Détection automatique des infections du site opératoire

Marine Quéroué

Master Sciences, Technologies, Santé
Mention Santé Publique

Promotion 2018-2019

Parcours
Systèmes d'Information et Technologies Informatiques pour la Santé (SITIS)

DÉTECTION AUTOMATIQUE DES INFECTIONS DU SITE OPÉRATOIRE

Marine QUÉROUÉ

Soutenu publiquement le 10.09.2019

Mémoire réalisé dans le cadre d'une mission effectuée
du 11.02.2019 au 09.09.2019

Unité d'Informatique et d'Archivistique Médicales,
Service d'Information Médicale, Pôle de Santé Publique,
CHU de Bordeaux
Place Amélie Raba Léon - 33000 BORDEAUX

Maitres de stage
Frantz THIESSARD, MCU-PH
Sébastien COSSIN, AHU

Jury de Soutenance
Gayo DIALLO, tuteur universitaire
Jean Noel NIKIEMA, lecteur

Remerciements

Merci à Sébastien Cossin et Frantz Thiessard de m'avoir encadrée pour ce projet.

Merci à mon tuteur universitaire Gayo Diallo.

Merci à l'ensemble de l'équipe de l'unité IAM de m'avoir accueillie, formée et conseillée.

Merci à l'équipe d'Hygiène Hospitalière pour leur collaboration.

Merci au corps enseignant du M2 SITIS pour leur formation.

Merci à mes collègues de bureau de m'avoir soutenue et supportée.

Merci à mes camarades du M2 SITIS avec qui j'ai passé de bon moments, réalisé de super projets (petite dédicace au PinkPoo et au CRAC) et qui pour certain(e)s sont devenu(e)s des ami(e)s.

Merci à mes co-internes de Santé Publique (actuels et anciens, même celui parti vivre au fin fond de la Normandie).

Merci à mes amis.

Merci à ma famille.

Sommaire

TABLE DES FIGURES

TABLE DES TABLEAUX

Liste des abréviations

ATC	Classification anatomique, thérapeutique et chimique des médicaments
CCAM	Classification commune des actes médicaux
CCLIN	Centre de coordination de lutte contre les infections nosocomiales
CHU	Centre hospitalier universitaire
CIM10	Classification internationale des maladies, 10ème révision
CPIAS	Centre d'appui pour la prévention des infections associées aux soins
DPI	Dossier patient informatisé
ETL	Extract Transform Load
IA	Intelligence artificielle
IAS	Infection associée aux soins
IN	Infection nosocomiale
ISO	Infection du site opératoire
LR	Logistic regression
NLP	Natural Language Processing
OR	Odds-Ratio
PMSI	Programme de médicalisation des systèmes d'information
PROPIAS	Programme national d'actions de prévention des infections associées aux soins
RAISIN	Réseau d'alerte, d'investigation et de surveillance des infections nosocomiales
RF	Random forest
SIH	Système d'information hospitalier
SPICMI	Surveillance et prévention du risque infectieux liés aux actes de chirurgie et de médecine interventionnelle
SVM	Support vector machine
TAL	Traitement automatique des Langues
TF-IDF	Term frequency-inverse document frequency
UIAM	Unité d'informatique et d'archivistique médicales
VPN	Valeur prédictive négative
VPP	Valeur prédictive positive

1. INTRODUCTION

L'amélioration de la surveillance et de la prévention des infections du site opératoire (ISO) fait partie de l'axe 3 du programme national d'actions de prévention des infections associées au soins (PROPIAS). Il propose notamment de « Développer la recherche concernant l'analyse à partir des entrepôts de données pour la surveillance automatisée des infections post-opératoires d'actes ciblés ».(1)

L'intelligence artificielle (IA) réunit un ensemble de domaines de recherche dont l'objectif est de créer des processus cognitifs comparables à ceux de l'être humain et dont les algorithmes d'apprentissage automatique (ou *Machine Learning*) font partie. Elle est au coeur de la réflexion actuelle visant à l'amélioration de la qualité des soins au bénéfice du patient et à la réduction de leurs coûts.(2)

Ce travail de Master 2 s'inclue dans cette dynamique et a été réalisé au sein du Pôle de Santé Publique du Centre Hospitalier Universitaire (CHU) de Bordeaux. L'équipe de l'Unité d'Informatique et d'Archivistique Médicales (UIAM) du Service d'Information Médicale s'est associée au Service d'Hygiène Hospitalière avec pour objectif la réalisation d'un outil de détection automatique des ISO.

1.1 Les infections du site opératoire en France

1.1.1 DÉFINITION

Une ISO est une *infection associée aux soins* (IAS), c'est-à-dire une infection qui survient durant la prise en charge d'un patient (diagnostique, thérapeutique, palliative, préventive, éducative, opératoire) par un professionnel de santé. Classiquement, l'infection doit survenir après un délai minimum de 48h (délai d'incubation de l'agent infectieux en cause) après le début de la prise en charge pour être qualifiée d'IAS.

Une ISO peut également être qualifiée d'*infection nosocomiale* (IN), cas particulier d'IAS contractée dans un établissement de santé.

Une ISO est donc une infection qui survient à la suite d'une prise en charge opératoire. Les Centers for Disease Control and prevention ont établi les critères qui permettent de poser le diagnostic (annexe 1). On distingue ainsi différents types d'ISO selon leur localisation. Les infections superficielles de la plaie opératoire touchent la peau et le tissu sous-cutané. Les infections profondes de la plaie opératoire touchent les tissus mous profonds. Enfin, les infections d'organe ou de cavité se trouvent à proximité ou à distance du site opératoire mais sont en lien avec ce dernier.

La présence de pus, de signes inflammatoires locaux et la mise en évidence de micro-organismes sont des critères qui permettent le diagnostic d'ISO. Enfin, il existe un délai maximal de survenue de l'infection. Il est de 30 jours à compter de la date d'intervention mais se voit prolongé à 1 an en cas de présence de matériel prothétique.

1.1.2 ÉPIDÉMIOLOGIE

D'après l'enquête nationale de prévalence des infections nosocomiales, entre 2012 et 2017, la proportion d'ISO est passée de 13,5 % à 15,92 % des IN. Elles sont donc classées en deuxième position après les infections urinaires (28,47 %) et dépassent désormais les pneumonies (15,63 %).

Plus précisément, ce sont les ISO profondes et au niveau de l'organe qui ont vu leurs proportions augmentées passant respectivement de 4,8 % à 5,77 % et de 5,5 % à 7,74 % entre 2012 et 2017. La proportion d'ISO superficielles a quant à elle diminuée entre 2012 et 2017 passant de 3,2 % à 2,41 %.[(3)]

Concernant l'incidence des ISO, le rapport 2017 de surveillance des ISO dans les établissements de santé déclare des taux d'incidence par ordre décroissant de 3,99 % pour la chirurgie réparatrice et reconstructive, 3,44 % pour la chirurgie coronaire, 2,60 % pour l'urologie, 2,32 % pour la chirurgie vasculaire, 1,97 % pour la chirurgie digestive, 1,88 % pour la gynécologie-obstétrique, 1,72 % pour la chirurgie bariatrique, 1,37 % pour la chirurgie orthopédique, 1,32 % pour la chirurgie thoracique, 1,10 % pour la traumatologie et 0,79 % pour la neurochirurgie.[(4)]

Connaître le taux d'incidence des ISO est une nécessité afin de maîtriser le risque infectieux opératoire et de prévenir de leurs conséquences.

1.1.3 CONSÉQUENCES

Si la réduction de l'incidence des ISO fait partie des objectifs du PROPIAS, cela est notamment dû aux lourdes conséquences que ces IAS peuvent entraîner. En effet, si certaines ne représentent qu'un inconfort passager pour le patient, d'autres présentent un coût humain inacceptable. Les infections aiguës disséminées peuvent notamment conduire au décès du patient.

De la simple reprise opératoire aux séquelles physiques et infections chroniques parfois invalidantes, au-delà du coût humain, les conséquences des ISO représentent un véritable coût financier et social. Arrêt maladie et perte de revenu (voir d'emploi) pour le patient. Allongement des durées d'hospitalisation et des frais de prise en charge pour les établissements de santé.

1.1.4 SURVEILLANCE

La surveillance a pour objectif de faire baisser le taux d'ISO. En effet, aux États-Unis, les programmes de surveillance des IN ont montré que l'ISO était la première IN évitable et que leur mise en place avait permis une réduction du taux d'ISO dans les établissements participants.(5)

En France, depuis 1999 et jusqu'en 2018, la surveillance nationale des ISO était organisée par le réseau d'alerte, d'investigation et de surveillance des IN (RAISIN), regroupant les 5 centres de coordination de lutte contre les IN (CCLIN) ainsi que Santé Publique France (anciennement l'institut de veille sanitaire).

Le réseau ISO-RAISIN en fournissant un outil standardisé de surveillance permettait aux établissements de santé volontaires de comparer leur taux d'ISO de façon temporelle et inter-services, et ceci afin de mettre en place une politique de lutte contre les IN adéquate. Deux protocoles de surveillance était ainsi proposés :

La *surveillance des interventions prioritaires* (niveau patient) était adressée aux services souhaitant surveiller des chirurgies prioritaires (dont la liste variait selon les années) et souhaitant pouvoir comparer les résultats du service à des données standardisées sur les caractéristiques du patient, du séjour et de l'intervention. Les participants devaient inclure 100 interventions consécutives de la même spécialité sur le premier semestre de l'année avec 1 mois de surveillance post-opératoire (3 mois pour les interventions avec implant et les ostéosynthèses).(6)

La *surveillance agrégée globale* (niveau service) était une surveillance de première intention. Elle était adressée aux établissements réalisant des chirurgies non inclues dans la surveillance des interventions prioritaires, ainsi qu'aux services ayant un très faible taux d'incidence des ISO. Elle permettait de disposer de données internes mais ne permettait aucune comparaison entre établissements. Les inclusions se faisaient pendant au moins 2 mois durant le premier semestre avec 1 mois de surveillance postopératoire (3 mois pour les interventions avec implant et les ostéosynthèses).(7)

Ces deux surveillances s'avéraient extrêmement chronophage de part le recueil de données qu'elles exigeaient. En effet, la surveillance niveau patient nécessitait la création de fiches pour toutes les interventions incluses avec des données concernant le patient, le séjour, l'intervention et l'infection. La surveillance niveau service nécessitait quant à elle la création d'une fiche uniquement en cas d'ISO. Ces surveillances étaient effectuées par les équipes des Services d'Hygiène Hospitalière des établissements volontaires. Le retour au dossier patient, la consultation des chirurgiens et le remplissage manuel des fiches de surveillance ne laissait que peu de temps pour les activités de prévention en lien avec la problématique des ISO.

Suite au remplacement des CCLIN par les centres d'appui pour la prévention des IAS (CPIAS), la surveillance nationale des ISO a été repensée :

En novembre 2018, le CPIAS Ile-de-France a été nommé par Santé Publique France pour le pilotage de la mission nationale « Surveillance et prévention du risque infectieux liés aux actes de chirurgie et de médecine interventionnelle » (SPICMI). Cette mission a pour vocation le remplacement du réseau ISO-Raisin. Elle a notamment pour objectif de passer à un nouveau système de surveillance semi-automatisée reposant sur l'utilisation des données des systèmes d'information hospitalier (SIH) des établissements de santé. Cette surveillance représenterait un gain en termes de ressources nécessaires pour la collecte des données dans chaque établissement.(8)

Cette nouvelle surveillance n'est pas encore effective mais le CPIAS Ile-de-France a déjà communiqué sur ce nouvel outil expliquant que le calcul des taux d'incidence des ISO aurait pour dénominateur le nombre d'actes réalisés extrait du programme de médicalisation des systèmes d'information (PMSI) local et pour numérateur le nombre d'ISO confirmées. Ces dernières serait le résultat d'un détection automatisée des ISO suspectées par prise en compte et croisement de différentes données et confirmées ensuite par validation du chirurgien.(8)

Le CPIAS Ile-de-France a pour rôle de guider les établissements souhaitant prendre cette orientation grâce aux retours d'expérience de ceux qui l'ont déjà mis en place. Un état des lieux de l'automatisation de la surveillance des ISO en France a été réalisé début 2019. Il doit également définir à l'aide d'un groupe de travail les données à recueillir, les sources à interfacer et les algorithmes à créer pour la détection des ISO. Enfin, à terme une plateforme informatique sera mise à disposition et permettra aux établissements engagés dans la démarche d'importer leurs données et d'éditer le rapport correspondant.(8)

En attendant la mise en place de cet outil, il a été conseillé aux établissements de santé d'entamer une réflexion sur l'automatisation de la surveillance locale. C'est le cas du CHU de Bordeaux où les équipes de l'UIAM et du Service d'Hygiène Hospitalière se sont associées début 2019 pour la création de l'outil de détection automatique des ISO qui fait l'objet de ce mémoire.

1.2 Utilisation secondaire des données de santé

La détection automatique des ISO fait partie des cas d'usage de l'utilisation secondaire des données de santé. En effet, depuis l'informatisation des hôpitaux et la mise en place du PMSI, la prise en charge d'un patient génère un très grand volume de données médico-administratives. Ces données ont pour utilisation première le soin, la gestion administrative et la tarification. Encore très peu exploitées, elles représentent une mine d'or pour une utilisation secondaire dans les domaines de l'épidémiologie et de la recherche, mais également pour l'amélioration de la qualité et de la sécurité des soins.

Au sein des établissements de santé ces données sont contenues dans le SIH et leur exploitation n'est pas sans poser quelques difficultés.

1.2.1 LE SYSTÈME D'INFORMATION HOSPITALIER

Chaque établissement de santé possède un SIH composé d'un ensemble d'applications distinctes (ou logiciels métiers) qui communiquent entre elles (interopérabilité) selon les besoins. Il existe notamment des applications dédiées au dossier patient, à la prescription, à la biologie, à l'imagerie, à l'anatomopathologie, à la gestion administrative des mouvements et à la tarification de l'activité (PMSI) **(figure 1)**.

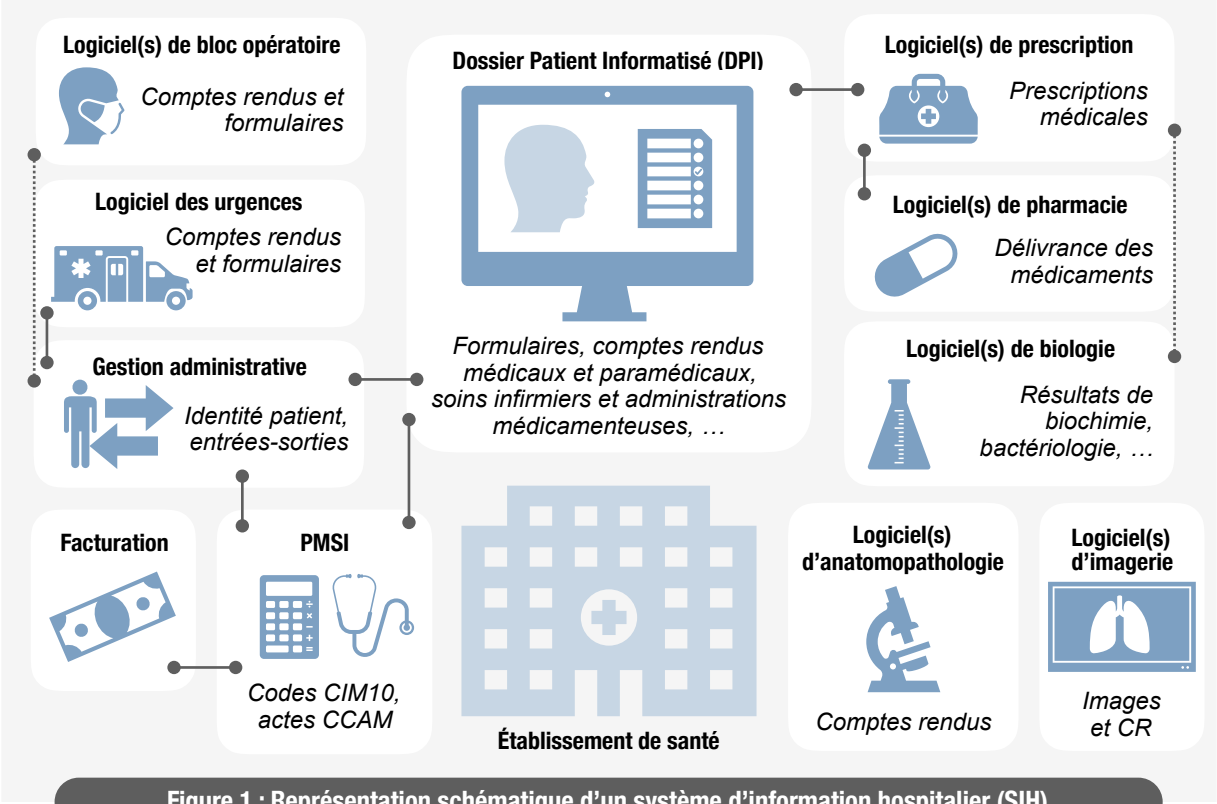

Figure 1 : Représentation schématique d'un système d'information hospitalier (SIH)

Le SIH présente donc l'avantage de disposer d'une grande richesse d'information mais dont les données sont très hétérogènes et dont l'exploitation est rendue difficile car il a été optimisé pour la production de soins et non pour une utilisation secondaire. Concernant l'hétérogénéité des données, on distingue classiquement 3 grandes catégories :

Les *données structurées et codées* sont les plus facilement exploitables. Elle font référence à une terminologie internationale, nationale ou parfois locale. Chaque concept dispose d'un code et d'un libellé qui identifient clairement l'information. Par exemple, au sein du PMSI, les maladies sont codées avec la Classification Internationale des Maladies, 10ème révision (CIM10)[(9)] et les actes médicaux sont codées avec la Classification Commune des Actes Médicaux (CCAM)[(10)]. Bien que simple d'utilisation, ces données présentent quelques limites. Un code de concept large pourra être utilisé de façon juste dans de nombreux cas mais sera peu informatif. Au contraire, un code de concept étroit sera très informatif mais peut-être utilisé à tort pour qualifier un concept voisin. Enfin, la qualité du codage est utilisateur dépendante et nécessite une certaine expertise.

Les *données semi-structurées* ne sont pas standardisées et sont établissements dépendantes voire services dépendantes. Au CHU de Bordeaux c'est via le logiciel DxCare que des formulaires sont créés pour alimenter le DPI. Chaque service dispose de formulaires personnalisés pour la collecte de données pertinentes. Ainsi, l'information est parfois redondante car codée de manière différente pour un même concept. De plus elle est complexe d'interprétation lors d'une utilisation secondaire car le caractère informatif se trouve bien souvent dans le libellé du champs à remplir/cocher, ce qui nécessite un traitement supplémentaire en terme d'extraction d'informations.

Les *données non structurées* (ou texte libre) correspondent le plus souvent à du texte utilisé pour décrire la prise en charge du patient comme par exemple les comptes rendus d'hospitalisation. L'information y est la plus riche sur le plan sémantique mais son exploitation est plus difficile. En effet, le plus souvent, il faut avoir recours aux méthodes de Traitement Automatique des Langues (TAL) pour en extraire de l'information *(cf. 1.3)*.

Au-delà de l'hétérogénéité des données, le SIH a une construction en silo. La séparation des données du fait des différents logiciels métiers qui le compose et la faible communication entre ces applications rendent difficiles les requêtes. En effet, si l'on souhaite interroger les données stockées de part et d'autres, l'étape de réconciliation des données augmente considérablement le temps de réponse. De plus il s'agit de requêtes complexes nécessitant l'intervention d'un expert du domaine.

Pour pallier aux problématiques rencontrées au sein du SIH et en vue d'une utilisation secondaire de ces données de santé, le CHU de Bordeaux a fait le choix de les intégrer dans un entrepôt de données biomédicales.

1.2.2 L'ENTREPÔT DE DONNÉES BIOMÉDICALES

Le principe d'un entrepôt de données biomédicales est d'extraire les données présentent au sein des différentes applications du SIH, de les transformer et de les charger dans une unique base de données appelée entrepôt. Ce principe est désigné par l'acronyme ETL pour Extract Transform Load **(figure 2)**.

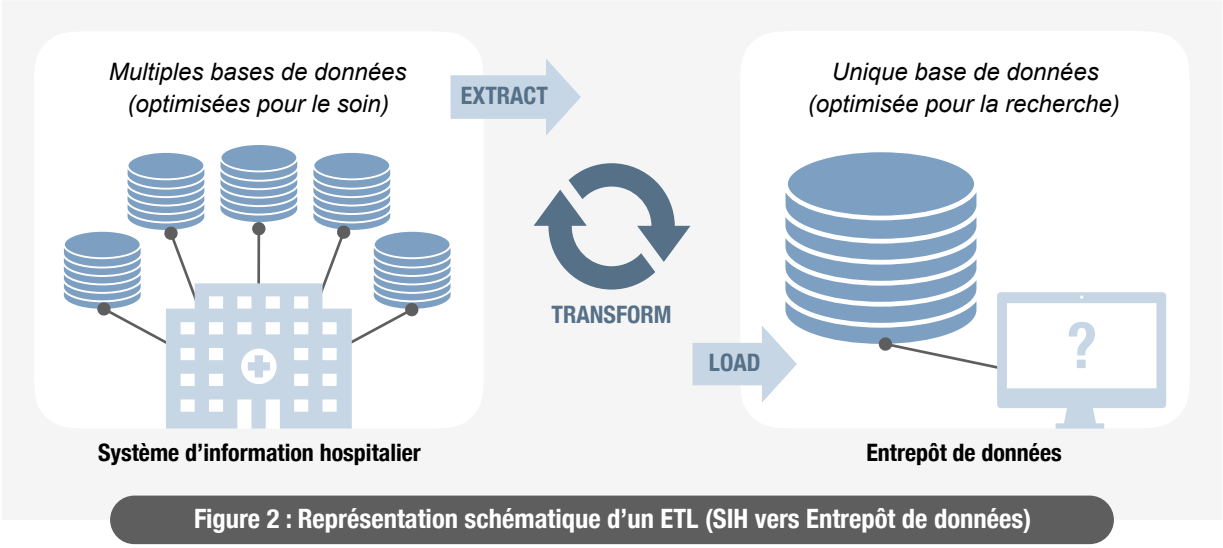

Figure 2 : Représentation schématique d'un ETL (SIH vers Entrepôt de données)

En novembre 2017, le CHU de Bordeaux a donc mis en oeuvre son entrepôt de données de santé basé sur la solution open source i2b2.[(11)] Développée à Harvard en 2004, ce modèle de données en étoile est optimisé pour la recherche de patients. Il comprend uniquement 6 tables permettant de couvrir l'ensemble des données utiles à une utilisation secondaire **(figure 3)**.

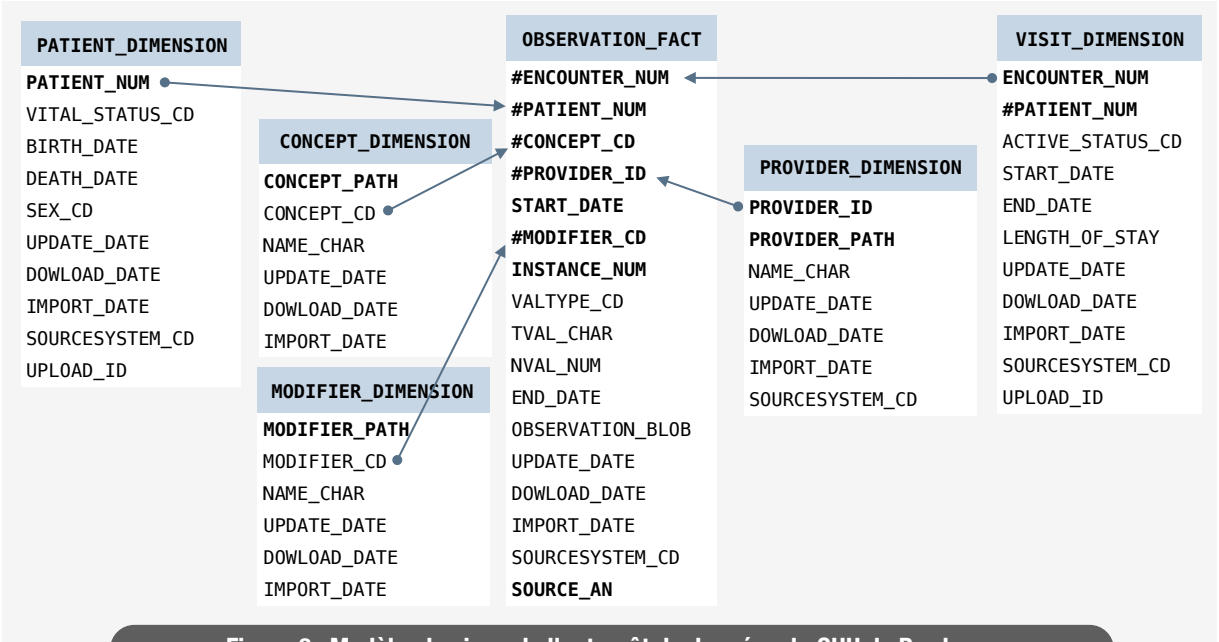

Figure 3 : Modèle physique de l'entrepôt de données du CHU de Bordeaux

La table **PATIENT_DIMENSION** contient les données socio-démographiques du patient (date de naissance, sexe, …).

La table **VISIT_DIMENSION** contient les données relatives aux venues du patient (date d'entrée, date de sortie, …).

La table **PROVIDER_DIMENSION** contient les informations relatives à l'ensemble des unités de prise en charge (numéro d'unité et son libellé).

La table **CONCEPT_DIMENSION** contient l'ensemble des concepts (identifiants et libellés) disponibles dans l'entrepôt. Il existe des concepts relatifs au PMSI avec les diagnostics CIM10 et les actes CCAM. Il existe également des concepts relatifs aux prescriptions et aux administrations médicamenteuses, ainsi que des concepts relatifs aux résultats de biologie et d'anatomopathologie. Enfin il existe des concepts spécifiques du CHU de Bordeaux correspondant aux formulaires et documents textuels disponibles.

La table **MODIFIER_DIMENSION** contient des concepts additionnels permettant d'apporter une information complémentaire aux premiers. On parle aussi de modificateurs.

La table **OBSERVATION_FACT** est la table centrale du modèle en étoile, elle reliée aux 5 autres. C'est ici que chaque élément unique de prise en charge est stocké. Par exemple un diagnostic CIM10 codé pour un patient x lors de sa venue y au sein d'une unité z avec l'information complémentaire qu'il s'agit d'un diagnostic principal. Cette table est de loin la plus volumineuse et peut contenir plusieurs milliards de faits.

Ce schéma i2b2 permet donc d'ajouter toute information qui serait jugée pertinente en vue d'une utilisation secondaire et ceci peu importe son format.

Inscrit au projet d'établissement 2016-2020, le projet d'entrepôt de données biomédicales piloté par le pôle de Santé Publique est toujours en cours en développement. C'est l'équipe de l'UIAM du Service d'Information Médicale qui en a en charge l'exploitation avec à terme un désir de mise à disposition d'outils pour l'ensemble des services du CHU de Bordeaux.

Plusieurs projets d'utilisation secondaire des données de santé sont actuellement en cours faisant notamment appel aux techniques de TAL et d'apprentissage automatique. La détection automatique des ISO en fait partie.

1.3 Traitement automatique des Langues

L'entrepôt de données de santé du CHU de Bordeaux contient des données non structurées sous forme de texte libre, notamment les comptes rendus d'hospitalisation, dont le traitement nécessite d'avoir recours à des méthodes de TAL ou *Natural Language Processing* (NLP) en anglais.

Le TAL désigne un ensemble de recherches et de développement dans les domaines de la linguistique, de l'informatique et de l'IA, afin comprendre et/ou de reproduire le langage naturel humain à l'aide d'une machine.

Le langage naturel présente deux difficultés majeures pour la machine. Il est ambigu car une entité linguistique peut faire l'objet d'une multitude d'interprétations (les homonymes en sont l'exemple le plus simple). D'autre part, il est implicite, c'est-à-dire que le contexte permettant l'interprétation n'est pas clairement énoncé mais fait appel à des connaissances du monde que la machine n'a pas.

La compréhension d'un énoncé par la machine nécessite donc la réalisation de plusieurs étapes successives [(12)] :

- Segmentation du texte en unités anatomiques (tokens).
- Traitement lexical : identification des composants des unités lexicales (mot) et de leurs propriétés.
- Traitement syntaxique : identification des constituants de plus haut niveau (groupe de mots) et des relations entre eux.
- Traitement sémantique : donner du sens à l'énoncé en associant à chaque concept évoqué un object ou une action du monde de référence.
- Traitement pragmatique : identification de la fonction de l'énoncé selon le contexte particulier dans lequel il a été produit.

Plus précisément, deux notions sont importantes à comprendre pour la suite du mémoire :

La *lemmatisation* est un traitement lexical. Certains mots prennent plusieurs formes selon qu'ils sont au singulier ou pluriel (pour les substantifs et les adjectifs), selon le genre (pour les adjectifs) et selon la conjugaison (pour les verbes). Le lemme d'un mot correspond à sa *forme canonique* c'est-à-dire son entrée lexicale commune (celle présente dans le dictionnaire de la langue). Ainsi, le lemme d'un substantif est son singulier (chevaux devient cheval), le lemme d'un adjectif est son masculin-singulier (petites devient petit) et le lemme d'un verbe est son infinitif (mangeons devient manger).

L'*étiquetage morpho-syntaxique* (ou part-of-speech tagging en anglais) consiste à associer à chaque mot une étiquette grammaticale (nom, adjectif, verbe …).

1.4 Apprentissage automatique (ou Machine Learning)

1.4.1 PRINCIPES GÉNÉRAUX

L'apprentissage automatique fait partie des champs d'études de l'IA. En effet, fondé sur des approches statistiques, il permet à la machine de s'entraîner et d'apprendre à partir d'un jeu de données pour ensuite effectuer une tâche sans avoir été explicitement programmée pour.

Avant de pouvoir être utilisé, un modèle construit via Machine Learning doit comme son nom l'indique passer par une phase d'apprentissage ou d'entraînement. Cette étape nécessite de disposer de données qui serviront de jeu d'apprentissage. Si les données sont étiquetées, c'est-à-dire si les données sont annotées de telle manière que la réponse à la tâche demandée est connue, alors on parle d'apprentissage supervisé. On demande ici au modèle d'effectuer une tâche de prédiction. Plus précisément, si la variable à prédire est qualitative on parle de tâche de classification. Lorsque la variable à prédire est quantitative on parle de tâche de régression. Enfin, lorsque la réponse à la tâche n'est pas connue et que le jeu d'apprentissage n'est pas étiqueté on parle d'apprentissage non supervisé.

Concernant la détection automatique des ISO, il s'agit d'une tâche de classification automatique par apprentissage supervisé **(figure 4)**. Dans le jeu de données d'apprentissage, chaque intervention chirurgicale est annotée pour la variable qualitative « ISO » avec 0 désignant l'absence d'ISO et 1 désignant la présence d'une ISO. C'est la variable Y à prédire. De plus chaque intervention possède également un ensemble de variables X_i dites prédictives qui contiennent les informations permettant à l'algorithme de faire un choix. Le modèle apprend grâce à ces données dont la réponse à la question est déjà connue. Il doit à terme être capable de classer de nouvelles données pour lesquelles la réponse à la question *« présence d'une ISO ? »* n'est pas connue.

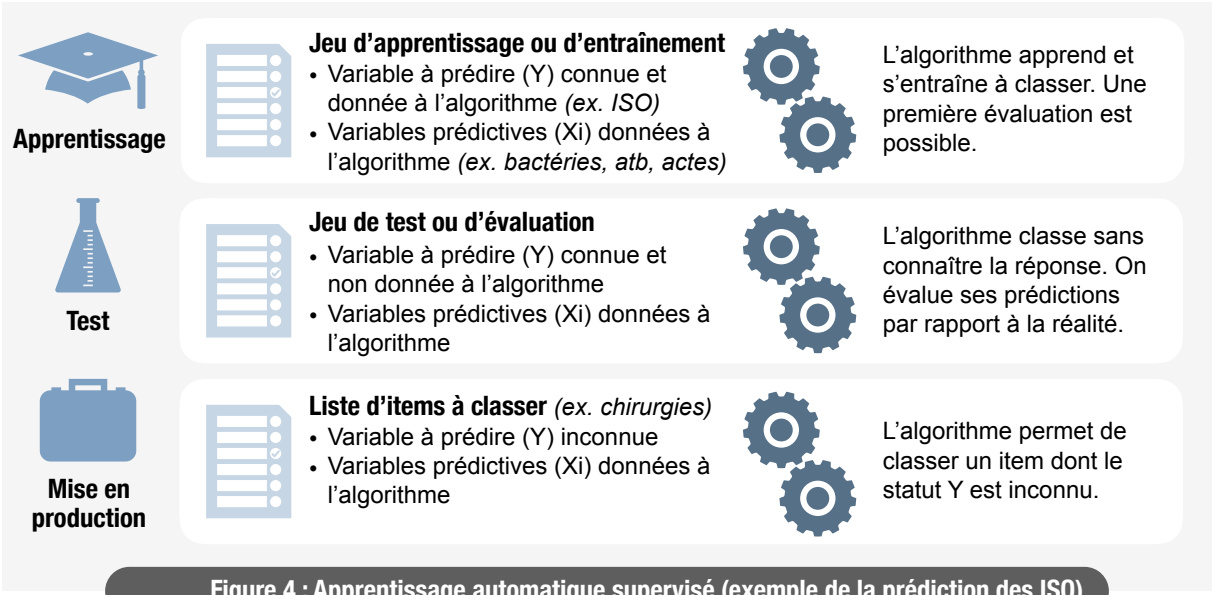

Figure 4 : Apprentissage automatique supervisé (exemple de la prédiction des ISO)

1.4.2 LES ALGORITHMES D'APPRENTISSAGE AUTOMATIQUE

L'apprentissage machine repose sur des méthodes statistiques. Parmi les algorithmes les plus utilisés on peut cité :

- la régression logistique (LR pour *Logistic Regression*)
- les arbres de décisions dont dérives les forêts d'arbres décisionnels (RF pour *Random Forest*)
- les machines à vecteurs de support (SVM pour *Support Vector Machine*)
- le boosting (ensemble de méthodes permettant d'augmenter les performances d'un classifieur faible par itérations)
- les réseaux de neurones, il s'agit alors d'un apprentissage profond ou Deep Learning, sous-domaine du Machine Learning

Concernant la détection automatique des ISO, il a été choisi d'utiliser des algorithmes de LR et de RF. Ces derniers sont détaillés ci-dessous.

La LR est un modèle de régression binaire, c'est-à-dire que la variable Y à prédire (ou variable expliquée) ne peut prendre que deux modalités (0 ou 1). Les variables X_i prédictives (ou explicatives) peuvent quant à elles être continues ou binaires. On fournit ainsi un vecteur de variables explicatives (X_1, X_2, …, X_n) pour lesquelles le modèle estime des coefficients de régression (b_i) et renvoie la probabilité P(Y) entre 0 et 1 en maximisant une fonction de vraisemblance. Classiquement, au-delà du seuil de 0,5 la variable à prédire Y est probable.

Régression logistique

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}}$$

L'algorithme de RF combine plusieurs arbre de décisions. Un arbre de décision dispose de plusieurs branches au bout desquelles les décisions finales possibles sont situées (ce sont les feuilles de l'arbre) et sont atteintes en fonction des décisions prises à chaque étape. Le nombre d'arbres est un paramètre important qui varie selon la question posée. Ces derniers sont construit aléatoirement et s'entraînent sur un sous-ensemble du jeu de données d'apprentissage (on parle de bagging) et sur un sous-ensemble de variables prédictives X_i (on parle de projections aléatoires). Les prédictions sont ensuite moyennées lorsque les données sont quantitatives ou utilisés pour un vote pour des données qualitatives.

1.4.3 ÉVALUATION DES PERFORMANCES

Les performances d'un algorithme de classification automatique peuvent être évaluées à partir d'une matrice de confusion **(figure 5)**.

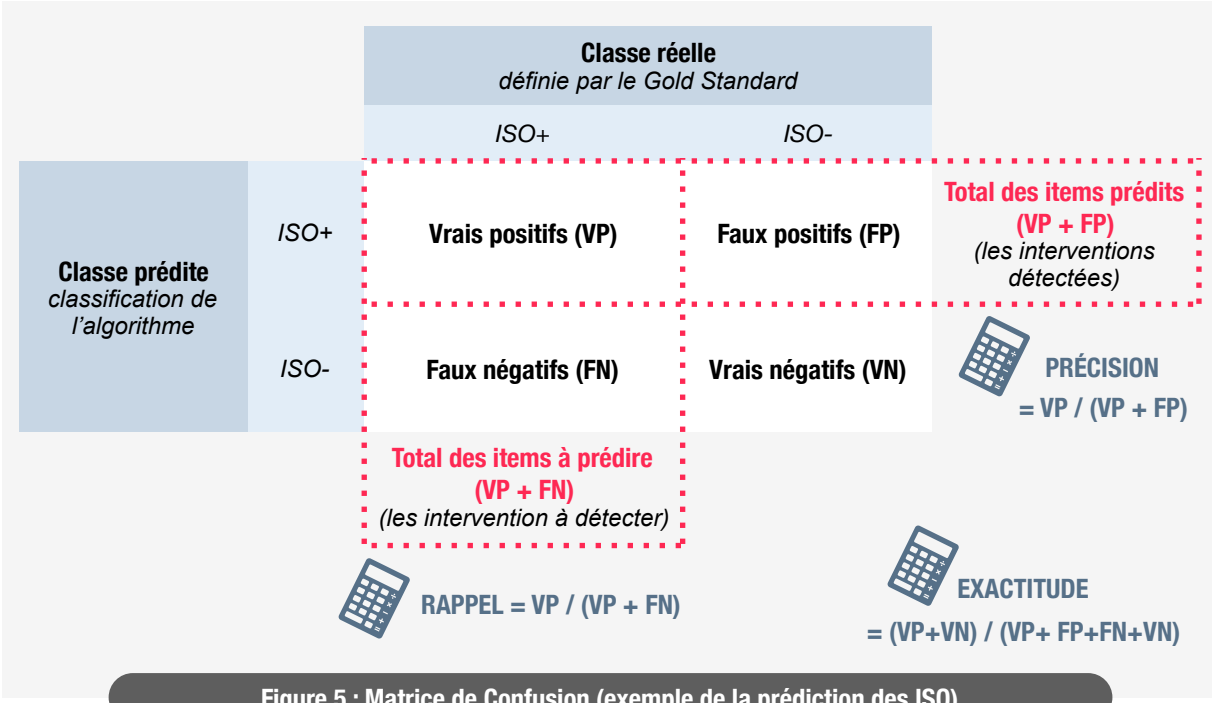

Figure 5 : Matrice de Confusion (exemple de la prédiction des ISO)

Le **RAPPEL** (ou sensibilité) correspond à la proportion d'ISO prédites parmi l'ensemble des ISO réelles. Un rappel à 100% signifie que l'algorithme a détecté *toutes* les interventions qu'il devait détecter. Il n'y a pas de faux négatifs.

La **PRÉCISION** (ou valeur prédictive positive (VPP)) correspond à la proportion d'ISO réelles parmi l'ensemble des ISO prédites par l'algorithme. Une précision à 100% signifie que l'algorithme n'a détecté *que* les interventions qu'il devait détecter. Il n'y pas de faux positifs.

L'**EXACTITUDE** correspond à la proportion de prédictions correctes sur l'ensemble des items. Une exactitude à 100% signifie que l'algorithme n'a fait aucune erreur.

La spécificité et la valeur prédictive négative (VPN) sont moins utilisées dans ce cadre et s'intéressent aux items à ne pas détecter (classe réelle ISO-). La spécificité correspond à la proportion d'ISO- correctement classées (VN) parmi les ISO- (VN+FP). La VPN correspond à la proportion d'ISO- (VN) parmi les items non détectés (VN+FN).

Ces paramètres peuvent être calculés suite à la prédiction sur jeu d'apprentissage mais ces derniers seront biaisés avec un risque de surestimation des réelles performances de l'algorithme. Pour ce faire l'idéal est de disposer d'un jeu de données de test dont la variable à prédire est également connue.

1.5 État de l'art

Plusieurs établissements de santé ont déjà étudié la possibilité d'automatiser la surveillance des ISO et plus généralement des IAS au sein de leur structure. Diverses sources de données peuvent être utilisées comme le PMSI, les données de biologie ou les documents textuels, seules ou de façon conjointe. Certains établissements ont eu recours au TAL et/ou à la classification par apprentissage supervisé.

Entre 2010 et 2011, le CHU de Nantes a comparé l'utilisation des données de PMSI seules à celles des données de bactériologie seules ainsi qu'à la combinaison des deux. Les résultats ont montrés que la combinaison de deux sources d'informations augmentait le rappel qui atteignait les 90% sur 4400 interventions étudiées. La précision demeurait néanmoins faible (21 à 25%). Selon eux, une surveillance assistée par ordinateur peut être mise en œuvre dans les hôpitaux français en utilisant les sources de données disponibles. Le gain de temps permettant aux professionnels de la prévention des IAS de consacrer plus de temps aux tâches de prévention et d'éducation. Ils soulignent cependant la nécessité d'une étude multicentrique pour évaluer la transposabilité de cette méthode. (13)

Entre 2009 et 2011, un outil sémantique d'analyse des documents textuels médicaux pour la détection d'IAS est développé (ALADIN). 1607 documents provenant de 4 hôpitaux universitaires ont été analysés et annotés par des expert en IAS. Les éléments pertinents ont été codés selon des terminologies sélectionnées par des spécialistes. Un dictionnaire des concepts relatifs aux IAS (environ 4000 termes) a ainsi été créé et intégré directement dans des outils de TAL. À partir de ces éléments, les linguistes ont construit un premier ensemble d'heuristiques visant à repérer les cas d'IAS. Les performances se sont révélées intéressantes avec une sensibilité de 88% et une spécificité de 97% toutes chirurgies confondues. (14)

En 2013, le CHU de Rennes décide d'utiliser les documents textuels présents dans son entrepôt de données pour détecter automatiquement les ISO suite à une neurochirurgie. Ils utilisent NOMINDEX un outil de TAL qui extrait les concepts MeSH du texte libre. Les concepts pertinents sont sélectionnés à la fois selon leur fréquence au sein d'un groupe et à la fois selon les relations sémantiques qu'ils entretiennent avec les concepts du groupe opposé. Par exemple un concept du groupe ISO+ dont le concept fils se trouve dans le

groupe ISO- ne sera pas retenu. Cette approche a été comparé au système de surveillance classique ainsi qu'à l'utilisation des codes diagnostiques CIM10. Elle est la plus performantes avec un rappel de 92% et une précision de 40%. (15)

En dehors de la France des recherches sont également menées dans ce sens. L'Université de Stockholm a étudié l'utilisation de techniques d'apprentissage automatique pour la détection des infections associées aux soins. Les chercheurs ont utilisé des données hospitalières recueillies lors d'une enquête de prévalence ponctuelle. Elles comprenaient des données textuelles, des codes CIM10, des administrations médicamenteuses, des résultats microbiologiques et la température corporelle. Les algorithmes d'apprentissage séparateurs à vaste marge et gradient tree boosting se sont révélés performants pour cette tâche avec un excellent rappel. (16)

Dans le cadre du projet national d'amélioration de la qualité chirurgicale (NSQIP) aux États-Unis, l'Université du Minnesota a également mis en place un outil automatisé de détection des événements indésirables liés à une intervention chirurgicale. Ces derniers ont utilisé les données cliniques de leur centre médical ainsi que celles du registre NSQIP. Parmi celles-ci, ils ont sélectionné des données démographiques (sexe, âge, race) et des données cliniques (diagnostics CIM-9, résultats biologiques, administrations médicamenteuses, demandes d'examens complémentaires et constantes). Les modèles utilisant la régression logistique ont montré les meilleures performances éliminant de manière fiable la grande majorité des patients sans ISO et réduisant ainsi de manière significative la charge des registres. (17)

Enfin, l'Université de l'Utah a quant à elle développé un système de TAL pour identifier automatiquement les mentions d'ISO dans les comptes rendus de radiologie. Ils ont travaillé sur la chirurgie gastro-intestinale à partir de la base de données MIMICIII Critical Care dont ils ont extrait les codes diagnostiques, les codes d'acte et les comptes rendus de scanner dans les 30 jours suivant la procédure. Ces derniers ont été annotés par deux chirurgiens afin de créer un lexique de termes. Ils ont ensuite développé un système de TAL afin d'identifier et classer automatiquement les preuves d'ISO à partir de chaque compte rendu en s'appuyant sur une adaptation de l'algorithme ConText qui gère la négation et la temporalité. Leur système de TAL s'est montré plus performant que les deux autres approches testées utilisant des données administratives uniquement ou des techniques d'apprentissage automatique SVM avec une représentation du texte en n-grammes. (18)

1.6 Détection automatique des ISO

1.6.1 PRÉSENTATION DU PROJET

L'équipe de l'UIAM du Service d'Information Médicale et le Service d'Hygiène Hospitalière du CHU de Bordeaux ont collaboré afin d'élaborer un outil de détection automatique des ISO.

Ce projet est né d'un constat, la surveillance des ISO au sein du CHU de Bordeaux se révèle extrêmement chronophage pour les équipes d'hygiène avec un retour au dossier et la création manuelle de fiches de surveillance pour les chirurgies ciblées. Cette lourde activité de recueil laisse donc peu de temps pour les activités de prévention dans les services.

L'équipe de l'UIAM dispose quant à elle de toutes les informations nécessaires à cette surveillance stockées au sein de l'entrepôt de données dont elle a en charge l'exploitation. Il a donc été décidé de mettre en commun les compétences de ces deux services afin d'optimiser la surveillance et la prévention des ISO au sein du CHU de Bordeaux avec pour objectif final l'amélioration de la qualité des soins.

1.6.2 MISSIONS ET OBJECTIF DU STAGE

Interne de Santé Publique effectuant mon stage de master 2 au sein de l'UIAM, j'avais pour objectif principal d'utiliser mes compétences médicales, épidémiologiques, informatiques et statistiques afin d'élaborer des algorithmes permettant la détection des ISO à partir des informations disponibles dans l'entrepôt de données.

Mes objectifs secondaires ont été l'aide à l'intégration des données de bactériologie au sein de l'entrepôt ainsi que l'amélioration de la qualité des données déjà présentes.

Enfin, en tant qu'interne du service j'ai également eu pour mission de contribuer au projet d'Automatisation du Recueil d'Indicateurs de Pertinence de Prescription d'Antibiotiques (ARIPPA) porté par le Dr Frantz Thiessard. J'ai notamment effectué plusieurs extractions concernant le codage de l'antibioprophylaxie au sein de l'entrepôt de données, ainsi que la réalisation d'une base de donnée et son interface pour le recueil des protocoles et recommandations concernant les antibiothérapies de plus de 7 jours.

2. MÉTHODE

L'élaboration d'un outil de détection automatique des ISO à partir des informations disponibles dans l'entrepôt de données biomédicales du CHU de Bordeaux fait appel aux technologies de l'IA. Plus précisément, il s'agit d'un apprentissage automatique supervisé afin d'effectuer la tâche de classification des interventions en suspectes ou non suspectes d'ISO.

Nous avons réalisé deux approches différentes pour répondre à cet objectif **(figure 6)** :

- Une approche utilisant l'ensemble des données disponibles pour l'apprentissage et que l'on nommera *méthode D* (pour données).
- Une approche utilisant uniquement les données de texte libre pour l'apprentissage et que l'on nommera *méthode T* (pour texte).

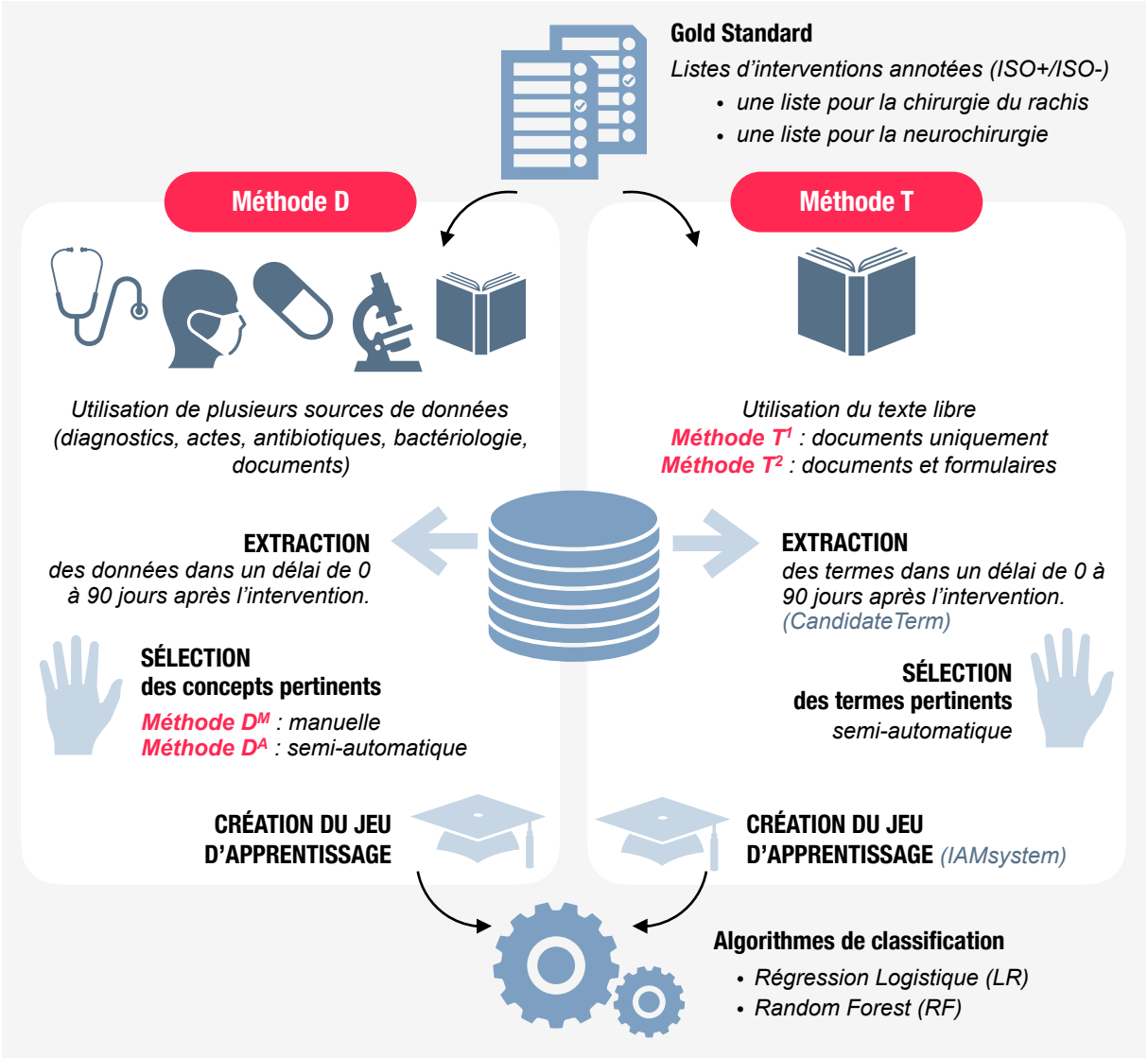

Figure 6 : Représentation schématique des approches méthodiques D & T

2.1 Gold Standard

Afin d'effectuer un apprentissage automatique supervisé nous devons disposer d'un jeu de données annotées c'est-à-dire d'une liste d'interventions pour lesquelles le statut concernant la présence d'ISO est connu. Cette liste de référence nous servira de Gold Standard pour évaluer les performances des algorithmes par la suite.

Le Service d'Hygiène Hospitalière réalise chaque année des enquêtes de surveillance des ISO concernant des chirurgies ciblées. C'est le cas de la chirurgie orthopédique du rachis qui a fait l'objet d'une surveillance en 2015, 2016 et 2017 sur une période de 3 mois consécutifs par an. La liste de l'ensemble des interventions sur ces périodes a été extraite par l'Unité de Coordination et d'Analyse de l'Information Médicale. Un retour auprès du chirurgien a été réalisé pour chacune de ces interventions, ce dernier ayant pour rôle d'annoter la présence ou l'absence d'ISO pour chacune d'entre elles. La création de fiches de surveillance a par la suite été effectuée par l'équipe d'Hygiène Hospitalière avec un retour au dossier patient pour les cas d'ISO déclarées. Ce sont ces fiches de surveillances annotées pour le statut vis-à-vis de la présence d'ISO qui constitueront notre gold standard et notre base pour la création d'un jeu de données d'apprentissage.

Nous disposions ainsi d'une liste de 2133 interventions concernant la chirurgie orthopédique du rachis et pour lesquelles 22 ISO avaient été déclarées. Plus précisément, 13 ISO ont été déclarées pour 662 interventions en 2015, 7 ISO déclarées pour 708 interventions en 2016 et 2 ISO déclarées pour 763 interventions en 2017 **(tableau 1)**.

Tableau 1 - Gold Standard de Chirurgie orthopédique du Rachis

	2015	2016	2017	Total
INTERVENTIONS	662	708	763	2133
ISO DÉCLARÉES	13	7	2	22

Dans un second temps, nous nous sommes intéressés à la neurochirurgie faisant également l'objet d'une surveillance entre 2010 et 2017. Nous avons ainsi constitué notre deuxième jeu de données d'apprentissage à partir des fiches de surveillances des années 2012 à 2017 disposant au total d'une liste de 2303 interventions de neurochirurgie avec 20 ISO déclarées **(tableau 2)**.

Tableau 2 - Gold Standard de Neurochirurgie

	2012	2013	2014	2015	2016	2017	Total
INTERVENTIONS	417	425	371	222	445	423	2303
ISO DÉCLARÉES	6	5	2	1	5	1	20

2.2 Méthode D

La première approche réalisée et testée est la *méthode D*, c'est-à-dire l'utilisation de l'ensemble des sources données présentes dans l'entrepôt pour la création de notre jeu d'apprentissage.

2.2.1 IDENTIFICATION DES PATIENTS

Une première étape d'identification des patients à partir des fiches de surveillance transmises a été nécessaire avant de procéder à l'extraction des données. En effet, il s'agit de fiches pseudonymisées c'est-à-dire avec uniquement la date de naissance, le sexe ainsi que les dates de séjours et d'intervention du patient.

Premièrement, il s'agit de récupérer le numéro d'identification personnel (NIP) du patient. Dans la majorité des cas, l'hygiène disposait d'un fichier annexe qu'il a suffit de croiser avec la fiche de surveillance correspondante. D'autres fichiers annexes ne disposaient pas du NIP mais des NOMS et PRÉNOM du patient nécessitant d'aller interroger les bases de données du CHU de Bordeaux afin de récupérer le NIP.

Secondairement et à partir du NIP, il s'agit de récupérer le PATIENT_NUM (identifiant patient spécifique de l'entrepôt de données) ainsi que le PATID (identifiant patient de la base de données du logiciel de bactériologie, en cours d'intégration dans l'entrepôt).

Certaines interventions n'ont pas été gardées dans le gold standard et n'ont pu être incluses dans les jeux de données d'apprentissage faute d'identifiant disponible.

2.2.2 INTERROGATION DE L'ENTREPÔT DE DONNÉES

Une fois récupérés les identifiants des patients, une extraction de leurs données disponibles au sein de l'entrepôt de données biomédicales est possible. La table interrogée est la table OBSERVATION_FACT. Pour rappel, elle contient tous les éléments de prise en charge dont ont bénéficié les patients au sein du CHU de Bordeaux. Elle contient notamment les diagnostics CIM10 codés, les actes CCAM réalisés, les prescriptions et administrations médicamenteuses, les résultats de biologie, les formulaires remplis dans les services et les documents textuels médicaux et paramédicaux tel que les comptes rendus d'hospitalisation.

Une notion importante à prendre en compte dans la survenue d'une ISO est son délai maximum d'apparition après la date d'intervention. Pour rappel, il est de 30 jours et se voit prolongé à 1 an en cas de présence de matériel prothétique. Concernant la chirurgie orthopédique du rachis, le délai de surveillance préconisé par les protocoles est de 90 jours après l'intervention.

Pour chaque patient du gold standard ayant subi une intervention de chirurgie du rachis à une date donnée, il a était extrait l'ensemble des éléments de prise en charge dans un délai allant de 0 à 90 jours après la date d'intervention. Concernant la neurochirurgie, le délai maximum d'apparition d'une ISO est de 30 jours, néanmoins, après retour au dossier patient, il existe généralement un retard au diagnostic et l'information apparaît tardivement dans le DPI. Il a donc été décidé de conserver le délai de 90 jours de surveillance pour cette spécialité chirurgicale.

Concernant la *méthode D* il a donc été extrait dans un délai de 90 jours **(figure 7)**:

- les diagnostics CIM10
- les actes CCAM
- les administrations médicamenteuses
- les documents textuels médicaux et paramédicaux

Les données de bactériologie ont fait l'objet d'une extraction propre (détaillée ci-après).

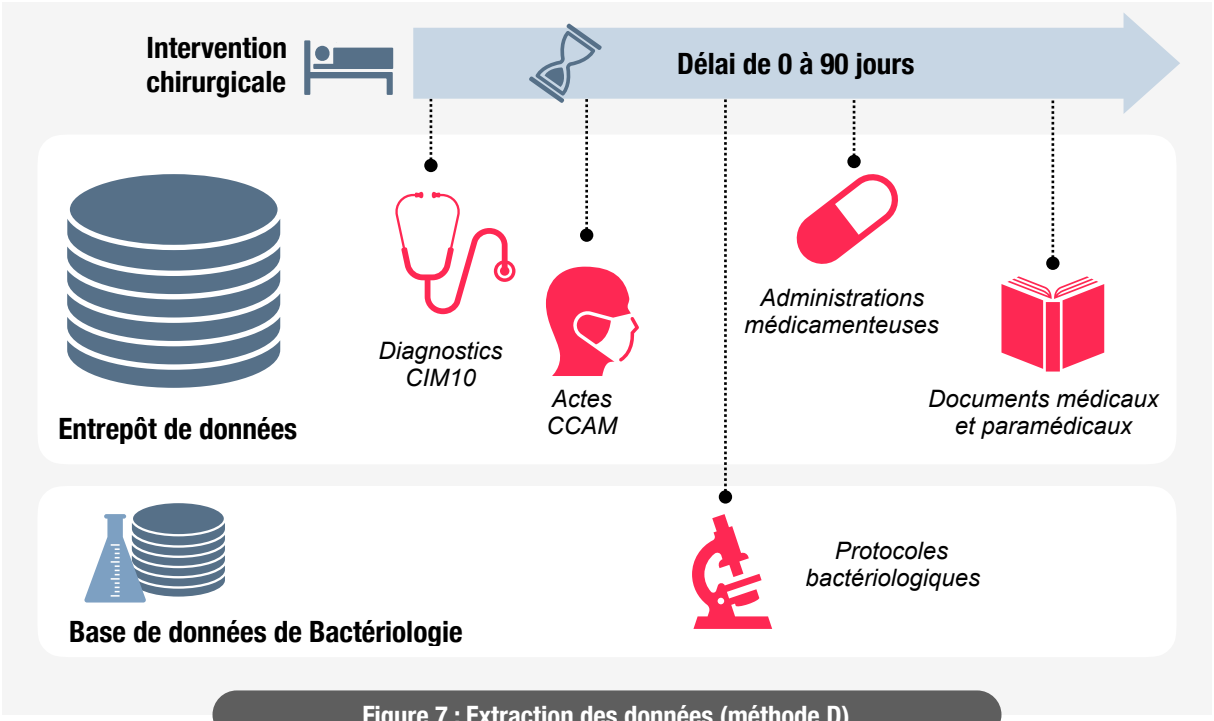

Figure 7 : Extraction des données (méthode D)

2.2.3 INTÉGRATION DES DONNÉES DE BACTÉRIOLOGIE

Actuellement, les données de bactériologie ne sont pas encore intégrées à l'entrepôt de données de santé i2b2. Leur intégration est notamment rendu difficile par le changement de logiciel métier de biologie médicale en cours durant cette année 2019. C'est donc un double travail d'intégration qui est actuellement réalisé par les équipes d'UIAM.

Les données bactériologie n'en demeurent pas moins importante pour ce projet de détection automatiques des ISO. Aussi, de manière conjointe à la réflexion menée concernant leur intégration, une compréhension de la structure de la base de données du logiciel métier en charge de la bactériologie a été nécessaire. Un concept intéressant a ainsi été mis en évidence, celui de protocole de bactériologie. En effet, suivant la demande (ou requête) effectuée par le clinicien au laboratoire, des protocoles sont mis en place et sont notamment spécifique de la source et de la nature du prélèvement. Ces derniers sont donc très informatifs sur le questionnement clinique à leur origine et sur le type d'infection en cours s'ils reviennent positifs.

Il existe par exemple un protocole nommé BISO pour examen bactériologique d'une plaie opératoire.

En attendant l'intégration de ces données de bactériologie, l'extraction des informations pour chaque patient s'est faite directement à partir de la base de données du logiciel métier via le PATID (identifiant patient). Un délai de 0 à 90 jours après la date d'intervention a également été appliqué **(figure X)**.

2.2.4 SÉLECTION DES VARIABLES D'INTÉRÊT

Nous avons donc extrait pour chaque patient un ensemble de concepts de l'entrepôt de données de santé ainsi qu'un ensemble de protocoles bactériologiques. Parmi eux certains semblent plus pertinents que d'autres pour la détection automatique d'une ISO.

En effet, comme vu précédemment, les algorithmes d'apprentissage automatique, pour prédire la variable Y, ont besoin qu'on leur fournisse un ensemble de variables prédictives X_i. Concernant la détection automatique des ISO, la variable a prédire est donc la présence ou l'absence d'ISO. Parmi l'ensemble des informations extraites certaines ont été sélectionnées pour devenir des variables prédictives X_i, ce sont nos variables d'intérêts.

Dans un premier temps, la sélection des variables d'intérêts c'est faite de façon manuelle. Pour simplifier nous parlerons de *méthode* D^M. Cette dernière n'a été réalisée que pour la chirurgie du rachis. Dans un second temps et dans un souci d'optimisation de la *méthode D*, nous avons partiellement automatisé la sélection des variables. Nous parlerons de *méthode* D^A. Cette méthode a été réalisée pour la chirurgie du rachis et pour la neurochirurgie.

2.2.4.1 Sélection manuelle (méthode D^M)

La sélection manuelle des variables d'intérêts pour nos algorithmes de classification fait appel à une expertise médicale pluridisciplinaire et à une revue de la littérature. Elle présente l'inconvénient d'être chronophage et peu reproductible. La chirurgie orthopédique du rachis est la première spécialité chirurgicale à laquelle nous nous sommes intéressés.

Concernant les diagnostics CIM10, c'est-à-dire les diagnostics codés à la fin du séjour du patient dans un objectif primaire de tarification de l'activité et secondaire d'épidémiologie, une liste a été retenue après revue de la littérature. Il s'agit de code CIM10 relatifs à la survenue d'une ISO suite à une chirurgie orthopédique quelle qu'elle soit **(tableau 3)**. Nous avons créé la variable **DIAG** codée 0 ou 1 (1 signifiant la présence d'au moins un de ces diagnostics).

Tableau 3 - Diagnostics CIM10 sélectionnés pour la chirurgie du rachis (méthode D^M)

T81.4	**Infection après un acte à visée diagnostique et thérapeutique, non classée ailleurs**
T84.5	**Infection et réaction inflammatoire dues à une prothèse articulaire interne**
T84.6	**Infection et réaction inflammatoire dues à un appareil de fixation interne [toute localisation]**
T84.7	**Infection et réaction inflammatoire dues à d'autres prothèses, implants et greffes orthopédiques internes**

Concernant les actes CCAM, un acte de reprise couramment utilisé lors de la présence d'une ISO nous a été conseillé par le Pr Jean-Marc Vital, chef du Service de Chirurgie Orthopédique et Traumatologique du CHU de Bordeaux et participant à la surveillance des ISO. Il s'agit du code AFPA001 de *Mise à plat de lésion infectieuse péridurale rachidienne et/ou paravertébrale postopératoire [sepsis], par abord direct*. Nous avons créé la variable **ACTE** codée 0 ou 1 (1 signifiant la présence de l'acte).

Concernant les prescriptions et administrations médicamenteuses, il a été décidé ne pas inclure les prescriptions souvent réalisées de façon anticipée par le clinicien et n'apportant aucune information fiable sur leur délivrance. Nous avons donc uniquement utilisé les administrations médicamenteuses et sélectionné les antibiotiques des classes J01 et J04 de la classification Anatomique, Thérapeutique et Chimique (ATC). Nous avons crée la variable **ATB** codée 0 ou 1 (1 signifiant l'administration d'au moins un antibiotique).

Concernant les protocoles de bactériologie, deux variables ont été créées. La variable **BAC** codée 0 ou 1 et pour laquelle 1 signifie la présence d'au moins un protocole revenu positif. La variable **PROT** codée 0 ou 1 et pour laquelle 1 signifie la présence d'au moins un protocole spécifique revenu positif. Par protocoles spécifiques **(tableau 4)** on entend protocoles sélectionnés manuellement parmi la liste des protocoles les plus fréquents chez les porteurs d'ISO de notre gold standard et parmi ceux qui semblaient médicalement pertinents.

Tableau 4 - Protocoles bactériologiques sélectionnés pour la chirurgie du rachis (méthode D^M)

BIOA	EXAMEN BACTERIOLOGIQUE D'UNE BIOPSIE AUTRE
BIOS	EXAMEN BACTERIOLOGIQUE D'UNE BIOPSIE OSTEO-ARTICULAIRE
BISO	EXAMEN BACTERIOLOGIQUE D'UNE PLAIE OPERATOIRE
BLQLA	EXAMEN BACTERIOLOGIQUE DE LIQUIDE DE LAME, REDON
BMAO	EXAMEN BACTERIOLOGIQUE D'UN MATERIEL ORTHOPEDIQUE
BPRO	EXAMEN BACTERIOLOGIQUE D'UN PUS PROFOND

Concernant les documents textuels, une liste de 12 termes médicalement pertinents a été réalisée manuellement parmi les termes dont les fréquences d'apparition chez les ISO+ étaient supérieures aux fréquences d'apparition chez les ISO- du Gold Standard. Chaque terme constituait une variable à part entière avec son nombre occurence. Ces termes étaient recherchés dans les documents sans faire appel aux méthodes de TAL telles que la lemmatisation. La liste des termes recherchés était la suivant : évènement, reprise, dommage, antibiothérapie, prélèvement, préjudice, cicatrice, écoulement, accident, sepsis, infection, effet indésirable.

Les formulaires et les résultats de laboratoire n'ont pas été utilisés.

2.2.4.2 Sélection semi-automatique (méthode D^A)

Cette méthode de sélection semi-automatique n'est autre qu'une automatisation de certaines étapes de la précédente. Elle se veut moins chronophage et plus reproductible.

L'automatisation concerne la sélection des diagnostics CIM10, des actes CCAM et des protocoles bactériologiques. Les administrations médicamenteuses restent celles des antibiotiques des classes J01 et J04 de la classification ATC. Les formulaires et les résultats de laboratoire ne sont toujours pas utilisés. Enfin, concernant les documents textuels nous verrons par la suite avec la *méthode T* les principes de la sélection semi-automatique des termes d'intérêts.

Nous nous sommes inspirés du TF-IDF (pour *Term Frequency-Inverse Document Frequency*) qui est une méthode de pondération de la fréquence d'une information relativement à sa fréquence en général. Parmi les concepts extraits pour nos listes d'interventions, nous souhaitions mettre en évidence les concepts à la fois fréquents dans notre population de patients ISO+ (TF) et peu fréquents parmi l'ensemble des patients de l'entrepôt de données (IDF).

Plusieurs méthodes de calcul du TF-IDF sont possibles. Prendre le logarithme du TF diminue par exemple la pondération du TF et donne donc plus d'importance à l'IDF. En testant plusieurs manières de calcul du TF-IDF, nous avons opté pour la formule suivante :

$$TF\text{-}IDF = \log(FqISO) * \log(FqEDS)$$

Avec **FqISO** la fréquence d'un concept au sein de la population ISO+ et **FdEDS** la fréquence du même concept au sein de l'entrepôt de données de santé.Parmi les concepts dont FqISO était supérieure à 20% et avec un TF-IDF élevé, nous avons sélectionné les concepts cliniquement pertinents **(tableau 5 & 6)**.

Tableau 5 - Diagnostics CIM10 sélectionnés de façon semi-automatique (méthode D^A)

	Chirurgie orthopédique du rachis
M46.26	Ostéomyélite vertébrale - Région lombaire
T84.6	Infection et réaction inflam dues à un appareil de fixation interne [toute localisation]
T81.38	Désunions dune plaie opératoire non classées ailleurs, autres et non précisées
Y83.4	Réactions - complications, sans accident cité - Autres chirurgies réparatrices
Y83.1	Réactions - complications, sans accident cité - Implantation prothèse interne
T81.4	Infection après un acte à visée diagnostique et thérapeutique, non classée ailleurs
G97.80	Perforation et déchirure accidentelle des méninges après un acte à visée diagnostique ou thérapeutique autre que rachicentèse
B95.6	Staphylococcus aureus, cause de maladies classées dans dautres chapitres
T81.8	Autres complications dun acte à visée diagnostique et thérapeutique, NCA
T84.7	Infection et réaction inflam dues à autres prothèses, implants et greffes orthopédiques internes
Y83.8	Réactions - complications, sans accident cité - Autres interventions chirurgicales
B95.7	Autres staphylocoques, cause de maladies classées dans dautres chapitres
	Neurochirurgie
T81.4	Infection après un acte à visée diagnostique et thérapeutique, non classée ailleurs
Y83.8	Réactions - complications, sans accident cité - Autres interventions chirurgicales

Tableau 6 - Actes CCAM sélectionnés de façon semi-automatique (méthode D^A)

	Chirurgie orthopédique du rachis
AFPA001	Mise à plat de lésion infectieuse péridurale rachidienne et/ou paravertébrale postopératoire [sepsis], par abord direct
	Neurochirurgie
LAPA015	Mise à plat de lésion infectieuse postopératoire de la voûte du crâne [calvaria], par reprise de l'abord précédent

Nous avons appliqué la même méthode concernant les protocoles de bactériologie positifs pondérant leur fréquence au sein du groupe d'interventions ayant été suivies d'une ISO à leur fréquence au sein du CHU de Bordeaux **(tableau 7)**.

Tableau 7 - Protocoles bactériologiques sélectionnés de façon semi-automatique (méthode D^A)

	Chirurgie orthopédique du rachis
BIOA	EXAMEN BACTERIOLOGIQUE D'UNE BIOPSIE AUTRE
BPRO	EXAMEN BACTERIOLOGIQUE D'UN PUS PROFOND
BHC	EXAMEN BACTERIOLOGIQUE D'UN PRELEVEMENT D'HEMOCULTURE
	Neurochirurgie
BSUP	EXAMEN BACTERIOLOGIQUE D'UN PUS SUPERFICIEL
BPRO	EXAMEN BACTERIOLOGIQUE D'UN PUS PROFOND

2.3 Méthode T

La *méthode T* consiste en une seconde approche utilisant uniquement le texte libre. Une extraction automatique de termes candidats est réalisée dans un premier temps. Une sélection semi-automatique des termes d'intérêts pour l'apprentissage est réalisée dans un second temps. Cette deuxième méthode se veut moins chronophage et plus reproductible que la *méthode D*. De ce fait elle peut-être plus facilement appliquée d'une spécialité chirurgicale à l'autre. D'autre part, utilisant uniquement du texte, elle n'est pas spécifique à l'entrepôt de données du CHU de Bordeaux et peut-être transposée à un autre établissement de santé.

2.3.1 EXTRACTION DES TERMES

Parmi les éléments de prise en charge extrait de l'entrepôt de données pour chaque patient, il existe des données non structurées sous forme de texte libre. C'est le cas des documents médicaux et paramédicaux tels que les comptes rendus d'hospitalisation et dont l'information se trouve dans la colonne `OBSERVATION_BLOB` de la table `OBSERVATION_FACT`. C'est également le cas des formulaires qui sont remplis dans les services, l'information se trouvant alors dans la colonne `TVAL_CHAR` de la même table. On nommera *méthode T^1* la méthode faisant uniquement appel aux termes retrouvés dans les documents. On nommera *méthode T^2* la méthode faisant appel aux termes retrouvés dans les documents et dans les formulaires.

Les données non structurées sont de traitement difficile et nécessitent d'avoir recours au TAL. *CandidateTerm* est un outil conçu par l'équipe de recherche en informatique appliquée à la santé (ERIAS). Il permet d'extraire de façon automatique les formes lemmatisées de groupes nominaux à partir de texte libre **(figure 8)**.

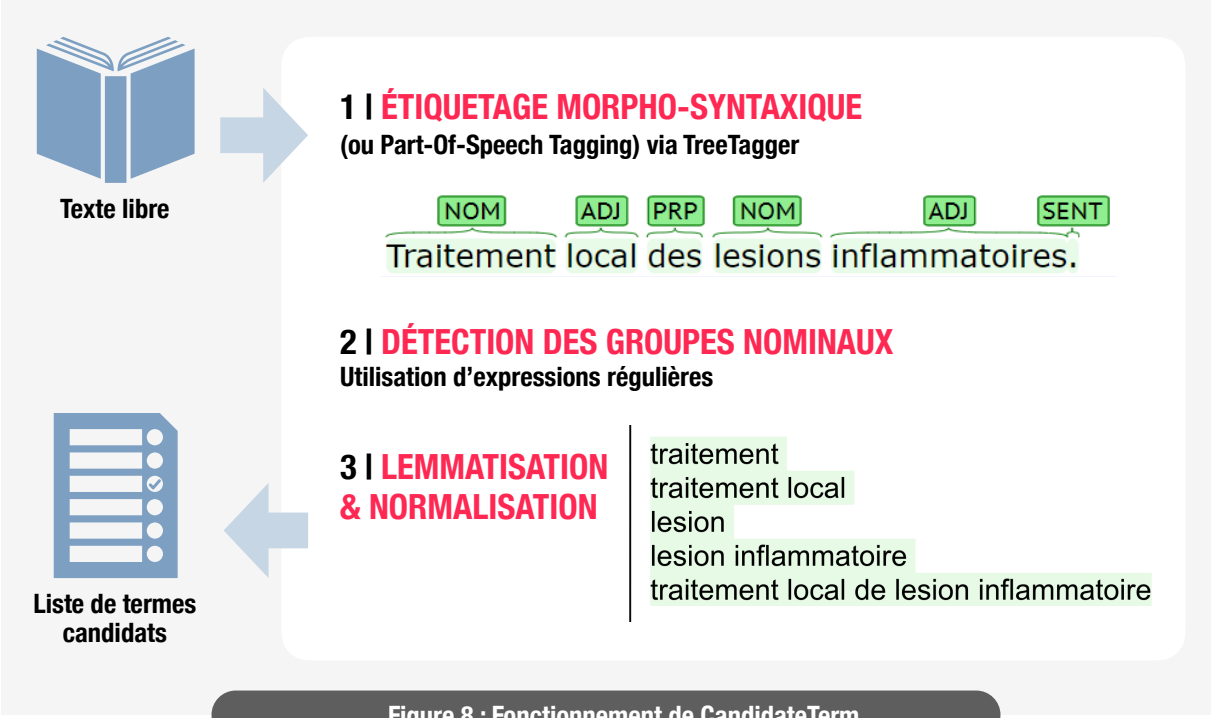

Figure 8 : Fonctionnement de CandidateTerm

Dans un premier temps, l'outil fait appel au logiciel TreeTagger[19] permettant l'étiquetage morpho-syntaxique d'énoncés en langue française. Dans un second temps, des groupes nominaux sont identifiés à l'aide d'une expression régulière. Enfin, ce sont les formes lemmatisées et normalisées de ces groupes nominaux identifiés et extraits du texte qui sont renvoyées à l'utilisateur comme termes candidats.

Par exemple, dans la phrase «reprise chirurgicale pour infection du site opératoire» les termes « reprise, reprise chirurgicale, reprise chirurgicale pour infection, infection, infection du site, infection du site opératoire, site, site opératoire » sont extraits automatiquement.

Le code de l'outil *CandidateTerm* est publié en open-source et est disponible ici :
https://github.com/scossin/CandidateTerm

2.3.2 SÉLECTION DES TERMES

CandidateTerm nous a permis d'extraire une liste de termes candidats à partir des données de texte libre. La deuxième étape consiste en la sélection de termes pertinents pour l'apprentissage et la détection des ISO. Pour cela nous avons procédé en plusieurs étapes **(figure 9)**.

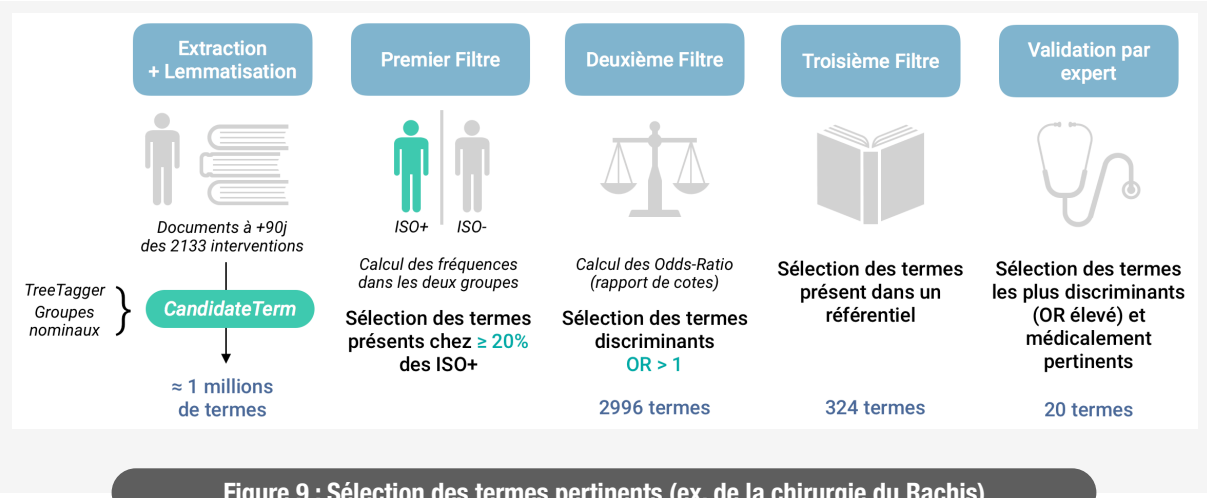

Figure 9 : Sélection des termes pertinents (ex. de la chirurgie du Rachis)

Après extraction des termes, chaque intervention possède un sac de termes. On sépare les interventions en deux groupes, celles avec la présence d'ISO (ISO+) et celles avec une absence d'ISO (ISO-), et on calcule pour chaque terme sa fréquence au sein des deux groupes. Un premier filtre consiste en la sélection des termes présent chez plus de 20% des interventions du groupe ISO+.

On calcule ensuite pour chaque terme le rapport de cotes (ou Odds-Ratio (OR)) de sa fréquence au sein du groupe ISO+ avec sa fréquence au sein groupe ISO-. Le deuxième filtre consiste en la sélection des termes dont l'OR est strictement supérieur à 1, c'est-à-dire ceux associés à la survenue d'une ISO.

On obtient ainsi une liste de termes dont ceux avec les OR les plus élevés sont les plus pertinents pour détecter une ISO suite à ses interventions. Néanmoins ces termes sont très spécifiques du CHU de Bordeaux et n'auraient pas la même signification dans un autre établissement de santé. Afin de permettre la réutilisation des termes d'intérêts sélectionnés, notamment dans le cadre de la mission nationale SPICMI, un troisième filtre est mis en place. La liste de termes pertinents est croisée avec les termes présents dans un référentiel. Ce référentiel est choisi selon la spécialité chirurgicale et les termes sont également extraits grâce à *CandidateTerm*. Les termes trop spécifiques du CHU de Bordeaux sont ainsi écartés de la liste de termes finale.

Enfin, la liste de termes candidats est soumise à l'expertise d'un médecin de santé publique, seule étape manuelle du processus de sélection. Elle permet de choisir les termes via leur pertinence clinique et pas uniquement statistique. On obtient ainsi une liste finale de 20 termes d'intérêts.

Chaque terme de la liste constitue une variable prédictive pour laquelle la présence du terme dans un délai de 90 jours suite à l'intervention est codée 0 ou 1. Pour constituer le jeu d'apprentissage, ces termes sont détectés dans le texte libre de l'entrepôt grâce à un deuxième outil nommé *IAMSystem* auquel on fourni la liste de toutes les formes possibles de nos 20 termes lemmatisés.

2.4 Algorithmes d'apprentissage automatique

2.4.1 PARAMÉTRAGE DES ALGORITHMES

Nous avons testés plusieurs algorithmes de classification pour nos deux approches. Dans les deux cas c'est la sensibilité (ou rappel) des algorithmes qui a été favorisée. En effet, l'outil de détection automatique des ISO a pour objectif le dépistage des ISO suspectes. Les ISO détectées seront ensuite présentées au chirurgien qui émettra le diagnostic final. Dans le cadre d'un dépistage par la machine on ne peut pas se permettre de manquer une ISO potentielle. Les faux négatifs ne sont pas tolérés et donc la sensibilité (et par la même occasion la VPN) a été paramétrée à 100%.

Pour ce faire, le seuil de probabilité au-dessus duquel l'algorithme prédit une ISO et classiquement situé à 50%, a été modifié en fonction des résultats obtenus sur le jeu d'apprentissage. Ainsi, le seuil de prédiction a été abaissé à la plus petite probabilité d'avoir une ISO parmi le groupe d'interventions connues ISO+.

Tout nos algorithmes présentent d'office un rappel (ou sensibilité) de 100% ainsi qu'une de VPN de 100%. Ils seront donc jugés sur leur précision (ou VPP) c'est-à-dire le nombre d'interventions connues ISO+ sur le nombre d'interventions détectées. De manière absolue, le nombre de faux positifs sera également un bon indicateur des performances des algorithmes.

2.4.2 MÉTHODE D

Les algorithmes de la *méthode D* apprennent sur l'ensemble des données disponibles dans l'entrepôt, c'est-à-dire les diagnostics CIM10, les actes CCAM, les administrations médicamenteuses, les protocoles bactériologiques et les documents textuels. Parmi elles des variables d'intérêts ont été dégagées dans un premier temps de manière manuelle (*méthode* D^M) et dans un second temps de manière semi-automatique (*méthode* D^A).

La *méthode D^M* a appris et a été testée uniquement avec le jeu de données d'apprentissage de la chirurgie orthopédique du rachis. L'ensemble des variables d'intérêts (17 au total) ont été fourni à un algorithme de LR. Afin d'optimiser les performances, certaines variables ont été retirées du modèle en fonction de leur pouvoir discriminant. Finalement ce sont les variables **DIAG**, **ACTE**, **ATB**, **PROT**, **DO** (pour « dommage »), **CI** (pour « cicatrice »), **EC** (pour « écoulement »), **SE** (pour « sepsis ») et **EI** (« effet indésirable ») qui ont été sélectionnées pour l'apprentissage dans le modèle D^M. Un algorithme de RF a également été testé avec les 17 variables.

La *méthode D^A* a été testée pour les deux spécialités chirurgicales. Concernant la chirurgie du rachis nous avons conservé le même type de variables que sélectionnées précédemment afin de comparer la méthode semi-automatique à la méthode manuelle.

Concernant la neurochirurgie, les variables conservées dans le modèle final sont **DIAG**, **ACTE**, **ATB**, **BAC**, **FER** (pour « fermeture »), **CIC** (pour « cicatrice »), **GER** (pour « germes »), **STA** (pour « staphylocoque ») et **AUR** (pour « aureus »).

2.4.3 MÉTHODE T

Les algorithmes de la *méthode T* on appris uniquement avec les données de texte libre. Plus précisément une liste de termes dont la présence (et non la fréquence) était codée 0 ou 1 pour chacun d'entre eux. Des algorithmes utilisant la LR et les RF ont été testés.

La *méthode T^1* utilise des termes sélectionnés et détectés à partir des seuls documents médicaux et paramédicaux. La *méthode T^2* utilise des termes sélectionnés et détectés à partir des documents et des formulaires. Ces deux variantes de la *méthode T* ont été testés pour la chirurgie orthopédique du rachis et pour la neurochirurgie.

La liste des termes sélectionnés selon l'approche et la spécialité chirurgicale est disponible en Annexe 2.

3. RÉSULTATS

Les résultats des différentes approches testées et de leurs variantes seront toujours présentés sous la forme d'une matrice de confusion représentant les prédictions de l'algorithme de manière absolue, ainsi qu'avec le calcul de la précision (ou VPP), de l'exactitude et de la spécificité. L'ensemble des algorithmes a été paramétré pour obtenir un rappel (ou sensibilité) de 100% et donc une VPN de 100%. Il est important de préciser que les résultats présentés concernent les performances des algorithmes testés sur leur jeu de données d'apprentissage.

3.1 Chirurgie du rachis : Méthode D

3.1.1 SÉLECTION MANUELLE DES VARIABLES

Concernant la chirurgie du rachis, la *méthode D^M* présente de meilleures performances avec un algorithme de LR **(tableau 8)** dont la précision est de 52% (avec 20 faux positifs) en comparaison à un algorithme de RF **(tableau 9)** dont la précision est de 20% (avec 87 faux positifs).

Tableau 8

Rachis : Matrice de confusion de la Méthode D^M
Algorithme de LR

		Classe réelle ISO+	Classe réelle ISO-	
Classe prédite	ISO+	22	20	42
	ISO-	0	2091	2091
		22	2111	2133

Performances :	Précision	Exactitude	Spécificité
	52,38 %	99,06 %	99,05 %

Tableau 9

Rachis : Matrice de confusion de la Méthode D^M
Algorithme de RF

		Classe réelle ISO+	Classe réelle ISO-	
Classe prédite	ISO+	22	87	109
	ISO-	0	2024	2024
		22	2111	2133

Performances :	Précision	Exactitude	Spécificité
	20,18 %	95,92 %	95,88 %

3.1.2 SÉLECTION SEMI-AUTOMATIQUE DES VARIABLES

Concernant la chirurgie du rachis, le passage d'une sélection manuelle à une sélection automatique des variables d'intérêt (*méthode D^A*) n'entraîne pas de perte de performances. Au contraire, la précision de l'algorithme de LR **(tableau 10)** passe de 52 à 54% (avec 19 faux positifs).

Nous n'avons pas testé l'algorithme de RF au vu des résultats précédents.

Tableau 10

Rachis : Matrice de confusion de la Méthode D^A
Algorithme de LR

		Classe réelle ISO+	Classe réelle ISO-	
Classe prédite	ISO+	22	19	41
	ISO-	0	2092	2092
		22	2111	2133

Performances :	Précision	Exactitude	Spécificité
	53,66 %	99,11 %	99,10 %

3.2 Chirurgie du rachis : Méthode T

3.2.1 UTILISATION DES DOCUMENTS

Concernant la chirurgie du rachis, les algorithmes utilisant uniquement les documents (*méthode T[1]*) se montrent moins performant avec une précision de 46% pour la LR **(tableau 11)** et une précision de 22% pour les RF **(tableau 12)**.

Néanmoins les performances de l'algorithme de LR **(tableau 11)** restent tout à fait correctes avec seulement 26 faux positifs pour 2133 interventions testées.

Tableau 11

Rachis : Matrice de confusion de la Méthode T[1]
Algorithme de LR

		Classe réelle ISO+	Classe réelle ISO-	
Classe prédite	ISO+	22	26	48
	ISO-	0	2085	2085
		22	2111	2133

Performances :	Précision	Exactitude	Spécificité
	45,83 %	98,78 %	98,77 %

Tableau 12

Rachis : Matrice de confusion de la Méthode T[1]
Algorithme de RF

		Classe réelle ISO+	Classe réelle ISO-	
Classe prédite	ISO+	22	78	100
	ISO-	0	2033	2033
		22	2111	2133

Performances :	Précision	Exactitude	Spécificité
	22,00 %	96,34 %	96,31 %

3.2.2 UTILISATION DES DOCUMENTS ET DES FORMULAIRES

Concernant la chirurgie du rachis, c'est l'algorithme de LR utilisant à la fois les documents et les formulaires (*méthode T[2]*) qui se montre le plus performant **(tableau 13)**. En effet, l'ajout des données des formulaires fait passer la précision de 46 à 61% (avec 14 faux positifs).

Tableau 13

Rachis : Matrice de confusion de la Méthode T[2]
Algorithme de LR

		Classe réelle ISO+	Classe réelle ISO-	
Classe prédite	ISO+	22	14	36
	ISO-	0	2097	2097
		22	2111	2133

Performances :	Précision	Exactitude	Spécificité
	61,11 %	99,34 %	99,34 %

Nous n'avons pas testé l'algorithme de RF au vu des résultats précédents.

3.3 Chirurgie du rachis : Correction du Gold Standard

Suite aux premiers résultats obtenus grâce à la *méthode D^M* appliqué à la chirurgie du rachis, une réflexion s'est engagée avec l'équipe d'hygiène concernant la qualité du Gold Standard. Il a été décidé d'entreprendre une vérification par retour au DPI pour les 20 faux positifs détectés via l'algorithme de LR **(tableau 8)**. Parmi eux, 8 ISO non déclarées par le chirurgien ont été mises en évidence. De plus, 9 faux positifs correspondaient à l'intervention de reprise (code AFPA001) d'une véritable ISO. Au final, seuls 3 faux positifs n'étaient réellement pas des ISO.

Le Gold Standard a donc été corrigé pour les 8 ISO non déclarées. Concernant les actes de reprises AFPA001 (soit 11 interventions) il a été décidé de les exclure du jeu d'apprentissage de part la confusion qu'ils entraînent pour la machine et la sous-estimation des performances de l'algorithme. Ces actes de reprises devraient faire l'objet d'un traitement à part entière.

Une fois corrigé, le jeu de données d'apprentissage pour la chirurgie orthopédique du rachis comportait 2122 interventions dont 30 avaient un statut ISO+.

3.3.1 MÉTHODE D

Suite à la correction du Gold Standard, les performances ont été améliorées concernant les modèles utilisant l'ensemble des données. La *méthode D^M* présente d'excellentes performances avec une précision de 81% et 7 faux positifs **(tableau 14)**.

Mais c'est la *méthode D^A* qui présente les meilleures performances (toutes méthodes confondues) en ce qui concerne la chirurgie du rachis. Elle obtient une excellente précision de 94% avec seulement 2 faux positifs **(tableau 15)**.

Tableau 14

Rachis : Matrice de confusion de la Méthode D^M
(GOLD STANDARD CORRIGÉ)

Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	30	7	37
	ISO-	0	2085	2085
		30	2092	2122

Performances :	Précision	Exactitude	Spécificité
	81,08 %	99,67 %	99,67 %

Tableau 15

Rachis : Matrice de confusion de la Méthode D^A
(GOLD STANDARD CORRIGÉ)

Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	30	2	32
	ISO-	0	2090	2090
		30	2092	2122

Performances :	Précision	Exactitude	Spécificité
	93,75 %	99,91 %	99,90 %

3.3.2 MÉTHODE T

Les performances de la *méthode T*[1] ont été légèrement améliorée suite à la correction du Gold Standard passant d'une précision de 46 à 50% **(tableau 16)**. Le nombre de faux positifs a cependant augmenté passant de 26 à 30. Les performances de la *méthode T*[2] ont quant à elles chutées de 61 à 46% (avec 35 faux positifs) **(tableau 17)**.

Tableau 16

Rachis : Matrice de confusion de la Méthode T[1]
(GOLD STANDARD CORRIGÉ)
Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	30	30	60
	ISO-	0	2062	2062
		30	2092	2122

Performances :	Précision	Exactitude	Spécificité
	50,00 %	98,59 %	98,57 %

Tableau 17

Rachis : Matrice de confusion de la Méthode T[2]
(GOLD STANDARD CORRIGÉ)
Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	30	35	65
	ISO-	0	2057	2057
		30	2092	2122

Performances :	Précision	Exactitude	Spécificité
	46,15 %	98,35 %	98,33 %

3.4 Neurochirurgie

Suite aux très bonnes performances des algorithmes d'apprentissage automatique pour la détection automatique des ISO en chirurgie orthopédique du rachis, il a été décidé de s'intéresser à la neurochirurgie.

3.4.1 MÉTHODE T

Concernant la neurochirurgie, ce sont les variantes de la *méthode T* qui ont été testées en premier lieu. En effet, l'utilisation du texte seul est la méthode la plus facilement transposable à d'une spécialité chirurgicale à l'autre.

Tableau 18

Neuro. : Matrice de confusion de la Méthode T[1]
Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	20	2222	2242
	ISO-	0	61	61
		20	2283	2303

Performances :	Précision	Exactitude	Spécificité
	0,89 %	3,52 %	2,67 %

Tableau 19

Neuro. : Matrice de confusion de la Méthode T[2]
Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	20	1194	1214
	ISO-	0	1089	1089
		20	2283	2303

Performances :	Précision	Exactitude	Spécificité
	1,65 %	48,15 %	47,70 %

Les performances des *méthodes T*[1] **(tableau 18)** et *T*[2] **(tableau 19)** ce sont avérées très mauvaises avec respectivement une précision de 0,89% et 1,65% et un nombre de faux positifs respectif de 2222 et 1194.

Un retour au dossier a été entrepris afin de comprendre une telle différence avec la chirurgie du rachis. Il s'est avéré que parmi les interventions du gold standard connues ISO+, celles qui avaient les probabilités les plus faibles selon l'algorithme ne disposaient en réalité soit d'aucunes informations textuelles permettant d'identifier une ISO, soit ces informations étaient bien trop tardives dépassant le délai de 90 jours de surveillance post-intervention.

Afin de permettre à l'algorithme d'apprendre correctement, il a été décidé de supprimer 4 interventions du jeu d'apprentissage, celles jugées comme présentant peu ou pas d'informations permettant de les classer. Au final, le nouveau jeu d'apprentissage comportait 2299 interventions avec 14 ISO déclarées.

Tableau 20

Neuro. : Matrice de confusion de la Méthode T[1]
(JEU D'APPRENTISSAGE MODIFIÉ)

Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	16	26	42
	ISO-	0	2257	2257
		16	2283	2299
Performances :		Précision	Exactitude	Spécificité
		38,10 %	98,87 %	98,86 %

Tableau 21

Neuro. : Matrice de confusion de la Méthode T[2]
(JEU D'APPRENTISSAGE MODIFIÉ)

Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	16	24	40
	ISO-	0	2259	2259
		16	2283	2299
Performances :		Précision	Exactitude	Spécificité
		40,00 %	98,96 %	98,95 %

Suite à la modification du jeu d'apprentissage, les performances ont été améliorées pour les deux variantes de la *méthode T*. L'ajout des formulaires dans la *méthode T*[2] **(tableau 21)** augmente les performances avec une précision de 40% (et 24 faux positifs) contre une précision de 38% (et 26 faux positifs) pour la *méthode T1* **(tableau 20)**. Il est important de préciser que suite au retrait de 4 interventions du jeu d'apprentissage le rappel ne peut plus être considéré à 100%.

3.4.2 MÉTHODE D

Concernant la neurochirurgie, il n'y a pas eu de sélection manuelle des variables d'intérêts. Seule la *méthode D^A* avec un algorithme de LR, très performante pour la chirurgie du rachis, a été testée.

La *méthode D^A* a été évaluée et optimisée sur le jeu d'apprentissage modifié précédemment.

Les performances sont moins bonnes qu'avec le *modèle T*. La précision est seulement de 13% avec 106 faux positifs parmi les 2299 interventions testés **(tableau 22)**.

Tableau 22

Neuro. : Matrice de confusion de la Méthode D^A
(JEU D'APPRENTISSAGE MODIFIÉ)

Algorithme de LR

		Classe réelle		
		ISO+	ISO-	
Classe prédite	ISO+	16	106	122
	ISO-	0	2177	2177
		16	2283	2299

Performances :	Précision	Exactitude	Spécificité
	13,11 %	95,39 %	95,36 %

4. DISCUSSION

4.1 Résultats

Concernant la chirurgie orthopédique du rachis, première spécialité chirurgicale à laquelle nous nous sommes intéressés, les résultats sont très encourageants pour le développement de méthodes semi-automatiques de détection des ISO. Le premier modèle conçu et testé est le modèle utilisant l'ensemble des données disponibles dans l'entrepôt et dont les variables d'intérêts ont été sélectionnées de façon manuelle par expertise médicale (*méthode D^M*). Grâce à un algorithme de régression logistique, pour un rappel (ou sensibilité) fixé à 100%, soit la totalité des ISO détectée, la précision atteignait 52%. Plus encourageant encore pour les équipes chargées de la surveillance, de manière absolue, sur 2133 interventions testées seulement 20 étaient classées de façon erronée en faux positifs.

Quatre démarches ont été entreprises de façon concomitante à la suite de ces premiers résultats prometteurs concernant la détection automatique des ISO pour la chirurgie du rachis. Ainsi, plusieurs réflexions ont été menées concernant l'automatisation de la *méthode D^M*, la conception d'un deuxième modèle uniquement à partir des données du texte libre (*méthode T*), la validité du Gold Standard de la chirurgie du rachis et enfin l'application à une autre spécialité chirurgicale telle que la neurochirurgie.

L'automatisation de la *méthode D^M* consiste en un passage d'une sélection entièrement manuelle des variables d'intérêts à une sélection semi-automatique avec proposition à l'expert d'une liste de variables pertinentes à sélectionner. Les performances ne sont pas diminuées avec cette nouvelle *méthode D^A* dont la précision passe de 52 à 54% (avec 19 faux positifs) pour la chirurgie du rachis. Ces résultats nous conforte dans l'idée de privilégier la *méthode D^A* à la *méthode D^M*.

La conception d'un deuxième modèle basé uniquement à partir des données de texte libre (*méthode T*) nous a parue intéressante dans un objectif de transposabilité de la méthode à d'autres établissements de santé. En effet, la *méthode D*, basée sur l'ensemble des informations disponibles dans notre entrepôt de données biomédicales, est très spécifique du CHU de Bordeaux. De plus la sélection des variables d'intérêts est chronophage et moins reproductible. Le premier modèle testé (*méthode T^1*) reposait uniquement sur le texte libre des documents médicaux et paramédicaux. Les performances sont inférieures à la *méthode D* avec une précision passant de 52 à 46%. Néanmoins ces résultats restent très intéressants avec seulement 26 faux positifs sur les 2133 interventions testées. Cela peut s'expliquer par le fait que la *méthode D* dispose de nombreuses sources de données (diagnostics, actes, antibiotiques, bactériologie et documents) ce qui augmente la probabilité de trouver l'information concernant la présence ou non d'une ISO.

Un deuxième variante de la *méthode T* a été testée utilisant à la fois le texte des documents médicaux et le texte présent dans les formulaires (*méthode T^2*). Les performances sont largement améliorées passant d'une précision de 46 à 61% et dépassant même la *méthode D* avec seulement 14 faux positifs. Néanmoins, la liste de termes sélectionnés avec cette variante de la *méthode T* pourrait perdre son caractère transposable à d'autres établissements de santé car les formulaires sont établissements dépendants. De plus, nous verrons par la suite que l'apport d'informations supplémentaires via les formulaires à la *méthode T^1* ne va pas toujours vers une amélioration de la précision.

Suite aux premiers résultats obtenus grâce à la *méthode D^M*, une réflexion a été menée avec l'équipe d'hygiène concernant les 20 faux positifs détectés par l'algorithme. En effet, le Gold Standard repose sur la déclaration du chirurgien vis-à-vis des 2133 interventions testées, or l'équipe d'hygiène n'est retournée au DPI que pour les 22 ISO déclarées. Il a donc été décidé de faire un retour au dossier pour les 20 faux positifs prédits par l'outil. Ce sont ainsi 8 ISO non déclarées qui ont été diagnostiquées par le Service d'Hygiène Hospitalière. D'autre part, 9 faux positifs étaient en réalité des interventions de reprise à la suite d'une véritable ISO. Ces actes de reprises, source d'ambiguïté pour l'apprentissage, ont été retirés du jeu d'entraînement et le Gold Standard a été corrigé pour les 8 ISO nouvellement diagnostiquées.

Suite à la correction du Gold Standard de la chirurgie du rachis, les performances ont été améliorées pour l'ensemble des modèles à l'exception de la *méthode T^2* qui a vu ses performances baisser. Au final, c'est la *méthode D^A* qui est la plus performante pour la détection des ISO suite à une chirurgie du rachis avec une excellente précision de 94% pour un rappel fixé à 100%. Seulement 2 interventions ont été classées à tord en faux positifs. La *méthode T^1*, la plus transposable, obtient des performances correctes avec une précision de 50% soit 30 interventions faussement positives pour 30 ISO détectées correctement.

Concernant la neurochirurgie, il a dans un premier temps été testé la *méthode T* de part sa facilité d'application d'une spécialité chirurgicale à l'autre. Les performances se révèlent d'abord très mauvaises avec une précision de 0,89% pour la *méthode T^1* et 1,65% pour la *méthode T^2*. Un retour au DPI a été réalisé afin de comprendre une telle différence entre la neurochirurgie et la chirurgie du rachis. Il s'est avéré que les cas ISO+ les moins probables selon l'algorithme étaient en réalité soit dépourvus d'informations textuelles concernant l'ISO, soit l'information était bien trop tardive dépassant le délai de surveillance défini à 90 jours. Nous avons ainsi ôté 4 interventions ISO+ du jeu d'apprentissage. Les performances se sont révélées par la suite bien meilleures rejoignant celles de la chirurgie du rachis avec une précision de 38% pour la *méthode T^1* et 40% pour le *méthode T^2* avec respectivement 26 et 24 faux positifs pour 2299 interventions testées. Néanmoins, cela signifie qu'un rappel de 100% ne peut-être atteint faute d'information.

Concernant la *méthode D^A* appliquée à la neurochirurgie, les performances sont mauvaises avec une précision très faible de 13% bien que l'exactitude (c'est-à-dire les interventions correctement prédites) soit de 95%. Cela démontre à quel point la *méthode D* est difficilement applicable d'une chirurgie à l'autre, même avec une sélection de variables d'intérêts différente. Une réflexion supplémentaire concernant les variables à intégrer ou non dans cette méthode fait partie des axes d'amélioration.

Les moins bonnes performances de nos *méthodes D et T* lorsqu'elles sont appliquées à la neurochirurgie pourraient être expliquées par l'incidence de cette dernière. En effet, la neurochirurgie est la spécialité chirurgicale avec le taux d'incidence (0,79%) le plus faible en 2017. Il est donc plus difficile de disposer d'un jeu d'apprentissage avec de nombreuses données concernants les interventions ISO+. Ce manque de données fait partie des limites de ce travail.

4.2 Limites

Dans un premier temps il est important de rappeler que les réelles performances d'un algorithme d'apprentissage automatique s'évaluent sur un jeu de données de test, c'est-à-dire un jeu de données avec lequel il ne s'est pas entraîné et qu'il ne connait donc pas. Or nous ne disposions pas d'un tel jeu de données. En effet, nos échantillons d'apprentissage étaient bien trop petits pour être divisés en deux. Nous avons néanmoins tenté cette approche pour le premier modèle testé (*méthode D^M*) en l'entraînement uniquement sur les années 2015 et 2016 de notre Gold Standard non corrigé et en le testant sur l'année 2017. Le rappel était de 100% et la précision de 33% avec les 2 ISO correctement détectées et 4 interventions détectées à tord sur les 763 interventions testées. Cependant ces chiffres sont trop faibles pour conclure. C'est pourquoi toutes les performances présentées dans la partie résultats de ce mémoire résultent d'un test sur le jeu de données d'apprentissage et sont de ce fait surévaluées. L'évaluation des performances sur un nouvel échantillon de test fait partie des axes d'amélioration.

La principale limite de notre projet est le faible nombre de cas d'ISO dans nos Gold Standard avec seulement 22 ISO déclarées sur 3 années de surveillance pour la chirurgie du rachis (30 ISO après correction) et 20 ISO déclarées sur 6 années de surveillance pour la neurochirurgie. Ce petit nombre de cas représente une limite à plusieurs niveaux. Premièrement, il est difficile pour un algorithme d'apprentissage automatique supervisé d'apprendre avec aussi peu de cas. Cela peu expliquer pourquoi la régression logistique a présenté de meilleurs résultats comparée à l'utilisation de forêts aléatoires (RF) plus adaptées à de gros volumes de données. Deuxièmement et comme vu précédemment, il n'a pas été possible de diviser nos échantillons en jeu d'entraînement et jeu d'évaluation. L'évaluation des performances est donc biaisée et nécessite la création d'un nouvel échantillon annoté sur l'année 2018. Nous y reviendrons.

Une autre limite majeure rencontrée lors de ce travail est la qualité de notre Gold Standard. En effet, ce dernier repose uniquement sur la déclaration des chirurgiens vis-à-dis des listes d'interventions présentées. Seules les ISO déclarées ont été à nouveau validées par l'équipe du Service d'Hygiène Hospitalière. Ce qui pose le problème des interventions déclarées à tord ISO- par le chirurgien, ce dernier pouvant avoir oublié la survenue d'une ISO pour certaines d'entre elles, on parle de biais de mémorisation. Ainsi, concernant la chirurgie du rachis, un retour au dossier patient a été entrepris par l'équipe d'hygiène afin de confirmer les 20 faux positifs prédits par la *méthode D^M*. Et ce sont donc finalement 8 ISO qui n'avaient pas été correctement déclarées dans notre Gold Standard de chirurgie du rachis. Ce travail de vérification n'a pas encore été effectué pour les autres faux positifs découverts à la suite de la correction du Gold Standard. Il n'a pas non plus été effectué pour la neurochirurgie. Les performances des algorithmes sont cette fois-ci possiblement sous-évaluées de part ces interventions faussement déclarées ISO-.

Toujours concernant le Gold Standard, certaines interventions déclarées ISO+ par le chirurgien ne disposent de peu ou pas d'informations dans le DPI concernant leur ISO. Parfois l'information est bien trop tardive compte tenu des délais de surveillance post-opératoire fixés. Ce problème a surtout été rencontré avec les cas de neurochirurgie. L'algorithme apprenait alors avec des cas ISO+ qui n'avaient pas de données informatisées. Les performances étaient bien évidemment très mauvaises. Ne pas les inclure dans le jeu d'apprentissage semblait le plus adéquat sachant que lors de la mise en production de l'outil ces ISO n'auraient quoi qu'il arrive pas été détectées. Un rappel de 100% ne peut donc pas être envisagé dans ces circonstances. Cela correspond à un problème plus général de présence de l'information au sein du DPI et plus globalement au sein du SIH. C'est un facteur limitant de nombreux projets faisant appel à des données informatisées, notamment pour de l'automatisation.

Concernant la conception des modèles et notamment la sélection des variables d'intérêts à inclure pour l'apprentissage, beaucoup d'étapes restent manuelles. Cela présente plusieurs inconvénients dont le caractère chronophage, peu reproductible et la nécessité d'une expertise médicale. Des améliorations ont déjà été entreprises pour rendre semi-automatique la sélection des variables dans la *méthode D*. De même, la *méthode T* ne dispose que d'une seule étape manuelle. Malgré tout, pour ces deux modèles, il existe une étape de sélection des concepts ou des termes par l'expert. Cela fait intervenir l'humain et rend la sélection cliniquement pertinente, ce qui n'est pas un mauvais point, notamment lorsqu'il faudra convaincre les chirurgiens de laisser le dépistage des ISO à la machine. Cependant, la sélection n'est que peu reproductible d'un expert à l'autre et cela influe sur les performances de l'algorithme. Une sélection via des méthodes uniquement statistiques fait partie des axes d'amélioration.

La *méthode D* est la méthode la moins facilement applicable d'une chirurgie à l'autre, car il faudra pour chaque variable (diagnostics pertinents, actes pertinents, protocoles bactériologiques pertinents et termes pertinents) refaire une sélection via l'expert. Il est également peu ou pas transposable à d'autres établissements de santé car il a été conçu à partir et pour une utilisation via l'entrepôt de données du CHU de Bordeaux. Au contraire, la *méthode T* qui ne se base que sur du texte peut être facilement utilisée dans un autre établissement et facilement appliquée à d'autres spécialités chirurgicales. L'application de la méthode à d'autres chirurgies fait partie de nos perspectives. La transposition de l'outil (et son évaluation) à d'autres établissement de santé en fait également partie.

Enfin, concernant l'outil d'extraction automatique des termes (*CandidateTerm*) utilisé pour la *méthode T*, ce dernier n'a pas fait l'objet d'une évaluation formelle. D'autre part, il est spécifique au français et pourrait être comparé aux outils similaires comme BioTex [(20)] et TermSuite [(21)]. Aucune détection de la négation ou de la temporalité n'est pour l'instant réalisée bien que des algorithmes en français existent [(22,23)].

4.3 Axes d'amélioration & Perspectives

Le premier axe d'amélioration constitue la création de nouveaux échantillons annotés pour l'évaluation des performances réelles de nos algorithmes sur des jeux de test. Dans un premier temps, nous nous intéressons aux interventions de chirurgie orthopédique du rachis sur l'année 2018. Grâce à la *méthode* D^A, méthode la plus performante pour détecter les ISO de cette spécialité chirurgicale, ce sont 48 ISO qui ont été détectées parmi les 2504 interventions réalisées en 2018. Une interface de validation est en cours de création pour permettre au chirurgien de valider ou d'invalider les ISO prédites par la machine. Nous obtiendrons ainsi une première évaluation des performances. De plus, une fois le jeu de données 2018 annoté, et de façon itérative, l'algorithme pourra de nouveau s'entraîner et s'améliorer. À terme, cette interface de validation pourrait être mise en production si les performances se révèlent à la hauteur de celles mesurées lors des phases d'apprentissage.

L'application de la *méthode D* à d'autres spécialités chirurgicales s'avère plus complexe que prévu. Une réflexion supplémentaire devrait être mise en oeuvre pour aller dans ce sens. De plus, une automatisation complète de la sélection des concepts pertinents via des méthodes uniquement statistique est à l'étude. Dans l'attente de meilleurs résultats, la *méthode T* représente une très bonne alternative avec des résultats qui restent corrects et qui sont compatibles avec une mise en production (l'effectif des interventions dépistées à tord restant faible). Cette méthode est rapidement applicable à une nouvelle spécialité chirurgicale et devrait être testée prochainement pour la pose de prothèses articulaires orthopédiques.

Ce projet s'inscrit dans la dynamique nationale de la mission SPICMI. Pour rappel, il a été demandé aux établissements de santé de débuter une réflexion concernant l'automatisation de la surveillance des ISO au sein de leur établissement. Nous espérons que notre outil de détection automatique des ISO et surtout la réflexion multidisciplinaire menée autour de sa création pourra guider le CPIAS Ile-de-France dans son travail d'élaboration ainsi que les autres établissements de santé. La *méthode T* semble l'outil le plus à même d'être reproduit à l'extérieur du CHU de Bordeaux.

Une fois les performances réelles des différents algorithmes mesurées sur des échantillons de test, ce travail devrait faire l'objet d'une publication scientifique. Un papier a déjà été soumis et accepté (annexe 3) pour la conférence MEDINFO (17ème *World Congress On Medical And Health Informatics*) qui a eu lieu à Lyon du 26 au 30 août 2019. Sébastien Cossin, AHU au sein de l'UIAM et encadrant de ce projet de master 2, a ainsi présenté notre travail dans le cadre d'une conférence sur le TAL.

5. CONCLUSION

La création d'un outil de détection automatique des ISO avait pour objectif de facilité la surveillance des ISO au sein du CHU de Bordeaux et de permettre un gain de temps aux équipes d'hygiène hospitalière afin de se consacrer aux activités de prévention. Deux approches utilisant des algorithmes d'apprentissage automatique ont été réalisées à partir des informations disponibles dans l'entrepôt de données biomédicales. La première utilise l'ensemble des données disponibles et se montre la plus performante lorsque les variables prédictives sont correctement sélectionnées. Ainsi, elle est capable de détecter toutes les ISO survenant à la suite à d'une chirurgie du rachis avec un très faible nombre d'interventions détectées à tord. La seconde approche utilise uniquement les données du texte libre et contrairement à la première, elle n'est pas spécifique du CHU de Bordeaux et peut être facilement transposée à un autre établissement. Moins performante concernant la chirurgie rachis, elle présente l'avantage de pouvoir être facilement appliquée à d'autres spécialités chirurgicales en gardant des performances correctes, ce qui n'est pas le cas de la première approche. Une interface de validation par le chirurgien des ISO dépistées par la machine est en cours de réalisation et servira à la fois d'outil d'évaluation et, à terme, d'outil pour la mise en production. 8 ISO ont déjà été diagnostiquées durant la réalisation de ce travail.

Bibliographie

1. Programme national d'actions de prévention des infections associées aux soins (Propias). Ministère des affaires sociales, de la santé et des droits des femmes; 2015. Disponible sur : https://solidarites-sante.gouv.fr/soins-et-maladies/qualite-des-soins-et-pratiques/securite/propias/article/programme-national-d-actions-de-prevention-des-infections-associees-aux-soins

2. Villani C. Donner un sens à l'intelligence artificielle, pour une stratégie nationale et européenne. Mission Villani sur l'intelligence artificielle; 2018. Disponible sur : https://www.aiforhumanity.fr

3. Enquête nationale de prévalence des infections nosocomiales et des traitements anti-infectieux en établissements de santé, France, mai-juin 2017. Santé Publique France; 2018. Disponible sur : https://invs.santepubliquefrance.fr/Publications-et-outils/Rapports-et-syntheses/Maladies-infectieuses/2018/Enquete-nationale-de-prevalence-des-infections-nosocomiales-et-des-traitements-anti-infectieux-en-etablissements-de-sante-France-mai-juin-2017

4. Surveillance des infections du site opératoire dans les établissements de santé. Réseau ISO-Raisin, France. Résultats 2017. Santé Publique France; 2019. Disponible sur : https://www.santepubliquefrance.fr/docs/surveillance-des-infections-du-site-operatoire-dans-les-etablissements-de-sante-reseau-iso-raisin-france.-resultats-2017

5. Haley RW, Quade D, Freeman HE, Bennett JV. The SENIC Project. Study on the efficacy of nosocomial infection control (SENIC Project). Summary of study design. American Journal of Epidemiology; 1980. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/6246798

6. Surveillance des Infections du Site Opératoire (Surveillance des interventions prioritaires), Protocole National. Réseau ISO-Raisin; 2018. Disponible sur : http://www.cpias-ile-de-france.fr/surveillance/reseau-iso.php

7. Surveillance globale agrégée des Infections du Site Opératoire, Protocole National. Réseau ISO-Raisin; 2018. Disponible sur : http://www.cpias-ile-de-france.fr/surveillance/reseau-iso.php

8. CPIAS Ile-de-France. Surveillance et prévention du risque infectieux en chirurgie et médecine interventionnelle (Spicmi) [Internet]. Disponible sur : http://www.cpias-ile-de-france.fr/surveillance/spicmi.php

9. World Health Organization (WHO). ICD-10 Version: 2010. Disponible sur : http://apps.who.int/classifications/icd10/browse/2010/en

10. Assurance Maladie. CCAM. Disponible sur : https://www.ameli.fr/accueil-de-la-ccam/index.php

11. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of Clinical and Genetic Data in the i2b2 Architecture. AMIA Annu Symp Proc. 2006; 2006:1040. Disponible sur : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839291/

12. Yvon F. Une petite introduction au Traitement Automatique des Langues Naturelles. Disponible sur : https://perso.limsi.fr/anne/coursM2R/intro.pdf

13. Leclère B, Lasserre C, Bourigault C, Juvin M-E, Chaillet M-P, Mauduit N, et al. Matching Bacteriological and Medico-Administrative Databases Is Efficient for a Computer-Enhanced Surveillance of Surgical Site Infections: Retrospective Analysis of 4,400 Surgical Procedures in a French University Hospital. Infect Control Hosp Epidemiol. 2014 Nov;35(11): 1330–5. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/25333426

14. Proux D, Hagège C, Gicquel Q, Kergourlay I, Pereira S, Rondeau G, et al. ALADIN: développement d'un outil sémantique d'analyse des documents textuels médicaux pour la détection d'infections associées aux soins. IRBM. 2012 Apr;33(2): 137–42. Disponible sur : https://www.em-consulte.com/en/article/702595

15. Boris C-G, Nicolas G, Pascal J, Marc CJ, Marc C. Full-text Automated Detection of Surgical Site Infections Secondary to Neurosurgery in Rennes, France. Studies in Health Technology and Informatics. 2013;572–575. Disponible sur : https://hal.archives-ouvertes.fr/hal-01142182/

16. Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalianis H. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. Health Informatics J. 2018 Mar;24(1): 24–42. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/27496862

17. Hu Z, Simon GJ, Arsoniadis EG, Wang Y, Kwaan MR, Melton GB. Automated Detection of Postoperative Surgical Site Infections Using Supervised Methods with Electronic Health Record Data. 2017;16. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/26262143

18. Chapman AB, Mowery DL, Swords DS, Chapman WW, Bucher BT. Detecting Evidence of Intraabdominal Surgical Site Infections from Radiology Reports Using Natural Language Processing. :10. Disponible sur : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977582/

19. Schmid H. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. (1995). Disponible sur : https://www.semanticscholar.org/paper/Improvements-in-Part-of-Speech-Tagging-with-an-to-Schmid/343d2492caa5265467884d7c172c658a680b7d3d

20. Lossio-Ventura JA, Jonquet C, Roche M, Teisseire M. BIOTEX: A System for Biomedical Terminology Extraction, Ranking, and Validation. Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272, CEUR-WS.org, Aachen, Germany, Germany, 2014: pp. 157–160. Disponible sur : https://hal.archives-ouvertes.fr/hal-01136531

21. Cram D, Daille B. Terminology Extraction with Term Variant Detection. Proceedings of ACL-2016 System Demonstrations, Association for Computational Linguistics, Berlin, Germany, 2016: pp. 13–18. doi:10.18653/v1/P16-4003. Disponible sur : https://www.aclweb.org/anthology/P16-4003/

22. Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. Détecter la négation, la temporalité et le sujet dans les textes cliniques Français, in: SIIM: Symposium Sur l'Ingénierie de l'Information Médicale, Toulouse, France, 2017. Disponible sur : https://hal-lirmm.ccsd.cnrs.fr/lirmm-01656834

23. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse, J Am Med Inform Assoc. 24 (2017) 607–613. doi:10.1093/jamia/ocw144. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/28339516

ANNEXE 1.

Critères diagnostiques des différents types d'ISO selon les Centers for Disease Control and prevention.

Infection superficielle de la plaie chirurgicale
L'infection se manifeste jusqu'à 30 jours après l'opération
ET
l'infection concerne uniquement la peau ou les tissus sous-cutanés de l'incision
ET au moins un des critères suivants :
- sécrétion purulente, avec ou sans confirmation microbiologique, de l'incision superficielle
- Isolement d'organismes dans une culture de liquide ou de tissu de l'incision superficielle obtenue sous asepsie
- Au moins un des signes ou symptômes d'infection suivants : douleurs spontanées ou à la palpation, tuméfaction localisée, rougeur, ou chaleur ET incision superficielle délibérément ouverte par le chirurgien, à moins que la culture des prélèvements microbiologiques de l'incision soit négative
- Diagnostic d'une infection superficielle de la plaie chirurgicale établi par un chirurgien ou le médecin impliqué

Ne pas considérer comme infection de plaie superficielle :
- un abcès des points de suture (c'est-à-dire inflammation minimale et sécrétion limitées aux points de suture)
- Une infection d'une épisiotomie ou d'un site de circoncision d'un nouveau-né
- Une infection d'une plaie de brûlure
- Une infection de la plaie chirurgicale qui s'étend jusque dans le fascia et les couches musculaires (voir infection de plaie chirurgicale profonde)

Infection profonde de la plaie chirurgicale
L'infection se manifeste jusqu'à 30 jours après l'intervention (si pas d'implant) ou jusqu'à un an (si présence d'implant) et l'infection semble liée à l'opération
ET
l'infection implique les tissus mous profonds (par exemple, fascia, couches musculaires) de l'incision
ET au moins un des points suivants :
- sécrétion purulente de l'incision profonde, mais ne provenant pas d'un organe ou d'une cavité profonde qui font partie du site opératoire
- Déhiscence spontanée d'une incision profonde ou ouverture délibérée par le chirurgien si le patient présente au moins un des signes ou symptômes suivants : fièvre (>38°C), douleur localisée spontanée ou à la palpation, à moins que la culture des prélèvements microbiologiques du site chirurgical ne soit négative (prélèvement stérile)
- Abcès ou autre évidence d'infection qui implique l'incision profonde à l'évaluation directe, lors de réintervention, ou à l'examen histopathologique ou radiologique
- Diagnostic d'une infection profonde de la plaie chirurgicale établi par un chirurgien ou le médecin impliqué

Infection du site opératoire d'organe ou cavité
L'infection se manifeste jusqu'à 30 jours après l'intervention (si pas d'implant) ou jusqu'à un an (si présence d'implant) et l'infection semble liée à l'opération
ET
l'infection implique n'importe quelle partie du site chirurgical (par exemple, organe ou cavité), en dehors de l'incision, qui a été ouverte ou manipulée durant l'opération
ET au moins un des points suivants :
- sécrétion purulente par un drain à travers la peau dans un organe ou une cavité
- Présence d'organismes dans une culture de liquide ou de tissu d'un organe ou d'une cavité obtenu de manière aseptique
- Abcès ou autre évidence d'infection impliquant l'organe ou la cavité détecté lors d'une évaluation directe, une réintervention ou par un examen histopathologique ou radiologique
- Diagnostic d'une infection d'organe ou de cavité du site opératoire établi par un chirurgien ou le médecin impliqué

ANNEXE 2.

Listes des termes sélectionnés selon la spécialité chirurgicale et la méthode d'extraction.

Chirurgie orthopédique du Rachis			
Documents (T^1)		Documents et Formulaires (T^2)	
Avant correction	Après correction*	Avant correction	Après correction*
antibiotherapie	antibiotherapie	antibiotherapie	antibiotherapie
antibiotique	antibiotique	cicatrice	antibiotique
cicatrice	cicatrice	crp	cicatrice
crp	cicatrice inflammatoire	desunion	cicatrice inflammatoire
desunion	crp	desunion de le cicatrice	crp
desunion de le cicatrice	curetage	ecoulement	desunion
ecoulement	desunion	ecoulement au niveau	desunion de le cicatrice
ecoulement purulent	desunion de le cicatrice	ecoulement purulent	ecoulement
fils	ecoulement	infectiologues	ecoulement purulent
infectiologues	ecoulement purulent	infection	infectiologues
lavage	fils	lavage	infection
lesion	infectiologues	lesion	lavage
niveau de le cicatrice	lavage	parage	lesion
parage	lesion	reprise	parage
plaie	parage	reprise pour lavage	reprise
rifadine	plaie	rifadine	reprise pour lavage
rifampicine	rifadine	rifampicine	rifadine
sepsis	sepsis	sepsis	sepsis
site operatoire	site operatoire	site operatoire	site operatoire
staphylocoque	staphylocoque	staphylocoque	staphylocoque
Neurochirurgie			
Documents (T^1)		Documents et Formulaires (T^2)	
Avant modification	Après modification**	Avant modification	Après modification**
ablation du volet	ablation du volet	ablation du volet	ablation du volet
antibiogramme	antibiotique	antibiogramme	antibiogramme
antibiotique	aureus	antibiotique	antibiotique
aureus	cicatrice	aureus	aureus
cicatrice	culture	culture	cicatrice
culture	epidermidis	epidermidis	culture
epidermidis	examen direct	examen direct	epidermidis
examen direct	fermeture	fistule	examen direct
fistule	fosfomycine	fosfomycine	fermeture
fosfomycine	germe	germe	fosfomycine
germe	glycorachie	pl	germe
hernie	hypoglycorachie	plaie	pl
pl	infection du site	raideur	plaie
raideur	pl	reprise chirurgical	reprise
reprise chirurgical	reprise	rifampicine	rifampicine
rifampicine	rifampicine	sd inflammatoire	sd inflammatoire
sd inflammatoire	sd inflammatoire	staph	staph
staph	staph	staphylocoque	staphylocoque
staphylocoque	staphylocoque	tdm	tdm
vancomycine	tdm	vancomycine	vancomycine

* 8 ISO non déclarées ont été reclassées en ISO+, 11 interventions de reprises post-ISO ont été exclues.

** 4 ISO avec peu ou pas d'informations exploitables ont été exclues.

ANNEXE 3.

Détection automatique des infections du site opératoire à partir d'un entrepôt de données

Marine Quéroué[a], Agnès Lashéras-Bauduin[b], Vianney Jouhet[ac], Frantz Thiessard[ac], Jean-Marc Vital[d], Anne-Marie Rogues[be], Sébastien Cossin[ac]

[a] CHU de Bordeaux, Pôle de Santé Publique, Service d'Information Médicale, Informatique et Archivistique Médicales (IAM), Bordeaux, F-33000, France
[b] CHU de Bordeaux, Pôle de Santé Publique, Service d'Hygiène Hospitalière, Bordeaux, F-33000, France
[c] Université de Bordeaux, Bordeaux Population Health Research Center, équipe ERIAS, UMR 1219, F-33000 Bordeaux, France
[d] CHU de Bordeaux, Pôle de chirurgie, Service de chirurgie orthopédique et traumatologique, unité de chirurgie du rachis, Bordeaux, F-33000, France
[e] Université de Bordeaux, Bordeaux Population Health Research Center, ISPED, UMR 1219, F-33000 Bordeaux, France

Résumé

La réduction de l'incidence des infections du site opératoire (infection associée aux soins survenant à la suite d'un acte de chirurgie) fait partie des objectifs du programme national de lutte contre les infections nosocomiales. Pour ce faire une surveillance manuelle est réalisée chaque année par l'équipe d'hygiène hospitalière et les chirurgiens du CHU de Bordeaux. Notre objectif était de développer un algorithme de détection automatique des ISO à partir des données du système d'information hospitalier. Les fiches de surveillance de la chirurgie du rachis des années 2015, 2016 et 2017 ont servi de gold standard pour classifier les actes de chirurgie et entraîner des algorithmes d'apprentissage automatique. Notre jeu d'apprentissage comprenait 22 ISO parmi 2133 actes de chirurgie du rachis. Des variables prédictives d'ISO ont été extraites à partir de l'entrepôt de données i2b2. Nous avons comparé deux approches différentes. La première utilise plusieurs sources de données et offre les meilleures performances mais est difficilement généralisable à d'autres établissements. La seconde est basée uniquement sur le texte libre avec extraction semi-automatique des termes discriminants puis apprentissage automatique. Les algorithmes réussissent à identifier l'ensemble des ISO avec 20 et 26 faux positifs respectivement sur le jeu d'apprentissage. Une évaluation sur des nouvelles données est en cours. Ces résultats sont encourageants pour le développement de méthodes de surveillance semi-automatisée.

Mots-clés:
Surgical Wound Infection
Datawarehouse
Natural Language Processing

Introduction

Entre 2012 et 2017, la proportion d'infections du site opératoire (ISO) est passée de 13,5 % à 15,92 % des infections associées aux soins (IAS) les classant en deuxième position après les infections urinaires d'après l'enquête nationale de prévalence des infections nosocomiales [1]. Durant cette période, les proportions d'ISO profondes et au niveau de l'organe ont augmenté (respectivement 4,8 % vs 5,77 % et 5,5 % vs 7,74 %). Seules les ISO superficielles ont vu leur proportion baisser (3,2 % vs 2,41 %). Dans le rapport 2017 de surveillance des ISO dans les établissements de santé, les taux d'incidence étaient en 2017 de 1,37 % pour la chirurgie orthopédique, 1,97 % pour la chirurgie digestive, 1,88 % pour la gynécologie-obstétrique, 1,10 % pour la traumatologie, 2,60 % pour l'urologie, 0,79 % pour la neurochirurgie, 1,72 % pour la chirurgie bariatrique, 3,44 % pour la chirurgie coronaire, 3,99 % pour la chirurgie réparatrice et reconstructive, 1,32 % pour la chirurgie thoracique et 2,32 % pour la chirurgie vasculaire [2].

Améliorer la surveillance et la prévention des ISO fait partie de l'axe 3 du programme national d'actions de prévention des IAS (Propias) [3]. Il a pour objectif de « disposer d'outils de surveillance des ISO graves (profondes ou nécessitant une reprise chirurgicale), d'évaluation de leur prévention et de gestion adaptés dans les 3 secteurs de l'offre de soins ».

Jusqu'en 2018, la surveillance des ISO en France était proposée aux établissements de santé volontaires par le réseau ISO-Raisin selon 2 modalités. La surveillance dite prioritaire concerne une liste d'interventions sentinelles. Il s'agit d'un recueil d'informations avec création de fiches concernant au moins 100 interventions consécutives de la même spécialité pendant les 6 premiers mois de l'année. La surveillance dite globale ou agrégée se fait quant à elle sur une période d'au moins deux mois au cours du premier semestre, concerne des interventions incluses ou non dans la liste prioritaire et nécessite la création de fiches seulement pour les ISO déclarées [2]. Cette surveillance s'avère être extrêmement chronophage du fait du recueil de données qu'elle impose et de la validation des ISO par les chirurgiens.

En novembre 2018, le CPias Ile-de-France a été nommé par Santé Publique France pour le pilotage de la mission nationale « Surveillance et prévention du risque infectieux liés aux actes de chirurgie et de médecine interventionnelle » (Spicmi). Cette mission a pour vocation le remplacement du réseau actuel ISO-Raisin [4]. Elle a notamment pour objectif de passer à un nouveau système de surveillance reposant principalement sur l'utilisation des données des systèmes d'information hospitalier (SIH) des établissements de santé. Cette surveillance « semi-automatisée » représenterait un gain en termes de ressources nécessaires pour la collecte des données dans chaque établissement.

Plusieurs structures ont déjà étudié la possibilité d'automatiser la surveillance des IAS et notamment des ISO. C'est le cas du CHU de Nantes qui a évalué l'utilisation des données du programme de médicalisation des systèmes d'information (PMSI) et de la bactériologie pour détecter les ISO [5]. Selon eux, une

surveillance assistée par ordinateur peut être mise en œuvre dans les hôpitaux français en utilisant des sources de données disponibles. Le gain de temps d'une détection semi-automatisée permettra aux professionnels de la prévention des infections associées aux soins de consacrer plus de temps aux tâches de prévention et d'éducation. Ils soulignaient cependant la nécessité d'une étude multicentrique pour évaluer la transposabilité de cette méthode.

En dehors de la France des recherches sont également menées dans ce sens. L'Université de Stockholm a étudié l'utilisation de techniques d'apprentissage automatique pour la détection des infections associées aux soins. Les chercheurs ont utilisé des données hospitalières recueillies lors d'une enquête de prévalence ponctuelle. Elles comprenaient des données textuelles, des codes CIM10 (classification internationale des maladies, 10ème révision), des administrations médicamenteuses, des résultats microbiologiques et la température corporelle. Les algorithmes d'apprentissage séparateurs à vaste marge (SVM) et gradient tree boosting (GTB) se sont révélés performants pour cette tâche avec un excellent rappel [6].

Dans le cadre du projet national d'amélioration de la qualité chirurgicale (NSQIP) aux États-Unis, l'Université du Minnesota a également mis en place un outil automatisé de détection des événements indésirables liés à une intervention chirurgicale. Ces derniers ont utilisé les données cliniques de leur centre médical ainsi que celles du registre NSQIP. Parmi celles-ci, ils ont sélectionné des données démographiques (sexe, âge, race) et des données cliniques (diagnostics CIM-9, résultats biologiques, administrations médicamenteuses, demandes d'examens complémentaires et constantes). Les modèles utilisant la régression logistique ont montré les meilleures performances éliminant de manière fiable la grande majorité des patients sans ISO et réduisant ainsi de manière significative la charge des registres [7].

Enfin, l'Université de l'Utah a quant à elle développé un système de traitement automatique du langage naturel (TAL) pour identifier automatiquement les mentions d'ISO dans les comptes rendus de radiologie. Ils ont travaillé sur la chirurgie gastro-intestinale à partir de la base de données MIMICIII Critical Care dont ils ont extrait les codes diagnostiques, les codes d'acte et les comptes rendus de scanner dans les 30 jours suivant la procédure. Ces derniers ont été annotés par deux chirurgiens afin de créer un lexique de termes. Ils ont ensuite développé un système de TAL afin d'identifier et classer automatiquement les preuves d'ISO à partir de chaque compte rendu en s'appuyant sur une adaptation de l'algorithme ConText qui gère la négation et la temporalité. Leur système de TAL s'est montré plus performant que les deux autres approches testées utilisant des données administratives uniquement ou des techniques d'apprentissage automatique SVM avec une représentation du texte en n-grammes [8].

Le CHU de Bordeaux a mis en œuvre en novembre 2017 un entrepôt de données de santé basé sur la solution open source i2b2 [9] pour faciliter la réutilisation des données à des fins de recherche, de prévention et d'amélioration de la qualité des soins. L'un des cas d'usage de l'entrepôt est de développer des méthodes de détection automatique des ISO en collaboration avec le service d'hygiène hospitalière chargé de leur surveillance et de leur prévention. Cet article présente les premiers résultats de l'implémentation d'une méthode de détection automatique des ISO de la chirurgie du rachis au CHU de Bordeaux.

Méthodologie

Dans le cadre de la surveillance nationale organisée par le réseau ISO-Raisin, et afin d'évaluer l'incidence des ISO au sein du CHU de Bordeaux, le service d'hygiène hospitalière réalise des enquêtes chaque année sur des chirurgies ciblées. C'est le cas de la chirurgie du rachis pour laquelle l'ensemble des actes réalisés sur une période de 3 mois consécutifs par an est étudié. Parmi une liste des actes extraits, il est demandé à chaque chirurgien de signaler lesquels ont conduit à une ISO. Pour chaque ISO déclarée, les membres du service d'hygiène hospitalière réalisent alors un retour au dossier patient informatisé et une analyse des causes pour dégager des mesures de prévention.

Les fiches de surveillance ainsi créées constituent un gold standard avec respectivement 13 ISO déclarées sur 662 actes, 7 sur 708 actes et 2 sur 763 actes pour les années 2015, 2016 et 2017. Au total, 2133 actes de chirurgie du rachis ont été classés par les chirurgiens ayant réalisé ces actes et 22 ISO ont été déclarées sur ces trois années.

L'entrepôt de données du CHU de Bordeaux contient des données structurées (diagnostics CIM10, actes de la Classification Commune des Actes Médicaux - CCAM, prescriptions et administrations médicamenteuses, biologie et bactériologie), semi-structurées (formulaires) et des données non structurées (documents en texte libre). Cet entrepôt est alimenté quotidiennement par les données du SIH. L'ensemble des données de notre cohorte est mobilisable pour développer des méthodes d'apprentissage automatique afin de prédire automatiquement des nouveaux cas d'ISO et ainsi mettre en place une surveillance semi-automatisée au CHU de Bordeaux.

Utiliser l'ensemble des données disponibles limite la transposabilité de l'algorithme à d'autres établissements. En effet, les établissements de santé ont chacun des applications différentes et certaines données du CHU de Bordeaux, par exemple les résultats de bactériologie, ne sont pas structurées de la même façon dans un autre établissement. Deux algorithmes ont été développés pour prendre en compte ce compromis entre transposabilité et quantité d'information, le premier utilise l'ensemble des données disponibles, il est donc spécifique au CHU de Bordeaux ; le second se limite aux données en texte libre et pourrait être transposé à d'autres établissements.

Les données PMSI pour la tarification à l'activité, les administrations médicamenteuses, les données de la bactériologie ainsi que l'ensemble du texte libre ont été utilisées par le premier algorithme. Les variables du modèle ont été sélectionnées manuellement à partir de connaissances expertes des ISO de la chirurgie du rachis. Concernant le PMSI, des codes CIM10 relatifs à une ISO et pertinents dans le cadre d'une chirurgie orthopédique du rachis ont été sélectionnés, ainsi qu'un acte CCAM de reprise post-ISO validé par les chirurgiens (tableau 1). Parmi les administrations médicamenteuses, seules les antibiothérapies ont été prises en compte (codes J01 et J04 de la classification ATC). Parmi les données de la bactériologie, des protocoles dont la demande était faite lors d'une suspicion d'ISO (plaie opératoire, pus profond, matériel orthopédique, biopsie ostéo-articulaire ou autre, liquide de lame ou redon) ont été sélectionnés.

Tableau 1 – Codes CIM10 et codes CCAM en rapport avec une infection du site opératoire de la chirurgie orthopédique du rachis sélectionnés par l'expert

Code CIM10	**Libellé**
T81.4	Infection après un acte à visée diagnostique et thérapeutique, non classée ailleurs
T84.5	Infection et réaction inflammatoire dues à une prothèse articulaire interne
T84.6	Infection et réaction inflammatoire dues à un appareil de fixation interne [toute localisation]
T84.7	Infection et réaction inflammatoire dues à d'autres prothèses, implants et greffes orthopédiques internes
Code CCAM	**Libellé**
AFPA001	Mise à plat de lésion infectieuse péridurale rachidienne et/ou paravertébrale postopératoire [sepsis], par abord direct

Enfin concernant le texte libre, 12 termes spécifiques à l'ISO ont été sélectionnés par un expert du domaine.

Ainsi pour chaque acte de notre cohorte, dans un délai de 90 jours après la date d'intervention (période de surveillance requise pour la chirurgie du rachis) nous avons relevé la présence ou non d'un diagnostic relatif à une ISO signalée par le chirurgien, la présence ou non d'un acte de reprise, la présence ou non d'une administration antibiotique, la présence ou non d'un protocole bactériologique en rapport avec une suspicion d'ISO et pour chaque terme d'intérêt sa fréquence dans le texte libre.

Le second algorithme était uniquement basé sur les données non structurées en texte libre de l'entrepôt de données, notamment les comptes rendus opératoires, de consultation et d'hospitalisation dans un délai de 90 jours après l'intervention. Contrairement au premier algorithme, la sélection des termes d'intérêt a commencé par une étape de sélection automatique de termes. Un étiquetage morpho-syntaxique était d'abord réalisé par le logiciel TreeTagger [10] puis les groupes nominaux étaient extraits à l'aide d'une expression régulière. Les formes lemmatisées de ces groupes nominaux correspondaient aux termes d'intérêt. Par exemple, dans la phrase « reprise chirurgicale pour infection du site opératoire » les termes « reprise, reprise chirurgicale, reprise chirurgicale pour infection, infection, infection du site, infection du site opératoire, site, site opératoire » étaient extraits automatiquement. Le code de cet outil est publié en open-source[1].

Un premier filtre a permis d'exclure les termes non fréquents chez les patients atteints d'ISO. Un seuil arbitraire de 20 % a été fixé, si un terme survenait chez moins de 5 patients atteints d'ISO parmi les 22, le terme était exclu. Un odds ratio (rapport de cotes), une mesure utilisée en classification automatique de textes pour quantifier le lien entre un terme et une catégorie [12], a été calculé pour chacun des termes. Un deuxième filtre excluait les termes qui pourraient être trop spécifiques au CHU de Bordeaux. Le chapitre d'un livre sur les infections postopératoires de la chirurgie du rachis [11] a été utilisé : les termes ont été extraits par la même méthode décrite plus haut et les termes des dossiers patients non présents dans cette référence ont été exclus. Finalement, les 20 termes les plus associés aux ISO ont été conservés après validation par un expert du domaine. Cette méthode de sélection semi-automatique des termes est résumée dans la figure 1. Dans la matrice fournie aux algorithmes d'apprentissage, chaque patient était une ligne et chaque terme une colonne. Si le terme était retrouvé dans les 90 jours après la date d'intervention, l'information était codée 1 et 0 sinon.

Deux algorithmes de classification ont été testés : la régression logistique et les forêts aléatoires. Pour la détection semi-automatique des ISO, il est souhaitable d'obtenir une parfaite sensibilité. Les algorithmes ont été évalués dans ce sens : ils ont été comparés en fonction de leur spécificité pour une sensibilité fixée à 100%. Pour obtenir cette sensibilité, le seuil de prédiction a été fixé à la plus faible probabilité prédite parmi nos cas d'ISO.

Résultats

Premier algorithme

La régression logistique offre les meilleurs résultats avec l'ensemble des 22 ISO détectées sur les 2133 actes pour 20 faux positifs (tableau 2).

Tableau 2 –Prédictions de la régression logistique avec l'ensemble des données sur jeu d'apprentissage (années 2015, 2016 et 2017). M : ISO, T : prédiction

Algorithme 1	**M+**	**M-**	
T+	22	20	42
T-	0	2091	2091
	22	2111	2133

La spécificité est de 99,05 %, la valeur prédictive positive de 52,38 % et l'exactitude de 99,06 %. Un retour au dossier a été effectué concernant les 20 faux positifs : 8 étaient des ISO non déclarées (dont 2 superficielles), 9 étaient des actes de reprise post-ISO et 3 n'étaient pas des ISO. Le modèle utilisant les forêts aléatoires présente de moins bonnes performances avec une précision de 20,18 % (87 faux positifs) pour une sensibilité à 100 %.

Enfin nous avons évalué les performances de la régression logistique en scindant notre échantillon de départ. L'apprentissage se faisant sur les années 2015 et 2016, nous avons testé le modèle sur l'année 2017 (Tableau 3). Les 2 ISO signalées cette année ont correctement été détectées ainsi que 4 faux positifs sur 763 actes étudiés (valeur prédictive positive : 33,3 %).

[1] https://github.com/scossin/CandidateTerm

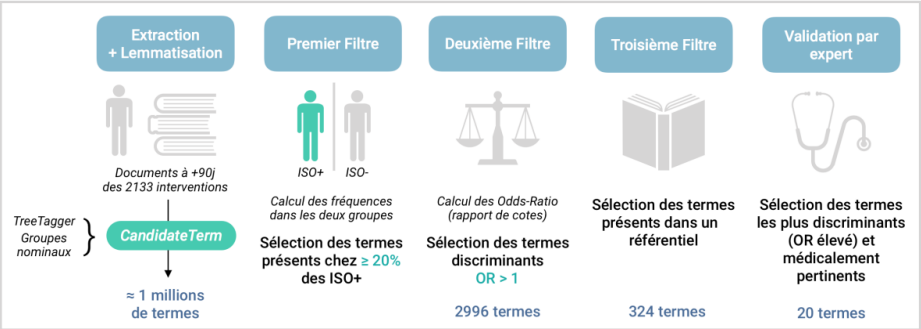

Figure 1 – Sélection semi-automatique des termes (algorithme 2). Un programme extrait automatiquement les termes (groupes nominaux) des documents. Les filtres conservent les termes fréquents chez les cas d'ISO, discriminants et spécifiques à la chirurgie du rachis. Enfin, un expert valide et conserve 20 termes cliniquement pertinents.

Tableau 3 – Prédictions de la régression logistique avec l'ensemble des données sur un jeu de test (année 2017) après apprentissage (année 2015 et 2016). M : ISO, T : prédiction

Algorithme 1	**M+**	**M-**	
T+	2	4	6
T-	0	757	757
	2	761	763

Tableau 4 –Prédictions de la régression logistique avec données textuelles uniquement, sur jeu d'apprentissage (années 2015, 2016 et 2017). M : ISO, T : prédiction

Algorithme 2	**M+**	**M-**	
T+	22	26	48
T-	0	2085	2085
	22	2111	2133

Deuxième algorithme

Environ un million de termes ont été extraits du texte libre et 2 996 termes conservés après le premier filtre qui excluait les termes non fréquents chez les patients atteints. Le terme « site opératoire » est survenu chez la totalité des patients atteints d'ISO et seulement chez 5 % des patients non atteints. Ce terme avait l'odds ratio (OR) le plus élevé.

Parmi ces 2 996 termes, seulement 324 termes ont été retrouvés dans l'ouvrage de référence sur les ISO de la chirurgie du rachis et ont donc été conservés par le deuxième filtre. Parmi les termes exclus, le terme « code AFPA001 » avait un OR très élevé (1302). Ce code d'acte CCAM a pour libellé « mise à plat de lésion infectieuses péridurale rachidienne et/ou paravertébrale postopératoire ». La présence de ce code, notamment dans les comptes rendus opératoires, est probablement trop spécifique au CHU de Bordeaux car les codes CCAM sont systématiquement présents.

Les 20 termes les plus associés aux ISO en termes d'odds ratio ont été conservés après validation par l'expert, dans l'ordre : site opératoire, antibiothérapie, sepsis, écoulement, parage, désunion de la cicatrice, lésion, plaie, infectiologues, lavage, écoulement purulent, rifadine, cicatrice, fils, staphylocoque, CRP, niveau de la cicatrice, rifampicine, point et antibiotique.

La régression logistique offrait les meilleurs résultats et détectait l'ensemble des cas d'ISO au prix de 26 faux positifs (tableau 4).

Discussion

Résultats

Ces premiers résultats, permettant de repérer l'ensemble des ISO pour très peu de faux positifs, sont très encourageants pour le développement de méthodes semi-automatiques.

Le premier algorithme utilise l'ensemble des données disponibles et est très spécifique aux données du CHU de Bordeaux. Sa meilleure capacité à séparer les ISO et les non ISO par rapport au deuxième algorithme utilisant uniquement le texte libre peut s'expliquer par le nombre élevé de sources de données utilisées : codes diagnostics, code d'acte, prélèvements bactériologiques, antibiothérapie et mention de l'ISO dans le texte libre. Utiliser plusieurs sources de données augmente la probabilité de trouver de l'information concernant une ISO chez un patient.

Bien que les résultats de cet article ne soient pas comparables avec d'autres travaux, les performances ont un ordre de grandeur similaire. L'algorithme développé il y a quelques années à Nantes reposant uniquement sur le PMSI et la bactériologie [5] a obtenu un rappel de 90 % avec une précision de 20 %. L'algorithme utilisant le TAL développé par l'Université de l'Utah présente quant à lui de très bonnes performances avec un rappel de 93 % pour une précision de 82 %.

Notre deuxième algorithme présente aussi des résultats intéressants. Il démontre que l'information est présente dans le texte libre pour la détection des ISO. Les termes détectés automatiquement dans le texte sont pertinents pour la détection des ISO et offrent des résultats similaires à une sélection manuelle des variables par un expert du domaine. Cette approche de sélection de variables pourrait être réutilisée pour la détection des ISO des autres chirurgies. La méthode proposée utilise une approche linguistique pour l'extraction de groupes nominaux. Une mesure statistique et des ressources externes

sont utilisées pour sélectionner et filtrer les termes associés aux ISO. Dans un souci de transposabilité, filtrer les termes à partir d'une ressource externe est une étape importante pour retirer des termes trop spécifiques à notre établissement bien que ceux-ci pourraient améliorer la prédiction. Cet algorithme pré-entraîné au CHU de Bordeaux présente l'avantage d'être plus facilement transposable à d'autres établissements.

Limites

Les performances des algorithmes sont surévaluées en l'absence d'un échantillon d'évaluation. Bien que fixée à 100 %, la sensibilité nécessite d'être mesurée sur un nouveau jeu de données.

La principale limite de cette étude est le petit nombre de cas (22) d'ISO dans le gold-standard et la qualité de ce dernier. Une enquête des faux positifs a permis de montrer des erreurs dans le gold-standard qui étaient soupçonnées. Ce premier modèle avec une parfaite sensibilité a été privilégiée pour détecter l'ensemble des ISO et explorer les faux-positifs. Certains cas d'absence d'ISO étaient en réalité des ISO non déclarées par les chirurgiens probablement à cause d'un biais de mémorisation (oubli du clinicien).

Une limite du premier algorithme est la sélection manuelle des variables d'intérêt par un expert du domaine, celle-ci est relativement chronophage et reste très spécifique à la chirurgie du rachis. Aussi un seul expert a procédé à la sélection. L'algorithme 2 est plus facilement généralisable car les étapes manuelles sont moins chronophages. Il serait intéressant d'automatiser la tâche de sélection des variables cependant valider manuellement la pertinence clinique des variables sélectionnées favorise l'acceptabilité de l'algorithme par les chirurgiens.

Une autre limite de ce travail est l'absence d'évaluation formelle de l'outil d'extraction automatique des termes des documents. Celui-ci est spécifique au français et pourrait être comparé aux outils similaires comme BioTex [12] et TermSuite [13].

Aucune détection de la négation ou de la temporalité n'est pour l'instant réalisée bien que des algorithmes en français existent [14,15].

Perspectives

Ce premier algorithme va permettre de détecter automatiquement des nouveaux cas d'ISO au-delà de la période d'étude (3 mois pendant 3 années). Ces nouveaux cas d'ISO détectés et validés permettront d'améliorer le gold-standard et la robustesse de l'algorithme.

Une interface de validation est en cours de développement pour permettre aux chirurgiens de visualiser les prédictions de l'algorithme afin de confirmer ou d'infirmer la présence d'une ISO chez un patient. Cette méthode semi-automatisée permettra de gagner du temps pour la déclaration des ISO, d'éviter un biais de mémorisation et de permettre une couverture temporelle plus large pour la surveillance.

Un travail similaire est en cours pour la détection des ISO en neurochirurgie et pour les poses de prothèses articulaires. La méthode de TAL est plus facilement généralisable à d'autres chirurgies et moins chronophage que la première approche de sélection manuelle des variables dans plusieurs sources de données.

Ce travail s'inscrit dans la démarche actuelle d'utilisation des outils informatiques et des algorithmes d'apprentissage automatique pour améliorer la prévention et la qualité des soins. Les algorithmes proposés pourraient entrer dans le cadre de la mission Spicmi qui a pour objectif de passer à un nouveau système de surveillance national reposant principalement sur l'utilisation des données des SIH. La possibilité d'utiliser le même algorithme pour l'ensemble des établissements reste à démontrer.

Conclusion

Dans cet article, nous montrons la faisabilité de détecter automatiquement des infections du site opératoire à partir d'un entrepôt de données au CHU de Bordeaux. Un outil de validation semi-automatique pourra permettre un gain de temps aux équipes en charge de la surveillance des infections du site opératoire. Deux algorithmes ont été développés, le premier utilise l'ensemble des données disponibles tandis que le second utilise seulement les données en texte libre et est donc plus facilement transposable à un autre établissement. Les premiers résultats sont prometteurs avec de bonnes performances répondant aux exigences de sensibilité d'un outil de détection. Bien que les résultats soient meilleurs en utilisant l'ensemble des données, l'algorithme utilisant le texte libre seul fournit des résultats très proches. Ces résultats sont pour l'instant limités à la chirurgie du rachis et les algorithmes nécessitent d'être évalués et améliorés à partir de nouvelles données annotées.

Références

[1] C. Daniau, L. Léon, H. Blanchard, C. Bernet, and E. Caillet-Vallet, Enquête nationale de prévalence des infections nosocomiales et des traitements anti-infectieux en établissements de santé, France, mai-juin 2017. Santé Publique France, 2018.

[2] CPIAS, Surveillance des infections du site opératoire dans les établissements de santé. Réseau ISO-Raisin, France. Résultats 2017, Santé Publique France, 2019.

[3] DGOS, Ministère des affaires sociales, de la santé et des droits des femmes. Programme national d'actions de prévention des infections associées aux soins, Ministère de la santé, 2015.

[4] Spicmi Réseau surveillance Iso, (n.d.). http://www.cpias-ile-de-france.fr/surveillance/spicmi.php (accessed June 11, 2019).

[5] B. Leclère, C. Lasserre, C. Bourigault, M.-E. Juvin, M.-P. Chaillet, N. Mauduit, J. Caillon, M. Hanf, D. Lepelletier, and SSI Study Group, Matching bacteriological and medico-administrative databases is efficient for a computer-enhanced surveillance of surgical site infections: retrospective analysis of 4,400 surgical procedures in a French university hospital, *Infect Control Hosp Epidemiol*. **35** (2014) 1330–1335. doi:10.1086/678422.

[6] C. Ehrentraut, M. Ekholm, H. Tanushi, J. Tiedemann, and H. Dalianis, Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting, *Health Informatics J*. **24** (2018) 24–42. doi:10.1177/1460458216656471.

[7] Z. Hu, G.J. Simon, E.G. Arsoniadis, Y. Wang, M.R. Kwaan, and G.B. Melton, Automated Detection of Post-operative Surgical Site Infections Using Supervised Methods with Electronic Health Record Data, *Stud Health Technol Inform*. **216** (2015) 706–710.

[8] A.B. Chapman, D.L. Mowery, D.S. Swords, Wendy.W. Chapman, and B.T. Bucher, Detecting Evidence of Intra-abdominal Surgical Site Infections from Radiology Reports Using Natural Language Processing, *AMIA Annu Symp Proc*. **2017** (2018) 515–524.

[9] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, and H.C. Chueh, Integration of clinical and genetic data

in the i2b2 architecture, *AMIA Annu Symp Proc*. (2006) 1040.
[10] H. Schmid, Improvements in Part-of-Speech Tagging with an Application to German., *Proceedings of the ACL SIGDAT-Workshop*. (1995).
[11] M. Tadié, Complications de la chirurgie du rachis : de l'identification à la prévention. 2015.
[12] J.A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire, BIOTEX: A System for Biomedical Terminology Extraction, Ranking, and Validation, in: Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272, CEUR-WS.org, Aachen, Germany, Germany, 2014: pp. 157–160.
[13] D. Cram, and B. Daille, Terminology Extraction with Term Variant Detection, in: Proceedings of ACL-2016 System Demonstrations, Association for Computational Linguistics, Berlin, Germany, 2016: pp. 13–18. doi:10.18653/v1/P16-4003.
[14] A. Abdaoui, A. Tchechmedjiev, W. Digan, S. Bringay, and C. Jonquet, French ConText: Détecter la négation, la temporalité et le sujet dans les textes cliniques Français, in: SIIM: Symposium Sur l'Ingénierie de l'Information Médicale, Toulouse, France, 2017. https://hal-lirmm.ccsd.cnrs.fr/lirmm-01656834 .
[15] N. Garcelon, A. Neuraz, V. Benoit, R. Salomon, and A. Burgun, Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse, *J Am Med Inform Assoc*. **24** (2017) 607–613. doi:10.1093/jamia/ocw144.

Adresse de correspondance

marine.queroue@chu-bordeaux.fr

DÉTECTION AUTOMATIQUE DES INFECTIONS DU SITE OPÉRATOIRE

Marine QUÉROUÉ

RÉSUMÉ

Introduction L'amélioration de la surveillance et de la prévention des infections du site opératoire (ISO) fait partie du programme national de lutte contre les infections nosocomiales. Notre objectif était de mettre en place un outil de détection automatique par apprentissage supervisé afin de remplacer le système de surveillance actuel.
Méthode Deux approches ont été menées pour détecter les ISO suite à une chirurgie du rachis et une neurochirurgie correspondant respectivement à 2133 et 2303 interventions. La première approche utilise les multiples sources d'information disponibles dans l'entrepôt de données du CHU de Bordeaux. La seconde approche utilise uniquement le texte libre. Pour chaque approche, nous avons comparé la précision de deux algorithmes, à savoir la régression logistique et les forêts aléatoires, avec un rappel fixé à 100%.
Résultats Le modèle final utilisant toutes les données a obtenu les meilleures performances pour la chirurgie du rachis avec une précision de 94%. Le modèle utilisant des données de texte libre a obtenu des résultats corrects et était meilleur pour la neurochirurgie.
Discussion L'utilisation du texte libre présente l'avantage d'être transposable à d'autres établissements de santé et facilement applicable à diverses spécialités chirurgicales avec des performances stables. Les performances de nos algorithmes doivent être évaluées sur un jeu de données test.

MOTS-CLÉS
Infection de plaie opératoire
Entreposage de données
Traitement du langage naturel
Apprentissage machine supervisé

ABSTRACT

Introduction Improving monitoring and prevention of surgical site infections (SSI) is part of the national nosocomial infection control program. Our goal was to implement an automated detection tool with a supervised machine learning to replace the current manual system.
Method Two approaches were conducted to detect SSI following spine surgery and neurosurgery corresponding to 2133 and 2303 procedures respectively. The first approach uses the multiple sources of information available in the data warehouse of Bordeaux University Hospital. The second approach uses only free text. For each approach, we compared the precision of two algorithms namely logistic regression and random forests algorithms for a fixed recall value of 100%.
Results The final model using all the data achieved the best performance for spine surgery with a precision of 94%. The model using free text data obtained correct results and was better for neurosurgery.
Discussion The use of free text has the advantage of being replicable to other health institutions and applicable to various surgical specialties with stable performance. The performance of our algorithms needs to be evaluated on a test dataset.

KEYWORDS
Surgical Wound Infection
Data Warehousing
Natural Language Processing
Supervised Machine Learning