition types are kept.

1) A drug administration.
2) A drug discontinuation.
3) A laboratory value that is implicitly due to a drug administration (e.g., lithium blood level > 0, INR > 1, etc.).

D. Expert Validation and Reorganization of the Rules

It is mandatory to filter, validate, and organize the rules that are obtained from the data mining: as the rules have to be used by physicians, they must provide simple, validated, and unquestionable knowledge. Several meetings are organized with external experts (physicians, pharmacologists, pharmacists, and statisticians) to filter and reorganize the set of rules. The rules are examined and validated against the SPCs and scientific references. During the review, the experts may ask for complementary queries on the potential ADE cases. At this step, the experts may manually add a few rules that are considered as mandatory although they were not discovered by the data mining process, for instance because the conditions of the rules never occur (e.g., absolute contraindication) or because the conditions occur but do not lead to any outcome.

In every rule, there is a set of conditions; the experts are asked to characterize each condition according to one of the following types.

1) *"Segmentation" conditions* are conditions that do not explain *why* an outcome occurs, but deeply change its probability. This kind of condition enables us to reduce overalerting. An example of "segmentation" condition is underlined in Fig. 2.
2) *"Subgroup" conditions* are fixed when, for some medical reasons, it does not make sense to consider the rules for all the patients in the same time. They are used to define

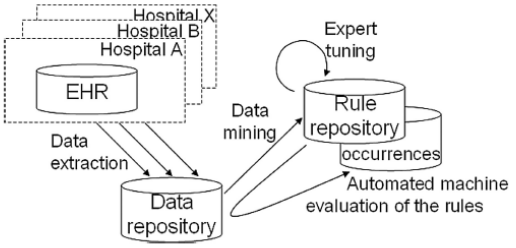


Fig. 3. Rules are stored in a rule repository. A machine evaluation automatically computes various statistics (occurrences) of the rules in every medical department.

the sample before computing the statistics. The following subgroups are systematically defined.

a) The INR deviations or vitamin K administrations are only explored for VKA-treated patients.
b) The increase of activated partial thromboplastin time is only explored for heparin-treated patients.
c) The hyperkalemia is explored separately for patients suffering from renal insufficiency or not.

3) *"Basic" conditions* group together all the other conditions.

E. Automated Computation of Contextualized Statistics About the Rules

Validated ADE detection rules are obtained from the phase of expert validation. Those rules are then evaluated from a statistical point of view to provide the users with the classical parameters such as confidence and relative risk. The statistics are contextualized, i.e., they are computed separately in each medical department or each hospital.

For that purpose, the validated rules are stored in a central rules repository [41], using an XML schema. Then, in a few minutes, an automated machine evaluation of the rules can be performed using the 60 907 stays of the evaluation set (see Fig. 3). A rule is a set of conditions leading to an outcome, such as C_1 & ... & C_k => O. Several statistics are computed for each rule, separately in every medical department.

1) Support = $P(O \cap C_1 \cap \ldots \cap C_k)$.
2) Confidence = $P(O \mid C_1 \cap \ldots \cap C_k)$.
3) Relative risk $\frac{RR=P(O/C_1 \cap \ldots \cap C_k)}{P(O/\overline{(C_1 \cap \ldots \cap C_k)}).}$
4) p value of the Fisher's exact test for independency between the outcome (O) and the set of conditions($C_1 \cap \ldots \cap C_k$).
5) Median delay between t1 (all the conditions are met) and t2 (the outcome occurs).

The cases that match the conditions of a rule and have the outcome are considered as potential ADE cases. Additional statistics are computed to describe them: number, average age, death rate, average length of stay, proportion of renal insufficiency, etc.

F. Preliminary Evaluation

An independent medical expert is asked to review the cases detected by the rules, i.e., the cases that match the conditions and the outcome of the rules with respect to temporal constraints. He is asked to assess whether each case is an ADE (the conditions are responsible from the occurrence of the outcome) or not (there is another explanation). He is also asked to review all the cases that do not match any rule but present the outcome on the track of false negatives.

V. Results

A. Overview of the Rules Obtained in This Paper

In this paper, 56 different outcomes enable to trace the potential consequences of ADEs. The supervised rule induction generates rules that predict each outcome. The rules are always filtered, validated, and tuned by the expert committee. 236 validated rules are obtained. The experts also add some rules that appear to be important in the academic knowledge and are not discovered by the data mining (e.g., the conditions never occur, or occur but not lead to the outcome). Over the 56 outcomes, we have the following.

1) Twenty-seven kinds of outcomes are observed and enable to discover ADE detection rules.
2) Ten outcomes are never or too rarely observed in the data, so that no rule is discovered. Data mining will be performed on larger datasets to get results.
3) Eighteen outcomes are observed but cannot be explained by the use of drugs in the available dataset: the medical background of the patient is a sufficient explanation, so that no rule is discovered.

The 236 rules that are obtained can be classified through the outcome they enable to predict (see Table II). Those rules can also be classified into several categories.

1) One hundred and twenty-seven rules have been discovered by data mining and confirmed by the SPCs and they bring new knowledge such as additional segmentation conditions.
2) Forty-four rules have been discovered by data mining and are not present in the SPCs but can be indirectly explained using academic knowledge. Those rules bring new knowledge in ADE detection.
3) Twenty-five rules have been discovered by data mining and already exist as is in the SPCs.
4) Forty rules have not been discovered by data mining but are important in the SPCs. They have been enforced by the experts and do not produce significant statistics. The contribution of this paper is to compute statistics about those rules and quantify their usefulness.

B. Example of an ADE Detection Rule and Related Statistics

The following rule is generated by data mining:
Vitamin K antagonist & anti-diarrheal drug → INR ≥ 5.
The rule can be explained as follows: in case of diarrhea, the VKA absorption is decreased, which is probably balanced by an increase of the VKA dose. Once an antidiarrheal drug is administered, the VKA absorption is restored. In the absence of dose adjustment, this leads to a VKA overdose detected by an INR value over 5, which can lead to a hemorrhage. The statistics that are related to that rule for the year 2009 are described

TABLE II
OUTCOMES AND NUMBER OF ADE DETECTION RULES

Outcome	Rules
Coagulation disorders	
Hemorrhage (detected by the administration of haemostatic)	7
Heparin overdose (activated partial thromboplastin time>1.23)	5
VKA overdose (INR>4.9)	57
VKA overdose (detected by the administration of vitamin K)	2
VKA underdose (INR<1.6)	18
Thrombocytosis (count>600,000)	5
Thrombopenia (count<75,000)	24
Nosocomial infections	
Bacterial infection (detected by the administration of antibiotic)	4
Fungal infection (detected by the administration of a systemic antifungal)	8
Fungal infection (detected by the admin. of local antifungal)	2
Ionic and renal disorders	
Hyperkalemia (K^+>5.3 mmol/l)	63
Hypocalcemia (Ca^{++} <2.2 mmol/l)	1
Hypokalemia (K^-<3.0 mmol/l)	1
Hyponatremia (Na^+<130 mmol/l)	2
Renal failure (creatinine>135 µmol/l or urea>8 mmol/l)	8
Others	
Acetaminophen overdose (detected by the administration of N-acetyl-cystein)	1
Anemia (Hb<10g/dl)	2
Diarrhea (detected by the administration of an anti-diarrheal)	1
Diarrhea (detected by the administration of an antipropulsive)	1
Hepatic cholestasis (alkaline phosphatase>240 UI/l or bilirubins>22 µmol/l)	3
Hepatic cytolysis (alanine transaminase>110 UI/l or aspartate transaminase>110 UI/l)	4
High CPK level (CPK>195 UI/l)	2
Hypereosinophilia (eosinophilocytes>10^9/l)	4
High level of pancreatic enzymes (amylase>90 UI/l or lipase>90 UI/l)	7
Lithium overdose (to high a lithium rate)	1
Neutropenia (count<1,500/mm^3)	2
Pancytopenia	1
Total	*236*

TABLE III
EXAMPLE OF A RULE: VKA & ANTI-DIARRHEAL → VKA OVERDOSE (INR≥5)

Hospital	Confidence	Support	Median delay	Relative risk	Fisher's exact test
N°1	9/41=22%	9/6110=1.5‰	3 days	16.45	p=0
N°2	2/9=22.2%	2/11923=0.2‰	2 days	75.64	p=0.0003
N°3	0/2=0%	0/1022=0‰		0	p=1
N°4	0/8=0%	0/7685=0‰		0	p=1
N°5	0/1=0%	0/1816=0‰		0	P=1
N°6	0/2=0%	0/6880=0‰		0	p=1

Confidence	Year 2007	Year 2008	Year 2009	Year 2010
Hospital N°1	6/43=14.0%	8/46=17.4%	9/41=22.0%	6/36=16.7%

in Table III, as well as the follow-up of one hospital during four years. Each line of the table displays the results obtained in each of the studied hospitals. In fact, the results are available for each medical department of those hospitals. In both hospitals 1 and 2, the probability of VKA overdose once the conditions are met is around 20%, with a significant increase of the risk, and a median delay of two or three days. In hospitals 3, 4, and 5, no case of VKA overdose is observed when the same conditions are matched. It is interesting to notice that even for a validated rule, its confidence may vary a lot with respect to the hospital or medical department. This is probably due to the fact that the patients (demographic and medical background), the medication processes, and the monitoring policies are different. For instance, in some departments, the nurses are quite self-powered for "comfort" drug administration. In other departments, the INR is not frequently monitored even in case of change in the medication. An immediate consequence of the results of Table III is that the use of such a rule in a clinical decision support system (CDSS) would lead to about 78% of false alerts in hospitals 1 and 2, which is acceptable, but also to 100% of false alerts in hospitals 3, 4, and 5, which is not acceptable.

TABLE IV
EVALUATION OF ADE DETECTION IN THE FIELD OF HYPERKALEMIA

Measure	Value
Number of stays	14,747
Number of hyperkalemia cases	117 (7.93‰)
Recall	39/41=95.1%
Precision	39/75=52.0%
F-Measure	67.2%
Number of cases reported	0 (0%)
Cases above 6 mmol/l	11/41=26.8%
Administration of Kayexalate	12/41=29.3%

C. Preliminary Evaluation

A preliminary evaluation has been conducted exhaustively on the hyperkalemia cases of hospital n°1 during the year 2010 (14 747 stays). The results are reported in Table IV. None of the cases detected in the review had been reported to the patient safety unit or to official agencies, although the potassium was above 6 mmol/l in 26.8% cases, and there was an administration of Kayexalate, a potassium chelator, in 29.3% cases. This review is being continued in all the fields covered by the rules.

VI. DISCUSSION

A. Overview

In this study, data about 115 447 past hospital stays are collected and prepared for data mining. A list of potential outcomes is obtained from the SPCs and 56 of them are traced in the data. By means of decision trees and association rules, decision rules are extracted from a training set (34% of the stay). An expert committee filters and validates those rules: 236 validated ADE detection rules are obtained, and statistics are automatically computed in an evaluation set (66% of the stays).

B. Discussion of the Method

This method is able to automatically discover ADE detection rules. Some are already known and validated. In addition, the method enables to discover new knowledge, such as segmentation conditions or unknown rules. The academic knowledge does not provide any probability of the ADEs. In this paper, we are able to sort the rules by confidence and to prioritize the knowledge. Each one of the 236 ADE detection rules is automatically complemented with contextualized statistics, i.e., statistics computed separately in every hospital or medical department. As shown in the example in Table III, the confidence

often varies a lot with respect to the place a rule is applied. Those differences might be due to latent variables that are not observed in the data, such as the risk monitoring policies or the medical background of the patient.

A drawback of the method is that only the data that are recorded can be mined. In this paper, we are not able to detect clinical events that are not registered in routinely collected data, e.g., rash, nausea, stomach pain, etc. The patient's weight and known drug allergies could have been used, but this information was not sufficiently present in the dataset. The drugs prescribed shortly before the hospitalization were not available and could not be analyzed. Finally, as the rule induction is data mining based, events that never or too rarely occur do not enable to discover rules, which is the case here for 11 outcomes. For that reason, the experts were allowed to add some important rules that never occur into the rule base, such as absolute contraindications. In that case, this paper contributes to compute the statistics about those academic rules. The same method could provide interesting results using other data where the outcomes occur as soon as they are available in databases, such as electrocardiographic records or oxymetry records.

For the data mining phase, the data have to be simplified. For instance, the duration and dose of medications have been ignored, as well as the numeric value of the laboratory results. However, the rules so obtained can be enriched by such parameters later, for instance, in a CDSS, for prospective ADE prevention.

Producing ADE detection rules by data mining is complex. Indeed, the ADE cases are not flagged in the data: when hyperkalemia can be observed, we do not simply know if it is an ADE or not. However, the supervised rule induction methods are used to get some rules that predict hyperkalemia, and in this paper, we try to obtain rules that predict hyperkalemia *in the frame of an ADE*. Yet most of the outcomes are principally due to the patients' diseases, and occasionally due to drugs. For that reason, an automated filtering and an expert filtering and reorganization of the rules are performed. As the decision trees are launched several times in different department and on different periods, their instability is not a problem and provides experts with several partially redundant rules, as the association rules do. Once the rules have been filtered and modified by the experts, they are automatically evaluated in all the medical departments using the evaluation set.

Some authors have developed specific rule-induction methods that deal with temporal aspects [42]–[45]. These methods try to discover some events that, in a given order, lead to an outcome. Regarding ADEs, those methods appear not to be relevant because the order of appearance of the conditions is not overriding, but the conditions have to be active simultaneously. It is not a problem of order of appearance, but a problem of concomitant presence and delay up to the condition. In addition, the discontinuation of a drug itself is a kind of event. For all these reasons, the temporal conditions are analyzed and filtered before the rule induction to ensure that all the events that are candidate to explain an outcome are compatible with the outcome regarding time. Then, the same constraints are applied for the rule automated evaluation.

C. Discussion of the ADE Detection Rules Discovered

This study enables us to automatically discover ADE detection rules by means of data mining techniques; the rules are then filtered and validated by experts. The rules consist of a set of conditions that lead to an effect, those conditions being related to demographic characteristics, drug administrations (without dose), laboratory results, or diagnoses. The number of conditions is not constrained by the method, and the output provides more complex rules than in other studies. In addition, this study takes into account the effects of drug discontinuation. Some previous works have involved segmentation conditions, such as the age, the renal function, the hepatic function, and the patient's weight [18], [24]. This study does so (53% of the 236 rules), except that the patient's weight is not available in the data.

One of the main risks of the systems developed to detect or prevent ADEs is overalerting. This is easily understandable when the official SPCs describe situations at risk of ADEs by means of thousands of rules. This leads to low positive predictive values and makes the system unreliable. Other authors describe sets of rules [29] but most of them lead to overalerting, notably because the rules are too simple and rarely involve segmentation conditions, i.e., conditions that are not directly responsible from the outcome but change its probability. In addition, those works do not support contextualization, i.e., the fact that the confidence of a rule varies deeply with respect to the place.

The rules discovered in this study mainly deal with the effects of anticoagulant drugs (35% of the rules) and hyperkalemia (27% of the rules). This paper also highlights the importance of pharmacokinetic drug-to-drug interactions (25% of the rules) that are often underestimated. Contrary to the accepted wisdom, many "important" ADE detection rules are not discovered by data mining because either the conditions never occur or, when the conditions are present, the outcome never occurs. This is probably because those rules are well known and, consequently, the risk is well monitored. However, it is possible to input the corresponding rules and enforce their automated evaluation.

D. Exploitation of the Results

Except the rules filtering performed by the experts, the whole process is fully automated. In order to analyze the data of a new hospital, about 1 h is required for 10^5 stays. The 236 rules and the related files (description of the mappings, lexicon, statistics computed in the automated evaluation, and textual explanations) consist of a set of XML files. The format of those files is well documented, so that it is easy to use them in any ADE detection or prevention application.

VII. Conclusion

This paper brings innovative and semiautomated solutions for ADE detection. The method is quite generic and could be applied to other kinds of data as soon as they are available in the EHR, such as structured results of electrocardiograms. The results of the method used here bring an important contribution to ADE knowledge. The rules that are obtained are versatile and can be used either as detection rules on past hospital stays, or

as prevention rules in a CDSS context. Those rules are already loaded in several prototypes that are developed in the frame of the PSIP Project.

1) A tool designed for retrospective ADE detection and follow-up in past hospitalizations: the Scorecards [46].
2) A knowledge-based system for prospective ADE prevention during the medication process, which is used by three CDSS: one embedded in a computerized physician order entry, another embedded in an EHR, and a prescription simulation tool that is available even without any Hospital information system.

REFERENCES

[1] L. T. Kohn, J. M. Corrigan, and M. S. Donaldson, *To Err Is Human: Building a Safer Health System*. Washington, DC: Natl. Acad., 1999.

[2] Institute of Medicine, *Preventing Medication Errors*. Washington, DC: Natl. Acad., 2007.

[3] T. K. Gandhi, D. L. Seger, and D. W. Bates, "Identifying drug safety issues: From research to practice," *Int. J. Qual. Health Care*, vol. 12, no. 1, pp. 69–76, Feb. 2000.

[4] D. W. Bates, R. S. Evans, H. Murff, P. D. Stetson, L. Pizziferri, and G. Hripcsak, "Detecting adverse events using information technology," *J. Am. Med. Inf.*, vol. 10, no. 2, pp. 115–128, Mar./Apr. 2003.

[5] T. Morimoto, T. K. Gandhi, A. C. Seger, T. C. Hsieh, and D. W. Bates, "Adverse drug events and medication errors: Detection and classification methods," *Qual. Saf. Health Care*, vol. 13, no. 4, pp. 306–314, Aug. 2004.

[6] R. Amalberti, C. Gremion, Y. Auroy, P. Michel, R. Salmi, P. Parneix, J. L. Quenon, and B. Hubert, "Typologie et méthode d'évaluation des systèmes de signalement des accidents médicaux et des événements indésirables," DRESS, Paris, France, Report, 2006.

[7] H. J. Murff, V. L. Patel, G. Hripcsak, and D. W. Bates, "Detecting adverse events for patient safety research: A review of current methodologies," *J. Biomed. Inf.*, vol. 36, no. 1–2, pp. 131–143, Feb./Apr. 2003.

[8] M. N. Cantor, H. J. Feldman, and M. M. Triola, "Using trigger phrases to detect adverse drug reactions in ambulatory care notes," *Qual. Saf. Health Care*, vol. 16, no. 2, pp. 132–134, Apr. 2007.

[9] M. Gysbers, R. Reichley, P. M. Kilbridge, L. Noirot, R. Nagarajan, W. C. Dunagan, and T. C. Bailey, "Natural language processing to identify adverse drug events," in *Proc. AMIA Annu. Symp. Proc.*, 2008, p. 961.

[10] G. B. Melton and G. Hripcsak, "Automated detection of adverse events using natural language processing of discharge summaries," *J. Am. Med. Inf. Assoc.*, vol. 12, no. 4, pp. 448–457, Jul./Aug. 2005.

[11] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe, "Extraction of adverse drug effects from clinical records," *Stud. Health Technol. Inf.*, vol. 160 (Pt 1), pp. 739–743, 2010.

[12] J. Almenoff, J. M. Tonning, A. L. Gould, A. Szarfman, M. Hauben, R. Ouellet-Hellstrom, R. Ball, K. Hornbuckle, L. Walsh, C. Yee, S. T. Sacks, N. Yuen, V. Patadia, M. Blum, M. Johnston, C. Gerrits, H. Seifert, and K. Lacroix, "Perspectives on the use of data mining in pharmacovigilance," *Drug Safety*, vol. 28, no. 11, pp. 981–1007, 2005.

[13] J. S. Almenoff, E. N. Pattishall, T. G. Gibbs, W. DuMouchel, S. J. Evans, and N. Yuen, "Novel statistical tools for monitoring the safety of marketed drugs," *Clin. Pharmacol. Ther.*, vol. 82, no. 2, pp. 157–166, Aug. 2007.

[14] A. Bate and I. R. Edwards, "Data mining in spontaneous reports," *Basic Clin. Pharmacol. Toxicol.*, vol. 98, no. 3, pp. 324–330, Mar. 2006.

[15] C. L. Bennett, J. R. Nebeker, P. R. Yarnold, C. C. Tigue, D. A. Dorr, J. M. McKoy, B. J. Edwards, J. F. Hurdle, D. P. West, D. T. Lau, C. Angelotta, S. A. Weitzman, S. M. Belknap, B. Djulbegovic, M. S. Tallman, T. M. Kuzel, A. B. Benson, A. Evens, S. M. Trifilio, D. M. Courtney, and D. W. Raisch, "Evaluation of serious adverse drug reactions: A proactive pharmacovigilance program (RADAR) vs safety activities conducted by the Food and Drug Administration and pharmaceutical manufacturers," *Arch. Int. Med.*, vol. 167, no. 10, pp. 1041–1049, May 2007.

[16] D. M. Coulter, A. Bate, R. H. Meyboom, M. Lindquist, and I. R. Edwards, "Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: Data mining study," *BioMed. J.*, vol. 322, no. 7296, pp. 1207–1209, May 2001.

[17] M. Hauben, V. Patadia, C. Gerrits, L. Walsh, and L. Reich, "Data mining in pharmacovigilance: The need for a balanced perspective," *Drug Safety*, vol. 28, no. 10, pp. 835–842, 2005.

[18] D. W. Bates, A. C. O'Neil, D. Boyle, J. Teich, G. M. Chertow, A. L. Komaroff, and T. A. Brennan, "Potential identifiability and preventability of adverse events using information systems," *J. Am. Med. Inf. Assoc.*, vol. 1, no. 5, pp. 404–411, 1994.

[19] A. K. Jha, G. J. Kuperman, J. M. Teich, L. Leape, B. Shea, E. Rittenberg, E. Burdick, D. L. Seger, M. Vander Vliet, and D. W. Bates, "Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report," *J. Am. Med. Inf. Assoc.*, vol. 5, no. 3, pp. 305–314, 1998.

[20] G. D. Fiol, B. H. Rocha, G. J. Kuperman, D. W. Bates, and P. Nohama, "Comparison of two knowledge bases on the detection of drug-drug interactions," in *Proc AMIA Symp.*, 2000, pp. 171–175.

[21] T. K. Gandhi, S. N. Weingart, A. C. Seger, J. Borus, E. Burdick, E. G. Poon, L. L. Leape, and D. W. Bates, "Outpatient prescribing errors and the impact of computerized prescribing," *J. Gen. Int. Med.*, vol. 20, no. 9, pp. 837–841, 2005.

[22] J. Judge, T. S. Field, M. DeFlorio, J. Laprino, J. Auger, P. Rochon, D. W. Bates, and J. H. Gurwitz, "Prescribers' responses to alerts during medication ordering in the long term care setting," *J. Am. Med. Inf. Assoc.*, vol. 13, no. 4, pp. 385–390, 2006.

[23] M. D. Paterno, S. M. Maviglia, P. N. Gorman, D. L. Seger, E. Yoshida, A. C. Seger, D. W. Bates, and T. K. Gandhi, "Tiering drug-drug interaction alerts by severity increases compliance rates," *J. Am. Med. Inf. Assoc.*, vol. 16, no. 1, pp. 40–46, 2009.

[24] A. Schedlbauer, V. Prasad, C. Mulvaney, S. Phansalkar, W. Stanton, D. W. Bates, and A. J. Avery, "What evidence supports the use of computerized alerts and prompts to improve clinicians' prescribing behavior?," *J. Am. Med. Inf. Assoc.*, vol. 16, no. 4, pp. 531–538, 2009.

[25] J. M. Teich, J. P. Glaser, R. F. Beckley, M. Aranow, D. W. Bates, G. J. Kuperman, M. E. Ward, and C. D. Spurr, "The Brigham integrated computing system (BICS): Advanced clinical systems in an academic hospital environment," *Int. J. Med. Inf.*, vol. 54, no. 3, pp. 197–208, 1999.

[26] G. J. Kuperman, J. M. Teich, D. W. Bates, F. L. Hiltz, J. M. Hurley, R. Y. Lee, and M. D. Paterno, "Detecting alerts, notifying the physician, and offering action items: A comprehensive alerting system," in *Proc. AMIA Annu. Fall Symp.*, 1996, pp. 704–708.

[27] B. Honigman, J. Lee, J. Rothschild, P. Light, R. M. Pulling, T. Yu, and D. W. Bates, "Using computerized data to identify adverse drug events in outpatients," *J. Am. Med. Inf. Assoc.*, vol. 8, no. 3, pp. 254–266, 2001.

[28] T. S. Field, J. H. Gurwitz, L. R. Harrold, J. M. Rothschild, K. Debellis, A. C. Seger, L. S. Fish, L. Garber, M. Kelleher, and D. W. Bates, "Strategies for detecting adverse drug events among older persons in the ambulatory setting," *J. Am. Med. Inf. Assoc.*, vol. 11, no. 6, pp. 492–498, 2004.

[29] S. M. Handler, R. L. Altman, S. Perera, J. T. Hanlon, S. A. Studenski, J. E. Bost, M. I. Saul, and D. B. Fridsma, "A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting," *J. Am. Med. Inf. Assoc.*, vol. 14, no. 4, pp. 451–458, Jul./Aug. 2007.

[30] E. Chazard, B. Merlin, G. Ficheur, J. C. Sarfati, and R. Beuscart, "Detection of adverse drug events: Proposal of a data model," *Stud. Health Technol. Inf.*, vol. 148, pp. 63–74, 2009.

[31] *Patient Safety by Intelligent Procedures in Medication*. (Dec. 2010). [Online]. Available at http://www.psip-project.eu

[32] L. Breiman, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[33] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Knowledge Discovery and Data Mining*. Menlo Park, CA: Amer. Assoc. Artif. Intell., 1996.

[34] N. Lavrac, "Selected techniques for data mining in medicine," *Artif. Intell. Med.*, vol. 16, no. 1, pp. 3–23, May 1999.

[35] R. Quilan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kauffman, 1993.

[36] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[37] H. P. Zhang, J. Crowley, H. Sox, and R. A. Olshen, *Tree Structural Statistical Methods. Encyclopedia of Biostatistics*. Chichester, U.K: Wiley, 2001, pp. 4561–4573.

[38] T. M. Therneau, B. Atkinson, and B. Ripley, *Rpart: Recursive Partitioning*, 2007.

[39] R_Development_Core_Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation Statistical Comput., 2008.

[40] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Washington, DC, May1993, pp. 207–216.

[41] E. Chazard, G. Ficheur, B. Merlin, E. Serrot, and R. Beuscart, "Adverse drug events prevention rules: Multi-site evaluation of rules from various sources," *Stud. Health Technol. Informat.*, vol. 148, pp. 102–111, 2009.
[42] M. Campos, J. Palma, and R. Marìn, "Temporal data mining with temporal constraints," in *Proceedings of the 11th Conference on Artificial Intelligence in Medicine*. Amsterdam, The Netherlands: Springer-Verlag, 2007, pp. 67–76.
[43] L. Sacchi, C. Larizza, C. Combi, and R. Bellazzi, "Data mining with temporal abstractions: Learning rules from time series," *Data Mining Knowl. Discov.*, vol. 15, pp. 217–247, 2007.
[44] J. M. Ale and G. H. Rossi, "An approach to discovering temporal association rules," in *Proc. ACM Symp. Appl. Comput.*, 2000, pp. 294–300.
[45] M. Stacey and C. McGregor, "Temporal abstraction in intelligent clinical data analysis: A survey," *Arif. Intell. Med.*, vol. 39, pp. 1–24, 2007.
[46] R. Marcilly, E. Chazard, M. C. Beuscart-Zephir, W. Hackl, A. Baceanu, A. Kushniruk, and E. Borycki, "Design of adverse drug events-scorecards," *Stud. Health Technol. Informat.*, vol. 164, pp. 377–381, 2011.

Emmanuel Chazard was born in Agen, France, in 1977. He received the M.D. and Ph.D. degrees in 2007 and 2011, respectively, from the Lille University of Law and Health Sciences, Lille, France.

He is currently a Physician at the Lille University Hospital, Lille, France. He is also a Teacher and Researcher at the Lille University, Lille, in the fields of public health, biostatistics, and medical informatics.

Grégoire Ficheur, photograph and biography not available at the time of publication.

Stéphanie Bernonville, photograph and biography not available at the time of publication.

Michel Luyckx, photograph and biography not available at the time of publication.

Régis Beuscart, photograph and biography not available at the time of publication.

Routine use of the "ADE Scorecards", an application for automated ADE detection, in a general hospital

Emmanuel Chazard[a], Michel Luyckx[b], Jean-Baptiste Beuscart[c], Laurie Ferret[a], Régis Beuscart[a]

[a] *Public Health Department, Lille University Hospital; UDSL EA 2694; Univ Lille Nord de France; F-59000 Lille, France.*
[b] *Hospital Pharmacy, Lille University Hospital; UDSL EA 4481; Univ Lille Nord de France; F-59000 Lille, France.*
[c] *Geriatrics Department, Lille University Hospital; UDSL EA 2694; Univ Lille Nord de France; F-59000 Lille, France.*

Abstract

Retrospective detection of Adverse Drug Events (ADEs) is challenging, notably because ADEs result from complex interactions between many factors. Data mining techniques have recently emerged in the field of automated retrospective ADE detection. The "ADE Scorecards" are a research application based on data-mining that has been built in the frame of the PSIP European Project, and enables for automated potential ADE retrospective detection. The objective of this paper is to evaluate the use of the ADE Scorecards in real-life healthcare situation. For that purpose, the ADE Scorecards have been implemented in a French general hospital and have been used by the physicians and pharmacists during three years (corresponding to 73,000 inpatient stays). According to the results, 2% of the analyzed inpatient stays have a potential ADE with hyperkalemia, and 1% of them have a potential ADE with vitamin K antagonist overdose. In practice, the application, which was first designed to be a standalone web-based application for the physicians, has been used as a part of a more global quality improvement approach led by the pharmacists.

Keywords:

Adverse Drug Events, Adverse Drug Reactions, Data Mining, Data Reuse, Electronic Health Records.

Introduction

Adverse Drug Events (ADEs) can be defined as "injur(ies) due to medication management rather than the underlying condition of the patient" [1]. That definition emphasizes that ADEs are due to a combination of causes, including drugs (drug administration, dose variations, and drug discontinuations) and characteristics of the patient (such as age, diseases, renal and hepatic functions) [2]. That complexity explains why a certain skill is required to properly detect ADE cases.

Retrospective ADE detection consists in analyzing past hospital stays to discover cases where ADEs occurred. Several approaches have been developed in that field [3-4], and can be grouped into 2 categories: expert-operated methods and automated methods. The first ones mainly consist of retrospective medical chart reviews and reporting systems. The development of automated methods is more recent and tries to address the under-declaration and under-detection biases. Those methods are natural language processing of discharge summaries [5-8], and data mining of electronic health records (EHRs) [9].

Based on data mining, an application has been developed within the PSIP European Project (Patient Safety through Intelligent Procedures in medication). This application, named "ADE Scorecards" [10], is a surveillance tool that enables to automatically detect potential past ADE cases by highlighting the potential causes (drugs, biological context, demographics, etc) and the outcome. Those potential ADE cases can then be confirmed by experts and used for physicians' training. This application has been installed in five hospitals (2 Danish, 2 French and 1 Bulgarian) as a proof of concept. It has been routinely used by the physicians and pharmacists of a French general hospital during three years.

The objective of this paper is to present the application and show the results of its use in real-life situation.

Materials and Methods

EHRs from the Denain General Hospital

A structured description of past hospital stays is automatically extracted from the EHRs of the Denain General Hospital, in the North of France. Those records fit a data model that has been designed previously [11], and only uses routinely-collected data: no additional data has to be specifically recorded. The data model includes medical and administrative information (e.g. age, gender, admission date), diagnoses (ICD10 codes), medical procedures, drugs administered daily to the patient (ATC codes), laboratory results (IUPAC codes), and free-text records anonymized using the FASDIM procedure [12].

Adverse Drug Events detection rules

The knowledge about ADEs is generally described using ADE detection rules. An ADE detection rule is made of one or several Boolean conditions that may lead to an outcome, with a given probability, such as C_1 & ... & $C_k \rightarrow O$. This is a simplified notation, as in addition the rule implicitly requires that time constraints are respected: the condition must precede the outcome, and still be active when the outcome occurs. That representation is widely used either for prospective ADE prevention or retrospective ADE detection [13]. In this work we use a set of 236 rules that have been discovered in a previous work by data mining of EHRs [9]. In that work, routinely-collected inpatient data have been used to identify potential outcomes. Then, by mean of data-mining techniques (such as decision trees and association rules), conditions statistically associated have been identified. Finally, the rules have been

filtered, reorganized and validated by pharmacology experts. Those rules involve 1 to 4 conditions (demographic characteristics, drug administrations or discontinuations, laboratory results, or diagnoses) and an outcome that can be detected in the data (e.g. "Hyperkalemia", "International Normalized Ratio elevation", etc.). The rules include 56 kinds of outcomes. They are described as a set of structured XML files, including:

- mappings that enable to transform the raw data into Boolean variables with temporal attributes,
- the set of rules,
- a lexicon for automated translation into English, French, Danish or Bulgarian, and
- a set of free-text explanations, that explain each rule and provide with bibliographic references in each language.

The ADE Scorecards, a retrospective ADE detection tool

The ADE Scorecards are a web-based application that enables to detect and display potential past ADE cases in a user-friendly interface. The process relies on two steps (Figure 1).

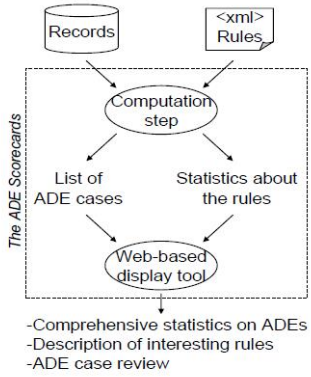


Figure 1 – The 2 parts of the ADE Scorecards

Rule: $C_1 \cap ... \cap C_k \rightarrow O$

Time constraints:
For $(C_1 \cap ... \cap C_k)$: all the conditions are present in the same time (they can start at different times)
For $(O \cap C_1 \cap ... \cap C_k)$: the same as above, and all the conditions are present before the outcome starts.

Support: $Sup = P(O \cap C_1 \cap ... \cap C_k)$

Confidence: $Conf = P(O / C_1 \cap ... \cap C_k)$

Risk ratio: $RR = \frac{P(O / C_1 \cap ... \cap C_k)}{P(O / \overline{(C_1 \cap ... \cap C_k)})}$

P value: p value of the Fisher's exact test for independency between the outcome O and the set of conditions $(C_1 \cap ... \cap C_k)$

Delay: median delay between $Time(C_1 \cap ... \cap C_k)$ and $Time(O)$ (when both events occur).

Figure 2 – Contextualized statistics (underlined)

The *computation step* consists in running the ADE detection rules onto the inpatient stays that are extracted from the EHRs. Several contextualized statistics are computed for each rule in each medical department (Figure 2). They enable a contextualized behavior of the application. Indeed, previous works have demonstrated the need for such contextualization in the field of ADE detection [14]. As a result of that first step, several inpatient stays are flagged as "potential ADEs" (those cases are not always real ADEs, they have to be confirmed).

The second step, *web-based display tool*, consists in displaying the potential ADE cases, the related ADE detection rules, and statistics. For that purpose, a web-based application has been developed using a Human-centered design process [15]. It preserves the anonymity of the patients. Finally, the users are provided with contextualized information: the potential ADE cases are detected in their medical unit, the statistics are contextualized, and only the ADE detection rules that are useful in their medical unit are displayed. A free demonstration of the application is available on the Web [16]. Finally, a case facility enables to visualize each potential ADE case, enabling the physicians to making their own opinion on the detected cases.

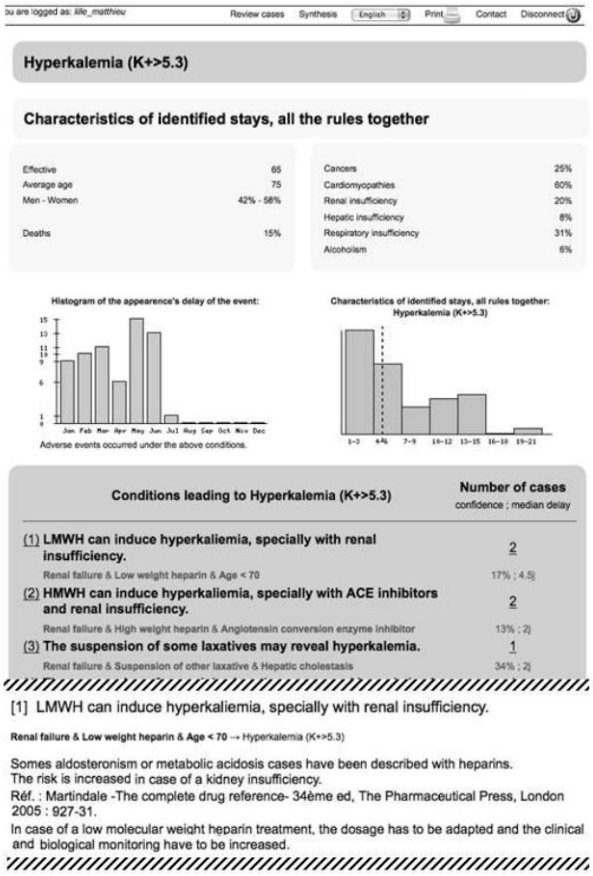


Figure 3 – Scorecard of potential ADEs with hyperkalemia (hatched lines denote a truncation of the screenshot)

Quantitative observation of the application

The data are analyzed from January 2007 to August 2012. Some statistics are computed to describe the patients and their medical background, especially the characteristics that are known to have a strong interference with medications. Comparisons are performed using Chi-2 tests and Student's t-tests with a 5% alpha risk. Confidence intervals are computed with a 5% alpha risk.

The number of potential ADE cases detected by the application is computed in the medical departments and year after year. A focus is made over the most frequent outcomes:

- Hyperkalemia: it is defined as K^+>5.5mmol/l in this case. This ionic trouble may induce lethal cardiac rhythm troubles.
- INR increase: it is defined as INR>5 in this case (INR=international normalized ratio of the prothombin time). Such a disorder could induce a severe hemorrhage. A frequent cause is a VKA overdose or a VKA biological availability increase (VKA=vitamin K antagonists).

Qualitative preliminary evaluation of the use of the application in real-life

The daily use of the Scorecards is observed from January 2010 to December 2012 by a human-factors specialist and a pharmacist. They observe how the application is used in practice by analyzing log files, interviewing users and analyzing staff meetings that use the outputs of the tool. This observation is the preliminary study of a more complete and structured evaluation.

Results

Description of the inpatient stay database

The number of inpatient stays analyzed is in Table 2. Only the inpatient stays that present the following characteristics are analyzed: there is at least one drug administration, the patient is hospitalized during at least 2 days, and the patient is hospitalized in a medical department where the ADE Scorecards are implemented. The general characteristics of the inpatient stays are displayed in Table 1. Those characteristics are compared between the medical departments in Table 3.

Table 1 – Description of the inpatients stays analyzed

Parameter	Estimated value
General characteristics	
Age (years)	60.2
Length of stay (days)	8.01
Proportion of men	40.8%
Abnormal laboratory results	
INR increase	2.46%
Hyperkalemia	5.43%
Chronic diseases (ICD10 codes)	
Renal insufficiency	2.02%
Hepatic insufficiency	4.90%
Administered drugs	
VKA	8.34%
Diuretics	23.3%
Main medical department	
Cardiology	24.4%
Geriatrics	3.75%
Gynecology Obstetrics	10.0%
Internal medicine	18.4%
Pneumology	15.8%
Surgery	27.7%

Table 2 – Number of inpatient stays and stays analyzed

Year	Total number of stays	Number of stays analyzed
2007	10,244	6,084
2008	11,338	6,271
2009	12,469	6,215
2010	14,747	6,490
2011	15,042	6,274
2012 (Jan-Aug)	9,996	4,301
TOTAL	73,836	35,635

Table 3 – Comparison of the inpatients stays between medical departments ($p<0.001$ in each line of the table)

Department	Cardiology	Geriatrics	Gyn. Obs.	Int. med.	Pneumo.	Surgery
Age (years)	67.6	82.4	28.0	69.7	67.8	57.6
Length of stay (days)	8.19	11.6	6.56	10.5	11.8	8.42
Men	42.8%	28.8%	0.00%	39.4%	63.2%	37.8%
Renal insufficiency	3.04%	4.83%	0.04%	4.20%	2.04%	0.60%
Hepatic insufficiency	13.7%	2.42%	0.04%	6.26%	2.34%	1.47%
VKA	15.5%	12.9%	0.00%	13.7%	14.8%	1.67%
Diuretics	41.1%	30.1%	0.00%	31.6%	35.4%	12.9%

Irrespectively from being ADEs or not, many abnormal laboratory results are observed during the hospitalizations as defined in the "Material and Methods" section. Their incidence rate is displayed in Table 4.

Table 4 – Proportion of stays with an abnormal laboratory result detected during the hospitalization (being ADEs or not)

Parameter	Estimated value
INR increase	2.46% [2.30% ; 2.62%]
Hyperkalemia	5.43% [5.19% ; 5.67%]

Estimated number of ADEs

This section presents the number of potential ADE cases detected by the ADE Scorecards (without expert validation).

Table 5 displays the number and proportion of potential ADE cases with INR increase year after year (see also Figure 4). Those proportions are detailed by medical department in Table 6.

Table 5 – Potential ADE cases with INR increase (2012: from January to August)*

Year	Number	Proportion
2007	67	1.10% [0.84%;1.36%]
2008	60	0.96% [0.72%;1.20%]
2009	71	1.14% [0.88%;1.41%]
2010	49	0.76% [0.54%;0.97%]
2011	61	0.97% [0.73%;1.22%]
2012*	45	1.05% [0.74%;1.35%]
TOTAL	*353*	*0.99% [0.89%;1.09%]*

Table 6 – Potential ADE cases with INR increase by medical department (comparison: p<0.001)

Department	Proportion
Cardiology	1.36% [1.07%;1.65%]
Geriatrics	0.00% [0.00%;0.00%]
Gynecology Obstetrics	0.00% [0.00%;0.00%]
Internal medicine	1.63% [1.27%;1.99%]
Pneumology	2.12% [1.67%;2.56%]
Surgery	0.14% [0.05%;0.23%]

Table 7 displays the number and proportion of potential ADE cases with hyperkalemia year after year (see also Figure 4). Those proportions are detailed by medical department in Table 8.

Table 7 – Potential ADE cases with Hyperkalemia (2012: from January to August)*

Year	Number	Proportion
2007	145	2.38% [2.00%;2.77%]
2008	146	2.33% [1.95%;2.70%]
2009	125	2.01% [1.66%;2.36%]
2010	108	1.66% [1.35%;1.98%]
2011	120	1.91% [1.57%;2.25%]
2012*	81	1.88% [1.48%;2.29%]
TOTAL	*725*	*2.03% [1.89%;2.18%]*

Table 8 – Potential ADE cases with Hyperkalemia by medical department (comparison: p<0.001)

Department	Proportion
Cardiology	2.65% [2.25%;3.05%]
Geriatrics	3.36% [2.22%;4.51%]
Gynecology Obstetrics	0.00% [0.00%;0.00%]
Internal medicine	2.68% [2.22%;3.14%]
Pneumology	2.77% [2.26%;3.27%]
Surgery	0.88% [0.66%;1.10%]

Figure 4 – incidence rates of potential ADEs ("+":INR increase; "o": Hyperkalemia; "---": 95% confidence intervals)

Qualitative preliminary evaluation

The ADE Scorecards are routinely used by two ways. Some of the physicians or head nurses use to connect monthly on the application to view the statistics and review some cases. But the most important use is made by the pharmacists of the hospital. Every month in every department, the pharmacists use to plan a meeting with all the physicians. Since the ADE Scorecards are installed, some interesting cases detected by the application are used to support the discussion. The pharmacists report that, formerly, their recommendations could sometimes be perceived as "too theoretical". By means of the ADE Scorecards, they can now support their recommendations with visual displaying of real ADE cases from the department they meet. Moreover, the ADE Scorecards are able to detect complex pharmacokinetic drug interactions that are rarely known and that were not discussed with the physicians formerly. According to the users of the Scorecards, about half the detected cases are real ADE cases with a cause-to-effect relationship between the potential causes highlighted by the application and the outcome. According to them, this ratio is sufficient to use it as a support tool for morbidity and mortality reviews, after an expert filtering.

Discussion

Initially designed in a research project, the ADE Scorecards have demonstrated after three years of daily use that they could also support real-life healthcare. The application enables to detect past ADE cases and highlights the causal conditions that are linked with some outcomes. It also enables to compute longitudinal statistics about ADEs.

In this study we estimate a 2% incidence rate of ADEs with hyperkalemia and a 1% incidence rate of ADEs with INR increase. Those figures, provided for two specific outcomes, are consistent with the literature. According to the literature, ADEs occur in 2.4 to 5.2 per 100 hospitalized adult patients [17-21]. In [22], 2.8 ADEs occur for 100 patients*days, this could correspond to 5-10% of the stays.

However, the incidence rates that are displayed in this study are related to *potential* ADEs and not *confirmed* ADEs. The ADE cases should be validated by means of an expert review. The qualitative evaluation suggests that the accuracy of the detection should be around 50%. A quantitative expert-operated review performed on a limited sample showed previously a precision (positive predictive value) of 52% in the field of hyperkalemia [9]. A more complete quantitative evaluation is still in progress: it consists in a case review (detected and undetected cases) performed by pharmacology experts.

The curves on Figure 4 suggest a small decrease of the ADE incidence rates of Hyperkalemia along the observation period. As the patients may differ from a year to another, a simple statistical comparison of proportions would not be sufficient. A more complete study is also in progress, in order to adjust incidence rates with the patients' medical background, by means of propensity scores.

The application was initially designed for a wide use by the physicians as a standalone tool. The qualitative evaluation suggests that the ADE Scorecards are more likely to be used as a support to a more global quality improvement approach, leaded by pharmacists. For instance, the ADE Scorecards enable to quickly find ADE cases, and those cases -as well as their user-friendly scrollable representations- are easy to validate or not, and to use in morbidity and mortality reviews. The quali-

tative evaluation that is still in progress principally aims at highlighting usability lacks and suggesting improvements of the tool.

Conclusion

The ADE Scorecards have demonstrated they could be used in real-life healthcare, especially as a support to more traditional quality improvement approaches. Complementary studies must be led to compute the precision of the ADE detection, and to assess whether the use of the application is associated with a change in the ADE incidence rates.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under Grant Agreement n°216130 - the PSIP project.

References

[1] Institute Of Medicine. Preventing Medication Errors. Washington, DC: The National Academic Press; 2007.

[2] Nebeker JR. Clarifying Adverse Drug Events: A Clinician's Guide to Terminology, Documentation, and Reporting. Ann Intern Med. 2004;140:795-801.

[3] 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(2):115-28.

[4] 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(4):306-14.

[5] Cantor MN, Feldman HJ, Triola MM. Using trigger phrases to detect adverse drug reactions in ambulatory care notes. Qual Saf Health Care. 2007 Apr;16(2):132-4.

[6] Gysbers M, Reichley R, Kilbridge PM, Noirot L, Nagarajan R, Dunagan WC, et al. Natural language processing to identify adverse drug events. AMIA Annu Symp Proc. 2008:961.

[7] Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):448-57.

[8] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, Ohe K. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform. 2010;160(Pt 1):739-43.

[9] Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data Mining to generate Adverse Drug Events detection rules. IEEE Trans Inf Technol Biomed. 2011 Nov;15(6):823-30. Epub 2011 Aug 22.

[10]Chazard E, Baceanu A, Ferret L, Ficheur G. The ADE Scorecards: a tool for Adverse Drug Event detection in Electronic Health Records. Stud Health Technol Inform. 2011.

[11]Chazard E, Merlin B, Ficheur G, Sarfati JC, PSIP Consortium, Beuscart R. Detection of adverse drug events: proposal of a data model. Stud Health Technol Inform. 2009;148:63-74.

[12]Chazard E, Mouret-Kubiak C, Ficheur G, Beuscart R. Déidentification automatisée de courriers médicaux : la méthode FASDIM. Revue d'Épidémiologie et de Santé Publique, Volume 60, Supplement 1, March 2012, Page S18.

[13]Handler SM, Altman RL, Perera S, Hanlon JT, Studenski SA, Bost JE, et al. A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):451-8.

[14]Chazard E, Bernonville S, Ficheur G, Beuscart R. A Statistics-based Approach of Contextualization for Adverse Drug Events Detection and Prevention. Stud Health Technol Inform. 2012;180:766-70.

[15]ISO 13407:1999, Human-centered design processes for interactive systems.

[16]Free demonstration of the ADE Scorecards. [cited 2012 December 10]. Available from: http://psip.univ-lille2.fr/prototypes/public/

[17]Bates DW, O'Neil AC, Boyle D, Teich J, Chertow GM, Komaroff AL, Brennan TA. Potential identifiability and preventability of adverse events using information systems. J Am Med Inform Assoc, 1994. 1(5): p. 404-11.

[18]Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, Laffel G, Sweitzer BJ, Shea BF, Hallisey R, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. JAMA. 1995;274:29-34.

[19]Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. JAMA 1997.

[20]Nebeker JR, Hoffman JM, Weir CR, Bennett CL, Hurdle JF. High rates of adverse drug events in a highly computerized hospital. Arch Int Med 2005.

[21]Senst BL, Achusim LE, Genest RP, Cosentino LA, Ford CC, Little JA, Raybon SJ, Bates DW. Practical approach to determining costs and frequency of adverse drug events in a health care network. Am J Health-Sys Pharm 2001.

[22]Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, Burdick E, Seger DL, Vander Vliet M, Bates DW. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inform Assoc, 1998. 5(3): p. 305-14.

Address for correspondence

Corresponding Author: Emmanuel Chazard; Clinique de Santé Publique, CHRU de Lille, 150 rue Yersin, 59037 Lille Cedex, France; emmanuel.chazard@univ-lille2.fr

Preventive Medicine 93 (2016) 121–127

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

The risks of pulmonary embolism and upper gastrointestinal bleeding beyond 35 days after total hip replacement for coxarthrosis among middle-aged patients: A cross-over cohort

Grégoire Ficheur, MD, PhD [a,*], Alexandre Caron, MD [a], Jean-Baptiste Beuscart, MD, PhD [b], Laurie Ferret, PharmD, PhD [c], Sophie Putman, MD [d], Régis Beuscart, MD, PhD [a], Emmanuel Chazard, MD, PhD [a]

[a] Univ. Lille EA 2694, CHU Lille, Department of Public Health, F-59000 Lille, France
[b] Univ. Lille EA 2694, CHU Lille, Department of Geriatric Medicine, F-59000 Lille, France
[c] CHU Lille, Department of Pharmacy, F-59000 Lille, France
[d] CHU Lille, Department of Orthopedic Surgery, F-59000 Lille, France

ARTICLE INFO

Article history:
Received 5 May 2016
Received in revised form 3 August 2016
Accepted 5 September 2016
Available online 6 September 2016

Keywords:
Patient safety
Venous thromboembolic event
Bleeding event
Total hip arthroplasty
Total hip replacement

ABSTRACT

Prophylactic anticoagulation is recommended up to 35 days after total hip replacement (THR). Although several observational studies have assessed the incidence of thrombotic events or bleeding events after THR, the corresponding measures of association have never been studied concomitantly. Here, we evaluated the duration of the elevated risks (relative to the baseline risk) of both venous thromboembolic events and bleeding events after THR for coxarthrosis among middle-aged patients.

This was a population-based, cross-over cohort study of data extracted from the French national inpatient database between 2007 and 2013. We included middle-aged patients (aged 45 to 69) having undergone THR for coxarthrosis. We compared the numbers of pulmonary embolisms (PEs) (respectively upper gastrointestinal bleedings (UGIBs)) following the THR with the numbers occurring during three unexposed periods one year later. This enabled us to estimate the odds ratio (OR) [95% confidence interval (CI)] for each of six successive 35-day intervals.

The study included 108,099 patients. The ORs for PE were respectively 12.4 (95% CI, 8.6–17.8) (absolute risk difference rate per 100,000 (ARD/100,000) = 130) and 5.0 (95% CI, 3.4–7.4) (ARD/100,000 = 52) for the first two 35-day intervals, and the risk was close to 1 thereafter. The risk of UGIB fell quickly, with an OR of 6.5 (95% CI, 4.6–9.1) (ARD/100,000 = 83) and 0.8 (95% CI, 0.4–1.6) for the first two 35-day intervals, respectively. The majority of UGIBs occurred during the inpatient stay for THR.

Among middle-aged patients, the risk of a PE remains elevated beyond 35 days after THR for coxarthrosis, whereas the risk of a UGIB remains elevated for the first 35 days only.

1. Introduction

The American College of Chest Physicians' guidelines (Falck-Ytter et al., 2012; Guyatt et al., 2012) recommend the administration of antithrombotic agents for at least 10 to 14 days after major orthopedic surgery (Grade 1B). The continuation of anticoagulation therapy on an outpatient basis is recommended for the 35 days following surgery (Grade 2B). These guidelines are based on randomized clinical trials (RCTs), the results of which must be complemented by population-based studies (Rothwell, 2005) for three main reasons: (i) population-based studies evaluate the effectiveness and safety on real data; (ii) the rarity of the event means that a population-based study (with greater statistical power than an RCT) is required to estimate the corresponding risk - especially since RCTs in similar populations report different major bleeding rates (Dahl et al., 2010); and (iii) RCTs have limited follow-up periods and cannot calculate the risk relative to the baseline risk (outside the context of surgery and prophylactic anticoagulation).

To the best of our knowledge, three population-based studies (Lalmohamed et al., 2013a; Pedersen et al., 2012; Sweetland et al., 2009) have assessed the risk (relative to baseline risk) of a venous thromboembolic event (VTE) for >35 days after total hip replacement (THR), and only one study has assessed the risk (relative to baseline risk) of a bleeding event (Lalmohamed et al., 2013b). However, none of the studies cited above simultaneously assessed the corresponding measures of association for both types of events. In contrast, several studies have assessed the incidence of both thrombotic events and bleeding events (Guijarro et al., 2011; Lanes et al., 2011; Pedersen et al., 2014).

* Corresponding author at: Department of Public Health, Lille University Hospital, 2 avenue Oscar Lambret, F-59037 Lille cedex, France.
E-mail address: gregoire.ficheur@univ-lille2.fr (G. Ficheur).

http://dx.doi.org/10.1016/j.ypmed.2016.09.010
0091-7435/

Historically, THR was only performed on elderly, relatively inactive patients. Progressively, this procedure has been extended to middle-aged patients, who now account for over half of all THRs. The increasing proportion of middle-aged patients is also due to aging of the "baby-boomer generation", and is forecast to continue until 2030 (Kurtz et al., 2009). Even though THR in these patients raises specific issues (including the fact that the prosthesis will be present for a longer period and that they are at higher risk for revision (Malchau et al., 2002)), there are currently very few literature data on this population.

In terms of methodological aspects, the Observational Medical Outcomes Partnership (Ryan et al., 2013; Simpson et al., 2013) (OMOP, an empirical, strict, systematic evaluation of study designs) has shown that cross-over designs (Maclure, 1991; Maclure and Mittleman, 2000) (such as cross-over cohorts and case-crossover studies) are superior in the field of pharmacoepidemiology. These designs were not used in the above-cited population-based studies.

1.1. Objective

The primary objective of the present study was to simultaneously assess the duration of the elevated risks (relative to the baseline risk, outside the context of surgery) of a VTE and of a bleeding event following THR for coxarthrosis in a population of middle-aged patients.

2. Material and methods

2.1. Data sources

For the period from 2007 to 2013, the "acute care" section of the French national inpatient database contained information on 171,556,421 inpatient stays. Collection of these data has been approved by the French National Data Protection Commission (CNIL authorization number 1,754,053). The database contains a summary of each inpatient stay in France, including the ICD-10 diagnostic code ("WHO | International Classification of Diseases (ICD)"), the medical procedures performed (coded according to the French CCAM classification) and the patient's age, gender, and unique identifier. With regard to the quality of the data, 153 tests (Agence Technique de l'Information sur l'hospitalisation, 2015) are performed routinely when the information on the inpatient stay is sent to the French public health insurance agency. These include checks on the chronology of the inpatient stays, the format (missing, incorrect or imprecise values) of the demographic characteristics (gender, age, date and mode of entry, date and discharge mode), the format of procedure codes and diagnostic codes, and the agreement between procedure codes, diagnostic codes, the length of stay, age, and gender.

2.2. Definitions

For the avoidance of doubt concerning the terms used in the Methods and Results section, it should be noted that the term "venous thromboembolism (VTE)" includes deep vein thrombosis (DVT) and pulmonary embolism (PE). The term "bleeding event" encompasses upper gastrointestinal bleeding (UGIB) and intracranial bleeding.

2.3. Study design

We carried out a population-based, crossover cohort study by analyzing the French national inpatient database from 2007 to 2013. Each patient served as his/her own control, which enabled us to control for certain personal, time-constant confounding factors. The patient is analyzed when he/she is "exposed" to the THR and is used as his/her own control 12 months later (when he/she is no longer exposed to the THR).

For the study population as a whole (Fig. 1), we compared the likelihood of a PE after THR with the likelihood of a PE during three unexposed periods around 12 months after THR. We then defined six successive, 35-day, high-risk intervals. The use of 35-day intervals allows to precisely address the risks beyond the recommended 35-day period of anticoagulation.

2.4. Patients

2.4.1. Inclusion criteria (definition of exposure)

We included middle-aged patients (aged 45 to 69) having undergone THR between July 1st, 2007, and March 31st, 2012. Inpatient stays with THR were identified by the procedure code NEKA020 (corresponding to "Hip joint replacement with total hip replacement") and one of the following ICD-10 diagnostic codes: M16.0, M16.1 or M16.9 ("osteoarthritis of hip").

2.4.2. Non-inclusion criteria

We did not include patients meeting one or more of the following criteria: those with THR and an S72.x diagnostic code ("fracture of femur") or a T84.x diagnostic code ("complications of internal orthopedic prosthetic devices, implants and grafts"), and those with a history (as recorded during previous hospitalizations) of thromboembolism: VTE (the I80.x diagnostic code "phlebitis and thrombophlebitis" or the I26.x diagnostic code "pulmonary embolism"), myocardial infarction (the I21.x code "ST elevation and non-ST elevation myocardial infarction" or the I22.x code "Subsequent ST elevation and non-ST elevation myocardial infarction", available from So (So et al., 2006)), ischemic stroke (the I63.x code "cerebral infarction"), "nonpyogenic thrombosis of intracranial venous system" (code I67.6) or "central retinal artery occlusion" (code H34.1)).

2.4.3. Exclusion criteria

We excluded patients having undergone another THR during the 21 months following the PE, since we wanted to assess the risk relative to an unexposed period.

2.5. Measurements

2.5.1. Outcome definition

We determined whether the included patients had suffered a PE in the 210 days following THR. If more than one PE was detected during a given exposed period or unexposed period, the first event in each period was selected.

Several algorithms for tracking VTEs within claims data have been developed and evaluated (Tamariz et al., 2012). Many of these refer to ICD-9. An evaluation of ICD-10 (using a database similar to our own) revealed that it was more difficult to identify inpatient stays with DVT than stays with PE (Casez et al., 2010). Furthermore, DVT does not necessarily require hospitalization.

In summary, we decided to use PE as a marker of the risk of a VTE in the present cross-over study. The ICD-10 codes used to identify PE are I26.0 ("pulmonary embolism with acute cor pulmonale") and I26.9 ("pulmonary embolism without acute cor pulmonale").

2.6. Statistical analysis

Firstly, we performed a descriptive analysis of the demographic characteristics of all inpatient stays meeting the inclusion criteria. Categorical data were expressed as the number (frequency). Quantitative data were expressed as the mean ± standard deviation (SD) or the median (interquartile range (IR)).

Secondly, we calculated the odds ratio (OR) [95% confidence interval (CI)] for each 35-day interval. We assessed the likelihood of a PE occurring from 0 to 34 days after THR, relative to the likelihood of a PE occurring during three 35-day unexposed intervals (respectively 330, 365 and 400 days later). A similar analysis was performed for the five other 35-day-long intervals. We used conditional logistic regression to calculate the OR [95% CI] for each interval. Then, we calculated the

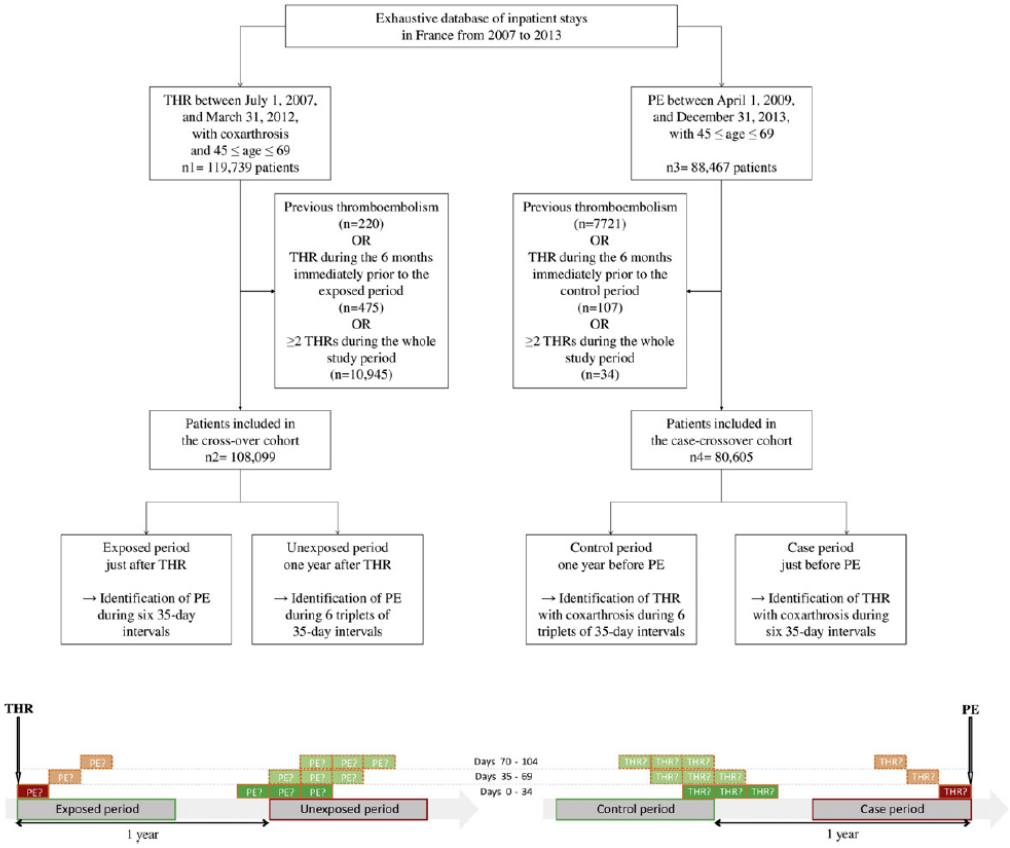


Fig. 1. A flow diagram of patient inclusion and a schematic representation of the study designs (left panel: cross-over cohort; right panel: case-crossover).

absolute risk differences [95% CI] between the exposed period and the unexposed periods for each interval.

We calculated the number of PEs for each 35-day interval in the exposed period and in the unexposed periods. Furthermore, we computed the PE rate for 100,000 THRs, using the number of THRs included between July 1st, 2007, and March 31st, 2012, as the denominator.

All statistical analyses were performed using R statistical software (version 3.1.2) (R Core Team, 2013).

2.7. Sensitivity analysis

Our sensitivity analysis was based on a case-crossover design, i.e. the presence of exposure to THR among patients with a PE. We included all patients with a first PE between April 1st, 2009, and December 31st, 2013. Three control periods were chosen: 400, 365 and 330 days before the case period. This sensitivity analysis was valuable because the French national inpatient database does not contain data on deaths occurring outside hospital. The use of a case-crossover design (with a control period in the past) ensured that the patients were alive throughout the study period.

Furthermore, we used post-hoc analyses to confirm the results of the cohort cross-over analysis. Firstly, we changed the definition of the outcome: the PEs (respectively UGIBs) were considered to be present if the CCAM procedure code (see eMethod1 and eMethod2 in the appendices) and a consistent ICD-10 diagnostic code were found. Secondly, we performed a post-hoc conservative analysis by computing the odds ratio after the exclusion of (i) patients with a length of stay (for THR) of at least 15 days, and (ii) patients with an intercurrent admission between the time of the inpatient stay for THR and the time of the inpatient stay for PE.

Lastly, we performed a negative control by assessing exposure that was not expected to increase the risk of a PE. To this end, we analyzed the exposure defined by the AHPA009 procedure ("release of the median nerve in the carpal tunnel, using a direct approach") and the diagnostic code G56.0 ("carpal tunnel syndrome") over the six successive 35-day intervals.

2.8. Bleeding risk

Our analysis of the risk of a bleeding event was similar to that performed for PEs: we built a cross-over cohort retrospectively and then performed a case-crossover study as a sensitivity analysis. The inclusion criteria were the same as those for the thromboembolic risk in the cross-over cohort. The main non-inclusion criterion was a history of bleeding. We used the codes suggested in Hippisley-Cox et al.'s study (Hippisley-Cox and Coupland, 2014) in which a bleeding risk score was built for patients initiating anticoagulant treatment. The ICD-10 codes described in the latter publication correspond to severe upper gastrointestinal bleeding and intracranial bleeding.

As with the thromboembolic risk, bleeding events should (i) always result in hospitalization and (ii) not be serious enough to irreversibly change the patient's life. This is why we decided to track UGIBs but not intracranial bleeding (which is always major bleeding (Schulman et al., 2005)). We therefore chose to use the ICD-10 codes for UGIB described in Hippisley-Cox et al.'s (Supplemental eMethod3). The diagnoses were defined as "any upper gastrointestinal ulcer with perforation, bleed, or both; melena; haematemesis; laceration with bleed; varices

with bleed; haemorrhagic gastritis; and other unspecified gastrointestinal bleeds". These bleeding events do not include diagnoses that may be directly related to surgery itself.

3. Results

3.1. Description of the population

For the main cross-over cohort analysis, we included 108,099 patients with an inpatient stay for THR from July 1st, 2007, to March 31st, 2012. A descriptive analysis showed that patients' mean ± SD age was 60.9 ± 5.9 and the proportion of females was 47.1%. The mean ± SD and median (IR) length of stay were respectively 8.5 ± 2.9 and 8.0 (7.0 to 9.0).

For the case-crossover study, 80,605 patients with PE were included from April 1st, 2009, to December 31st, 2013. The flow diagram in Fig. 1 provides information on the patients included in the cross-over cohort and the case-crossover study.

3.2. The risk of a PE, relative to baseline risk

Table 1 summarizes the results for the risk of a PE for each 35-day interval after THR (left panel: cross-over cohort; right panel: case-crossover study), together with the number of events per 100,000 THRs. The corresponding ORs [95% CI] are graphically presented on Fig. 2. In the cross-over cohort, significantly more PEs occurred 0–34 days after THR (154 events, or 142 events per 100,000 THRs) than during unexposed intervals one year later (12 events, or 11 events per 100,000 THRs); this corresponds to an OR of 12.4 (95% CI, 8.6 to 17.8). For days 35–69 after THR, there was still a significant increase in the number of PEs (71 events, or 65 events per 100,000 THRs), relative to the control intervals one year later (14 events, or 12 events per 100,000 THRs). This corresponds to an OR of 5.0 (95% CI, 3.4 to 7.4). After this time point, the odds ratio was close to 1 and was no longer significantly elevated.

The results obtained in the case-crossover study were similar to those obtained in the cross-over cohort. The OR calculated for the first 35-day interval was very similar, and the ORs computed for the two following 35-day intervals were slightly greater.

These results were confirmed by changing the definition of the event, albeit with slightly greater ORs for the first two 35-day intervals (see eTable 2 in the appendices). The conservative post-hoc analysis also confirmed the above-mentioned results: for days 35–69 and days 70–104, the ORs were respectively 4.0 (95% CI, 2.7 to 6.1), and 1.5 (95% CI, 0.8 to 2.5). The results for the negative control were as expected (see eTable 1 in the appendices): the ORs computed for the six 35-day intervals were never significantly different from 1.

3.3. The risk of an UGIB, relative to baseline risk

108,268 patients with THR were included in the cross-over cohort, and 110,731 patients with UGIB were included in the case-crossover. Table 2 shows the results for the UGIB risk for each 35-day interval after THR (left panel: cross-over cohort; right panel: case-crossover), together with the number of events per 100,000 THRs, the corresponding ORs [95% CI] are presented on Fig. 2.

In both analyses, the odds ratio fell between the first and second 35-day intervals (from 6.5 (95% CI, 4.6 to 9.1) to 0.8 (95% CI, 0.4 to 1.6) in the cross-over cohort and from 8.1 (95% CI, 5.6 to 11.7) to 1.0 (95% CI, 0.5 to 1.9) in the case-crossover study). It is noteworthy that the majority of UGIBs occurred during the inpatient stay with THR. In both analyses, the risk was no longer significantly elevated after the first interval. The results obtained by changing the definition of the event were very similar (see eTable 3 in the appendices).

Lastly, only two patients experienced both PE and UGIB after THR.

4. Discussion

The risk (relative to the baseline risk, outside the context of surgery) of a PE after THR for coxarthrosis in French middle-aged patients was estimated for six successive 35-day intervals. The odds ratio was found to be significantly elevated for the first two intervals, i.e. for 70 days. The risk of a UGIB fell quickly, from an OR of 6.5 (95% CI, 4.6 to 9.1) for days 0 to 34 to 0.8 (95% CI, 0.4 to 1.6) for days 35 to 69.

These values can be compared with those reported by the population-based studies cited in the Introduction. The risks (relative to baseline risk) of a VTE reported by Lalmohamed et al. (Lalmohamed et al., 2013a), were similar to our values: for the first three successive intervals ("<2 weeks, 2–6 weeks, 6–12 weeks"), the hazard ratios (HRs) were respectively 15.4 (95% CI, 11.6 to 20.4), 10.6 (95% CI, 8.21 to 13.6), and 2.84 (95% CI, 2.02 to 4.00). Lalmohamed et al. found that the HR was close to 1 beyond 12 weeks, which is consistent with our results. For the first 90 days, the relative risk (RR) of a VTE reported by Pedersen et al. (Pedersen et al., 2012) (RR of 10.65 (95% CI, 7.68 to 14.78)) also agrees with our results. However, in contrast to the present study, Pedersen et al. found that the risk remained elevated beyond this period (from 91 to 365 days) albeit with a value close to 1 (RR of 2.01 (95% CI, 1.46 to 2.75)). Lastly, our values are substantially lower than those reported by Sweetland et al. (Sweetland et al., 2009). We also compared the risk (relative to baseline risk) of an UGIB with the only other study, to our knowledge, having estimated the corresponding measure of association after THR (Lalmohamed et al., 2013b): our results are similar for the first 35 days but differ for the following 35 days. Lalmohamed et al. found that the risk of UGIB remained elevated for 6–12 weeks (with a HR of 2.39 (95% CI, 1.79 to 3.20)), which was not the case in our study.

The analysis of medical claims databases always raises the issue of data quality; it is well known that hospitals with higher VTE screening

Table 1
The risk of a PE, as a function of the time interval (in days) after THR (French inpatient database, 2007–2013).

	Cross-over cohort			Case-crossover	
	Exposed period	Unexposed periods	Absolute risk difference	Case period	Control periods
Interval in days	Number of events[a,b] (rate of events per 100,000 THRs)		Rate of events per 100,000 THRs[a] [95% CI]	Number of events[a,b] (rate of events per 100,000 THRs)	
0–34	154[c] (142)	12 (11)	130 [107 to 153]	138 (127)	11 (10)
35–69	71 (65)	14 (12)	52 [36 to 68]	71 (65)	11 (10)
70–104	25 (23)	14 (12)	9 [0 to 19]	42 (38)	13 (12)
105–139	24 (22)	15 (13)	8 [−1 to 17]	18 (16)	13 (12)
140–174	21 (19)	16 (14)	4 [−4 to 13]	30 (27)	16 (14)
175–209	21 (19)	17 (15)	3 [−5 to 12]	22 (20)	15 (14)

[a] Truncated to an integer.
[b] The mean for 3 unexposed (or control) periods.
[c] Including 76 events (70 per 100,000) during the inpatient stay with THR.

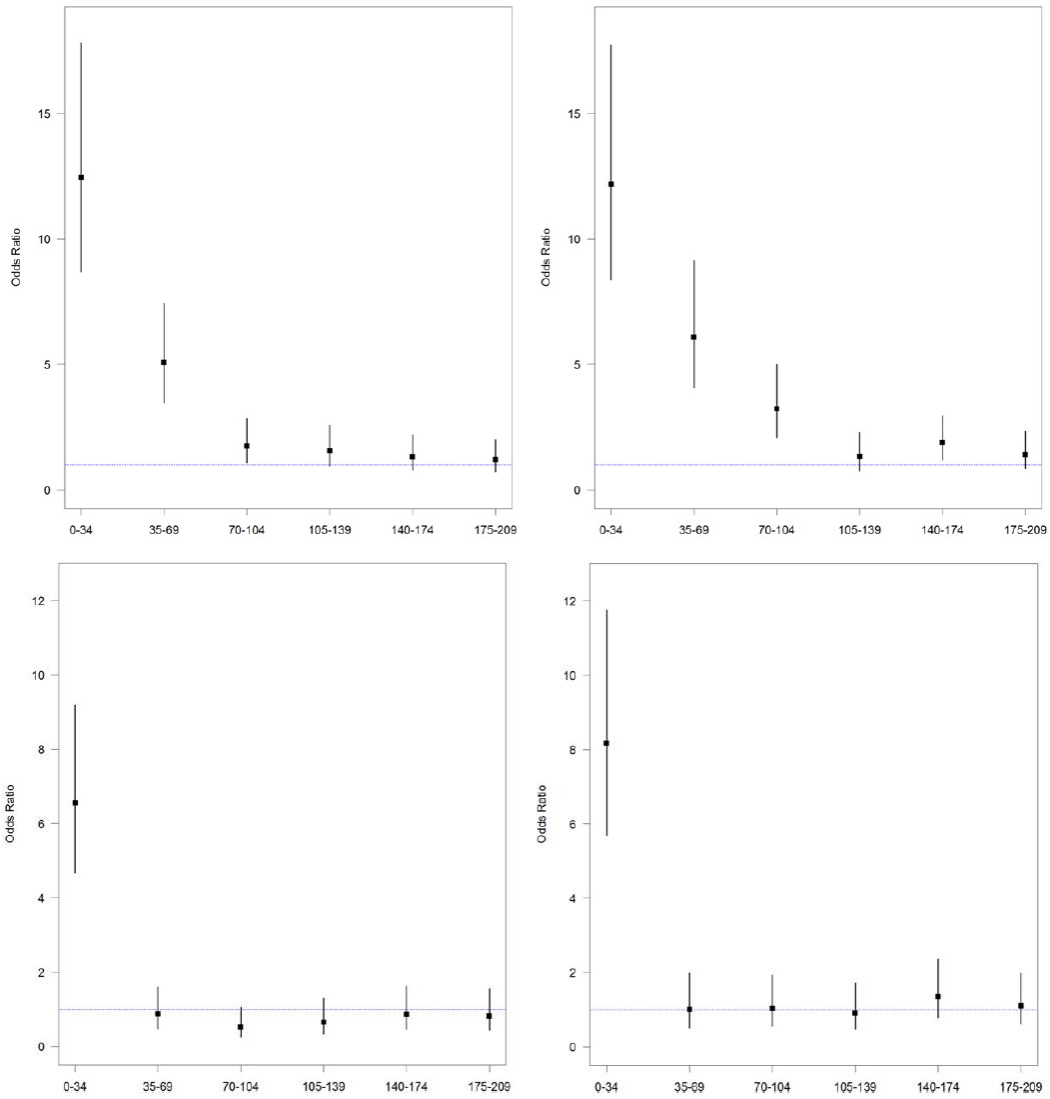


Fig. 2. The risks of a pulmonary embolism or an upper gastrointestinal bleeding event as a function of the time interval (in days) after THR (upper panel: the risk of a pulmonary embolism; bottom panel: the risk of a upper gastrointestinal bleeding event; left panel: cross-over cohort; right panel: case-crossover study) (French inpatient database, 2007–2013). The error bars indicate the OR's 95% CI and the horizontal blue line corresponds to an OR of 1.

Table 2
The risk of an UGIB event, as a function of the time interval (in days) after THR (French inpatient database, 2007–2013).

	Cross-over cohort			Case-crossover	
	Exposed period	Unexposed periods		Case period	Control periods
Interval in days	Number of events[a,b] (rate of events per 100,000 THRs)		Absolute risk difference rate of events per 100,000 THRs[a] [95% CI]	Number of exposures[a,b] (rate of events per 100,000 THRs)	
0–34	107[c] (98)	16 (15)	83 [64 to 102]	109 (100)	13 (12)
35–69	13 (12)	15 (13)	−1 [−9 to 5]	11 (10)	11 (10)
70–104	9 (8)	17 (16)	−7 [−14 to 0]	13 (12)	12 (11)
105–139	10 (9)	15 (14)	−4 [−11 to 2]	12 (11)	13 (12)
140–174	12 (11)	14 (12)	−1 [−9 to 5]	18 (16)	13 (12)
175–209	12 (11)	14 (13)	−2 [−9 to 4]	15 (13)	13 (12)

[a] Truncated to an integer.
[b] Mean for 3 unexposed (or control) periods.
[c] Including 58 events (53 per 100,000) during the inpatient stay with THR.

rates also have higher VTE rates (Bilimoria et al., 2013). In the case of PE, we were able to use diagnostic codes with good recall and high precision (Casez et al., 2010). This metrological quality was confirmed by our sensitivity analysis, in which the definition of an event required a diagnostic code and a compatible procedure code.

The list of codes selected for upper gastrointestinal bleeding came from the literature (Hippisley-Cox and Coupland, 2014) but were not evaluated against medical records which is a major study limitation. Also, we did not perform subgroup analyses as a function of anticoagulant prescription because we did not have this information. Even though this information was not required for addressing the study's primary objective, it would have been very interesting to have performed subgroup analyses by stratifying by anticoagulation prophylaxis.

Furthermore, the analysis of events that only account for a proportion of the total events of interest constitutes another study limitation, and raises the question of whether our findings can be generalized to the entire set of events concerned. In the case of VTEs, there are discrepancies in the literature as to whether DVT and PE are significantly associated (Parvizi et al., 2014, 2010). In terms of bleeding events, the UGIBs after THR have been studied several times alone (Hallas et al., 2006; Lalmohamed et al., 2013b), but, to the best of our knowledge, the possible association between UGIB and intracranial bleeding has not been studied.

Lastly, the use of such a cross-over cohort indirectly implies that the patient did not die during the study period. There were only 3 cases of fatal PE and 2 cases of fatal UGIB in our study; these occurred only during the inpatient stay with THR and never during the readmissions. This number is too small to significantly affect the value of the OR for the first 35-day interval.

At present, the guidelines issued by the French Society of Anesthesiology and Intensive Care (*Société Française d'Anesthésie et de Réanimation*) recommend the administration of low-molecular-weight heparin (LMWH) for 42 days, i.e. 7 days longer than the US recommendation (Guyatt et al., 2012). Furthermore, novel oral anticoagulants were rarely prescribed in France during the entire inclusion period for the crossover cohort (Drouet, 2014).

This risk of a PE remained elevated for 70 days and decreased between the first and second intervals. This seems to be consistent with data on laboratory markers of coagulation after THR; the prothrombotic state remains active for at least 5 weeks after major orthopedic surgery (Dahl et al., 1995; Dindo et al., 2009; Wilson et al., 2001) and probably recedes gradually.

It is likely that the measures of association computed for the first 35-day interval are modified by the use of physical and anticoagulation prophylaxies in our population. Indeed, even if the recommended prophylaxis have not proven its effectiveness regarding the rate of fatal PE and the overall mortality (Falck-Ytter et al., 2012; Freedman et al., 2000; Tasker et al., 2010), it has proven its effectiveness regarding the rate of non-fatal VTE (Eikelboom et al., 2001; Falck-Ytter et al., 2012; Freedman et al., 2000; Sobieraj et al., 2012). Contrariwise, it seems reasonable to think that our estimation beyond 35 days is less modified by the recommended prophylaxis, and that an extension of the prophylaxis beyond 35 days (for some patients in our analyzed sample) would be conservative regarding the results obtained for the interval 35–69 days, as it would decrease the risk of a PE and would increase the risk of a UGIB. Finally, our results show that the OR for the whole second interval (days 35–69) was 5.0 (95% CI, 3.4 to 7.4) but the value was probably higher than this at the start of the interval, i.e. at day 35. For these reasons, we think that our result could justify the assessment of the benefit–risk scale of a prophylaxis extended beyond 35 days.

5. Conclusion

Among middle-aged patients having undergone THR for coxarthrosis, the risk of a PE remains elevated >35 days after surgery, whereas the risk of a UGIB is elevated for the first 35 days only.

Authors' contributions

Grégoire Ficheur, Sophie Putman and Emmanuel Chazard contributed to the conception and design of the study and planned the statistical analysis; Grégoire Ficheur collected and abstracted the data; Grégoire Ficheur and Alexandre Caron performed the statistical analysis; Grégoire Ficheur, Laurie Ferret and Emmanuel Chazard drafted the manuscript. Grégoire Ficheur, Régis Beuscart and Jean-Baptiste Beuscart performed the literature search. All authors contributed to the interpretation of findings, revising the manuscript for important intellectual content, and approved the final version to be published. All authors had full access to all the data, including statistical reports and tables.

Conflict of interest statement

None.

Funding source

The authors are employed by the University of Lille, or the Lille University Hospital. These funding organizations did not suggest the subject of this study, and did not have access to the results before publication. This study received no external funding.

Ethical approval

In France, retrospective, registry-based studies do not require approval by an investigational review board. The data were structured so that individual patients could not be identified. Collection of these data was approved by the French National Data Protection Commission (authorization: CNIL 1754053).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ypmed.2016.09.010.

References

Agence Technique de l'Information sur l'hospitalisation, 2015. Manuel des Groupes Homogènes de Malades.

Bilimoria, K.Y., Chung, J., Ju, M.H., et al., 2013. Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure. JAMA 310, 1482–1489. http://dx.doi.org/10.1001/jama.2013.280048.

Casez, P., Labarère, J., Sevestre, M.-A., Haddouche, M., Courtois, X., Mercier, S., Lewandowski, E., Fauconnier, J., François, P., Bosson, J.-L., 2010. ICD-10 hospital discharge diagnosis codes were sensitive for identifying pulmonary embolism but not deep vein thrombosis. J. Clin. Epidemiol. 63, 790–797. http://dx.doi.org/10.1016/j.jclinepi.2009.09.002.

Core Team, R., 2013. R: A Language and Environment for Statistical Computing. R Foundationfor Statistical Computing, Vienna, Austria.

Dahl, O.E., Aspelin, T., Arnesen, H., Seljeflot, I., Kierulf, P., Ruyter, R., Lyberg, T., 1995. Increased activation of coagulation and formation of late deep venous thrombosis following discontinuation of thromboprophylaxis after hip replacement surgery. Thromb. Res. 80, 299–306. http://dx.doi.org/10.1016/0049-3848(95)00180-Y.

Dahl, O.E., Quinlan, D.J., Bergqvist, D., Eikelboom, J.W., 2010. A critical appraisal of bleeding events reported in venous thromboembolism prevention trials of patients undergoing hip and knee arthroplasty. J. Thromb. Haemost. 8, 1966–1975. http://dx.doi.org/10.1111/j.1538-7836.2010.03965.x.

Dindo, D., Breitenstein, S., Hahnloser, D., Seifert, B., Yakarisik, S., Asmis, L.M., Muller, M.K., Clavien, P.-A., 2009. Kinetics of D-dimer after general surgery. Blood Coagul. Fibrinolysis Int. J. Haemost. Thromb. 20, 347–352. http://dx.doi.org/10.1097/MBC.0b013e32832a5fe6.

Drouet, L., 2014. Rapport de l'ANSM sur les anticoagulants en France en 2014: état des lieux, synthèse et surveillance. Sang Thromb. Vaiss. 26, 225–229.

Eikelboom, J.W., Quinlan, D.J., Douketis, J.D., 2001. Extended-duration prophylaxis against venous thromboembolism after total hip or knee replacement: a meta-analysis of the randomised trials. Lancet 358, 9–15.

Falck-Ytter, Y., Francis, C.W., Johanson, N.A., Curley, C., Dahl, O.E., Schulman, S., Ortel, T.L., Pauker, S.G., Colwell, C.W., 2012. Prevention of VTE in orthopedic surgery patients. Chest 141, e278S–e325S. http://dx.doi.org/10.1378/chest.11-2404.

Freedman, K.B., Brookenthal, K.R., Fitzgerald, R.H., Williams, S., Lonner, J.H., 2000. A meta-analysis of thromboembolic prophylaxis following elective total hip arthroplasty*. J. Bone Jt. Surg. 82, 929.

Guijarro, R., Montes, J., San Roman, C., Ignacio Arcelus, J., Barillari, G., Granero, X., Monreal, M., 2011. Venous thromboembolism and bleeding after total knee and hip

arthroplasty: findings from the Spanish National Discharge Database. Thromb. Haemost. 105, 610–615.
Guyatt, G.H., Akl, E.A., Crowther, M., Gutterman, D.D., Schuünemann, H.J., 2012. Executive summary: antithrombotic therapy and prevention of thrombosis, 9th ed: American college of chest physicians evidence-based clinical practice guidelines. Chest 141, 7S–47S. http://dx.doi.org/10.1378/chest.1412S3.
Hallas, J., Dall, M., Andries, A., Andersen, B.S., Aalykke, C., Hansen, J.M., Andersen, M., Lassen, A.T., 2006. Use of single and combined antithrombotic therapy and risk of serious upper gastrointestinal bleeding: population based case-control study. BMJ 333, 726. http://dx.doi.org/10.1136/bmj.38947.697558.AE.
Hippisley-Cox, J., Coupland, C., 2014. Predicting risk of upper gastrointestinal bleed and intracranial bleed with anticoagulants: cohort study to derive and validate the QBleed scores. BMJ 349, g4606. http://dx.doi.org/10.1136/bmj.g4606.
Kurtz, S.M., Lau, E., Ong, K., Zhao, K., Kelly, M., Bozic, K.J., 2009. Future young patient demand for primary and revision joint replacement: national projections from 2010 to 2030. Clin. Orthop. Relat. Res. 467, 2606–2612. http://dx.doi.org/10.1007/s11999-009-0834-6.
Lalmohamed, A., Vestergaard, P., Jansen, P.a.F., Grove, E.L., de Boer, A., Leufkens, H.G.M., van Staa, T.P., de Vries, F., 2013a. Prolonged outpatient vitamin K antagonist use and risk of venous thromboembolism in patients undergoing total hip or knee replacement. J. Thromb. Haemost. JTH 11, 642–650. http://dx.doi.org/10.1111/jth.12158.
Lalmohamed, A., Vestergaard, P., Javaid, M.K., de Boer, A., Leufkens, H.G.M., van Staa, T.P., de Vries, F., 2013b. Risk of gastrointestinal bleeding in patients undergoing total hip or knee replacement compared with matched controls: a nationwide cohort study. Am. J. Gastroenterol. 108, 1277–1285. http://dx.doi.org/10.1038/ajg.2013.108.
Lanes, S., Fraeman, K., Meyers, A., Ives, J.W., Huang, H.-Y., 2011. Incidence rates for thromboembolic, bleeding and hepatic outcomes in patients undergoing hip or knee replacement surgery. J. Thromb. Haemost. 9, 325–332. http://dx.doi.org/10.1111/j.1538-7836.2010.04155.x.
Maclure, M., 1991. The case-crossover design: a method for studying transient effects on the risk of acute events. Am. J. Epidemiol. 133, 144–153.
Maclure, M., Mittleman, M.A., 2000. Should we use a case-crossover design? Annu. Rev. Public Health 21, 193–221. http://dx.doi.org/10.1146/annurev.publhealth.21.1.193.
Malchau, H., Herberts, P., Eisler, T., Garellick, G., Söderman, P., 2002. The Swedish total hip replacement register. J Bone Jt. Surg. Am. 84, S2–S20.
Parvizi, J., Jacovides, C.L., Bican, O., Purtill, J.J., Sharkey, P.F., Hozack, W.J., Rothman, R.H., 2010. Is deep vein thrombosis a good proxy for pulmonary embolus? J. Arthroplasty, American Association of Hip and Knee Surgeons (AAHKS) Supplement 25, 138–144. http://dx.doi.org/10.1016/j.arth.2010.05.001.
Parvizi, J., Parmar, R., Raphael, I.J., Restrepo, C., Rothman, R.H., 2014. Proximal deep venous thrombosis and pulmonary embolus following total joint arthroplasty. J. Arthroplast. 29, 1846–1848. http://dx.doi.org/10.1016/j.arth.2014.04.023.
Pedersen, A.B., Johnsen, S.P., Sørensen, H.T., 2012. Increased one-year risk of symptomatic venous thromboembolism following total hip replacement: a nationwide cohort study. J. Bone Joint Surg. (Br.) 94-B, 1598–1603. http://dx.doi.org/10.1302/0301-620X.94B12.29358.
Pedersen, A.B., Mehnert, F., Sorensen, H.T., Emmeluth, C., Overgaard, S., Johnsen, S.P., 2014. The risk of venous thromboembolism, myocardial infarction, stroke, major bleeding and death in patients undergoing total hip and knee replacement: a 15-year retrospective cohort study of routine clinical practice. Bone Jt. J. 96-B, 479–485. http://dx.doi.org/10.1302/0301-620X.96B4.33209.
Rothwell, P.M., 2005. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 365, 82–93. http://dx.doi.org/10.1016/S0140-6736(04)17670-8.
Ryan, P.B., Stang, P.E., Overhage, J.M., Suchard, M.A., Hartzema, A.G., DuMouchel, W., Reich, C.G., Schuemie, M.J., Madigan, D., 2013. A comparison of the empirical performance of methods for a risk identification system. Drug Saf. Int. J. Med. Toxicol. Drug Exp. 36 (Suppl. 1), S143–S158. http://dx.doi.org/10.1007/s40264-013-0108-9.
Schulman, S., Kearon, C., the subcommittee on control of anticoagulation of the scientific and standardization committee of the international society on thrombosis and haemostasis, 2005. Definition of major bleeding in clinical investigations of antihemostatic medicinal products in non-surgical patients. J. Thromb. Haemost. 3, 692–694. http://dx.doi.org/10.1111/j.1538-7836.2005.01204.x.
Simpson, S.E., Madigan, D., Zorych, I., Schuemie, M.J., Ryan, P.B., Suchard, M.A., 2013. Multiple self-controlled case series for large-scale longitudinal observational databases. Biometrics 69, 893–902. http://dx.doi.org/10.1111/biom.12078.
So, L., Evans, D., Quan, H., 2006. ICD-10 coding algorithms for defining comorbidities of acute myocardial infarction. BMC Health Serv. Res. 6, 161. http://dx.doi.org/10.1186/1472-6963-6-161.
Sobieraj, D.M., Lee, S., Coleman, C.I., Tongbram, V., Chen, W., Colby, J., Kluger, J., Makanji, S., Ashaye, A.O., White, C.M., 2012. Prolonged versus standard-duration venous thromboprophylaxis in major orthopedic surgery: a systematic review. Ann. Intern. Med. 156, 720–727. http://dx.doi.org/10.7326/0003-4819-156-10-201205150-00423.
Sweetland, S., Green, J., Liu, B., González, A.B.d., Canonico, M., Reeves, G., Beral, V., 2009. Duration and magnitude of the postoperative risk of venous thromboembolism in middle aged women: prospective cohort study. BMJ 339, b4583. http://dx.doi.org/10.1136/bmj.b4583.
Tamariz, L., Harkins, T., Nair, V., 2012. A systematic review of validated methods for identifying venous thromboembolism using administrative and claims data. Pharmacoepidemiol. Drug Saf. 21, 154–162. http://dx.doi.org/10.1002/pds.2341.
Tasker, A., Harbord, R., Bannister, G., 2010. Meta-analysis of low molecular weight heparin versus placebo in patients undergoing total hip replacement and post-operative morbidity and mortality since their introduction. Hip Int. 20, 64–74.
WHO | International Classification of Diseases (ICD) [WWW Document], 2016D. WHO. URL. http://www.who.int/classifications/icd/en/ (accessed 09.12.16).
Wilson, D., Cooke, E.A., McNally, M.A., Wilson, H.K., Yeates, A., Mollan, R.A.B., 2001. Changes in coagulability as measured by thrombelastography following surgery for proximal femoral fracture. Injury 32, 765–770. http://dx.doi.org/10.1016/S0020-1383(01)00139-5.

Références

[1] Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Mag 1996;17:37.

[2] Clifton C. Encyclopedia Britannica: Definition of Data Mining n.d. https://www.britannica.com/technology/data-mining (accessed August 29, 2016).

[3] Piatetsky-Shapiro G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Mag 1990;11:68.

[4] Coan T. Business intelligence: using insight to improve the value and performance of your practice. J Med Pract Manag MPM 2007;23:34–6.

[5] Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform 2017;26. doi:10.15265/IY-2017-007.

[6] Dan Corlan A. Medline trend: automated yearly statistics of PubMed results for any query 2004. http://dan.corlan.net/medline-trend.html (accessed August 29, 2016).

[7] Data Curation - MeSH - NCBI n.d. http://www.ncbi.nlm.nih.gov/mesh/68066289 (accessed August 31, 2016).

[8] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Acm Sigmod Rec., vol. 22, ACM; 1993, p. 207–216.

[9] Jung M, Hoerbst A, Hackl WO, Kirrane F, Borbolla D, Jaspers MW, et al. Attitude of physicians towards automatic alerting in computerized physician order entry systems. A comparative international survey. Methods Inf Med 2013;52:99–108. doi:10.3414/ME12-02-0007.

[10] Safran C. Using routinely collected data for clinical research. Stat Med 1991;10:559–64.

[11] Safran C. Reuse of clinical data. Yearb Med Inform 2014;9:52–4. doi:10.15265/IY-2014-0013.

[12] Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306:848–55. doi:10.1001/jama.2011.1204.

[13] Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc JAMIA 2012;19:859–66. doi:10.1136/amiajnl-2011-000535.

[14] Bates DW, O'Neil AC, Boyle D, Teich J, Chertow GM, Komaroff AL, et al. Potential identifiability and preventability of adverse events using information systems. J Am Med Inform Assoc JAMIA 1994;1:404–11.

[15] Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Trans Inf Technol Biomed Publ IEEE Eng Med Biol Soc 2011;15:823–30. doi:10.1109/TITB.2011.2165727.

[16] Windal F, Jeribi K, Ficheur G, Degoul S, Martinot A, Beuscart R, et al. Pediatric emergency department crowding: survival tree clustering for length of patient stay. Stud Health Technol Inform 2014;205:1095–9.

[17] Barak-Corren Y, Israelit SH, Reis BY. Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow. Emerg Med J EMJ 2017;34:308–14. doi:10.1136/emermed-2014-203819.

[18] Chazard E, Beuscart R. Graphical representation of the comprehensive patient flow through the hospital. AMIA Annu Symp Proc AMIA Symp AMIA Symp 2007:110–4.

[19] Evans RS, Burke JP, Classen DC, Gardner RM, Menlove RL, Goodrich KM, et al. Computerized identification of patients at high risk for hospital-acquired infection. Am J Infect Control 1992;20:4–10.

[20] Köpcke F, Prokosch H-U. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014;16:e161. doi:10.2196/jmir.3446.

[21] Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int 2015;2015. doi:10.1155/2015/639021.

[22] ATIH. Accès aux bases PMSI | Publication ATIH n.d. http://atih.sante.fr/bases-de-donnees/commande-de-bases (accessed May 18, 2017).

[23] Mallon WJ. Big data. J Shoulder Elb Surg Am Shoulder Elb Surg Al 2013;22:1153. doi:10.1016/j.jse.2013.07.034.

[24] Salcido RS. Big data and disruptive innovation in wound care. Adv Skin Wound Care 2013;26:344. doi:10.1097/01.ASW.0000432244.36301.fc.

[25] Ketchersid T. Big data in nephrology: friend or foe? Blood Purif 2013;36:160–4. doi:10.1159/000356751.

[26] Hovenga EJS, Grain H. Health data and data governance. Stud Health Technol Inform 2013;193:67–92.

[27] Müller H, Hanbury A, Al Shorbaji N. Health information search to deal with the exploding amount of health information produced. Methods Inf Med 2012;51:516–8.

[28] Porche DJ. Men's Health Big Data. Am J Mens Health 2014;8:189. doi:10.1177/1557988314529838.

[29] Callebaut W. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. Stud Hist Philos Biol Biomed Sci 2012;43:69–80. doi:10.1016/j.shpsc.2011.10.007.

[30] Fan J, Liu H. Statistical analysis of big data on pharmacogenomics. Adv Drug Deliv Rev 2013;65:987–1000. doi:10.1016/j.addr.2013.04.008.

[31] Lupşe O-S, Crişan-Vida M, Stoicu-Tivadar L, Bernard E. Supporting diagnosis and treatment in medical care based on big data processing. Stud Health Technol Inform 2014;197:65–9.

[32] Berger ML, Doban V. Big data, advanced analytics and the future of comparative effectiveness research. J Comp Eff Res 2014;3:167–76. doi:10.2217/cer.14.2.

[33] Chute CG, Ullman-Cullere M, Wood GM, Lin SM, He M, Pathak J. Some experiences and opportunities for big data in translational research. Genet Med Off J Am Coll Med Genet 2013;15:802–9. doi:10.1038/gim.2013.121.

[34] Ola O, Sedig K. The challenge of big data in public health: an opportunity for visual analytics. Online J Public Health Inform 2014;5:223. doi:10.5210/ojphi.v5i3.4933.

[35] Dereli T, Coşkun Y, Kolker E, Güner O, Ağırbaşlı M, Ozdemir V. Big data and ethics review for health systems research in LMICs: understanding risk, uncertainty and ignorance-and catching the black swans? Am J Bioeth AJOB 2014;14:48–50. doi:10.1080/15265161.2013.868955.

[36] Harnessing big data. How to achieve value. Hosp Health Netw AHA 2014;88:61–71.

[37] Jee K, Kim G-H. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res 2013;19:79–85. doi:10.4258/hir.2013.19.2.79.

[38] Van Horn JD, Toga AW. Human neuroimaging as a “Big Data” science. Brain Imaging Behav 2013. doi:10.1007/s11682-013-9255-y.

[39] Moore KD, Eyestone K, Coddington DC. The big deal about big data. Healthc Financ Manag J Healthc Financ Manag Assoc 2013;67:60–6, 68.

[40] O’Driscoll A, Daugelaite J, Sleator RD. “Big data”, Hadoop and cloud computing in genomics. J Biomed Inform 2013;46:774–81. doi:10.1016/j.jbi.2013.07.001.

[41] Buyer’s brief: cognitive computing in the age of big data. Healthc Financ Manag J Healthc Financ Manag Assoc 2014;68:35–6.

[42] Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. Harv Bus Rev 2012;90:70–6, 128.

[43] Khoury MJ, Lam TK, Ioannidis JPA, Hartge P, Spitz MR, Buring JE, et al. Transforming epidemiology for 21st century medicine and public health. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol 2013;22:508–16. doi:10.1158/1055-9965.EPI-13-0146.

[44] Bonney S. HIM’s role in managing big data: Turning data collected by an EHR into information. J AHIMA Am Health Inf Manag Assoc 2013;84:62–4.

[45] Jayapandian CP, Chen C-H, Bozorgi A, Lhatoo SD, Zhang G-Q, Sahoo SS. Cloudwave: distributed processing of “big data” from electrophysiological recordings for epilepsy clinical research using hadoop. AMIA Annu Symp Proc AMIA Symp AMIA Symp 2013;2013:691–700.

[46] Schadt EE. The changing privacy landscape in the era of big data. Mol Syst Biol 2012;8:612. doi:10.1038/msb.2012.47.

[47] Aji A, Wang F, Saltz JH. Towards Building a High Performance Spatial Query System for Large Scale Medical Imaging Data. Proc ACM SIGSPATIAL Int Conf Adv Geogr Inf Syst ACM GIS ACM SIGSPATIAL Int Conf Adv Geogr Inf Syst 2012;2012:309–18.

[48] Matheson GO, Klügl M, Engebretsen L, Bendiksen F, Blair SN, Börjesson M, et al. Prevention and management of noncommunicable disease: the IOC Consensus Statement, Lausanne 2013. Clin J Sport Med Off J Can Acad Sport Med 2013;23:419–29. doi:10.1097/JSM.0000000000000038.

[49] Afendi FM, Ono N, Nakamura Y, Nakamura K, Darusman LK, Kibinge N, et al. Data Mining Methods for Omics and Knowledge of Crude Medicinal Plants toward Big Data Biology. Comput Struct Biotechnol J 2013;4:e201301010. doi:10.5936/csbj.201301010.

[50] Litman RS. Complications of laryngeal masks in children: big data comes to pediatric anesthesia. Anesthesiology 2013;119:1239–40. doi:10.1097/ALN.0000000000000016.

[51] Ward JC. Oncology reimbursement in the era of personalized medicine and big data. J Oncol Pract Am Soc Clin Oncol 2014;10:83–6. doi:10.1200/JOP.2014.001308.

[52] Lindenmayer DB, Likens GE. Analysis: don’t do big-data science backwards. Nature 2013;499:284. doi:10.1038/499284d.

[53] Toh S, Platt R. Big data in epidemiology: too big to fail? Epidemiol Camb Mass 2013;24:939. doi:10.1097/EDE.0b013e3182a71390.

[54] Castellanos FX, Di Martino A, Craddock RC, Mehta AD, Milham MP. Clinical applications of the functional connectome. NeuroImage 2013;80:527–40. doi:10.1016/j.neuroimage.2013.04.083.

[55] Markowetz A, Błaszkiewicz K, Montag C, Switala C, Schlaepfer TE. Psycho-Informatics: Big Data shaping modern psychometrics. Med Hypotheses 2014;82:405–11. doi:10.1016/j.mehy.2013.11.030.

[56] Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35:332–8. doi:10.1016/j.genhosppsych.2013.03.008.

[57] Klingström T, Soldatova L, Stevens R, Roos TE, Swertz MA, Müller KM, et al. Workshop on laboratory protocol standards for the Molecular Methods Database. New Biotechnol 2013;30:109–13. doi:10.1016/j.nbt.2012.05.019.

[58] Maclean D, Kamoun S. Big data in small places. Nat Biotechnol 2012;30:33–4. doi:10.1038/nbt.2079.

[59] Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA J Am Med Assoc 2013;309:1351–2. doi:10.1001/jama.2013.393.

[60] Hamilton B. Impacts of big data. Potential is huge, so are challenges. Health Manag Technol 2013;34:12–3.

[61] Mohammed Y, Mostovenko E, Henneman AA, Marissen RJ, Deelder AM, Palmblad M. Cloud parallel processing of tandem mass spectrometry based proteomics data. J Proteome Res 2012;11:5101–8. doi:10.1021/pr300561q.

[62] Karlsson J, Trelles O. MAPI: a software framework for distributed biomedical applications. J Biomed Semant 2013;4:4. doi:10.1186/2041-1480-4-4.

[63] DiLeo MV, Strahan GD, den Bakker M, Hoekenga OA. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PloS One 2011;6:e26683. doi:10.1371/journal.pone.0026683.

[64] Greene CS, Tan J, Ung M, Moore JH, Cheng C. Big Data Bioinformatics. J Cell Physiol 2014. doi:10.1002/jcp.24662.

[65] Marx V. Biology: The big challenges of big data. Nature 2013;498:255–60. doi:10.1038/498255a.

[66] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet 2010;11:647–57. doi:10.1038/nrg2857.

[67] Ansermino JM. From the Journal archives: Improving patient outcomes in the era of Big Data. Can J Anaesth J Can Anesth 2014. doi:10.1007/s12630-014-0146-5.

[68] Bower MR, Stead M, Brinkmann BH, Dufendach K, Worrell GA. Metadata and annotations for multi-scale electrophysiological data. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Conf 2009;2009:2811–4. doi:10.1109/IEMBS.2009.5333570.

[69] Ranganathan S, Schönbach C, Kelso J, Rost B, Nathan S, Tan TW. Towards big data science in the decade ahead from ten years of InCoB and the 1st ISCB-Asia Joint Conference. BMC Bioinformatics 2011;12 Suppl 13:S1. doi:10.1186/1471-2105-12-S13-S1.

[70] Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct 2012;7:43; discussion 43. doi:10.1186/1745-6150-7-43.

[71] Cole JB, Newman S, Foertter F, Aguilar I, Coffey M. Breeding and Genetics Symposium: really big data: processing and analysis of very large data sets. J Anim Sci 2012;90:723–33. doi:10.2527/jas.2011-4584.

[72] Finding correlations in big data. Nat Biotechnol 2012;30:334–5. doi:10.1038/nbt.2182.

[73] Kolker E, Stewart E, Ozdemir V. Opportunities and challenges for the life sciences community. Omics J Integr Biol 2012;16:138–47. doi:10.1089/omi.2011.0152.

[74] Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med 2014. doi:10.1136/bjsports-2014-093546.

[75] Feldmann E, Liebeskind DS. Developing Precision Stroke Imaging. Front Neurol 2014;5:29. doi:10.3389/fneur.2014.00029.

[76] Green DE, Rapp EJ. Can big data lead us to big savings? Radiogr Rev Publ Radiol Soc N Am Inc 2013;33:859–60. doi:10.1148/rg.333135035.

[77] Huberman BA. Sociology of science: Big data deserve a bigger audience. Nature 2012;482:308. doi:10.1038/482308d.

[78] Özdemir V, Badr KF, Dove ES, Endrenyi L, Geraci CJ, Hotez PJ, et al. Crowd-funded micro-grants for genomics and “big data”: an actionable idea connecting small (artisan) science, infrastructure science, and citizen philanthropy. Omics J Integr Biol 2013;17:161–72. doi:10.1089/omi.2013.0034.

[79] Mavandadi S, Dimitrov S, Feng S, Yu F, Yu R, Sikora U, et al. Crowd-sourced BioGames: managing the big data problem for next-generation lab-on-a-chip platforms. Lab Chip 2012;12:4102–6. doi:10.1039/c2lc40614d.

[80] Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. Harv Bus Rev 2012;90:70–6, 128.

[81] Lynch C. Big data: How do your data grow? Nature 2008;455:28–9. doi:10.1038/455028a.

[82] Maps, “Big Data,”and Case Reports. Glob Adv Health Med Improv Healthc Outcomes Worldw 2012;1:5–7. doi:10.7453/gahmj.2012.1.3.001.

[83] Hoffman S, Podgurski A. Big bad data: law, public health, and biomedical databases. J Law Med Ethics J Am Soc Law Med Ethics 2013;41 Suppl 1:56–60. doi:10.1111/jlme.12040.

[84] Cockfield J, Su K, Robbins KA. MOBBED: a computational data infrastructure for handling large collections of event-rich time series datasets in MATLAB. Front Neuroinformatics 2013;7:20. doi:10.3389/fninf.2013.00020.

[85] Martin SF, Falkenberg H, Dyrlund TF, Khoudoli GA, Mageean CJ, Linding R. PROTEINCHALLENGE: crowd sourcing in proteomics analysis and software development. J Proteomics 2013;88:41–6. doi:10.1016/j.jprot.2012.11.014.

[86] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

[87] Therneau TM. Modeling Survival Data: Extending the Cox Model. Springer Science & Business Media; 2000.

[88] Chazard E, Merlin B, Ficheur G, Sarfati J-C, PSIP Consortium, Beuscart R. Detection of adverse drug events: proposal of a data model. Stud Health Technol Inform 2009;148:63–74.

[89] Djennaoui M, Ficheur G, Beuscart R, Chazard E. Improvement of the quality of medical databases: data-mining-based prediction of diagnostic codes from previous patient codes. Stud Health Technol Inform 2015;210:419–23.

[90] Merlin B, Chazard E, Pereira S, Serrot E, Sakji S, Beuscart R, et al. Can F-MTI semantic-mined drug codes be used for adverse drug events detection when no CPOE is available? Stud Health Technol Inform 2010;160:1025–9.

[91] Ficheur G, Chazard E, Schaffar A, Genty M, Beuscart R. Interoperability of medical databases: construction of mapping between hospitals laboratory results assisted by automated comparison of their distributions. AMIA Annu Symp Proc AMIA Symp AMIA Symp 2011;2011:392–401.

[92] Harmoniser et promouvoir l’informatique médicale. Wikipédia 2015.

[93] Kent W. A simple guide to five normal forms in relational database theory. Commun ACM 1983;26:120–125.

[94] Coatrieux G, Chazard E, Beuscart R, Roux C. Lossless watermarking of categorical attributes for verifying medical data base integrity. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Conf 2011;2011:8195–8. doi:10.1109/IEMBS.2011.6092021.

[95] Franco Contreras J, Coatrieux G, Chazard E, Cuppens F, Cuppens-Boulahia N, Roux C. Robust lossless watermarking based on circular interpretation of bijective transformations for the protection of medical databases. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Conf 2012;2012:5875–8. doi:10.1109/EMBC.2012.6347330.

[96] Coatrieux G, Maitre H, Sankur B, Rolland Y, Collorec R. Relevance of watermarking in medical imaging. Inf. Technol. Appl. Biomed. 2000 Proc. 2000 IEEE EMBS Int. Conf. On, IEEE; 2000, p. 250–255.

[97] Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inf 2014;83:303–12. doi:10.1016/j.ijmedinf.2013.11.005.

[98] Kleene SC. Representation of events in nerve nets and finite automata. DTIC Document; 1951.

[99] Chazard E, Ficheur G, Merlin B, Genin M, Preda C, PSIP consortium, et al. Detection of adverse drug events detection: data aggregation and data mining. Stud Health Technol Inform 2009;148:75–84.

[100] Chazard E, Ficheur G, Merlin B, Serrot E, PSIP Consortium, Beuscart R. Adverse drug events prevention rules: multi-site evaluation of rules from various sources. Stud Health Technol Inform 2009;148:102–11.

[101] Caron A, Chazard E, Muller J, Perichon R, Ferret L, Koutkias V, et al. IT-CARES: an interactive tool for case-crossover analyses of electronic medical records for patient safety. J Am Med Inform Assoc JAMIA 2016. doi:10.1093/jamia/ocw132.

[102] Ficheur G, Caron A, Beuscart J-B, Ferret L, Jung Y-J, Garabedian C, et al. Case-crossover study to examine the change in postpartum risk of pulmonary embolism over time. BMC Pregnancy Childbirth 2017;17:119. doi:10.1186/s12884-017-1283-y.

[103] Ficheur G, Caron A, Beuscart J-B, Ferret L, Putman S, Beuscart R, et al. The risks of pulmonary embolism and upper gastrointestinal bleeding beyond 35days after total hip replacement for coxarthrosis among middle-aged patients: A cross-over cohort. Prev Med 2016. doi:10.1016/j.ypmed.2016.09.010.

[104] Chazard E, Luyckx M, Beuscart J-B, Ferret L, Beuscart R. Routine use of the “ADE scorecards”, an application for automated ADE detection in a general hospital. Stud Health Technol Inform 2013;192:308–12.

[105] Chazard E, Bernonville S, Ficheur G, Beuscart R. A statistics-based approach of contextualization for adverse drug events detection and prevention. Stud Health Technol Inform 2012;180:766–70.

[106] Chazard E, Băceanu A, Ferret L, Ficheur G. The ADE scorecards: a tool for adverse drug event detection in electronic health records. Stud Health Technol Inform 2011;166:169–79.

[107] Chazard E, Preda C, Merlin B, Ficheur G, PSIP consortium, Beuscart R. Data-mining-based detection of adverse drug events. Stud Health Technol Inform 2009;150:552–6.

[108] Mehta RL, Kellum JA, Shah SV, Molitoris BA, Ronco C, Warnock DG, et al. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care Lond Engl 2007;11:R31. doi:10.1186/cc5713.

[109] Frachon I, Etienne Y, Jobic Y, Le Gal G, Humbert M, Leroyer C. Benfluorex and unexplained valvular heart disease: a case-control study. PloS One 2010;5:e10128. doi:10.1371/journal.pone.0010128.

[110] Etienne Y, Jobic Y, Frachon I, Fatemi M, Castellant P, Quintin-Roué I. Mitral and aortic valvular disease associated with benfluorex use. J Heart Valve Dis 2011;20:348–50.

[111] Ficheur G, Chazard E, Merlin B, Ferret L, Luyckx M, Beuscart R. Supervised analysis of drug prescription sequences. Stud Health Technol Inform 2013;192:293–7.

[112] 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;13:306–14. doi:10.1136/qhc.13.4.306.

[113] Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inform Assoc JAMIA 1998;5:305–14.

[114] Kuperman GJ, Teich JM, Bates DW, Hiltz FL, Hurley JM, Lee RY, et al. Detecting alerts, notifying the physician, and offering action items: a comprehensive alerting system. Proc Conf Am Med Inform Assoc AMIA Annu Fall Symp AMIA Fall Symp 1996:704–8.

[115] Del Fiol G, Rocha BH, Kuperman GJ, Bates DW, Nohama P. Comparison of two knowledge bases on the detection of drug-drug interactions. Proc AMIA Annu Symp AMIA Symp 2000:171–5.

[116] Gandhi TK, Weingart SN, Seger AC, Borus J, Burdick E, Poon EG, et al. Outpatient prescribing errors and the impact of computerized prescribing. J Gen Intern Med 2005;20:837–41. doi:10.1111/j.1525-1497.2005.0194.x.

[117] Judge J, Field TS, DeFlorio M, Laprino J, Auger J, Rochon P, et al. Prescribers' responses to alerts during medication ordering in the long term care setting. J Am Med Inform Assoc JAMIA 2006;13:385–90. doi:10.1197/jamia.M1945.

[118] Honigman B, Lee J, Rothschild J, Light P, Pulling RM, Yu T, et al. Using computerized data to identify adverse drug events in outpatients. J Am Med Inform Assoc JAMIA 2001;8:254–66.

[119] Chazard E, Puech P, Gregoire M, Beuscart R. Using Treemaps to represent medical data. Stud Health Technol Inform 2006;124:522–7.

[120] Shneiderman B. Tree Visualization with Tree-maps: 2-d Space-filling Approach. ACM Trans Graph 1992;11:92–99. doi:10.1145/102377.115768.

[121] The PHP Group. PHP: Hypertext Preprocessor n.d. http://php.net/ (accessed February 3, 2017).

[122] World Wide Web Consortium. W3C SVG Working Group n.d. https://www.w3.org/Graphics/SVG/ (accessed February 3, 2017).

[123] Bellamy-Royds JDE Amelia. SVG Essentials. n.d.

[124] Bruls M, Huizing K, Van Wijk JJ. Squarified treemaps. Data Vis. 2000, Springer; 2000, p. 33–42.

[125] Chazard E, Ficheur G, Beuscart J-B, Preda C. How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests. Value Health J Int Soc Pharmacoeconomics Outcomes Res 2017;20:992–8. doi:10.1016/j.jval.2017.02.009.

[126] Chazard E, Dumesnil C, Beuscart R. How much does hyperkalemia lengthen inpatient stays? About methodological issues in analyzing time-dependant events. Stud Health Technol Inform 2015;210:835–9.

[127] Nestrigue C, Or Z. Excess Costs of Adverse Events in Hospitals in France. Issues Health Econ IRDES 2011.

[128] Ficheur G, Chazard E, Beuscart J-B, Merlin B, Luyckx M, Beuscart R. Adverse drug events with hyperkalaemia during inpatient stays: evaluation of an automated method for retrospective detection in hospital databases. BMC Med Inform Decis Mak 2014;14:83. doi:10.1186/1472-6947-14-83.

[129] Hackl WO, Ammenwerth E, Marcilly R, Chazard E, Luyckx M, Leurs P, et al. Clinical evaluation of the ADE scorecards as a decision support tool for adverse drug event analysis and medication safety management. Br J Clin Pharmacol 2013;76 Suppl 1:78–90. doi:10.1111/bcp.12185.

[130] Koutkias V, Kilintzis V, Stalidis G, Lazou K, Collyda C, Chazard E, et al. Constructing Clinical Decision Support Systems for Adverse Drug Event Prevention: A Knowledge-based Approach. AMIA Annu Symp Proc AMIA Symp AMIA Symp 2010;2010:402–6.

[131] Marcilly R, Chazard E, Beuscart-Zéphir M-C, Hackl W, Băceanu A, Kushniruk A, et al. Design of Adverse Drug Events-Scorecards. Stud Health Technol Inform 2011;164:377–81.

[132] Ferret L, Luyckx M, Ficheur G, Chazard E, Beuscart R. Evaluation of a Computer Application for Retrospective Detection of Vitamin K Antagonist Treatment Imbalance. J Patient Saf 2016. doi:10.1097/PTS.0000000000000182.

[133] Ferret L, Luyckx M, Merlin B, Ficheur G, Chazard E, Beuscart R. Evaluation of a computerized tool allowing retrospective detection of potential vitamin K antagonist overdoses in complex contexts. Stud Health Technol Inform 2013;192:553–6.

[134] Koutkias VG, McNair P, Kilintzis V, Skovhus Andersen K, Niès J, Sarfati J-C, et al. From adverse drug event detection to prevention. A novel clinical

decision support framework for medication safety. Methods Inf Med 2014;53:482–92. doi:10.3414/ME14-01-0027.

[135] Perichon R, Chazard E, Beuscart R. Patients drug exchange forum corpus: toward drug safety signals detection. Stud Health Technol Inform 2015;210:1023.

[136] Băceanu A, Atasiei I, Chazard E, Leroy N, PSIP Consortium. The expert explorer: a tool for hospital data visualization and adverse drug event rules validation. Stud Health Technol Inform 2009;148:85–94.

[137] Leroy N, Chazard E, Beuscart R, Beuscart-Zephir MC, Psip Consortium. Toward automatic detection and prevention of adverse drug events. Stud Health Technol Inform 2009;143:30–5.

[138] Kohn LT, Corrigan JM, Donaldson MS, others. To err is human: building a safer health system. vol. 6. National Academies Press; 2000.

[139] European Community. Directive 2001/83/EC on the Community code relating to medicinal products for human use. 2001. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2001L0083:20090720:EN:PDF (accessed August 29, 2016).

[140] Aspden P, Wolcott J, Bootman L, Cronenwelt L. Institute of Medicine, Preventing Medication Errors, Quality Chasm Series. Washington, DC: The National Academies Press; 2007.

[141] Handler SM, Wright RM, Ruby CM, Hanlon JT. Epidemiology of medication-related adverse events in nursing homes. Am J Geriatr Pharmacother 2006;4:264–72. doi:10.1016/j.amjopharm.2006.09.011.

[142] Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. JAMA 1995;274:29–34.

[143] Gurwitz JH, Field TS, Avorn J, McCormick D, Jain S, Eckler M, et al. Incidence and preventability of adverse drug events in nursing homes. Am J Med 2000;109:87–94.

[144] Nebeker JR, Barach P, Samore MH. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. Ann Intern Med 2004;140:795–801.

[145] Begaud B. Standardized assessment of adverse drug reactions: the method used in France. Special workshop--clinical. Drug Inf J 1984;18:275–81.

[146] Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36:131–43.

[147] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010;160:739–43.

[148] Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf 2005;28:981–1007.

[149] Bate A, Edwards I. Data mining in spontaneous reports. Basic Clin Pharmacol Toxicol 2006;98:324–330.

[150] Chazard E, Beeler PE, Ficheur G, Dalleur O, Beuscart R, Bates DW. Drug-drug interactions with Vitamin K antagonists: are we focusing on the right rules? [IN PRESS] 2017.

[151] van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc JAMIA 2006;13:138–47. doi:10.1197/jamia.M1809.

[152] Phansalkar S, Desai A, Choksi A, Yoshida E, Doole J, Czochanski M, et al. Criteria for assessing high-priority drug-drug interactions for clinical decision support in electronic health records. BMC Med Inform Decis Mak 2013;13:65. doi:10.1186/1472-6947-13-65.

[153] van der Sijs H, Aarts J, van Gelder T, Berg M, Vulto A. Turning off frequently overridden drug alerts: limited opportunities for doing it safely. J Am Med Inform Assoc JAMIA 2008;15:439–48. doi:10.1197/jamia.M2311.

[154] Smithburger PL, Buckley MS, Bejian S, Burenheide K, Kane-Gill SL. A critical evaluation of clinical decision support for the detection of drug-drug interactions. Expert Opin Drug Saf 2011;10:871–82. doi:10.1517/14740338.2011.583916.

[155] van der Sijs H, Mulder A, van Gelder T, Aarts J, Berg M, Vulto A. Drug safety alert generation and overriding in a large Dutch university medical centre. Pharmacoepidemiol Drug Saf 2009;18:941–7. doi:10.1002/pds.1800.

[156] Lin C-P, Payne TH, Nichol WP, Hoey PJ, Anderson CL, Gennari JH. Evaluating clinical decision support systems: monitoring CPOE order check override rates in the Department of Veterans Affairs' Computerized Patient Record System. J Am Med Inform Assoc JAMIA 2008;15:620–6. doi:10.1197/jamia.M2453.

[157] Isaac T, Weissman JS, Davis RB, Massagli M, Cyrulik A, Sands DZ, et al. Overrides of medication alerts in ambulatory care. Arch Intern Med 2009;169:305–11. doi:10.1001/archinternmed.2008.551.

[158] Yeh M-L, Chang Y-J, Wang P-Y, Li Y-CJ, Hsu C-Y. Physicians' responses to computerized drug-drug interaction alerts for outpatients. Comput Methods Programs Biomed 2013;111:17–25. doi:10.1016/j.cmpb.2013.02.006.

[159] Taegtmeyer AB, Kullak-Ublick GA, Widmer N, Falk V, Jetter A. Clinical usefulness of electronic drug-drug interaction checking in the care of cardiovascular surgery inpatients. Cardiology 2012;123:219–22. doi:10.1159/000343272.

[160] Fritz D, Ceschi A, Curkovic I, Huber M, Egbring M, Kullak-Ublick GA, et al. Comparative evaluation of three clinical decision support systems: prospective screening for medication errors in 100 medical inpatients. Eur J Clin Pharmacol 2012;68:1209–19. doi:10.1007/s00228-012-1241-6.

[161] Slight SP, Seger DL, Nanji KC, Cho I, Maniam N, Dykes PC, et al. Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. PloS One 2013;8:e85071. doi:10.1371/journal.pone.0085071.

[162] Caruba T, Colombet I, Gillaizeau F, Bruni V, Korb V, Prognon P, et al. Chronology of prescribing error during the hospital stay and prediction of pharmacist's alerts overriding: a prospective analysis. BMC Health Serv Res 2010;10:13. doi:10.1186/1472-6963-10-13.

[163] Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in outpatients. J Am Med Inform Assoc JAMIA 2014;21:487–91. doi:10.1136/amiajnl-2013-001813.

[164] van der Sijs H, van Gelder T, Vulto A, Berg M, Aarts J. Understanding handling of drug safety alerts: a simulation study. Int J Med Inf 2010;79:361–9. doi:10.1016/j.ijmedinf.2010.01.008.

[165] Holbrook AM, Pereira JA, Labiris R, McDonald H, Douketis JD, Crowther M, et al. Systematic overview of warfarin and its drug and food interactions. Arch Intern Med 2005;165:1095–106. doi:10.1001/archinte.165.10.1095.

[166] Chazard E, Ficheur G, Beuscart R. Risque hémorragique sous anti-vitamines K : quelles sont réellement les interactions prioritaires ? Rev DÉpidémiologie Santé Publique 2016;64, Supplement 1:S11–2. doi:10.1016/j.respe.2016.01.041.

[167] Petit A-E, Mangeard H, Chazard E, Puisieux F. [Changes in drug management of Alzheimer's disease in nursing homes: Impact of the media campaign against specific drugs for Alzheimer's disease]. L'Encephale 2016. doi:10.1016/j.encep.2015.03.006.

[168] Ferret L, Beuscart J-B, Ficheur G, Beuscart R, Luyckx M, Chazard E. Evaluation of compliance with recommendations of prevention of thromboembolism in atrial fibrillation in the elderly, by data reuse of electronic health records. Stud Health Technol Inform 2015;210:394–8.

[169] Frély A, Chazard E, Pansu A, Beuscart J-B, Puisieux F. Impact of acute geriatric care in elderly patients according to the Screening Tool of Older Persons' Prescriptions/Screening Tool to Alert doctors to Right Treatment criteria in northern France. Geriatr Gerontol Int 2015. doi:10.1111/ggi.12474.

[170] Ficheur G, Schaffar A, Caron A, Balcaen T, Beuscart J-B, Chazard E. Elderly Surgical Patients: Automated Computation of Healthcare Quality Indicators by Data Reuse of EHR. Stud Health Technol Inform 2016;221:92–6.

[171] Chazard E, Babaousmail D, Schaffar A, Ficheur G, Beuscart R. Process assessment by automated computation of healthcare quality indicators in hospital electronic health records: a systematic review of indicators. Stud Health Technol Inform 2015;210:867–71.

[172] Beeler PE, Bates DW, Hug BL. Clinical decision support systems. Swiss Med Wkly 2014;144:w14073. doi:10.4414/smw.2014.14073.

[173] Nigam A. Changing health care quality paradigms: the rise of clinical guidelines and quality measures in American medicine. Soc Sci Med 1982 2012;75:1933–7. doi:10.1016/j.socscimed.2012.07.038.

[174] McGory ML, Kao KK, Shekelle PG, Rubenstein LZ, Leonardi MJ, Parikh JA, et al. Developing quality indicators for elderly surgical patients. Ann Surg 2009;250:338–47. doi:10.1097/SLA.0b013e3181ae575a.

[175] Chazard E, Marcolino MS, Dumesnil C, Caron A, Palhares DMF, Ficheur G, et al. One Million Electrocardiograms of Primary Care Patients: A Descriptive Analysis. Stud Health Technol Inform 2015;216:69–73.

[176] Rosier A, Mabo P, Temal L, Van Hille P, Dameron O, Deléger L, et al. Personalized and automated remote monitoring of atrial fibrillation. Eur Eur Pacing Arrhythm Card Electrophysiol J Work Groups Card Pacing Arrhythm Card Cell Electrophysiol Eur Soc Cardiol 2015. doi:10.1093/europace/euv234.

[177] Macfarlane PW, Devine B, Clark E. The university of glasgow (Uni-G) ECG analysis program. Comput. Cardiol. 2005, 2005, p. 451–4. doi:10.1109/CIC.2005.1588134.

[178] Macfarlane PW, Devine B, Latif S, McLaughlin S, Shoat DB, Watts MP. Methodology of ECG interpretation in the Glasgow program. Methods Inf Med 1990;29:354–61.

[179] Physio-Control, Inc., Medtronic B.V. Statement of Validation and Accuracy for the Glasgow 12-Lead ECG Analysis Program 2009.

[180] Chazard E, Dumesnil C, Marcolino MS, Caron A, Alkmim MB, Pinho-Ribeiro AL. Exploitation automatisée des données électrocardiographiques pour le codage : mise en place et évaluation. Rev DÉpidémiologie Santé Publique 2014;62, Supplement 3:S76. doi:10.1016/j.respe.2014.01.017.

[181] Ficheur G, Ferreira Careira L, Beuscart R, Chazard E. EpiHosp: A web-based visualization tool enabling the exploratory analysis of complications of implantable medical devices from a nationwide hospital database. Stud Health Technol Inform 2015;210:409–13.

[182] By the American Geriatrics Society 2015 Beers Criteria Update Expert Panel. American Geriatrics Society 2015 Updated Beers Criteria for Potentially Inappropriate Medication Use in Older Adults. J Am Geriatr Soc 2015;63:2227–46. doi:10.1111/jgs.13702.

[183] Guerra MTE, Viana RD, Feil L, Feron ET, Maboni J, Vargas AS-G. One-year mortality of elderly patients with hip fracture surgically treated at a hospital in Southern Brazil. Rev Bras Ortop 2016;52:17–23. doi:10.1016/j.rboe.2016.11.006.

[184] Le dispositif Paerpa. Ministère Solidar Santé 2017. http://social-sante.gouv.fr/systeme-de-sante-et-medico-social/parcours-des-patients-et-des-usagers/le-parcours-sante-des-aines-paerpa/article/le-dispositif-paerpa (accessed June 2, 2017).

[185] LOI n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé. 2016.

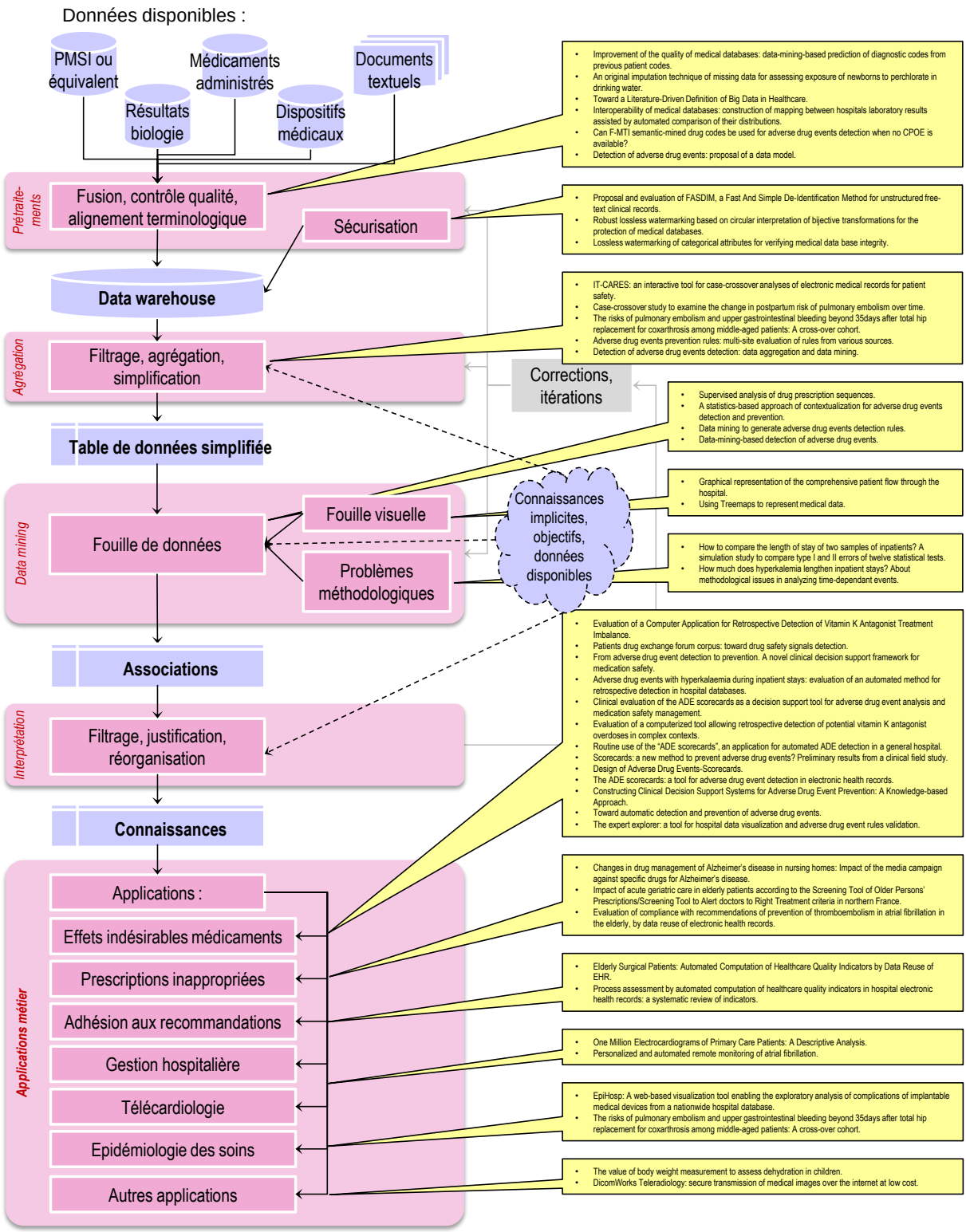
[186] Mise en œuvre du système national des données de santé et nouveau cadre d'accès aux données de santé - Mise en œuvre du système national des données de santé et nouveau cadre d'accès aux données de santé - Ministère des Solidarités et de la Santé n.d. http://drees.social-sante.gouv.fr/etudes-et-statistiques/acces-aux-donnees-de-sante/mise-en-oeuvre-du-systeme-national-des-donnees-de-sante-et-nouveau-cadre-d/article/mise-en-oeuvre-du-systeme-national-des-donnees-de-sante-et-nouveau-cadre-d (accessed June 3, 2017).

[187] GIE SESAM-Vitale. Parts de télétransmission n.d. http://www.sesam-vitale.fr/web/giesv/chiffres-parts-de-marche (accessed June 11, 2017).

[188] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

Annexes

1 Positionnement des publications PUBMED



2 Curriculum vitae

2.1 Profil, état civil

Emmanuel Chazard,
Nationalité française,
né le 18/10/1977 à Agen (47), 39 ans,
célibataire, une fille

Discipline :
Discipline académique : CNU 4604 biostatistiques, informatique médicale et technologies de communication
Spécialité hospitalière : DES Santé Publique et médecine sociale

2.2 Curriculum vitae

2.2.1 Postes occupés

Depuis 2011 :	Maître de Conférence des Universités - Praticien Hospitalier ➢ dans le Service Méthodologie, Biostatistiques Gestion de données Archives, pôle Santé Publique, Pharmacie Pharmacologie du CHU de Lille. ➢ au département de Biostatistiques et d'Informatique Médicale de la faculté de médecine de l'Université de Lille (secteur Droit et Santé) ➢ au Centre d'Etudes et de Recherche en Informatique Médicale, membre de l'équipe d'accueil EA 2694 « Santé Publique : épidémiologie et qualité des soins »
2014	Mobilité de recherche au Brigham and Women's Hospital (Harvard Medical School, Boston, MA, USA) (voir ci-après)
2014-2016	Mobilité de recherche à Inria (France) (voir ci-après)
2013	Mobilité de recherche au Service de télésanté de l'hôpital de l'université fédérale du Minas Gerais (Belo Horizonte, MG, Brésil) (voir ci-après)
2007-2011	Assistant hospitalo-universitaire. Mêmes affectations que le poste de MCU-PH.
2003-2007	Interne en Médecine, spécialité Santé publique (CHU de Lille, DRASS, Institut Pasteur de Lille)
1994-2004	Activités diverses durant les études médicales : encadrement de cours de catamaran particuliers ou collectifs (tous niveaux, tous âges, sur 10 ans), stages hospitaliers d'externat (mi-temps 48 mois),

encadrement médical d'enfants diabétiques (temps plein 2 mois), cours particuliers de mathématiques niveau lycée (sur 5 ans)

2.2.2 Techniques, langues

Statistiques : utilisation courante des logiciels de statistiques R / S+, SAS, SPSS

Informatique : programmation orientée objet en PHP 5 et JAVA ; notions PERL et C
Maîtrise des syntaxes SQL, HTML, XML, SVG…
Maîtrise des logiciels usuels : Business Objects®, Word®, Excel®, Powerpoint®, Access®, Publisher®…

Information médicale : Maîtrise des aspects techniques et tarifaires du PMSI et de la T2A en MCO, SSR, HAD. Exploitation des fichiers PMSI aux formats natifs.

Langues : Anglais professionnel parlé et écrit couramment, rudiments d'Allemand

2.2.3 Activités diverses

Activités web : auteur du cours « PMSI T2A Facturation » disponible en ligne (le plus consulté et cité nationalement)
Auteur d'un cours théorique de catamaran (n°1 sur Google depuis 2001)

Voile : Moniteur fédéral de voile (spécialité catamaran) à l'école des Glénans

Musique : Pratique sérieuse du piano classique, grand amateur de musique russe postromantique

Mandats : Elu en tête de liste : conseil scientifique de l'université Lille II (2005), CA du CROUS Midi-Pyrénées (2000), conseil d'UFR Médecine - Purpan, Toulouse III (1999), CA de l'Université Toulouse III (1998)

Permis : Permis B, Permis côtier

2.3 Diplômes

2011 **Thèse d'Université** intitulée « Automated detection of Adverse Drug Events by Data Mining of Electronic Health Records », soutenue en Anglais le 9 février 2011 dans la sous-section CNU 46-04, à l'école Doctorale Biologie et Santé de l'Université de Lille.
Directeur de thèse : Pr R. Beuscart.

2008 **Master 2 Méthodologie Statistique** en Recherche biomédicale, Faculté de Médecine Kremlin-Bicêtre, Université Paris XI.

2007 Diplôme d'Etat de **Docteur en Médecine**,
DES de Santé Publique et médecine sociale
Faculté de Médecine Henri Warembourg, Université de Lille.

2006 **Master 2 Marketing et Management** des Entreprises du Secteur de la Santé, Institut d'Administration des Entreprises, Université Lille 1 (major).

2005 **DESS Informatique** et Information dans les Réseaux de Soins, Faculté de Médecine Henri Warembourg, Université de Lille (major).

2003 Maîtrise des sciences biomédicales
Certificats de génétique, biologie moléculaire, biologie cellulaire,
Facultés de Médecine Purpan et Rangueil, Université Toulouse III.

1995 Baccalauréat S spécialité Mathématiques, Agen

2.4 Mobilités de recherche

J'ai réalisé 4 mobilités de recherche de 3 mois chacune.

2.4.1 Service de télésanté UFMG, Belo Horizonte, Brésil

Du **15 septembre 2013 au 18 décembre 2013**, j'ai travaillé dans le service de télésanté de l'hôpital de l'UFMG (université fédérale du Minas Gerais), à Belo Horizonte, capitale du Minas Gerais, un des 26 états du Brésil.

Le travail mené a été explicité dans la section 3.4 en page 80. Durant cette mission, j'ai analysé les données disponibles dans une base de données qui avait permis l'interprétation à distance de plus d'un million d'ECG. J'ai également étudié la concordance entre l'interprétation automatique réalisée par le Glasgow Program, et un gold standard issu d'une triple interprétation par des cardiologues. J'ai également tenté d'améliorer la détection des ECG normaux en incorporant les données cliniques, les traitements et les résultats de l'interprétation automatisée, afin d'en diminuer les faux négatifs et permettre une réorganisation du processus de revue des nouveaux ECG.

Ces travaux ont donné lieux aux publications suivantes :

- Chazard E, Marcolino MS, Dumesnil C, Caron A, Palhares DMF, Ficheur G, et al. One Million Electrocardiograms of Primary Care Patients: A Descriptive Analysis. Stud Health Technol Inform. 2015;216:69–73.
- Chazard E, Dumesnil C, Marcolino MS, Caron A, Alkmim MB, Pinho-Ribeiro AL. Exploitation automatisée des données électrocardiographiques pour le codage : mise en place et évaluation. Revue d'Épidémiologie et de Santé Publique. 2014 Mar;62, Supplement 3:S76.

2.4.2 Equipe MODAL, Inria

Du **3 janvier 2014 au 4 avril 2014**, puis **du 1er mai 2016 au 31 juillet 2016**, j'ai travaillé dans l'équipe de recherche MODAL (MOdel for Data Analysis and Learning, dirigée par Christophe Biernacki) d'Inria (anciennement INRIA, institut national de recherche en informatique et automatique). L'objectif de cette équipe est notamment de développer et évaluer des méthodes de data mining qui nécessitent le moins possible de connaissance experte sur les données.

Mes travaux précédents m'avaient amené à rechercher des associations entre des motifs incluant des médicaments, et des événements péjoratifs potentiellement imputables à ces médicaments. Le problème plus général de recherche d'association est fréquemment retrouvé dans les travaux touchant aux données hospitalières. Néanmoins, le caractère temporel de cette association est insuffisamment pris en compte dans la plupart des méthodes. J'ai donc réalisé des recherches qui ont permis de répondre à deux questions.

Question 1 : quel tests statistique choisir pour comparer les durées de séjour de deux groupes de patients ?

Question 2 : est-il possible de comparer les durées de séjour d'un groupe de patient présentant un événement avec un groupe de patient ne le présentant pas, lorsque cet événement est temps-dépendant ?

Ces travaux ont donné lieux aux publications suivantes :

- Chazard E, Ficheur G, Beuscart J-B, Preda C. How to Compare the Length of Stay of Two Samples of Inpatients? A Simulation Study to Compare Type I and Type II Errors of 12 Statistical Tests. Value in Health 2017. doi:10.1016/j.jval.2017.02.009.
- Chazard E, Dumesnil C, Beuscart R. How much does hyperkalemia lengthen inpatient stays? About methodological issues in analyzing time-dependant events. Stud Health Technol Inform. 2015;210:835-9.
- Chazard E, Preda C, Beuscart R. Comparer la durée de séjour de deux groupes de patients : quel test choisir ? Comparaison des risques alpha et bêta de douze tests statistiques. Revue d'Épidémiologie et de Santé Publique. 2016 Mar;64, Supplement 1:S32–3.

2.4.3 Brigham and Women's Hospital, Boston, USA

Du **7 avril 2014 au 7 juillet 2014**, j'ai travaillé à la Division de Médecine générale et de médecine interne du Brigham and Women's Hospital (BWH), dirigée par David W. Bates, en relation avec la Harvard Medical School, à Boston, Massachussets, USA.

David W. Bates a écrit ou co-écrit à ce jour plus de 680 articles référencés PUBMED. La plupart de ces articles portent sur la détection et la prévention des effets indésirables du médicament, faisant de lui une référence mondiale du domaine. Son équipe a en outre mis au point un logiciel de prévention prospective des EIM : au moment de la prescription, une alerte peut être envoyée au médecin en cas de situation à risque. Cet outil est aujourd'hui utilisé au Brigham & Women's Hospital. Cependant, il a été observé que les alertes étaient trop nombreuses (over-alerting) et parfois jugées non-pertinentes par les médecins, dont l'avis est enregistré à chaque alerte. Néanmoins, ce problème n'a pas été investigué au-delà du constat. De mon côté, j'avais déjà travaillé sur la réduction de l'over-alerting à l'aide de méthodes mixant la fouille statistique de données et la prise en compte des facteurs humains.

Durant cette mobilité, j'ai repris les règles visibles dans la littérature, et notamment dans une revue de référence [2]. Tirant profit des données recueillies dans le cadre du projet européen PSIP, j'ai évalué le risque lié à chaque règle de manière empirique à l'aide d'un modèle de Cox à covariables temps-dépendantes, afin de proposer une « alerte graduée » utilisant des moyens plus ou moins interruptifs en fonction de la probabilité de survenue d'un EIM. Ceci a permis, pour chaque médicament, de calculer la modification du risque d'observer un surdosage ou un sous-dosage en AVK (voir section 3.2 en page 72). Les résultats de cette recherche sont assez frappants, et sont en cours de publication.

Ces travaux ont donné lieux aux publications suivantes :

- Chazard E, Beeler PE, Ficheur G, Dalleur O, Beuscart R, Bates DW. Drug-drug interactions with Vitamin K antagonists: are we focusing on the right rules? J Am Med Inform Assoc [IN PRESS] 2017.

[2] Holbrook AM, Pereira JA, Labiris R, McDonald H, Douketis JD, Crowther M, et al. Systematic overview of warfarin and its drug and food interactions. Arch Intern Med. 2005 May 23;165(10):1095–106.

- Chazard E, Ficheur G, Beuscart R. Risque hémorragique sous anti-vitamines K : quelles sont réellement les interactions prioritaires ? Revue d'Épidémiologie et de Santé Publique. 2016 Mar;64, Supplement 1:S11–2.

2.5 Contrats de recherche (n=8)

J'ai participé aux projets de recherche financés suivants :

- **ANR CLINMINE**
  Optimisation de la prise en Charge des Patients à l'Hôpital
  Projet ANR, 2014-2017
  http://www.agence-nationale-recherche.fr/?Projet=ANR-13-TECS-0009
  http://www.lifl.fr/ClinMine/pmwiki/index.php
  Responsable de workpackage, modèles de données, acquisition de données nationales et aspects règlementaires.
- **PREPS PERFHU**
  Jugement partagé sur la performance des CHU
  Projet PREPS 2013-2017
  Participation.
- **PREPS EVALSI**
  Evaluation des systèmes d'information
  Projet PREPS 2013-2017
  Participation.
- **ANR HOST**
  Hôpital sous tension
  Projet ANR, 2012-2015
  http://www.agence-nationale-recherche.fr/?Projet=ANR-11-TECS-0010
  Participation à la fouille de données statistique.
- **ANR AKENATON**
  Automated Knowledge Extraction from medical records iN Association with a Telecardiology Observation Network
  Projet ANR 2007-2010
  http://www.agence-nationale-recherche.fr/?Projet=ANR-07-TECS-0001
  Définition du modèle de données.
- **FP7 PSIP, PSIP+ et PSIPEVAL**
  Patient Safety Through Intelligent Procedures in medication
  suivi de deux extensions
  Projet européen (ERC, FP7), 2008-2011
  http://cordis.europa.eu/project/rcn/85437_en.html
  Participant, responsable du modèle de données et de la fouille statistique de données.

2.6 Coopérations nationales et internationales

Dans le cadre des activités décrites ci-dessus, je collabore avec plusieurs partenaires :

- **des partenaires académiques :**
  - des chercheurs du Brigham and Women's Hospital et de la Harvard Medical School, à Boston, USA
  - des chercheurs de l'Université Fédérale du Minas Gerais, au Brésil
  - les chercheurs du CIC-IT 807 « Biocapteurs et e-santé, innovation et usages » à Lille, spécialisés en facteurs humaines et évaluation de l'utilisabilité des systèmes d'information d'une part, et en traitement du signal et dispositifs médicaux d'autre part
  - les membres de l'équipe MODAL à Inria, Villeneuve d'Ascq
  - des enseignants-chercheurs de l'Université Lille 1 : du laboratoire de mathématiques Paul Painlevé, de l'école Polythch'Lille, de l'Institut d'Administration des Entreprises
  - le Département Image et Traitement de l'Information, LaTIM, unité INSERM U650, Brest (M G. Coatrieux)
  - des enseignants-chercheurs du LTSI UMR1099 et de l'Université Rennes 1
- **des partenaires hospitaliers :**
  - les hôpitaux de Copenhague (Danemark)
  - le CHU de Rouen
  - le CH de Denain
  - le CH de Dunkerque
  - le CH de Valencienne
  - le GHICL (groupement des hôpitaux de l'institut catholique de Lille)
- **des partenaires industriels :**
  - Vidal SA, éditeur de contenus sur le médicament
  - IBM Denmark, éditeur informatique
  - Oracle, éditeur informatique
  - Medasys, éditeur informatique
  - Alicante, éditeur informatique
  - Sorin Biomedica, fabriquant de dispositifs médicaux cardiologiques

2.7 Soutien à la recherche clinique

Indépendamment de mes activités propres de recherche, je participe au soutien méthodologique dans le cadre de l'UF 2554 « Méthodologie, Biostatistiques et Data Management » dirigée par le Pr Duhamel. Cette partie sera détaillée dans les travaux hospitaliers (voir section 3.2).

2.8 Organisation de manifestations

J'ai participé à l'organisation de manifestations locales ou nationales :

- **Journée de recherche en Santé Publique**
  le 11 décembre 2008 à Lille
  manifestation scientifique régionale
  principal organisateur scientifique et logistique
- **Journée RNTS** (Réseau National Technologies pour la Santé)
  le 29 novembre 2005 à Lille Grand Palais
  réunion technique nationale
  organisation logistique
- **Journée CITH** (Centres d'Innovation Technologique Hospitaliers)
  le 28 novembre 2005 à Lille Grand Palais
  réunion technique nationale
  organisation logistique
- **Congrès JFIM 2005** (Journées Francophones d'Informatique Médicale)
  12 et 13 mai 2005 à Lille
  congrès scientifique international
  organisation logistique

2.9 Participation à des congrès

J'ai participé aux congrès suivants (par date décroissante) :

- **MIE 2017** (Medical Informatics in Europe)
  24-26 avril 2017 à Manchester (Royaume Uni)
  Congrès scientifique international
- **EMOIS 2017** (évaluation, management, organisation, information, santé)
  23-24 mars 2017 à Nancy (France)
  Congrès scientifique national, présentation orale
- **CSH 2016** (Congrès Convergences Santé Hôpital)
  29 septembre - 1 octobre 2016 à Avignon (France)
  Congrès scientifique national, présentation orale (conférence invitée)
- **MIE 2016 / HEC 2016** (Medical Informatics in Europe)
  28 août - 1 septembre 2016 à Munich (Allemagne)
  Congrès scientifique international
- **Big data: modelling, estimation and selection** (Ecole Centrale de Lille)
  9 juin - 10 juin 2016 à Lille (France)
  Congrès scientifique national, présentation orale (conférence invitée)
- **STC 2016** (Transforming Healthcare with the Internet of Things)
  17-19 avril 2016 à Paris (France)
  Congrès scientifique international, présentation orale
- **EMOIS 2016** (évaluation, management, organisation, information, santé)
  10-11 mars 2016 à Dijon (France)
  Congrès scientifique national, présentation orale
- **FIC 2016** (Forum international de la cybersécurité)
  25-26 janvier 2016 à Lille (France)
  Congrès technique national, présentation orale

- **MEDINFO 2015** (Medical Informatics)
  20-27 août 2015 à Sao Paulo (Brésil)
  Congrès scientifique international, présentation orale
- **MIE 2015** (Medical Informatics in Europe)
  26-29 mai 2015 à Madrid (Espagne)
  Congrès scientifique international, présentation orale
- **EMOIS 2015** (évaluation, management, organisation, information, santé)
  26-27 mars 2015 à Nancy (France)
  Congrès scientifique national, présentation orale
- **RIPPS 2014** (methodological approaches to paediatric pharmacoepidemiology & pharmacovigilance)
  5 décembre 2014 à Lyon (France)
  Congrès scientifique national, présentation orale
- **MIE 2014** (Medical Informatics in Europe)
  30 août - 3 septembre 2014 à Istanbul (Turquie)
  Congrès scientifique international, présentation orale
- **EMOIS 2014** (évaluation, management, organisation, information, santé)
  3-4 avril 2015 à Paris (France)
  Congrès scientifique national, présentation orale
- **MEDINFO 2013** (Medical Informatics)
  20-23 août 2013 à Copenhague (Danemark)
  Congrès scientifique international, présentation orale
- **DDTWC 2013** (Drug Discovery and Therapy World Congress)
  3-7 juin 2013 à Boston (Massachussets, USA)
  Congrès scientifique international, présentation orale
- **Séminaire de Télémédecine**
  13-17 mai 2013 à Belo Horizonte (Minas Gerais, Brésil)
  Séminaire de recherche franco-brésilien, présentation orale
- **EMOIS 2013** (évaluation, management, organisation, information, santé)
  21-22 mars 2013 à Nancy (Fr)
  Congrès scientifique national, présentation orale
- **Journées Usage du Contexte en Santé**
  27 septembre 2012 à Paris (Fr)
  Journée de travail nationale, présentation orale
- **Congrès ADELF 2012** (Association des Epidémiologistes de Langue Française)
  12-14 septembre 2012 à Bruxelles (Belgique)
  Congrès scientifique international francophone, poster
- **MIE 2012** (Medical Informatics in Europe)
  26-29 août 2012 à Pise (Italie)
  Congrès scientifique international, présentation orale
- **EMOIS 2012** (évaluation, management, organisation, information, santé)
  12-13 mars 2012 à Dijon (Fr)
  Congrès scientifique national, présentation orale
- **JFIM 2011** (Journées Francophones d'Informatique Médicale)
  23-24 septembre 2011
  Congrès scientifique international, présentation orale
- **SFAR 2011** (société française en anesthésie réanimation)
  21-24 septembre 2011
  Congrès scientifique national, présentation orale

- **MIE 2011** (Medical Informatics in Europe)
  28-31 août 2011 à Oslo (Norvège)
  Congrès scientifique international, présentation de poster
- **Second international PSIP Workshop**
  16-17 mai 2011 à Paris (Fr)
  Congrès scientifique international, présentation orale
- **EMOIS 2011** (évaluation, management, organisation, information, santé)
  17-18 mars 2011 à Nancy (Fr)
  Congrès scientifique national, présentation orale
- **Congrès MEDINFO 2010**
  12-15 septembre 2010 au Cap (Afrique du Sud)
  Congrès scientifique international, présentations orales
- **EMOIS-ADELF 2010** (évaluation, management, organisation, information, santé)
  22-23 avril 2010 à Bordeaux (Fr)
  Congrès scientifique national, présentation orale
- **International PSIP Workshop**
  23-26 septembre 2009 à Belgirate (Italie)
  Congrès scientifique international, présentations orales
- **Congrès RITS 2009** (recherche en imagerie et technologies de la santé)
  18, 19 et 20 mars 2009 à Lille
  congrès scientifique national, présentation orale
- **EMOIS 2009** (évaluation, management, organisation, information, santé)
  5-6 mars 2009 à Nancy
  congrès scientifique national, présentation orale
- **Congrès MIE 2008** (Medical Informatics in Europe)
  25-28 mai 2008 à Göteborg (Suède)
  congrès scientifique international, présentation orale dans un workshop
- **Congrès AMIA 2007** (American Medical Informatics Association)
  10-14 novembre 2007 à Chicago (USA)
  congrès scientifique international, présentation orale d'un article référencé
- **Congrès MIE 2006** (Medical Informatics in Europe)
  27-30 août 2006 à Maastricht (Pays-Bas)
  congrès scientifique international, présentation orale d'un article référencé

2.10 Expertises et peer-reviews

Je réalise des expertises :

- ANR TECSAN
- DGOS PREPS
- ID2Santé

Je réalise des revues en tant que pair pour des journaux internationaux indexés :

- APCI Applied Clinical Informatics
- BMC Bioinformatics
- BMC Medical informatics and decision making
- CAG Computers & Graphics
- CMPB Computer Methods and Programs in Biomedicine
- CTWO Cognition Technology & Work
- GGI Geriatrics & Gerontology International
- IJMI International Journal of Medical Informatics
- IMIA Yearbook
- IEEE-JBHI Journal of Biomedical and Health Informatics
- IRBM Ingénierie et Recherche Biomédicale
- JGEM Journal de Gestion et d'économie médicale

Je n'ai pas de responsabilité éditoriale à proprement parler.

3 Activités hospitalières

3.1 Le service Méthodologie, Biostatistiques Gestion de données Archives

Au CHU de Lille, le pôle S3P comprend plusieurs services :

- la Pharmacie centrale, Pr P. Odou
- la Pharmacologie, Pr R Bordet
- le Service Méthodologie, Biostatistiques Gestion de données Archives, Pr Beuscart
- le Département d'Information Médicale, Dr D Theis
- le Service d'Epidémiologie Régionale, Pr P Amouyel
- le service de Médecine du Travail, Pr A Sobaszek
- l'unité de nutrition artificielle à domicile, Dr D Lescut

Je suis affecté au Service Méthodologie, Biostatistiques Gestion de données Archives, qui comprend 3 UF :

- l'UF 2554 Méthodologie, Biostatistiques et Data Management, Pr A Duhamel
- l'UF 0041 Archives Médicales, Dr JM Renard
- l'UF 0690 Analyse des données dossier patient, Dr E Chazard

Je réalise mes activités dans deux de ces UF.

3.2 Conseil méthodologique et statistique

L'UF Méthodologie, Biostatistiques et Data Management, anciennement « plateforme d'aide méthodologique », propose son soutien méthodologique aux études de recherche clinique. La recherche clinique s'est fortement structurée avec notamment la création de la Fédération de Recherche Clinique et celle de la Maison Régionale de la Recherche Clinique (MRRC), afin d'atteindre un niveau d'excellence débouchant sur des publications de haut niveau. Les biostatistiques représentent un facteur clé pour atteindre ces objectifs.

Dans le cadre de cette UF, sous la direction du Pr Duhamel, j'apporte un soutien statistique et épidémiologique aux cliniciens du CHU de Lille et aux chercheurs du CIC-IT. Je réalise cette activité avec l'aide de notre AHU et de nos internes de santé publique. Cette activité nous permet de former nos internes, d'améliorer le niveau des travaux de recherche, et de rencontrer les cliniciens. Avec le Dr Grégoire Ficheur, nous avons mis en place un cadre d'administration du flux de travail et des documents basé sur des systèmes de collaboration en ligne. En outre, nous avons développé un ensemble de fonctions programmées sous R, permettant l'accélération des tâches de faible valeur ajoutée (analyses univariées et bivariées) mais permettant également une amélioration de la qualité de rendu.

3.3 Mesure et amélioration de la qualité du dossier patient

En janvier 2016, l'UF « analyse du dossier patient » a été créée. J'en suis le responsable.

Nous avons transféré dans cette UF la gestion des études IPAQSS. Ces études permettent de mesurer quantitativement la qualité de l'information médicale des dossiers médicaux et contribuent à l'amélioration de la qualité de l'information médicale. Je participe à ce titre à la certification.

A travers cette UF, je porte le projet de création d'un entrepôt de données cliniques, qui est évoqué dans la section 2.2.1, page 96.

3.4 Le service, lieu d'émergence de médecins à fort potentiel

Les efforts d'encadrement de jeunes internes réalisés au quotidien par le Pr Beuscart, le Pr Duhamel, le Dr Ficheur et moi-même ont donné au service une solide réputation pour la formation des internes. Dans la subdivision lilloise, il y a plus de postes ouverts au choix que d'internes à affecter. Pour autant, tous les postes du service sont pris sans discontinuer depuis plus de 20 semestres d'affilée. De plus, il existe une file d'attente pour accéder aux postes du service.

L'intérêt de se préoccuper de la formation des internes dépasse notre seule mission de formation. Cela permet effectivement de détecter les internes à fort potentiel, et de les accompagner dans un programme individuel de formation (formation sur projets et masters 2, voire thèse d'université) et de réalisation de travaux de haut niveau pendant 3 à 4 ans. Certains de ces internes pourront devenir à leurs tours enseignants-chercheurs ou professionnels de haut niveau.

Afin de mieux préparer cette filière au long cours, nous profitons maintenant de la notoriété acquise lors des enseignements de première ou deuxième année des études médicales pour recruter des externes dans le service. Je reçois en entretien tout étudiant intéressé par la spécialité. Nous avons accueilli trois externes dans le cadre du stage en libre choix du dernier trimestre d'externat, post-ECN. Nous accueillerons désormais 2 externes par trimestre. Ces externes participent aux consultations de statistiques et réalisent des analyses descriptives univariées. Ils sont également conviés aux formations destinées aux internes de santé publique selon le calendrier des cours.